ICEIS 2012

14th International Conference on Enterprise Information Systems

Proceedings

Volume 1

Wrocław, Poland
28 June - 1 July, 2012

Sponsored by:

INSTICC

Co-organized by:

Wrocław University of Economics

In Cooperation with:

aaai

eic

SWIM

acm In-Cooperation

sigmis

SIGCHI

ICEIS 2012

Proceedings of the
14th International Conference on
Enterprise Information Systems

Volume 1

Wrocław, Poland

28 June - 1 July, 2012

Sponsored by
INSTICC – Institute for Systems and Technologies of Information, Control and Communication

Co-organized by
Wrocław University of Economics

In Cooperation with
AAAI – Association for the Advancement of Artificial Intelligence
IEICE – The Institute of Electronics, Information and Communication Engineers
SWIM – IEICE Special Interest Group on Software Interprise Modelling
ACM – Association for Computing Machinery
ACM SIGMIS – ACM Special Interest Group on Management Information Systems
ACM SIGCHI – ACM Special Interest Group on Computer Human Interaction

Edited by Leszek Maciaszek, Alfredo Cuzzocrea and José Cordeiro

Printed in Portugal
ISBN: 978-989-8565-10-5
Depósito Legal: 344492/12

http://www.iceis.org/
iceis.secretariat@insticc.org

Brief Contents

INVITED SPEAKERS

Schahram Dustdar

Vienna University of Technology

Austria

Dimitris Karagiannis

University of Vienna

Austria

Steffen Staab

University of Koblenz-Landau

Germany

Pericles Loucopoulos

Loughborough University

U.K.

Yannis Manolopoulos

Aristotle University

Greece

ORGANIZING AND STEERING COMMITTEES

CONFERENCE CHAIR

Joaquim Filipe, Polytechnic Institute of Setúbal / INSTICC, Portugal

PROGRAM CO-CHAIRS

José Cordeiro, Polytechnic Institute of Setúbal / INSTICC, Portugal

Leszek Maciaszek, Wroclaw University of Economics, Poland / Macquarie University ∼ Sydney, Australia

Alfredo Cuzzocrea, ICAR-CNR and University of Calabria, Italy

PROCEEDINGS PRODUCTION

Helder Coelhas, INSTICC, Portugal

Andreia Costa, INSTICC, Portugal

Patrícia Duarte, INSTICC, Portugal

Bruno Encarnação, INSTICC, Portugal

Liliana Medina, INSTICC, Portugal

Raquel Pedrosa, INSTICC, Portugal

Vitor Pedrosa, INSTICC, Portugal

Cláudia Pinto, INSTICC, Portugal

José Varela, INSTICC, Portugal

CD-ROM PRODUCTION

Pedro Varela, INSTICC, Portugal

GRAPHICS PRODUCTION AND WEBDESIGNER

Daniel Pereira, INSTICC, Portugal

SECRETARIAT

Vitor Pedrosa, INSTICC, Portugal

WEBMASTER

Susana Ribeiro, INSTICC, Portugal

SENIOR PROGRAM COMMITTEE

Balbir Barn, Middlesex University, U.K.

Albert Cheng, University of Houston, U.S.A.

Jan Dietz, Delft University of Technology, The Netherlands

Schahram Dustdar, Vienna University of Technology, Austria

António Figueiredo, University of Coimbra, Portugal

Michel Léonard, CUI, University of Geneva, Switzerland

Kecheng Liu, University of Reading, U.K.

Pericles Loucopoulos, Loughborough University, U.K.

Andrea de Lucia, Università degli Studi di Salerno, Italy

Yannis Manolopoulos, Aristotle University, Greece

Masao Johannes Matsumoto, Solution Research Lab, Japan

Alain Pirotte, University of Louvain, Belgium

Klaus Pohl, University of Duisburg-Essen, Germany

Matthias Rauterberg, Eindhoven University of Technology, The Netherlands

Colette Rolland, Université Paris 1 Panthéon-Sorbonne, France

Narcyz Roztocki, State University of New York at New Paltz, U.S.A.

Abdel-Badeeh Mohamed Salem, Ain Shams University, Egypt

Bernadette Sharp, Staffordshire University, U.K.

Alexander Smirnov, SPIIRAS, Russian Academy of Sciences, Russian Federation

Ronald Stamper, Measur Ltd, U.K.

Merrill Warkentin, Mississippi State University, U.S.A.

PROGRAM COMMITTEE

Miguel Angel Martinez Aguilar, University of Murcia, Spain

Patrick Albers, ESEO - Ecole Superieure D'Electronique de L'Ouest, France

Rainer Alt, University of Leipzig, Germany

Vasco Amaral, CITI FCT/UNL, Portugal

Andreas S. Andreou, Cyprus University of Technology, Cyprus

Wudhichai Assawinchaichote, King Mongkut's University of Technology Thonburi, Thailand

Cecilia Baranauskas, State University of Campinas - Unicamp, Brazil

Senén Barro, University of Santiago de Compostela, Spain

Rémi Bastide, ISIS - CUFR Jean-François Champollion, France

Bernhard Bauer, University of Augsburg, Germany

Lamia Hadrich Belguith, ANLP Research Group, MIRACL, University of Sfax, Tunisia

Jorge Bernardino, Institute Polytechnic of Coimbra - ISEC, Portugal

Felix Biscarri, University of Seville, Spain

Oliver Bittel, HTWG Konstanz - University of Applied Sciences, Germany

Danielle Boulanger, IAE- Université Jean Moulin Lyon 3, France

Jean-Louis Boulanger, CERTIFER, France

Peter Busch, Macquarie University $\sim$ Sydney, Australia

Miguel Calejo, Declarativa, Portugal

Coral Calero, University of Castilla - La Mancha, Spain

Luis M. Camarinha-Matos, New University of Lisbon, Portugal

Olivier Camp, ESEO, France

Roy Campbell, University of Illinoiis at Urbana-Champaign, U.S.A.

Gerardo Canfora, University of Sannio, Italy

Manuel Isidoro Capel-Tuñón, University of Granada, Spain

Angélica Caro, University of Bio-Bio, Chile

Jose Jesus Castro-schez, Escuela Superior de Informatica, Spain

Luca Cernuzzi, Universidad Católica "Nuestra Señora de la Asunción", Paraguay

Sergio de Cesare, Brunel University, U.K.

Ming-Puu Chen, National Taiwan Normal University, Taiwan

Shiping Chen, CSIRO ICT Centre Australia, Australia

Shu-Ching Chen, Florida International University, U.S.A.

Max Chevalier, Institut de Recherche en Informatique de Toulouse UMR 5505, France

Witold Chmielarz, Warsaw University, Poland

Daniela Barreiro Claro, Universidade Federal da Bahia (UFBA), Brazil

Cesar Collazos, Universidad del Cauca, Colombia

Jose Eduardo Corcoles, Castilla-La Mancha University, Spain

Antonio Corral, University of Almeria, Spain

Karl Cox, University of Brighton, U.K.

Sharon Cox, Birmingham City University, U.K.

Alfredo Cuzzocrea, ICAR-CNR and University of Calabria, Italy

Maria Damiani, University of Milan, Italy

Vincenzo Deufemia, Università di Salerno, Italy

Kamil Dimililer, Near East University, Cyprus

José Javier Dolado, University of the Basque Country, Spain

Dulce Domingos, Faculty of Science - University of Lisbon, Portugal

César Domínguez, Universidad de La Rioja, Spain

Ming Dong, Shanghai Jiao Tong University, China

PROGRAM COMMITTEE (CONT.)

Juan C. Dueñas, Universidad Politécnica de Madrid, Spain

Hans-Dieter Ehrich, Technische Universitaet Braunschweig, Germany

João Faria, FEUP - Faculty of Engineering of the University of Porto, Portugal

Antonio Fariña, University of A Coruña, Spain

Jamel Feki, University of Sfax - Faculté Des Sciences Economiques et de Gestion de Sfax, Tunisia

Antonio Fernández-Caballero, Universidad de Castilla-la Mancha, Spain

Edilson Ferneda, Catholic University of Brasília, Brazil

Maria João Silva Costa Ferreira, Universidade Portucalense, Portugal

Paulo Ferreira, INESC-ID / IST, Portugal

Filomena Ferrucci, Università di Salerno, Italy

Rita Francese, Università degli Studi di Salerno, Italy

Bogdan Franczyk, University of Leipzig, Germany

Ana Fred, Technical University of Lisbon / IT, Portugal

Lixin Fu, University of North Carolina, Greensboro, U.S.A.

Mariagrazia Fugini, Politecnico di Milano, Italy

Jose A. Gallud, University of Castilla-la Mancha, Spain

Matjaz Gams, Jozef Stefan Institute, Slovenia

Maria Ganzha, SRI PAS and University of Gdansk, Poland

Juan Garbajosa, Technical University of Madrid, Spain

Mouzhi Ge, Universitaet der Bundeswehr Munich, Germany

Marcela Genero, University of Castilla-La Mancha, Spain

Joseph Giampapa, Carnegie Mellon University, U.S.A.

Raúl Giráldez, Pablo de Olavide University of Seville, Spain

Pascual Gonzalez, Universidad de Castilla-la Mancha, Spain

Robert Goodwin, Flinders University of South Australia, Australia

Feliz Gouveia, University Fernando Pessoa / Cerem, Portugal

Luis Borges Gouveia, Universidade Fernando Pessoa, Portugal

Virginie Govaere, INRS, France

Janis Grabis, Riga Technical University, Latvia

Maria Carmen Penadés Gramaje, Universitat Politècnica de València, Spain

Sven Groppe, University of Lübeck, Germany

Wieslawa Gryncewicz, Wroclaw University of Economics, Poland

Nuno Guimarães, Lasige / Faculty of Sciences, University of Lisbon, Portugal

Maki K. Habib, The American University in Cairo, Egypt

Yaakov Hacohen-Kerner, Jerusalem College of Technology (Machon Lev), Israel

Sylvain Hallé, Université du Québec à Chicoutimi, Canada

Slimane Hammoudi, ESEO, France

Wahab Hamou-Lhadj, Concordia University, Canada

Christian Heinlein, Aalen University, Germany

Markus Helfert, Dublin City University, Ireland

Suvineetha Herath, Richard Stockton State College of New Jersey, U.S.A.

Orland Hoeber, Memorial University of Newfoundland, Canada

Wladyslaw Homenda, Warsaw University of Technology, Poland

Jun Hong, Queen's University Belfast, U.K.

Wei-Chiang Hong, Oriental Institute of Technology, Taiwan

PROGRAM COMMITTEE (CONT.)

Miguel J. Hornos, University of Granada, Spain

Kai-I Huang, Tunghai University, Taiwan

Akram Idani, Grenoble INP - Grenoble Institute of Technology, France

Marta Indulska, The University of Queensland, Australia

François Jacquenet, University of Saint-Étienne, France

Arturo Jaime, Universidad de La Rioja, Spain

Marijn Janssen, Delft University of Technology, The Netherlands

Wassim Jaziri, Higher Institute of Computer and Multimedia of Sfax, Tunisia

Paul Johannesson, Royal Institute of Technology, Sweden

Jan Jürjens, TU Dortmund & Fraunhofer ISST, Germany

Michail Kalogiannakis, University of Crete, Greece

Nikos Karacapilidis, University of Patras, Greece

Nikitas Karanikolas, Technological Educational Institute of Athens (TEI-A), Greece

Stamatis Karnouskos, SAP, Germany

Andrea Kienle, University of Applied Sciences, Dortmund, Germany

Marite Kirikova, Riga Technical University, Latvia

Alexander Knapp, Universität Augsburg, Germany

Fotis Kokkoras, TEI of Larisa, Greece

Ryszard Kowalczyk, Swinburne University of Technology, Australia

John Krogstie, NTNU, Norway

Subodha Kumar, Texas A&M University, U.S.A.

Rob Kusters, Eindhoven University of Technology & Open University of the Netherlands, The Netherlands

Halina Kwasnicka, Wroclaw University of Technology, Poland

Alain Leger, France Telecom Orange Labs, France

Kauko Leiviskä, University of Oulu, Finland

Daniel Lemire, UQAM - University of Quebec at Montreal, Canada

Da-Yin Liao, Applied Wireless Identifications, U.S.A.

Luis Jiménez Linares, University of de Castilla-La Mancha, Spain

Panos Linos, Butler University, U.S.A.

Stephane Loiseau, Leria, France

João Correia Lopes, Faculdade de Engenharia da Universidade do Porto/INESC Porto, Portugal

Maria Filomena Cerqueira de Castro Lopes, Universidade Portucalense Infante D. Henrique, Portugal

María Dolores Lozano, University of Castilla-la Mancha, Spain

Miguel R. Luaces, Universidade da Coruña, Spain

André Ludwig, University of Leipzig, Germany

Vicente Luque-Centeno, Carlos III University of Madrid, Spain

Mark Lycett, Brunel University, U.K.

Lukasz Lysik, Wroclaw University of Economics, Poland

Cristiano Maciel, Universidade Federal de Mato Grosso, Brazil

Rita Suzana Pitangueira Maciel, Federal University of Bahia, Brazil

Mirko Malekovic, University of Zagreb, Croatia

Nuno Mamede, INESC-ID, Portugal

Paolo Maresca, Università Federico II, Italy

Pierre Maret, Université de Saint Etienne, France

Tiziana Margaria, University of Potsdam, Germany

Farhi Marir, London Metropolitan University, U.K.

Herve Martin, Grenoble University, France

Maria João Martins, Instituto Superior Tecnico, Portugal

PROGRAM COMMITTEE (CONT.)

Katsuhisa Maruyama, Ritsumeikan University, Japan

Viviana Mascardi, University of Genoa, Computer Science Department, Italy

David Martins de Matos, L2F / INESC-ID Lisboa / Instituto Superior Técnico, Portugal

Wolfgang Mayer, University of South Australia, Australia

Javier Medina, University of Granada, Spain

Jerzy Michnik, University of Economics in Katowice, Poland

Luo Ming, Southeastern Institute of Manufacturing and Technology, Singapore

Michele Missikoff, IASI-CNR, Italy

Ghodrat Moghadampour, Vaasa University of Applied Sciences, Finland

Pascal Molli, LINA, University of Nantes, France

Lars Mönch, FernUniversität in Hagen, Germany

Valérie Monfort, SOIE Tunis, Tunisia

Francisco Montero, University of Castilla-la Mancha, Spain

Carlos León de Mora, University of Seville, Spain

Paula Morais, Portucalense University, Portugal

Fernando Moreira, Universidade Portucalense, Portugal

Haralambos Mouratidis, University of East London, U.K.

Pietro Murano, University of Salford, U.K.

Ana Neves, knowman - Consultadoria em Gestão, Lda, Portugal

Matthias Nickles, Technical University of Munich, Germany

Ann Nosseir, British University in Egypt, Egypt

Jose Angel Olivas, Universidad de Castilla - La Mancha, Spain

David Olson, University of Nebraska, U.S.A.

Mehmet Orgun, Macquarie University $\sim$ Sydney, Australia

Andrés Muñoz Ortega, Catholic University of Murcia (UCAM), Spain

Samia Oussena, University of West London, U.K.

Sietse Overbeek, Delft University of Technology, The Netherlands

Tansel Ozyer, TOBB ETU, Turkey

Claus Pahl, Dublin City University, Ireland

Marcin Paprzycki, Polish Academy of Science, Poland

José R. Paramá, University of A Coruña, Spain

Eric Pardede, La Trobe University, Australia

Viviana Patti, University of Torino, Italy

Loris Penserini, FBK-IRST, Italy

Massimiliano Di Penta, University of Sannio, Italy

Laurent Péridy, IMA-UCO, France

Dana Petcu, West University of Timisoara, Romania

Yannis A. Phillis, Technical University of Crete, Greece

Josef Pieprzyk, Macquarie University, Australia

Selwyn Piramuthu, University of Florida, U.S.A.

José Pires, Escola Superior de Tecnologia e Gestão, Portugal

Luís Ferreira Pires, University of Twente, The Netherlands

Geert Poels, Ghent University, Belgium

Michal Polasik, Nicolaus Copernicus University, Poland

Srini Ramaswamy, ABB, India

T. Ramayah, Universiti Sains Malaysia, Malaysia

Pedro Ramos, Instituto Superior das Ciências do Trabalho e da Empresa, Portugal

Marek Reformat, University of Alberta, Canada

Francisco Regateiro, Instituto Superior Técnico, Portugal

Hajo A. Reijers, Eindhoven University of Technology, The Netherlands

PROGRAM COMMITTEE (CONT.)

Ulrich Reimer, University of Applied Sciences St. Gallen, Switzerland

Marinette Revenu, Greyc Ensicaen, France

Nuno de Magalhães Ribeiro, Universidade Fernando Pessoa, Portugal

Debbie Richards, Macquarie University ∼ Sydney, Australia

Michele Risi, University of Salerno, Italy

David Rivreau, Université Catholique de L'ouest, France

Alfonso Rodriguez, University of Bio-Bio, Chile

Daniel Rodriguez, University of Alcalá, Spain

Pilar Rodriguez, Universidad Autónoma de Madrid, Spain

Oscar Mario Rodriguez-Elias, Institute of Technology of Hermosillo, Mexico

Erik Rolland, University of California at Riverside, U.S.A.

Jose Raul Romero, University of Cordoba, Spain

David G. Rosado, University of Castilla-la Mancha, Spain

Gustavo Rossi, Lifia, Argentina

Artur Rot, Wroclaw University of Economics, Poland

Francisco Ruiz, Universidad de Castilla-La Mancha, Spain

Roberto Ruiz, Pablo de Olavide University, Spain

Belen Vela Sanchez, Rey Juan Carlos University, Spain

Luis Enrique Sánchez, Sicaman Nuevas Tecnologias S.L., Spain

Manuel Filipe Santos, University of Minho, Portugal

Jurek Sasiadek, Carleton University, Canada

Andrea Schaerf, Università di Udine, Italy

Daniel Schang, ESEO, France

Sissel Guttormsen Schär, Institute for Medical Education, Switzerland

Isabel Seruca, Universidade Portucalense, Portugal

Jianhua Shao, Cardiff University, U.K.

Hossein Sharif, University of Portsmouth, U.K.

Alberto Silva, INESC, Portugal

Sean Siqueira, Federal University of the State of Rio de Janeiro, Brazil

Spiros Sirmakessis, Technological Educational Institution of Messolongi, Greece

Hala Skaf-molli, Nantes University, France

Chantal Soule-Dupuy, Universite Toulouse 1, France

José Neuman de Souza, Universidade Federal do Ceará, Brazil

Martin Stanton, Manchester Metropolitan University, U.K.

Chris Stary, University of Linz, Austria

Dick Stenmark, Gothenburg University, Sweden

Stefan Strecker, University of Hagen, Germany

Vijayan Sugumaran, Oakland University, U.S.A.

Hiroki Suguri, Miyagi University, Japan

Lily Sun, University of Reading, U.K.

Raj Sunderraman, Georgia State University, U.S.A.

Jerzy Surma, Warsaw School of Economics, Poland

Gion K. Svedberg, Malmo University, Sweden

Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland

Vladimir Tarasov, Jönköping University, Sweden

Arthur Tatnall, Victoria University, Australia

Sotirios Terzis, University of Strathclyde, U.K.

Claudine Toffolon, Université du Maine, France

Theodoros Tzouramanis, University of the Aegean, Greece

Athina Vakali, Aristotle University, Greece

José Ângelo Braga de Vasconcelos, Universidade Fernando Pessoa, Portugal

PROGRAM COMMITTEE (CONT.)

Michael Vassilakopoulos, University of Central Greece, Greece

Christine Verdier, LIG - University Joseph Fourier Grenoble, France

François Vernadat, European Court of Auditors, France

Maria Esther Vidal, Universidad Simon Bolivar, Venezuela

Aurora Vizcaino, Escuela Superior de Informática, Spain

Bing Wang, University of Hull, U.K.

Dariusz Wawrzyniak, Wroclaw University of Economics, Poland

Hans Weghorn, BW Cooperative State University Stuttgart, Germany

Hans Weigand, Tilburg University, The Netherlands

Gerhard Weiss, University of Maasticht, The Netherlands

Graham Winstanley, University of Brighton, U.K.

Wita Wojtkowski, Boise State University, U.S.A.

Viacheslav Wolfengagen, Institute JurInfoR, Russian Federation

Andreas Wombacher, University of Twente, The Netherlands

Robert Wrembel, Poznan University of Technology, Poland

Stanislaw Wrycza, University of Gdansk, Poland

Min Wu, Oracle, U.S.A.

Wen-Yen Wu, I-Shou University, Taiwan

Mudasser Wyne, National University, U.S.A.

Haiping Xu, University of Massachusetts Dartmouth, U.S.A.

Sadok Ben Yahia, Faculty of Sciences of Tunis, Tunisia

Lili Yang, Loughborough University, U.K.

Ping Yu, University of Wollongong, Australia

Yugang Yu, Erasmus University, The Netherlands

Wei Zhou, ESCP Europe, France

Eugenio Zimeo, University of Sannio, Italy

Lin Zongkai, Chinese Academy of Sciences, China

AUXILIARY REVIEWERS

Ankica Barisic, FCT UNL, Portugal

Bruno Barroca, CITI FCT UNL, Portugal

Gabriele Bavota, University of Salerno, Italy

Paulo Carreira, University of Lisbon, Portugal

Jan Claes, Ghent University, Belgium

Rui Domingues, FCT/UNL, Portugal

Jessica Diaz Fernandez, Universidad Politécnica de Madrid (Technical U. of Madrid), Spain

Daniel Lopez Fernnadez, UPM, Spain

Rodrigo Garcia-Carmona, Universidad Politécnica de Madrid, Spain

Laura Sánchez González, University of Castilla La Mancha, Spain

Krzysztof Kania, University of Economics in Katowice, Poland

Sandra Lovrencic, University of Zagreb, Faculty of organization and informatics Varazdin, Croatia

Paloma Cáceres García de Marina, Rey Juan Carlos University, Spain

Luã Marcelo Muriana, UFMT, Brazil

David Musat, UPM, Italy

Álvaro Navas, Universidad Politécnica de Madrid, Spain

Annibale Panichella, University of Salerno, Italy

Ignazio Passero, University of Salerno, Italy

Jonas Poelmans, Faculty of Business and Economics, K.U. Leuven, Belgium

Hércules Antonio do Prado, Embrapa/Universidade Católica de Brasília, Brazil

Abdallah Qusef, University of Salerno, Italy

Federica Sarro, Università di Salerno, Italy

Markus Schatten, University of Zagreb, Croatia

Claudio Schifanella, Università degli Studi di Torino, Italy

Diego Seco, University of A Coruña, Spain

SELECTED PAPERS BOOK

A number of selected papers presented at ICEIS 2012 will be published by Springer-Verlag in a LNBIP Series book. This selection will be done by the Conference Chair and Program Co-chairs, among the papers actually presented at the conference, based on a rigorous review by the ICEIS 2012 Program Committee members.

FOREWORD

This volume contains the proceedings of the 14th International Conference on Enterprise Information Systems (ICEIS 2012), held in Wroclaw, Poland, sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC) and co-organized by the Wroclaw University of Economics.

The conference was held in cooperation with the Association for Advancement of Artificial Intelligence (AAAI), the Association for Computation Machinery (ACM) Special Interest Group on Management Information Systems (SIGMIS), the ACM Special Interest Group on Computer Human Interaction (SIGCHI) and the Institute of Electronics Information and Communication Engineers (IEICE) Special Interest Group on Software Interprise Modelling (SWIM).

This conference has become a major point of contact between research scientists, engineers and practitioners in the area of business applications of information systems, with six simultaneous tracks, covering different aspects related to enterprise computing, including: "Databases and Information Systems Integration", "Artificial Intelligence and Decision Support Systems", "Information Systems Analysis and Specification", "Software Agents and Internet Computing", "Human-Computer Interaction" and "Enterprise Architecture". Papers published in each track describe the cutting-edge research work that is often oriented towards real world applications and highlight the benefits of Information Systems and Technology for industries, thus making a bridge between the academia and the enterprise worlds.

Following the trend of previous editions, ICEIS 2012 had a number of satellite events, namely special sessions and workshops, related to the field of the conference, including the following workshops: the 9th International Workshop on Security in Information Systems (WOSIS), the 10th International Workshop on Modelling, Simulation, Verification and Validation of Enterprise Information Systems (MSVVEIS), the 9th International Workshop on Natural Language Processing and Cognitive Science (NLPCS), the 1st International Workshop on Web Intelligence (WEBI) and the 1st International Workshop on Interaction Design in Educational Environments (IDEE), and special sessions on Semantic Computing and Ontology Engineering (SCOE), New Tools, Techniques and Methodologies for Information System Testing (NTMIST) and Model Driven Development for Information Systems: Techniques, Tools and Methodologies (MDDIS).

ICEIS 2012 received 299 paper submissions from 49 countries and districts on all continents. 28 papers were published and presented as full papers, i.e. completed work (10 pages/30' oral presentation) and 75 papers, reflecting work-in-progress, were accepted and orally presented as short papers (6 pages/20' oral presentation). Furthermore, 56 contributions were accepted and presented as posters.

These numbers, lead to a "full-paper" acceptance ratio around 9%, and a total oral acceptance ratio of 34%. Additionally, as usual in the ICEIS conference series, a number of

invited talks, presented by internationally recognized specialists in different areas, have positively contributed to reinforce the overall quality of the Conference and to provide a deeper understanding of the Enterprise Information Systems field.

The program for this conference required the dedicated effort of many people. First, we must thank the authors, whose research and development efforts are recorded here. Second, we thank the members of the program committee and the additional reviewers for a valuable help with their expert reviewing of all submitted papers. Third, we thank the invited speakers for their invaluable contributions and the time for preparing their talks. Fourth, we thank the workshop and special session chairs whose collaboration with ICEIS was much appreciated. Finally, special thanks to all the members of the Wroclaw University of Economics and INSTICC, whose close coordination and cooperation was fundamental for the success of this conference.

Two best paper awards are given at the closing session to outstanding papers presented at the conference: an award for the top regular paper in the conference plus an award for the best student paper, overall. The selection is based on the classifications and comments provided by the Program Committee and also on the oral presentation quality, assessed by session chairs.

A final selection of papers, from those presented at the conference, will be done based on peer-assessment, i.e. on the classifications and comments provided by the Program Committee and on the assessment provided by session chairs. Extended and revised versions of these papers will be published in a book by Springer-Verlag.

We wish you all an exciting conference and an unforgettable stay in Wroclaw, Poland. We hope to meet you again next year for the 15th ICEIS, to be held in Angers, France, details of which will be readily available at http://www.iceis.org.

José Cordeiro
Polytechnic Institute of Setúbal / INSTICC, Portugal

Alfredo Cuzzocrea
ICAR-CNR and University of Calabria, Italy

Leszek Maciaszek
Wroclaw University of Economics, Poland / Macquarie University ~ Sydney, Australia

CONTENTS

POSTERS

ARTIFICIAL INTELLIGENCE AND DECISION SUPPORT SYSTEMS

FULL PAPERS

SHORT PAPERS

POSTERS

INVITED SPEAKERS

KEYNOTE SPEAKERS

Design by Units
A Novel Approach for Building Elastic Systems

Schahram Dustdar
Vienna University of Technology, Vienna, Austria

Abstract: Systems are built by utilizing resources. Resources can include infrastructure such as compute power, storage space, and bandwidth, but also nontechnical resources such as the financial budget available or the human (expert) manpower needed to skillfully operate the system, make decisions, or perform human-based computing tasks. The elasticity of a system through virtualized resources is thus a fundamental requirement of Web-scale systems; in system design, those resources must receive careful consideration. In this talk I will discuss the main principles of elasticity and present a fresh look at this problem, and examine how to integrate people in the form of human-based computing and software services into one composite system, which can be modeled, programmed, and instantiated on a large scale in an elastic way.

BRIEF BIOGRAPHY

Schahram Dustdar (ACM Distinguished Scientist), is Full Professor of Computer Science with a focus on Internet Technologies heading the Distributed Systems Group, Vienna University of Technology (TU Wien). From 1999 - 2007 he worked as the co-founder and chief scientist of Caramba Labs Software AG in Vienna (acquired by Engineering NetWorld AG), a venture capital co-funded software company focused on software for collaborative processes in teams. He is Editor in Chief of Computing (Springer), Associate Editor of IEEE Transactions on Services Computing, and on the editorial board of IEEE Internet Computing, as well as author of some 300 publications and several books. More info on his homepage: www.infosys.tuwien.ac.at/Staff/sd

Hybrid Modeling

Dimitris Karagiannis
University of Vienna, Vienna, Austria

Abstract: In the fast paced and complex world of new business models, powerful techniques for supporting business operations and next generation enterprise systems are widely sought-after. For this purpose, modeling methods have not only been discussed and elaborated from an academic perspective but have also been successfully deployed on an industrial scale. For taking into account the distinct requirements of individual users and organizations, the creation of new and the adaptation of existing modeling methods are today common requirements. This process, denoted as "hybrid modeling", will be presented in this talk, based on the foundations and current challenges for the conceptualization of modeling methods, their implementation and deployment. The approach will be illustrated by reverting to a number of recent examples from the Open Models Initiative that provides an open community platform for the exchange of know-how on modeling methods, models and tools. Thereby we will revert to a meta modeling framework that has been developed at the University of Vienna. Furthermore, results of successful applications based on the ADOxx® platform in research and industrial projects will be shown.

BRIEF BIOGRAPHY

Dimitris Karagiannis is head of the research group knowledge engineering at the University of Vienna. His main research interests include knowledge management, modelling methods and meta-modelling. Besides his engagement in national and EU-funded research projects Dimitris Karagiannis is the author of research papers and books on Knowledge Databases, Business Process Management, Workflow-Systems and Knowledge Management. He serves as expert in various international conferences and is presently on the editorial board of Business & Information Systems Engineering (BISE), Enterprise Modelling and Information Systems Architectures and the Journal of Systems Integration. He is member of IEEE and ACM and is on the executive board of GI as well as on the steering committee of the Austrian Computer Society and its Special Interest Group on IT Governance. Recently he started the Open Model Initiative (www.openmodels.at) in Austria. In 1995 he established the Business Process Management Systems Approach (BPMS), which has been successfully implemented in several industrial and service companies, and is the founder of the European software- and consulting company BOC (http://www.boc-group.com), which implements software tools based on the meta-modelling approach.

Managing Online Business Communities

Steffen Staab
University of Koblenz-Landau, Koblenz, Germany

Abstract: Online Business Communities constitute an asset to companies for various purposes such as open innovation, customer self-help or knowledge management. In this talk we will present challenges and opportunities that arise from actively monitoring and managing business communities.

BRIEF BIOGRAPHY

Steffen Staab is professor for databases and information systems at the University of Koblenz-Landau. He is director of the institute for Web Science and Technologies (West; http://west.uni-koblenz.de). He is programme chair of WWW 2012 and editor-in-chief of Elsevier's Journal of Web Semantics. His interests are related to many aspects of Web Science, such as Semantic Web, Web Retrieval, Social Web, Multimedia Web, Software Web and Interactive Web. Steffen is project coordinator for the EU Integrated Project "Robust - Risk and Opportunities Management of Huge-Scale Business Community Cooperation". Previously, Steffen held positions as researcher, project leader and lecturer at the University of Freiburg, the University of Stuttgart/Fraunhofer Institute IAO, and the University of Karlsruhe and he is a co-founder of Ontoprise GmbH.

Requirements Engineering
Panacea or Predicament?

Pericles Loucopoulos
Loughborough University, Loughborough, U.K.

Abstract: The genesis of Requirements Engineering (RE) research around the mid 1970's was motivated by practitioners, who noticed the urgent need for disciplined RE in software projects that had grown large and unmanageable. Much of RE research since then has focused on artifacts that maintain the intellectual discipline by helping capture, share, represent, analyze, negotiate, and prioritize requirements as a basis for design decisions and interventions. The field of RE is arguably one of the most sensitive areas in the development of not only software but more importantly in the development of systems and organisational structures and processes supported by such systems. The scope of this keynote talk is to examine the contextual and methodological factors underpinning much of the practice of RE, to critically examine the utility of current thinking, to identify a set of challenges that are likely to shape the filed of RE in the years to come and to map a set of research directions that are likely to play a significant role in addressing these challenges.

BRIEF BIOGRAPHY

Pericles Loucopoulos is Professor of Information Systems in the Business School, Loughborough University, UK. He began his career in the City of London where he was responsible for delivering systems for financial applications. He moved to Manchester in 1984 to take up an academic appointment at the University of Manchester Institute of Science & Technology (UMIST) where in 1990 he was elected to the post of Professor in Information Systems Engineering in the Department of Computation. He has taught at Université de Paris I – Sorbonne, the University of the Aegean, the Delhi Institute of Technology and the Athens University of Economics and Business and has acted as scientific expert for U.K., Greek, Italian, Austrian, and Swiss Governmental institutions. His research work focuses on supporting the transformation of large, complex and dynamic enterprise systems through the provision of information systems. Theoretical results derived from his research have been applied on industrial scale problems in a variety of domains, such as banking, utilities, large-scale sports events etc. For his work he has received the 2005 OR Society's President Medal and the Inform Society's Edelman Laureate Medal. He is the co-editor-in-chief of the Journal of Requirements Engineering, associate editor of Information Systems and of the Journal of Database Management and serves on the Editorial Board of 10 other journals.

Trends in Blog Preservation

Vangelis Banos[1], Nikos Baltas[2] and Yannis Manolopoulos[1]
[1]*Department of Informatics, Aristotle University, Thessaloniki 54124, Greece*
[2]*Department of Computing, Imperial College, London SW7 2AZ, U.K.*
vbanos@gmail.com, manolopo@csd.auth.gr, nb605@doc.ic.ac.uk

Keywords: Blogs, Blog Preservation, Web Archiving.

Abstract: Blogging is yet another popular and prominent application in the era of Web 2.0. According to recent measurements often considered as conservative, as of now worldwide there are more than 152 million blogs with content spanning over every aspect of life and science, necessitating long term blog preservation and knowledge management. In this talk, we will present a range of issues that arise when facing the task of blog preservation. We argue that current web archiving solutions are not able to capture the dynamic and continuously evolving nature of blogs, their network and social structure as well as the exchange of concepts and ideas that they foster. Furthermore, we provide directions and objectives that could be reached to realize robust digital preservation, management and dissemination facilities for blogs. Finally, we will introduce the BlogForever EC funded project, its main motivation and findings towards widening the scope of blog preservation.

1 INTRODUCTION

Blogs are types of websites regularly updated and intended for general public consumption. Their structure is defined as a series of pages in reverse chronological order. Blogs have become fairly established as an online communication and web publishing tool. The set of all blogs and their interconnections is referred to as the Blogosphere (Agarwal N.). The importance and the influence of the blogosphere are constantly rising and have become the subject of modeling and research (Java A.). For instance, a 2006 study of the importance of blogs in politics, and for US Congress in particular, concluded that blogs play "an increasingly powerful role in framing ideas and issues for legislators and leaders directly" (Sroka T.N.). Blogpulse, a blog trend discovery service, identified 126 million blogs in 2009 and over 152 million blogs in 2010; while Tumblr, a relatively new blogging service, reports that they host over 33 million blogs (Tumblr Numbers); statistics which undoubtedly prove the wide acceptance and dynamic evolution of weblogs. Moreover, they underline the importance of this novel electronic publication medium and exert its significance as part of contemporary culture.

But despite the fast growth of blogosphere, there is still no effective solution for ubiquitous semantic weblog archiving, digital preservation, management and dissemination. Current weblog archiving tools and methods are ineffective and inconsistent, disregarding volatility and content correlation issues, while preservation methods for weblog data have not yet been duly considered. Indeed, existing Web Archiving solutions provide no means of preserving constantly changing content, like the content of weblogs.

Furthermore, to the best of our knowledge, no current Web Archiving effort has ever developed a strategy for effective preservation and meaningful usage of Social Media. The inter-dependence aspect of those media, demonstrated by weblogs featuring shared or adversary opinions, as well as weblogs that support, imitate or revolve around more central ones, is profoundly neglected. Two reasons are mainly responsible for this: firstly, the occasional harvesting of web resources and, secondly, their treatment as unstructured pages, leave little margin for capturing the aforementioned communication perspective of weblogs.

In this work, we present the new challenges that have to be met when facing blog preservation, including information integrity, data management, content dynamics and network analysis. Furthermore, we present the BlogForever EC funded project, its main motivation, objectives and findings towards widening the scope of blog preservation.

2 RELATED WORK

Web preservation is defined as 'the capture, management and preservation of websites and web resources'. Web preservation must be a start-to finish activity, and it should encompass the entire lifecycle of the web resource (Ashley K.). The topic of web preservation was initially addressed in a large scale by the Internet Archive in 1996 (The Internet Archive). Subsequently, many national memory institutions understood the value of web preservation and developed special activities towards this goal. Table 1 displays all major national and international web archiving projects which are part of the International Internet Preservation Consortium (IIPC).

Table 1: International Internet Preservation Consortium Members

Organization	Year	Access Methods
Bibliotheca Alexandrina's Internet Archive, Egypt	1996	URL Search
Bibliothèque nationale de France - Archives de l'Internet	2002	URL Search, Keyword Search, Full-Text Search, Topical Collections
Government of Canada Web Archive	2005	URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search
Croatian Web Archive (HAW)	2004	URL Search, Keyword Search
The Internet Archive (International)	1996	URL Search, Topical Collections
The Icelandic Web Archive	2004	URL Search
Finnish Web Archive	2006	URL Search, Full-Text Search
Kulturarw3 - The Web Archive of the National Library of Sweden	1997	URL Search
Library of Congress Web Archive, USA	2000	URL Search, Alphabetic Browsing, Subject Browsing, Topical Collections
Royal Library and the State and University Library, Aarhus, Denmark	2005	URL Search
Nettarkivet Norge (WebArchive Norway)	2001	Keyword Search
New Zealand Web Archive	1999	URL Search, Keyword Search, Alphabetic Browsing, Subject Browsing

Table 1: International Internet Preservation Consortium Members (cont.).

Organization	Year	Access Methods
National Library of Korea	2005	URL Search, Keyword Search, Subject Browsing
PANDORA Australia's Web Archive	1996	URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search, Subject Browsing
Digital Heritage of Catalonia (PADICAT)	2005	URL Search, Keyword Search, Alphabetic Browsing, Subject Browsing, Topical Collections
Webarchive of Slovenia	2007	URL Search, Alphabetic Browsing
The UK Government Web Archive	1997	URL Search, Alphabetic Browsing
UK Web Archive	2005	URL Search, Alphabetic Browsing, Full-Text Search, Subject Browsing, Topical Collections
Web Archiving Project, Japan	2002	Keyword Search, Full-Text Search, Topical Collections
Web archive of The Netherlands	2007	URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search, Topical Collections
WebArchiv - archive of the Czech web	2007	URL Search, Subject Browsing
Web Archive Switzerland	2008	URL Search, Keyword Search, Full-Text Search, Subject Browsing, Topical Collections
Webarchive Austria	2008	URL Search, Topical Collections

As digital preservation techniques progress and awareness is raised on the matter, there is a continuous trend towards preserving more complex objects (Billenness C.). In the scope of web preservation, this means evolving from the preservation of simple web resources (i.e. html documents, images, audio and video files) towards preserving more complex web entities such as complete websites, dynamic web portals and social media. This trend is persisting with more social media content being considered for preservation. For instance, the Library of Congress has started preserving all Twitter content since 2010 (Campbell L.).

The European Commission has identified the growing need to keep digital resources available and usable over time. To support research in the field, the FP7 ICT Research Programme 2009-2010 and 2011-2012 included specific provisions for digital preservation and web preservation under objectives ICT-2009.4.1: Digital Libraries and Digital

Preservation and Objective ICT-2011.4.3 Digital Preservation (Commission, Information and Communications Technologies). A number of EC funded projects pursuing advanced web preservation are listed below:

- **LiWA (**Living Web Archives) aimed to extend the current state of the art and develop the next generation of Web content capture, preservation, analysis, and enrichment services to improve fidelity, coherence, and interpretability of web archives (LiWA).
- **ARCOMEM** (From Collect-All Archives to Community Memories) is about memory institutions like archives, museums and libraries in the age of the social web. Social media are becoming more and more pervasive in all areas of life. ARCOMEM's aim is to help to transform archives into collective memories that are more tightly integrated with their community of users and to exploit Web 2.0 and the wisdom of crowds to make web archiving a more selective and meaning-based process (Edelstein O.).
- **SCAPE** (Scalable Preservation Environments) project will address scalability of large-scale digital preservation workflows. The project aims to enhance the state of the art in three concrete and significant ways. First, it will develop infrastructure and tools for scalable preservation actions; second, it will provide a framework for automated, quality-assured preservation workflows; and, third, it will integrate these components with a policy-based preservation planning and watch system. These concrete project results will be driven by requirements from, and in turn validated within, three large-scale testbeds from diverse application areas: web content, digital repositories, and research data sets (Edelstein O.).
- **LAWA** (Longitudinal Analytics of Web Archive Data) project will build an Internet-based experimental test bed for large-scale data analytics. Its focus is on developing a sustainable infra-structure, scalable methods, and easily usable software tools for aggregating, querying, and analyzing heterogeneous data at Internet scale. Particular emphasis will be given to longitudinal data analysis along the time dimension for Web data that has been crawled over extended time periods (LAWA).

The topic of web preservation in general and blog preservation in particular has been also addressed by a number of private startup companies throughout the world. Pagefreezer (PageFreezer.com) is claiming to support web archiving and social media archiving. Another popular service is VaultPress (VaultPress), which provides security, backup and support for Wordpress blogs.

Despite the presented activities in the field of web preservation, we argue that there is still no effective solution for ubiquitous semantic weblog archiving, digital preservation and dissemination. Current web archiving tools and methods are not designed for the semantic web era and are ineffective and inconsistent, disregarding volatility and content correlation issues. Additionally, preservation methods for weblog have not yet been duly considered.

In the following section, we will present a number of issues that arise when dealing with blog preservation.

3 BLOG PRESERVATION ISSUES, OVERVIEW AND CONSIDERATIONS

Blog preservation activities can be divided into three main groups: (a) content aggregation, (b) archiving, and (c) management. Here, we present the blocking issues for each one of these groups of activities.

3.1 Blog Content Aggregation

Existing web archiving solutions provide no means of aggregating and preserving constantly changing content, like the content of weblogs. The following two broad technical approaches are usually followed.

Firstly, there are initiatives that select and replicate web sites on an individual basis, an approach exemplified by the Web Capture Initiative (Web Archiving) and by some projects developed by national archives. A second group of initiatives use crawler programs to automatically gather and store large sets of publicly available web sites. The Internet Archive follows this approach by taking periodic snapshots of the entire web since 1996. Other crawler-based initiatives have focused on national domains, e.g. the pioneering Swedish Royal Library's Kulturarw project, which is now discontinued (Arvidson A.). A complete list of national web archiving projects is shown on Table 1. These initiatives are usually complemented by deposit approaches, where owners or administrators of websites choose to deposit the web content they are publishing to the repository.

Regardless of the target content, current initiatives employ general purpose web harvesting to collect their material. This approach, although easy to implement, results in problematic and incorrect

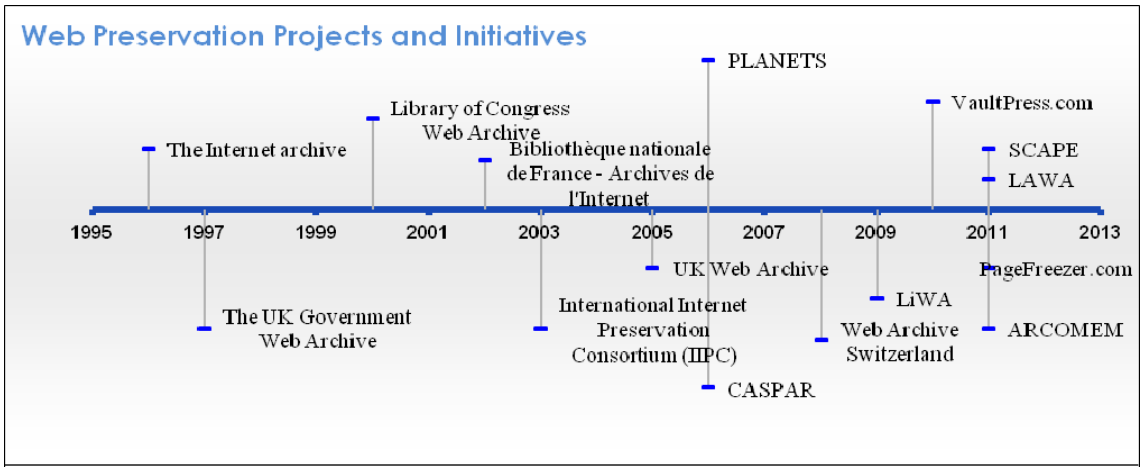


Figure 1: Timeline of important web preservation projects and initiatives.

web archives, especially for highly dynamic site types such as weblogs and wikis, which exhibit special characteristics. More precisely, current weblog content aggregation and digital preservation suffers from the following issues:

- **Web Content aggregation scheduling** is a common issue among web archiving projects, since all of them perform this task on regular intervals without considering web site updates. On the other hand, weblogs are extremely volatile and tend to be updated several times during the day, with new content from editors as well as user comments and discussions. As a result, a large amount of weblog content is not preserved, resulting in subsequent information loss and inconsistent web archives. For instance, Internet Archive's latest web preservation project, Archive-it (Archive-it), which uses the latest Heritrix web crawler (Heritrix), enables harvesting material from the Web as frequently as every 24 hours, once per week, once per month, once per quarter, annually or just once. The same also applies to the popular (Web Curator Tool Project), which is used by the Web Capture Initiative of the Library of Congress, the National Library of New Zealand and numerous other institutions worldwide.
- **Web content aggregation performance** is also a major issue for current web preservation initiatives. Most projects use brute-force methods to crawl through a domain or a set of URLs, retrieving each page, extracting links and visiting each one of them recursively, according to a set of predefined rules. This process is performed periodically without taking into account whether the target site has been modified since the previous content aggregation or which components of the weblog have actually been updated. Unlike regular web sites, weblogs support smart content aggregation by notifying third party applications in the event of content submission or modification. Two technologies supporting this are Blog Ping (Winer D.) and PubSubHub (Bhola S.). Nevertheless, they are not utilized by current web preservation initiatives, resulting in a waste of computing resources.
- **Quality assurance checking** is performed manually or in a semi-automatic way for most web preservation projects. The widely used Web Curator Tool requires the administrator to perform a "Quality Review Task" while the PANDORA Archive's quality checking process (McPhillips S.) also requires human supervision.

3.2 Blog Content Preservation

Preservation refers to the long-term storage and access of digital or digitised content. Existing generic web archiving solutions suffer from several preservation-related shortcomings that render them as poor choices for weblog archiving. These relate to both the long-term storage of a weblog as well as to the access and usage of the preserved content.

1. Current web preservation initiatives are geared towards aggregating and preserving **files** and not **information entities.** For instance, the Internet Archive aggregates web pages and stores them into WARC files (ISO 28500:2009), compressed files similar to zip which are assigned a unique identification number and stored in a distributed file system. Additionally, WARC supports some metadata such as provenance and HTTP protocol metadata. Implicit page elements, such as:

• Page title, headers, content, author information,

- Metadata such as Dublin Core elements,
- RSS feeds and other Semantic Web technologies such as Microformats (Khare R.) and Microdata (Ronallo J.) are completely ignored. This impacts greatly the way stored information is managed, reducing the utility of the archive and also hindering the creation of added-value services.

2. **Current web archiving efforts** disregard the preservation of Social Networks and of interrelations between the archived content. However, weblog interdependencies demonstrated by the identification of central actors and peripheral weblogs, as well as by the meme-effect that applies to them, need to be preserved, to provide meaningful features to the weblog repository.

3. **Current web archive scope is limited** to monolithic regions, subjects or events. There is no generic web archiving solution capable to implement arbitrary subjects and topic hierarchies. For instance, the National Library of Catalonia has initiated a web crawling and access project aiming to collect, process and provide permanent access to the entire cultural, scientific and general output of Catalonia in digital format (PADICAT).

Alternatively, the Library of Congress has developed online collections for isolated historical events such as September 11, 2001 (Library of Congress). There is an ongoing debate, about benefits or disadvantages of one or another long-term preservation methodology. Many papers have been written and many conferences dedicated to this issue have appeared. It is surprising however, how little has been done at practical level.

3.3 Blog Archive Management

Regardless of the way a weblog is archived, current solutions do not provide users with meaningful management features of the stored information. For example, the Internet Archive stores weblogs as generic documents, listing one post after another, an approach that hinders if not forbids further weblog management. Examining the list of national web archiving initiatives (Table 1) one can see that out of 23 projects, only 8 support Full text search (34%), 9 support Alphabetic Browsing (39%) and 8 support Topical Collections (34%). The most common feature available to all archives is URL Search.

Current solutions completely disregard the social aspect and interrelations of weblogs or other social media. Furthermore, due to the nature of periodic web crawling, users can only view the exact state of their weblog on prefixed dates or times. This solution cannot keep track of the evolving semantics and usage context of highly volatile hypertext pages like weblogs. For example, the Occasio News archive, which collects sites based on their relevance to social issues, only preserves specific snapshots from a certain newsgroup (Occasio News Archive Database). Articles do not follow a continuous timeline, a fact that renders their substantial analysis in the future impossible. This results in prolific loss of information with respect to recording the weblog's evolution.

Additionally, current weblog archives cannot preserve the information regarding how posts, relevance links or other weblogs affect the original content and how they led to its propagation or extinction. However, this process must be identified to be of high cultural and sociological value: it is essential to preserve the notions and reactions of contemporary society, the motivations and drives, the interactions between complementary and adversary approaches to certain topics.

Moreover, browsing the preserved Blogosphere through current Web Archiving solutions, like Internet Archive or PANDORA, remains a tentative if not impossible task. For example, within the framework of these solutions, weblog interrelations indicated in the form of Blogrolls are treated as regular hyperlinks of the retrieved Web page with no particular informational value. Not only does this approach lead to the risk of them being omitted during the harvesting stage, especially by domain specific web archives, but it also disregards the value of preserving how thematically correlated weblogs interact with each other.

Finally, though web archived content is generally classified into wide thematic, regional or temporal categories, there exists no robust categorization technique. Weblogs' topic metadata are omitted if they do not fall into the predefined categories. For example, inter-relational authorship information is rarely incorporated into the generic archive model. However, the authorship of electronic publication bears several interesting features, like identification of central actors with authority ranking, person searches and interrelations between authors and the role of anonymity. This has many channels of interest in text mining and the social networking and scientific communities, and would be a stronghold of web archives focusing on social network websites Moreover, the temporal aspect of each Web Archive merely relates to a specific web-snapshot acquired through harvesting. Our methods of real-time harvesting, result into a continuous observation of the lifecycle of a weblog and provide accurate representation for each weblog at any point in time.

As implied by the aforementioned facts, a large fraction of current weblogs lacks digital preservation or it is partially archived. Additionally, digital archives created by means of any of the above

mentioned solutions do not guarantee correctness and consistency, thus preventing their effectiveness and their proper usage.

4 DIRECTIONS TOWARDS ROBUST AND EFFECTIVE BLOG PRESERVATION

In this section, we present our approach towards robust and effective blog preservation. This is a challenge that the BlogForever project (BlogForever) is addressing from four different perspectives: modelling, aggregation, preservation and dissemination. The project's objectives are presented and then each one of the perspectives is outlined.

4.1 Objectives

The project's strategic objective is to provide complete and robust digital preservation, management and dissemination facilities for weblogs. Towards this end, the following scientific and technological objectives have been identified.

4.1.1 Study Weblog Structure and Semantics

BlogForever aims to analyse weblog structure and semantics to understand the unique and complex characteristics of weblogs and develop a generic data model as well as an ontology-based representation of the domain. To achieve this, weblogs are required to be understood and managed in 6 aspects:

1. As physical phenomena
2. As logical encodings
3. As conceptual objects with meaning to humans
4. As structural objects of networked discourse and collaboration for knowledge creation in large groups of humans
5. As sets of essential elements that must be preserved to offer future users the essence of the object
6. As ontologies created in a bottom-up manner by communities rather than specialists

Additionally, weblog aggregation heuristics will be developed to allow us to determine the best practices for efficient data extraction from weblogs.

4.1.2 Define a Robust Digital Preservation Policy for Weblogs

Developing a robust digital preservation policy for weblogs is one of the key objectives. The policy will include the following information:

1. Preservation strategy considerations for assessing risk, requirements for accessing deposited content and long-term accessibility of digital objects, as these factors are deemed to have enduring value. Furthermore, the preservation approach is to be described, including actions that are considered necessary for immediate, intermediate, and long-term preservation. In terms of depositing, it is important to have structures that allow for easy retrieval (and this relates to extracting structures and mapping to them; but also to predicting what and how queries of the future will look like – depending on the amount of flexibility that is required, the data storing can be simpler, or more complex).
2. The Assessment of Interoperability Prospects, which intends to address collaboration issues with existing generic European Web Archiving solutions. Moreover, means for reliable content transfer from the digital archive to other digital repositories, in the event of project termination are to be proposed.
3. The Digital Rights Management Policy, which addresses weblog copyright issues and controls the access level for each item and user in the digital archive.

4.1.3 Implement a Weblog Digital Repository

BlogForever aims to implement a digital repository web application, which will collect, archive, manage and disseminate weblogs. The platform will have the following 2 main components:

1. The weblog aggregation component, which will be capable of searching, harvesting and analysing large volumes of weblogs.
2. The digital repository component, which will be responsible for weblog data preservation. The digital repository will ensure weblog proliferation, safeguard their integrity, authenticity and long-term accessibility over time, and allow for better sharing and re-using of contained knowledge.

A detailed depiction of the BlogForever platform architecture can be seen on Figure 2.

4.1.4 Implement Specific Case Studies

BlogForever aims to design and implement specific case studies to apply and test the created infrastructure on extensive and diverse sets of weblogs. The case studies will be both generic (collecting weblogs from a wide array of topics) and domain specific (for example, a case study in

Figure 2: BlogForever Platform Architecture.

academic bloggers community). Thus the case studies will provide the required breadth and depth to validate the developed tools, and guarantee that the project's results could be successfully and widely replicated after the project ends. The impact of the digital repository will be also evaluated by monitoring system usage and gathering user feedback.

The case studies will begin in summer 2012 and are expected to be completed in August 2013. The largest case study will include 500.000 blogs.

4.2 Modelling

Working towards the objectives identified previously, we have already examined a number of tasks regarding modelling the blogosphere.

4.2.1 Weblog Survey

The BlogForever Weblog Survey report (Arango-Docio S.) outlines a principal investigation into:

1. the common practices of blogging and attitudes towards preservation of blogs;
2. the use of technologies, standards and tools within blogs; and finally,
3. the recent theoretical and technological advances for analysing blogs and their networks.

This investigation aims to inform the development of preservation and dissemination solutions for blogs within the context of BlogForever.

The objectives pursued in this study enabled discussion of:

- common weblog authoring practices;
- important aspects and types of blog data that should be preserved;
- the patterns in weblogs structure and data;
- the technology adopted by current blogs; and finally
- the developments and prospects for analysing blog networks and
- weblog dynamics.

To achieve the aims and objectives of this investigation, a set of review and evaluation exercises were conducted. The members of the BlogForever consortium jointly designed and implemented:

- an online survey involving 900 blog authors and readers;
- an evaluation of technologies and tools used in more than 200 thousand active blogs;
- a review of recent advances in theoretical and empirical research for analysing networks of blogs; and
- a review of empirical literature discussing dynamic aspects of blogs and blog posts.

4.2.2 Weblog Data Model

Our work on weblog data model (Stepanyan K.) identifies the data structures considered necessary for preserving blogs by revisiting the earlier inquiry summarised in the BlogForever Weblog Survey. The report includes an inquiry into

- the existing conceptual models of blogs,
- the data models of Open Source blogging systems, and
- data types identified from an empirical study of web feeds.

The report progresses to propose a data model intended to enable preservation of blogs and their individual components.

Pending work on weblog modeling includes an

exploration of ontologies' applications in the context of blog preservation.

4.2.3 User Requirements and Platform Specifications

Requirements descriptions for the BlogForever platform were thoroughly investigated and assembled from several sources including already completed work, semi-structured interviews with relevant stakeholders and a users' survey (Kalb H.). The report illustrates the method of interview conduction and qualitative analysis. It includes a description of relevant stakeholders and requirement categories.

The identified requirements were specified in a standardised template and modelled with the unified modelling language (UML). Thus, they can be easily explored and utilised by developers. Overall, the requirements are the foundation for the design phase because they represent the perspective of demand.

4.3 Aggregation

The first step to preserve blogs is to manage to achieve effective and complete blog content aggregation. This problem can be split down to two sub-problems, detecting blog updates and retrieving updated blog content.

4.3.1 Weblog Aggregation Prototypes

During our work on weblog aggregation techniques (Rynning M.), we evaluated available weblog data extraction methodologies and technologies. Additionally, a number of weblog data extraction prototypes were implemented to test the aforementioned techniques and evaluate alternative ways to implement the weblog spider component, one of the two key elements of the BlogForever platform. This work will be continued to articulate an optimal set of weblog aggregation techniques.

4.3.2 Spam Filtering

Our research on Spam filtering in the context of blog aggregation (Kim Y.) comprises a survey of weblog spam technology and approaches to their detection. While our work focused on identifying possible approaches to spam detection as a component within the BlogForever software, the discussion has been extended to include observations related to the historical, social and practical value of spam, and proposals of other ways of dealing with spam within the repository without necessarily removing them. It contains a general overview of spam types, ready-made anti-spam APIs available for weblogs, possible methods that have been suggested for preventing the introduction of spam into a blog, and research related to spam focusing on those that appear in the weblog context, concluding in a proposal for a spam detection workflow that might form the basis for the spam detection component of the BlogForever software.

4.4 Preservation

The process of digital preservation requires optimal retrieval and interpretation of the information to be preserved. As presented in the previous sections, our modelling and aggregation prototyping work will be the pillars upon which we will build an effective blog preservation platform.

4.4.1 Preservation Strategy

The preservation strategy will include information on assessing risk, requirements for accessing deposited content and long-term accessibility of digital objects, as these factors are deemed to have enduring value. Furthermore, the preservation approach is to be described, including actions that are considered necessary for immediate, intermediate, and long-term preservation. In terms of depositing, it is important to have structures which allow for easy retrieval (and this relates to extracting structures and mapping to them; but also to predicting what and how queries of the future will look like – depending on the amount of flexibility that is required, the data storing can be simpler, or more complex).

4.4.2 Interoperability Strategy

Our planned work on the interoperability prospects of the BlogForever platform intends to analyse the different facets of interoperability: syntactic, semantic and pragmatic (Papazoglou M.). Furthermore, we are planning to address collaboration issues with existing platforms as well as libraries, archives, preservation initiatives and businesses that might be in synergistic relationships with BlogForever archives.

4.4.3 Digital Rights Management

Our planned work on Digital Rights Management (DRM) will initially include the identification and analysis of open issues and relevant discussions on the topic of blog preservation. Our aims will be protecting public access to information, content creators and content managers.

4.5 Management and Dissemination

To facilitate weblog digital preservation, management and dissemination, the project will implement a digital repository specially tailored to weblog needs. BlogForever digital repository will have to facilitate not only the weblog content but also the extended metadata and semantics of weblogs, which have been accumulated by the weblog aggregator as presented in section 4.3.

The solution of creating a new software system as the basis of the weblogs repository has been considered and dismissed for this task, since many open-source repository back-ends are freely available on the Internet. In this respect, and taking into account the participation of CERN into the BlogForever consortium, the project will extend and adapt the globally acknowledged and widely used Invenio software (CERN). The technology offered Invenio covers all aspects of digital library management. It complies with the Open Archives Initiative metadata harvesting protocol (OAI-PMH) and uses MARC 21 as its underlying bibliographic standard. Its flexibility and performance make it a comprehensive solution for the management of document repositories of large size and render it as an ideal basis for the BlogForever platform.

Long term blog preservation will be one aspect of the BlogForever platform. The other will be providing facilities for various stakeholders (Kalb H.):

- **Content providers** are people or organisations, which maintain one or more blogs and, hence, produce blog content that can or should be preserved in the archive
- **Individual blog authors** are people that maintain their own blog.
- **Organisations** can serve as content providers if they maintain their own corporate blogs.
- **Content retrievers** are people or organisations which have an interest in the content stored in a blog archive and, therefore, they like to search, read, export, etc. that content.
- **Individual blog readers** are people who already read blogs for various reasons, e.g. family, hobbies, professional.
- In contrast, **libraries** operate more as a gatekeeper for individual retrievers. They provide access to various kinds of information sources, e.g. books, journals, movies, etc. Thereby, the access includes value added services like selecting and sorting the sources as well as adding metadata.
- **Businesses** also offer value added services based on the available information.

Each one of the aforementioned stakeholder has different blog preservation, archiving, management and dissemination requirements which have already been recorded and thoroughly documented, setting the priorities and work plan for the implementation of the BlogForever platform.

5 CONCLUSIONS

In this paper, we presented our perspective on the status of blog preservation and the blocking issues that arise when dealing with blog aggregation, preservation and management. Also, we identified a number of open issues that existing web archiving initiatives and platform face when dealing with blogs. Lastly, we presented an outline of the BlogForever EC funded project's current and future work towards creating a modern blog aggregation, preservation, management and dissemination platform.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Commission Framework Programme 7 (FP7), BlogForever project, grant agreement No.269963. We would also like to thank all BlogForever project partners for their invaluable contributions to the project.

REFERENCES

Agarwal N. and Liu H. (2008). Blogosphere: Research Issues, Tools and Applications. *ACM SIGKDD Explorations*, 10(1):18-31.

Arango-Docio S., Sleeman P. and Kalb H. (2011) BlogForever: D2.1 Survey Implementation Report. BlogForever WP2 Deliverable.

Archive-it, Web Archiving Services. 11 04 2012 http://www.archive-it.org/

Arvidson A. (2001). Kulturarw3. *Proceedings Preserving the Present for the Future*. Copenhagen, pages 101-104.

Ashley K., Davis R., Guy M., Kelly B., Pinsent E. and Farrell S. (2010) *A Guide to Web Preservation*.

Bhola S., Strom R., Bagchi S. and Zhao Y. (2002). Exactly-once Delivery in a Content-based Publish-Subscribe System. *Proceedings International Conference on Dependable Systems and Networks (DNS)*. Washington DC, pages 7-16.

Billenness C. (2011). The Future of the Past – Shaping New Visions for EU-research in Digital Preservation.

Proceedings Workshop, European Commission, Information Society and Media Directorate-General, Luxemburg.

BlogForever. BlogForever Project. 15 04 2012 http://blogforever.eu

Campbell L. and Dulabahn B. (2010). Digital Preservation: the Twitter Archives and NDIIPP. *Proceedings 7th International Conference Preservation of Digital Objects (iPRES),* Vienna.

CERN. Invenio. 09 04 2012 http://invenio-software.org/

Commission, European. (2011). Information and Communications Technologies.

Edelstein O., Factor M., King R., Risse T., Salant E. and Taylor P. (2011). Evolving Domains, Problems and Solutions for Long Term. *Proceedings 8th International Conference Preservation of Digital Objects (iPRES),* Singapore.

Heritrix (2012). IA Web Crawler. 14 04 2012 https://webarchive.jira.com/wiki/display/Heritrix/

IIPC (2012). International Internet Preservation Consortium. 10 04 2012. http://www.netpreserve.org

ISO 28500:2009 (2009), Information and Documentation – WARC File Format. Geneva: ISO.

Java A., Kolari P., Finin T. and Oates T. (2006). Modeling the Spread of Influence on the Blogosphere. *Proceedings 3rd WWW Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. Edinburgh.

Kalb H., Kasioumis N., García Llopis J., Postaci S and Arango-Docio S. (2011). BlogForever: D4.1 User Requirements and Platform Specifications Report. Blogforever WP4 Deliverable.

Khare R. and Celik T. (2006). Microformats: a Pragmatic Path to the Semantic Web. *Proceedings 15th International Conference on World Wide Web (WWW).* Edinburgh, pages 865-866.

Kim Y. and Ross S. (2012). BlogForever: D2.5 Weblog Spam Filtering Report and Associated Methodology. BlogForever WP2 Report.

LAWA. Longitudinal Analytics of Web Archive Data Project. 15 04 2012 http://www.lawa-project.eu/

Library of Congress, September 11, 2001, Web Archive. 10 04 2012. http://lcweb2.loc.gov/diglib/lcwa/html/sept11/

LiWA. Living Web Archives Project. 15 04 2012 http://liwa-project.eu

McPhillips S. (2012) PANDORA Archive Technical Details. 05 08 2004. 12 04 2012 http://pandora.nla.gov.au/pandoratech.html

Occasio News Archive Database. 10 04 2012 http://newsarchive.occasio.net/

PADICAT: The Digital Heritage of Catalonia. 10 04 2012. http://www.padicat.cat/

PageFreezer.com - Social Media and Website Archiving. 10 04 2012. http://pagefreezer.com

Papazoglou M.P. and Ribbers P.M.A. (2006). *E-business: Organizational and Technical Foundations*. John Wiley.

Ronallo J. (2012). HTML5 Microdata and Schema.org. *code4lib journal* (2012-02-03).

Rynning M., Banos V., Stepanyan K., Joy M. and Gulliksen M. (2011) BlogForever: D2.4 Weblog spider prototype and associated methodology." BlogForever WP2 Deliverable.

Sroka T.N. (2006). Understanding the Political Influence of Blogs: a Study of the Growing.

Stepanyan K., Joy M., Cristea A., Kim Y., Pinsent E. and Kopidaki S. (2011). BlogForever D2.2 Weblog Data Model. BlogForver WP2 Deliverable.

The Internet Archive (1996). http://archive.org

Tumblr Numbers: The Rapid Rise of Social Blogging. 14 04 2012, http://mashable.com/ 2011/11/14/tumblr-infographic/

VaultPress - Safeguard your site. 10 04 2012. http://www.vaultpress.com

Web Archiving, Library of Congress. 12 04 2012 http://www.loc.gov/webarchiving/

Web Curator Tool Project. 12 04 2012 http://webcurator.sourceforge.net/

Winer D. (2012) Original Announcement of Blog Ping. 12 04 2012 http://xmlrpc.scripting.com/weblogsCom.html

DATABASES AND INFORMATION SYSTEMS INTEGRATION

FULL PAPERS

Social Information Systems
The End of Shadow Applications?

Marc Quast[1] and Mark J. Handel[2]
[1]*University of Grenoble, Campus C207, 220 Rue de la Chimie, 38400 Saint-Martin d'Hères, France*
[2]*The Boeing Company, MC 7L-70, PO Box 3707, Seattle, WA 98124, U.S.A.*
marc.quast@imag.fr, mark.j.handel@boeing.com

Keywords: Information Systems, Business Applications, Enterprise Architecture, Social Software Engineering.

Abstract: In large corporations, line-of-business organizations frequently introduce unofficial "shadow" applications to work around the limitations of the established information system. This paper presents a software architecture designed to alleviate this phenomenon, and reuses examples from a recent industry experience report to demonstrate how shadow application proliferation could be avoided without sacrificing flexibility and reactivity. We present the initial results of our prototype, and discuss the possibility of a social information system designed to both reduce the present chaos and enable the cooperative design and evolution of business applications.

1 INTRODUCTION

Delivering the right information at the right time to the right persons is one of the most important requirements of today's business world (Spahn and Wulf, 2009). Nevertheless, corporate information systems are a widespread source of frustration (Newell et al., 2007). Business units do not accept the poor service provided by their IT departments and build up independent IT resources to suit their specific or urgent requirements (Zarnekow et al., 2006).

As a result, information systems of large corporations are a web of numerous applications. At the center we find a fairly small set of stable and robust enterprise applications. These are surrounded by a larger set of semi-official applications and a very large number of unofficial applications. We adopt the term of *shadow application* proposed by Handel and Poltrock (2010) for the last two categories, i.e. applications introduced by business units to satisfy requirements not met by official applications.

Though the benefit of "getting the job done" is sufficient to justify, and indeed pay for, their existence, shadow applications raise serious problems: duplicated and inconsistent data is commonplace, and having critical information and functionality scattered, unreachable and managed outside of standard IT processes is obviously not what comes to mind when envisioning a well-structured and robust information system.

Building upon our industry experience[1], this paper proposes a potential solution. After a short definition of shadow applications, their main characteristics and the causes of their emergence, we propose an alternative architecture for business applications which could prevent the systematic recourse to shadow applications in their vicinity, using two use cases from (Handel and Poltrock, 2010) to illustrate its effects. We present our prototype implementation and our first results, and discuss the possibility of a social information system designed to both reduce the present chaos and enable the cooperative design and evolution of business applications.

2 UNDERSTANDING SHADOW APPLICATIONS

Shadow applications are characterized by their purpose. If application A exists to work around the limitations of application B, or if A's features belong in B according to its users, A can be considered a

[1]The authors have a cumulated experience of over thirty years in the development and operation of business applications in industrial environments.

shadow application. This partial definition illustrates the subjective nature of the phenomenon.

Shadow applications are also characterized by their ownership. If it is owned by the IT department, it is an official application; otherwise it is a shadow application. The important distinction is not so much "IT or not IT" but "ownership by the actor effectively using the application". This allows the owner to quickly adapt the tool without consulting other parties or relying on the IT organization's priorities. It also provides him with full control over the visibility of the data and access to features.

Individual spreadsheets meet this definition. These are often used for simple data storage and manipulation, as a substitute for more robust business applications. This is a very common and possibly dominant use case since their introduction (Nardi and Miller, 1990), and Handel and Poltrock (2010) qualify such spreadsheets as shadow applications.

"Official" and "shadow" are relative concepts, and apply recursively at various levels of an organization. In other terms, multiple layers of shadow applications exist, the final one being personal applications.

Shadow applications are typically loosely integrated with some official and other shadow applications. However, manual synchronization is not uncommon (Hordijk and Wieringa, 2010).

We define a shadow application as an application which:

- works around another application's limitations and
- is both functionally and technically owned by the organization using it.

Shadow applications are usually considered a "necessary evil" (Hordijk and Wieringa, 2010). Organizations cannot work without them, but would prefer to avoid the data duplication they imply as well as the burden they represent in development and maintenance costs.

The benefits of shadow applications must outweigh the drawbacks; otherwise line-of-business organizations would not develop, deploy, and maintain them. We will refer to the main benefits of shadow applications as perceived by their owners as the "AVI capabilities."

- The owner has full Autonomy to implement new features.
- The owner decides about Visibility of the application to the larger organization.
- The owner can Integrate (manually or automatically) with other applications.

2.1 Examples of Shadow Applications

A recent experience report contributes observations about shadow applications in a 10 year engineering project (Handel and Poltrock, 2010). In this paper, we will use fictional examples derived from the information disclosed in this report.

- *"Luxury can report delays on process instances, but not the reasons for these delays which are managed by a shadow application."*
- *"Sometimes the tasks tracked by Luxury were informally decomposed into subtasks; (…) Luxury had no provisions for this kind of task decomposition."*

We make the assumption that Luxury tracks requests, a common use case in engineering environments. Figure 1 below shows a fictional central database of an official application and two of its shadow applications, managing delay analyses and subtasks respectively.

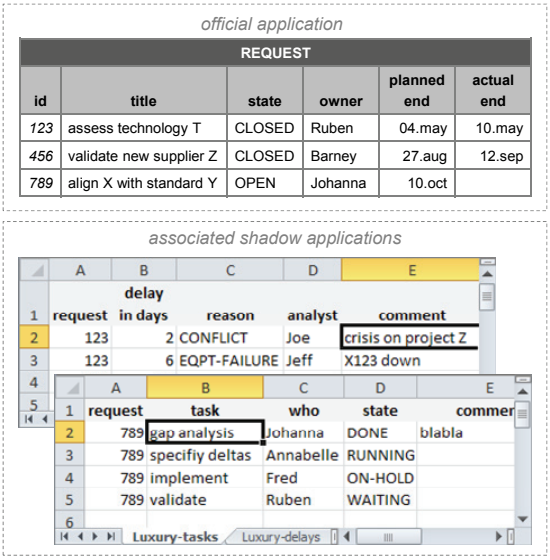


Figure 1: Example of fictional official database and associated shadow application data.

While spreadsheets are arguably the most common form, shadow application architectures are limited only by the owner's resources, including full-blown business applications and, more fashionably, third-party applications in the "stealth cloud", i.e. cloud services being consumed by business users without the knowledge, permission or support of the IT department (Gotts, 2010.).

2.2 Causes of Shadow Application Emergence

Shadow applications emerge to work around the shortcomings of official applications (Zarnekow et al., 2006). Thus we need to understand the causes for these problems.

Large organizations are not consistent and orderly systems. Referring both to groups and individuals, Kling (1991) describes working relationships as *"multivalent with and mix elements of cooperation, conflict, conviviality, competition, collaboration, control, coercion, coordination and combat (the c-words)"*. Requirements from different stakeholders are thus often divergent or conflicting, which explains why the difficulty of requirements engineering increases exponentially with the number and diversity of participants. Ackerman (2000) indicates that when there are hidden or conflicting goals, people will resist articulating these. Under such circumstances, it is a challenge to converge on a consistent set of requirements and deliver a working application at all. But widespread dissatisfaction with the result is almost guaranteed by construction.

As an aggravating factor, corporations are not static. They must adapt to changes in their environment like new markets, technologies or regulations. Though the aforementioned c-words impact is often obvious at the time of application introduction, the continuous evolution of business requirements turns this into a subtle though continuous problem. Any change in any stakeholder's universe can invalidate the initial compromises and demand new rounds of discussion, yielding further dissatisfaction.

Besides inter-organization conflicts, some c-words foster shadow application emergence by themselves. A successful shadow application and the knowledge it captures is usually highly visible within an organization, and its ownership provides recognition (competition) and power (control, coercion).

There are other contributing factors. The widespread practice of reducing IT costs lowers both reactivity and quality of IT support, inciting business units to help themselves (Hoyer and Stanoevska-Slabena, 2008). Technical obsolescence, a consequence of either respectable age or unfortunate choice of foundation technologies, can make it difficult to find the right skillset to implement changes. This paper focuses on the following factors leading to shadow application emergence.

- Business unit considers it impossible to converge on a single set of requirements fulfilling all stakeholders' requirements.
- Business unit does not want to rely on slow or expensive third parties.
- Business unit considers it in its best interest to produce a new system they own.

2.3 Preventing Shadow Application Proliferation

Our opinion is that with present software architectures, no matter how carefully official applications are crafted, over time they will spawn shadow applications whenever resourceful communities have urgent unsatisfied needs.

Our hypothesis is that if an application provides the AVI capabilities, the need for shadow applications is greatly reduced. Today's software architectures cannot provide these capabilities because the components of a business application (such as data elements, workflows, or forms) are *shared* among organizations. This sharing is both the main reason why business applications exist and the main reason for the emergence of their shadow counterparts. We therefore propose an application architecture with a fundamentally new and different sharing principle.

3 REQUIREMENTS FOR AN ALTERNATIVE APPLICATION ARCHITECTURE

In this section we attempt to express the AVI capabilities as a set of requirements for an application architecture, with the following definitions.

- *Actor* designates an individual or a group of individuals, for example the entire company, an organization, department, project team, or community.
- *Elements* are runtime application components, like business entities (in our previous example a "*Request*"), workflows, forms, reports, or even configuration entries. With this definition an *Application* is a collection of related Elements.

3.1 Functional Requirements

Our first two requirements cover the most central operations in shadow application development.

R1: *Actors can extend existing Elements*.
R2: *Actors can add new Elements.*

Example 1 below reuses an observation from (Handel and Poltrock, 2010) to illustrate how an application satisfying R1 and R2 could defuse the need for shadow applications.

Example 1 – *Luxury*[2]

The official application manages *Request* entities, with among others attributes *title*, *state* and *delay*.

The "Quality" department needs to record the reasons for delays when they occur. Using R2, they introduce a new Element *DelayAnalysis* with attributes like *reasonForDelay* and *analyst* and associations with existing Element *Request.* Behind the scene, this leverages R1 to extend the Request Element with the reverse association *delayAnalyses*. This blends the new Element and extensions with the original *Luxury* entities thus enabling intuitive bi-directional navigation.

Other operations are adding missing attributes to an existing business element or adding more detailed states in an existing workflow. Example 1 highlights a new problem: the extensions are of interest only to a subset of the application's users, and may be confidential. To avoid cross-Actor pollution and conflicts, both R1 and R2 imply that Actors are *isolated* from each other by default, which yields the requirement R3.

R3: *Actors have private spaces*.

Elements are hosted in such private spaces and are by default not visible outside of them. We call these spaces *Perspectives*. In a typical enterprise setting, today's official applications would be Perspectives providing 'scaffolding' Elements, i.e. skeletons of business entities and associated high-level rules and functionality. Organizations at various levels would have their own Perspectives, hosting the extensions and additional Elements reflecting their concerns and level of detail. Individuals could likewise replace their spreadsheets with private extensions and Elements hosted in a private Perspective. However, completely isolated Perspectives would defeat the purpose of enterprise applications, which yields R4.

R4: *Actors can share the Elements they own.*

[2]In the report, *Luxury* refers to both a business process and the supporting official application(s). We only refer to the latter here.

Perspectives can make selected Elements visible, either to everybody ("public") or to a restricted set of Perspectives. We call this operation *export*. Obviously the previously mentioned official Perspectives would export their Elements to all users. And business-unit-level Perspectives would export their Elements to the relevant Actors. Even individuals can share their Elements with others.

It is interesting to note here that this empowers the entire employee base to contribute to the overall information system, which we think provides significant benefits we will discuss later in this paper. The downside is that this could lead to cacophony through an overwhelming amount of available Elements, dictating R5.

R5: *Actors can select relevant Elements*.

Thus, a symmetrical *import* operation is necessary. An Actor must be able to select among all Elements available to him only the ones he considers relevant. Instead of building his environment from scratch an Actor would *inherit* the Elements from the groups he belongs to, but must be able to *unimport* these if not relevant for him. Example 2 below, again from Handel and Poltrock (2010), illustrates how R1-R5 could have avoided another real-life shadow application.

Example 2 – *Fallen*

"Official application *Fallen* had produced a shadow application which added translations into Japanese next to English data fields."

Extending existing entities with additional attributes is a typical use case of R1. Such extensions would be owned by the Japanese branch of the company, and hosted on their servers in a Perspective (R3) we can call *http://fallen.acme.co.jp/Translations*.

Employees of the Japanese branch would inherit these extensions, and some groups or Japanese employees could even choose to unimport the initial English attributes (R5). The extensions could be exported to other Japanese-spoken employees in other regions (R4).

Our previous use of the term Application encompassed a broad spectrum, from full-blown enterprise systems to private spreadsheets. Likewise, for Perspectives we envision a broad range from big Perspectives hosting self-sufficient third-party applications to tiny individual Perspectives with just a few extensions replacing spreadsheets. Some Perspectives may just factorize the optimal list of import and unimport declarations for a given organization or community.

3.2 Usability Requirements

A significant percentage of today's shadow applications are created by people without software engineering skills using office software like spreadsheets (Nardi and Miller, 1990). This observation makes usability a key requirement.

R6: *No programming is required for R1-R5.*

The last item may sound like reviving the dream of software without programmers. However, the data-centric nature of business applications makes it much less difficult for end-users to participate than more feature-centric software; significant contributions of entities, attributes, simple formulas and associations can be made through a forms-based interface, especially in the presence of example instances (Markl, Altinel, Simmen and Singh, 2008).

Contributions are not limited to what can be done by end-users through forms. A language-based representation of perspectives and elements is still necessary for professional software developers. Even for business units, contractors and interns have always been a means to get access to development skills beyond their internal competence to implement complex shadow applications. In a perspective-centric architecture, such expert contributions would still be possible, with the benefit of being better integrated with the rest of the information system.

4 CONCEPTUALIZATION

Figure 2 below shows the high-level meta-model of our proposal. It is centered on a classical enterprise directory component with *Users* and *Groups*. *Perspectives* are hosted by *Repositories*. Perspectives can define *Fragments*, which can be either self-sufficient (R2) or extensions of a Fragment from another Perspective (R1). Repositories can live on different servers.

At runtime, a User opens a *Session*, which determines a set of Perspectives – owned by the User or inherited from the Groups he is member of. This in turn determines a set of Fragments, which can be woven into *Elements*. The Session becomes the Application, tailored to the connected user's profile. We call this a *virtual private application*, private because it reflects the user's unique combination of elements, virtual because it does not exist outside of the session.

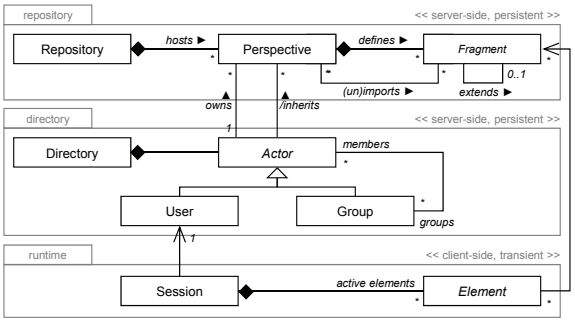


Figure 2: Meta-model of Perspective-centric architecture.

In a perspective-centric architecture, applications are thus dynamically composed at runtime. Today, commercial-off-the-shelf (COTS) applications need to suit the requirements of a variety of customers and provide some degree of flexibility through configuration and customization mechanisms (Brehm, Heinzl, and Markus, 2001). We consider our proposal a *generalization* of these mechanisms found in application platforms like issue-tracking, PLM and ERP systems.

5 PROTOTYPE

We have designed and implemented a first prototype of a perspective-centric system. Considering the complexity of the general case of extensible Elements, our prototype mainly focuses on data, i.e. business entities.

5.1 End-user Experience

Our main objective was to verify that the dynamic, perspective-centric nature of the system could be made transparent to end-users during normal use. The screenshots below show two different users connected to a Luxury-like application, both displaying a request object. The first user belongs to the quality group and thus sees DelayAnalysis objects, the second user is from the planning group and sees SubTask objects.

It is important to stress the *additive* nature of the system, as opposed to subtractive, i.e. filtering. In a filtering approach, somewhere an Element would exist with all attributes, which get filtered out depending on the users' profile. In our approach, Fragments exist in various places and get pulled together by the Session. The main visible difference with a regular system is the presence of edit buttons, which allow inspection and tailoring of the connected user's model as illustrated in the next

screenshot, which shows (1) the possibility to import another Entity "Customer", and (2) that Element "Request" is a composition of Fragments from three different perspectives.

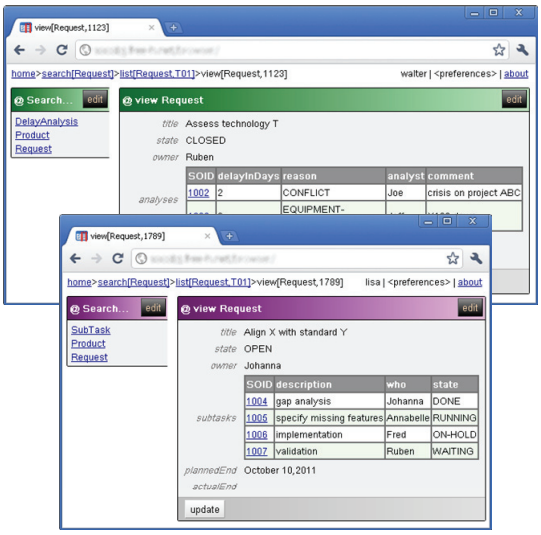


Figure 3: Two different users during normal use.

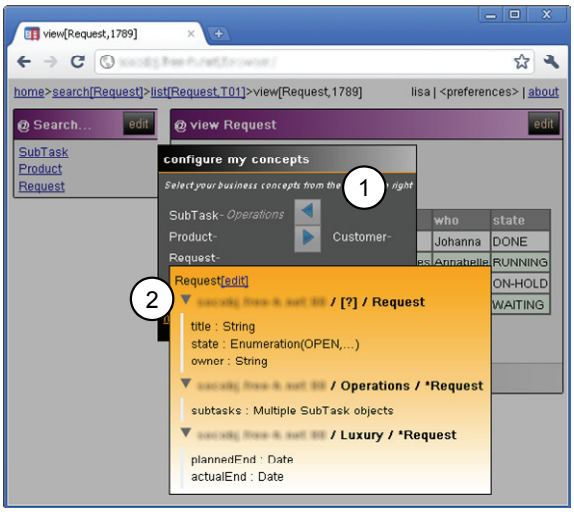


Figure 4: A user inspecting his model.

An ideal interface should have the intuitiveness of a spreadsheet, where filling an empty "header" cell transparently creates an extension with the new attribute, with default type and visibility. We believe the presence of actual records makes such example-centric modeling possible.

5.2 Architecture

It may appear natural to host extensions on the same server as their root Elements. However, to fully meet R1 and R2 any Actor must be able to provide his own storage and computing resources for extensions. Otherwise, although independent in functional terms, he is dependent from a physical resource point of view. This constraint dictates a *distributed* architecture, where Perspectives can be hosted on distinct servers and are pulled together at runtime by a client session.

It is important to guarantee that official systems cannot be disrupted or slowed down by extensions hosted on unreliable servers. No organization would accept an architecture with the potential for any unfortunate experiment by an employee to degrade access to central services. This constraint dictates *asynchronous* communication between components, allowing results from a high-reliability official system to be displayed without waiting for the extension results which may arrive later or never.

The prototype implementation is broken down in the following components.

- A central *directory* component, which in a real setting is the enterprise directory server where users and groups would just need to be annotated with references (URLs) to their associated perspectives.
- *Repository* components, which host Perspectives with entity definitions, extensions and associated instances, persisted in a database and exposed through web services.
- On the client-side, the *client session* component communicates with previous components to build a data model at runtime, and a *dynamic user interface* builds simple forms by inspecting this model.

Figure 5 below illustrates the main interactions between the components, at initialization-time and during regular use.

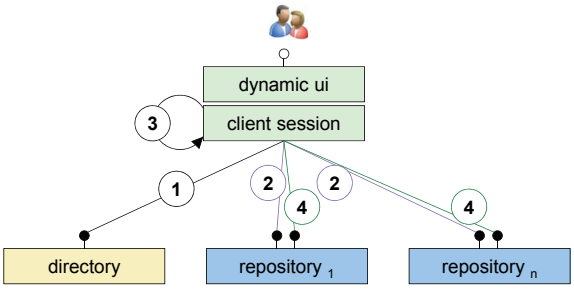


Figure 5: Architecture of the prototype.

In step (1), the client authenticates the user and gets as a reply the full graph of his groups and perspectives. The client then (2) requests all perspectives and the associated Fragment declarations from the various repositories involved. Receiving a Fragment triggers the (3) weaving

mechanism which composes Elements. Usage is then similar to any distributed system, i.e. accessing an object triggers several requests (4).

The communication between components is standard REST over HTTP. The protocol has been kept simple in order to enable integration of legacy systems in a perspective-centric landscape through the development of wrappers.

5.3 Limitation

The main conceptual limitation of our first prototype is the focus on data only. Considering the centrality of data in business applications, we think that the results presented in the next section still represent a significant contribution.

6 INITIAL RESULTS

From a technical point of view, the prototype has demonstrated the feasibility of asynchronous runtime composition of a data model, the transparency for end-users during normal use, and end-user update of the data model in their own perspective.

As first proof-of-concept, we have instantiated the prototype with a project tracking use case and a configuration of 3 groups and 5 individuals with different perspectives. The prototype has been able to compose the individual models on the fly, proving the validity of the concepts of Perspective and Fragment.

We have presented this prototype to 8 information system professionals from 6 different industrial and educational organizations. All of them have over 20 years of experience and have witnessed the emergence of numerous shadow applications. Though they did raise some concerns, covered in the discussion section of this paper, their reactions to the proposal varied from fairly positive to enthusiastic. 4 out of 8 subjects have volunteered for evaluating the prototype with real application data.

As a second proof-of-concept, we have instantiated the prototype with the Luxury-like use-case presented in previous sections of this paper. The “Luxury” perspective and its associated “quality” and “operations” perspectives have allowed a unified representation of the three points of view. We were able to walk through use cases of both the Luxury and Fallen shadow applications, and show that technically they would have been avoided with a mature perspective-centric implementation.

The highly dynamic nature of the proposal initially made all interviewed professionals uncomfortable, illustrating the fairly conservative attitude they adopt regarding the architecture of business applications, particularly the data layer. One manager has expressed a desire to restrict the perspective-centric nature of an application to the initial phases of its life, and to “freeze” the model once it has been collaboratively built and validated. This directly contradicted his earlier statements of continuously evolving and conflicting requirements, which he acknowledged. Experimentation with industry datasets is now required to validate our initial results.

7 DISCUSSION

Perspectives represent different, finer and more connected information system grains than applications. We think they allow an information system to evolve organically in a unified and more controlled way than today’s proliferation of shadow applications, without sacrificing the business units’ ownership of their specific application elements. The reactivity and autonomy their mission demands is thus preserved.

7.1 Towards Social Information Systems

A perspective-centric application architecture represents a major shift of responsibilities from IT departments towards the community of users, not unlike the freedom spreadsheets have provided (Nardi and Miller, 1990). An IT department’s main responsibility would be to provide the platform on which anyone (the IT department itself, but also business organizations and individual employees) could contribute elements in their area of expertise. We think this could leverage the collective intelligence (Surowiecki, 2004) and energy of employees to collaboratively build and maintain the corporate information system, in a form of internal crowd-sourcing.

Considering today’s mostly feudal management of information systems, this is a fairly disruptive proposal. Indeed, during our interviews most subjects have raised the concern that it could result in chaos. This concern typically takes the official applications as a reference, while in our opinion it only represents the tip of the information system iceberg. When including all shadow applications in the picture, information systems today can already not guarantee the overall consistency, and rely upon

humans to keep the whole together. However, as one architect interviewed observed, the chaos is often feared to be in core business attributes. But these are often the best-understood and least controversial of the data elements; uncertainty is greatest on the highly domain-specific attributes. By properly segregating these into the correct perspectives, overall uncertainty may actually be reduced.

Collecting all shadow application data in a unified infrastructure may seem to aggravate inconsistency, but in reality it just reveals the present state. We think a unified infrastructure would provide additional leverage to the previously mentioned human factor in at least two ways.

In the consumer-space, "social" mechanisms like tagging, rating, voting, and targeted sharing have proven effective in organizing huge repositories of consumer-contributed data (Surowiecki, 2008). In a business environment, users could organize application elements through similar mechanisms. We think dealing with authenticated professionals is an even more beneficial setting than the consumer space for social technologies to apply, and envision *social information systems* where elements are contributed from the bottom up, shared with other Actors, ranked and improved through social feedback mechanisms and eventually gradually "promoted" to more central perspectives.

This could result in the *democratic* (or *meritocratic*) evolution of a corporation's application landscape, a generalization of today's frequently requested transfer of shadow applications from business units to IT departments (Handel and Poltrock, 2010).

As opposed to today's situation where shadow applications are mostly disconnected from their parent applications and extremely heterogeneous in their implementation, a unified architecture would make the continuous evolution and divergence *observable*. Indicators could be envisioned (number of extensions, number of unimports…) and dashboards built to monitor application evolution. Pattern-matching techniques could be used to automatically detect convergence opportunities (Ahmadi, Jazayeri, Lelli and Nesic, 2008; Sabatzadeh, Finkelstein and Goedicke, 2010) and notify the owners of the candidate elements, fostering convergence discussions.

7.2 Impact on Collaboration

Although the goal of the proposed architecture is to make evolution a continuous process, introduction of significant chunks still require traditional projects.

From a functional point of view, the painful and hazardous process of elaborating the union of divergent requirements could be replaced by the identification of the intersection, containing only the elements all stakeholders agree on, and then spawn smaller groups to discuss the next level of detail, thus reducing the risk of conflict and communication overhead. We think Perspectives would thus contain the various layers of *boundary objects* (Star, 1990) around which people collaborate.

From a technical point of view, private spaces could help in integrating running development projects with live production environments, facilitating continuous integration and delivery (Fowler, 2010a). Boundaries between mockup, prototype, beta and production environments could be smoothened and concurrent development made easier, as well as quick experimentation encouraged.

7.3 Evaluation in the Real World

One of the challenges of this work is to find suitable ways to evaluate the underlying concepts of social information systems. A standard approach would be to deploy this with a small group of users, and study its usage. However, if it were deployed in this fashion, it would become just one more shadow application, and many of the benefits of a perspective-centric system would be lost. On the other hand, this approach is new and unfamiliar enough to both potential users and IT organizations that a major implementation would be difficult to accomplish. As illustrated by the aforementioned discomfort of the IT professionals, this requires a significant shift in thinking by IT and line-of-business managers about how crucial data is stored and managed. A successful perspective-centric system requires not only technological sophistication, but also a degree of organizational change that is not always present (the "c-words").

7.4 Challenges and Further Work

A real deployment of such a social architecture would almost certainly exhibit a high degree of coupling of its elements, making the system vulnerable to the evolution of central elements. However, since all dependencies are explicit, *evolution policies* could be defined. For example, if a high-level perspective deletes an element, it could be marked as orphan and be proposed to adoption to owners of perspectives which import or extend it.

We think a significant number of common business application features can be implemented in

a generic way in the form of *functional aspects* (Filman, Elrad, Clarke and Akşit, 2008) to be *applied* by an end-user while building his model. For example, if a particular attribute demands traceability this could be a single checkbox on the model's form, a simple boolean annotation on the model itself, and could tell a repository to produce history records with timestamp, user, and previous value. We are working on more complex aspects like lifecycle management and authorization.

The manipulation of model and instances through the same interface presents both the opportunity to leverage contributions from people without modeling skills and the risk to confuse them. Beyond the prototype's naïve forms for model manipulation, we consider usability for contributors with a broad spectrum of software skills a challenge. For contributors with software engineering skills, the development of robust application code on top of a dynamic foundation is not trivial, and needs appropriate programming language bindings.

Other challenges are not new but rather inherited from the present situation. As an example, a user could define an extension concatenating two attributes, and export this extension to colleagues who do not have permission to see the initial data. This is similar to what happens when people extract confidential data in today's shadow applications, breaking the initial authorization mechanism. A perspective-centric system would actually improve on this situation; by having a complete view of all the attributes, a system would be able to detect and warn about possible permission violations.

At a higher level, perspective-centric architectures present a number of interesting challenges, like monitoring and convergence mechanisms, and adapting the consumer-space social recommendation mechanisms to application elements in a business environment.

8 RELATED WORK

We consider the work presented in this paper a novel combination of existing approaches. Shadow applications are a widely known but widely accepted problem. They are frequently mentioned when studying information system agility (Desouza, 2007) or dissatisfaction with business applications (Hoyer and Stanoevska-Slabena, 2008), but not always considered as a problem (Handel and Poltrock, 2010).

Situational applications are enterprise applications built on-the-fly by business units to solve a specific business problem (Markl et al., 2008), and can be considered a superset of shadow applications. Situational applications have attracted recent interest from enterprise mashup researchers (Hoyer and Fischer, 2008) who aim at allowing end users to integrate and combine services, data and other content (Bitzer and Schumann, 2009) to bridge the IT/business gap. Mashups can be interpreted as an evolution of service-oriented architectures (Watt, 2007), which expose business functionality as standard and composable services.

Mashups are part of the broader topic of end-user development (Nestler, 2008); (Sutcliffe, 2005), which advocates the empowerment of end-users to implement their own specific requirements, and has intensively studied spreadsheets (Nardi and Miller, 1990); (Spahn and Wulf, 2009) and more recently collaborative and social aspects in enterprise settings (Ahmadi et al., 2009).

Model-Driven Engineering (Schmidt, 2006) elevates the level of abstraction at which software is developed, turning models into central and productive artifacts, with a specific models@runtime branch focusing on model interpretation. The Software Language Engineering (Kleppe, 2008) and Domain-Specific Languages (Fowler, 2010a) domains, related to MDE by the heavy reliance on meta-models, focus on domain expert involvement in software development and configuration through specific textual representations.

The Component-Based Software Engineering (McIlroy, 1968) community is actively researching robust dynamic systems, where components can appear and disappear during execution. It provides foundation concepts and technologies for making a social application cope with dynamic elements and services of variable reliability.

Linked Data (Bizer, Heath, and Berners-Lee, 2009) integrates distributed, loosely coupled and independently managed repositories of persistent entities, but targets an internet-wide database and mostly-read access.

Social Software Engineering focuses on the understanding of the human and social aspects of software engineering. It covers both the social aspects in the software engineering process and the engineering of social software (Ahmadi et al., 2008). In the Requirements Engineering domain, Lohmann et al. (2009) propose to apply social mechanisms like voting and commenting. Studies on ViewPoints (Sabetzadeh et al., 2010) have focused on capturing divergent concerns but aim at reconciling these at the specification and design level.

The tailoring of enterprise systems, from simple

configuration to the modification of commercial code, is a topic of sufficient complexity for (Brehm et al., 2001) to propose a typology. Recent interest in cloud computing has yielded research in multi-tenancy (Jansen et al., 2010), a way to configure the same software installation for various isolated organizations.

9 CONCLUSIONS

In this paper we have presented an alternative architecture for business applications designed to reduce shadow application proliferation. We have described the main characteristics of shadow applications, the causes of their emergence, and have proposed an architecture principle to defuse this phenomenon based on an isolation mechanism we call *perspectives*. We have presented our prototype, our first results on real-life use cases and the encouraging feedback it has received.

We have discussed our broader vision of a social information system leveraging the collective intelligence of an organization's employees, and the possibility of democratic evolution through the use of social mechanisms.

We have no silver bullet claim, rather a potentially interesting paradigm worth exploring. We have no revolution claim either, merely an original combination of existing approaches and a generalization of business application configuration mechanisms. This is enabled by continuously growing processing power versus fairly stable core requirements of business applications, a better understanding of distributed systems, and recent social technologies.

ACKNOWLEDGEMENTS

This work has been funded by Nano-2012 grant MoDeSI. The authors would like to thank the participating interviewees for their time and helpful comments.

REFERENCES

Ackerman, M. S., 2000. The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility, *Human-Computer Interaction*, Volume 15

Ahmadi, N., Jazayeri, M., Lelli, F. and Nesic, S., 2008. A Survey of Social Software Engineering, *IEEE/ACM ASE - Workshops*

Ahmadi, N., Jazayeri, M., Lelli, F. and Repenning, A., 2009. Towards the Web of Applications: Incorporating End User Programming into the Web 2.0 Communities, *Proc SoSEA 2009*, ACM

Bitzer, S. and Schumann, M., 2009. Mashups : An Approach to Overcoming the Business/IT Gap in Service-Oriented Architectures, *Value Creation in e-Business Management*, ISBN 978-3-642-03131-1

Bizer, C., Heath, T. and Berners-Lee, T., 2009. Linked Data - The Story So Far, *International Journal on Semantic Web and Information Systems*

Brehm, L., Heinzl, A. and Markus, M. L., 2001. Tailoring ERP Systems: A Spectrum of Choices and their Implications, *Proc HICSS*, IEEE

Desouza, K. C., ed., 2007. *Agile Information Systems : Conceptualization, Construction and Management*, ISBN 978-0-7506-8235-0

Filman, R. E., Elrad, T., Clarke, S. and Akşit, M., 2008. *Aspect-Oriented Software Development*, ISBN 0-321-21976-7, Addison-Wesley, 2008

Fowler, M., 2010. *Domain-Specific Languages*, ISBN 978-0-321-71294-3, Addison-Wesley

Fowler, M. *Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation*, ISBN 978-0-321-60191-9, Addison Wesley, 2010

Gotts, I., 2010. A New Cloud: The Stealth Cloud?, *http://www.cio.com/article/630164*, October 2010

Handel, M. and Poltrock, S., 2010. Working Around Official Applications : Experiences from a Large Engineering Project, *Proc CSCW,* ACM Press

Hoyer, V. and Fischer, M., 2008. Market Overview of Enterprise Mashup Tools, *Proc ICSOC*, Springer Verlag

Hordijk, W. and Wieringa, R., 2010 Rationality of Cross-System Data Duplication: A Case Study, *Proc CAiSE*, Springer Verlag

Hoyer, V. and Stanoevska-Slabena, K., 2008. The Changing Role of IT Departments in Enterprise Mashup Environments, *Proc SOC*, Springer-Verlag

Jansen, S., Houben, G-J. and Brinkkemper, S., 2010. Customization Realization in Multi-tenant Web Applications: Case Studies from the Library Sector, *Proc ICWE*, Springer-Verlag

Kleppe, A., 2008. *Software Language Engineering: Creating Domain-Specific Languages Using Metamodels*, ISBN 9780321553454, Addison-Wesley

Kling, R., 1991. Cooperation, Coordination and Control in Computer-Supported Work, *Communications of the ACM*, Volume 34 Issue 12

Lohmann, S., Dietzold, S., Heim, P., Heino, N., 2009. A Web Platform for Social Requirements Engineering, *Software Engineering Workshops 2009*

Markl, V., Altinel, M., Simmen, D. and Singh, A., 2008. Data Mashups for Situational Applications, *Proc MBSDI 2008*, Springer-Verlag

McIlroy, D., 1968. Mass-Produced Software Components, *Software Engineering, Report on a conference*

sponsored by the NATO Science Committee

Nardi, B. A. and Miller, J. R., 1990. An Ethnographic Study of Distributed Problem Solving in Spreadsheet Development, *Proc CSCW 1990*, ACM Press

Nestler, T., 2008. Towards a Mashup-driven End-User Programming of SOA-based Applications, *Proc iiWAS 2008*, ACM Press

Newell, S., Wagner, E. L. and David, G., 2007. Clumsy Information Systems: A Critical Review of Enterprise Systems, *Agile Information Systems*, Elsevier

Sabetzadeh, M., Finkelstein, A. and Goedicke, M., 2010. ViewPoints, *Encyclopedia of Software Engineering*, P. Laplante, Ed.

Schmidt, D. C., 2006. Model-Driven Engineering, *IEEE Computer Vol 39*

Spahn, M. and Wulf, V., 2009. End-User Development for Individualized Information Management: Analysis of Problem Domains and Solution Approaches, *Proc ICEIS 2009*, Springer Verlag

Star, S. L., 1990. The Structure of Ill-Structured Solutions : Boundary Objects and Heterogeneous Problem Solving, *Distributed artificial intelligence*, Vol. 2, Morgan Kaufmann

Surowiecki, J., 2004. The Wisdom of Crowds, ISBN 978-0385503860

Sutcliffe, A., 2005. Evaluating the costs and benefits of end-user development, *ACM SIGSOFT Software Engineering Notes*, Vol 30

Watt, S., 2007. Mashups - The evolution of the SOA: Situational applications and the Mashup ecosystem, *http://ibm.com/developerworks/webservices/library/,* Nov 2007

Zarnekow, R, Brenner, W. and Pilgram, U., 2006. *Integrated Information Management: Applying Successful Industrial Concepts in IT*, ISBN 978-3540323068

An Efficient Sampling Scheme for Approximate Processing of Decision Support Queries

Amit Rudra[1], Raj P. Gopalan[2] and N. R. Achuthan[3]
[1]*School of Information Systems, Curtin University, Kent Street, Bentley, WA 6155, Australia*
[2]*Department of Computing, Curtin University, Kent Street, Bentley, WA 6155, Australia*
[3]*Department of Mathematics and Statistics, Curtin University, Kent Street, Bentley, WA 6155, Australia*
{A.Rudra, R.Gopalan, N.R.Achuthan}@curtin.edu.au

Keywords: Sampling, Approximate Query Processing, Data Warehousing.

Abstract: Decision support queries usually involve accessing enormous amount of data requiring significant retrieval time. Faster retrieval of query results can often save precious time for the decision maker. Pre-computation of materialised views and sampling are two ways of achieving significant speed up. However, drawing random samples for queries on range restricted attributes has two problems: small random samples may miss relevant records and drawing larger samples from disk can be inefficient due to the large number of disk accesses required. In this paper, we propose an efficient indexing scheme for quickly drawing relevant samples for data warehouse queries as well as propose the concepts of database and sample relevancy ratios. We describe a method for estimating query results for range restricted queries using this index and experimentally evaluate the scheme using a relatively large real dataset. Further, we compute the confidence intervals for the estimates to investigate whether the results can be guaranteed to be within the desired level of confidence. Our experiments on data from a retail data warehouse show promising results. We also report the levels of accuracy achieved for various types of aggregate queries and relate them to the database relevancy ratios of the queries.

1 INTRODUCTION

Analytical queries containing aggregate functions such as sum and average on a data warehouse are used to gain a good sense of the business situation and to support business decisions. Most often we require timely retrieval of query results with an acceptable level of accuracy rather than absolute precision, and so approximate results within certain limits of accuracy will be acceptable to the user. Pre-computation with materialized views and sampling are two ways to handle such queries. However, it is impractical to maintain a large number of materialized views for all possible combinations of information retrieval (Hellerstein et al., 1997). In contrast, sampling can provide faster results that are accurate within given assured confidence levels.

The main motivation for use of sampling in processing queries on a large database or a data warehouse is to save time and resources. Even though random sampling is both efficient and effective as an approximation method, its use for database querying has attracted significant research interest only recently (Li et al, 2008); (Joshi and Jermaine, 2008); (Jin et al., 2006). Sampling has also been shown to be effective for aggregate queries (Hellerstein et al., 1997); (Jermaine, 2007; (Jin et al., 2006); (Jermaine, 2003); (Bernadino et al., 2002); (Speigel and Polyzotis, 2009; Jermaine et al., 2004). As sampling data may not be fully representative of the entire data in a data warehouse, it is desirable to return both the query result and the confidence intervals that indicate the reliability of the results (Li et al, 2008).

A significant problem with random sampling for database queries from a database on stored disk is that picking records at random requires almost the same amount of I/O as processing the query over the whole database (Olken and Rotem, 1990). To keep down the cost of sampling based query processing, a more efficient method of drawing samples is needed. Another problem is that a random sample drawn from a very large dataset may not contain relevant records that satisfy the range restrictions of a given query. To deal with this problem, we require a sampling scheme that will include in the sample

records satisfying the query predicates.

Joshi and Jermaine (2008) introduced the ACE Tree which is a binary tree index structure for efficiently drawing samples for processing database queries. They demonstrated the effectiveness of this structure for single and two attribute database queries, but did not deal with multi-attribute aggregate queries. For extending the ACE Tree to *k* key attributes, Joshi and Jermaine proposed binary splitting of one attribute range after another at consecutive levels of the binary tree starting from the root; from level *k*+1, the process is repeated with each attribute in the same sequence as before. This process could lead to an index tree of very large height for a data warehouse even if only a relatively small number of attributes are considered.

Li et al. (2008) proposed a sampling cube framework for answering analytical queries on a data warehouse which calculates confidence intervals for any multidimensional query. The sampling cube is constructed from a random sample of the data warehouse. After building the sampling cube, there is no further access to the original data records should a query require a different sample from the one already drawn. If a query has too few sample records in the sampling cube, they expand the query to gather more sample records from the sampling cube itself in an attempt to improve the quality of the query result.

In this paper, we propose the *k*-MDI Tree which extends the ACE Tree structure to deal with multi-dimensional data warehouse queries. Unlike the ACE Tree, the *k*-MDI tree allows non-binary splits of data ranges for key values that do not split evenly into 2^n distinct ranges. The number of levels in the *k*-MDI tree can be limited to the number of key attributes. The shallow tree structure resulting from multi-way branching also facilitates quicker retrieval of leaf nodes from disk storage. Unlike the sampling cube of Li et al. (2008), new samples that contain relevant records are drawn for each query. These records can be considered as drawn from a subset of the data warehouse that satisfies the query predicates. In estimating the query results, we take into account the proportion of relevant records for the query in the whole data warehouse. The sampling and estimation methods are evaluated experimentally using a real life data set.

The rest of the paper is organized as follows: In Section 2, we define some relevant terms and briefly describe the ACE Tree structure. In Section 3, our k-way multi-dimensional (*k*-MDI) indexing structure is described in detail. We also introduce the concept of relevancy ratios, both for the database and a specific sample. Section 4 reports the experimental results that evaluate the efficacy of our scheme. Section 5 is the conclusion of the paper.

2 TERMS, DEFINITIONS AND ACE TREE STRUCTURE

In this section, we define some terms pertaining to data warehousing, define confidence interval and then review briefly the ACE Tree structure (Joshi and Jermaine, 2008) that has preceded the *k*-MDI tree we propose in Section 3.

2.1 Dimensions and Measure

To support decision support queries, data is usually structured in large databases called data warehouses. Typically, data warehouses are relational databases with a large table in the middle called the fact table connected to other tables called dimensions. For example, consider the fact table Sales shown as Table 1. A *dimension* table Store linked to StoreNo in this fact table will contain more information on each of the stores such as store name, location, state, and country (Kimball and Moss, 2002). Other dimension tables could exist for items and date. The remaining attributes like quantity and amount are typically, but not necessarily, numerical and are termed *measures*. A typical decision support query aggregates a measure using functions such as Sum(), Avg() or Count(). The fact table Sales along with all its dimension tables form a *star schema*.

Table 1: Fact table Sales.

SALES				
Store No	**Date**	**Item**	**Quantity**	**Amount**
21	12-Jan-11	iPad	223	123,455
21	12-Jan-11	PC	20	24,800
24	11-Jan-11	iMac	11	9,990
77	25-Jan-11	PC	10	12,600

In decision support queries a measure is of interest for calculation of averages, totals and counts. For example, a sales manager may like to know the total sales quantity and amount for certain item(s) in a certain period of time for a particular store or even all (or some) stores in a region. This may then allow her to make decisions to order more or less stocks as appropriate at a point in time.

2.2 Confidence Interval

When estimating with samples we indicate the reliability of the estimate by its confidence interval. Consider a sample of records x with the mean of the sample denoted by $\bar{x}$, and the size of the sample n. For a desired confidence level (e.g. 95%) the confidence interval estimator of the population mean μ is given by:

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \ , \quad \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the critical t-value and s the standard deviation of the sample (Keller, 2009).

2.3 ACE Tree Structure

The ACE Tree is a balanced binary tree where the leaf nodes contain the randomized samples of key values and the internal nodes above them are the index nodes. Each internal node contains a range R of key values, a key value k that splits R into left and right sub-trees, pointers to the left and right branch (child) nodes, and counts of database records falling in the left and right sub-trees. Figure 1 shows the structure of an example ACE Tree. The root node $I_{1,1}$ with its range $I_{1,1}.R$ labeled as [0-64] signifies the key value range of the whole data set. The key of the root node partitions the range $I_{1,1}.R$ into $I_{2,1}.R$ = [0-32] and $I_{2,2}.R$ = [33-64]. This partitioning of ranges is propagated down the tree among the descendants of respective nodes. The ranges associated with a section of a leaf node are determined by the ranges associated with each internal node on the path from the root node to the leaf. If we look at the path from $I_{1,1}$ i.e. the root node down to the leaf node L_4, we come across the following ranges 0-64, 0-32 and 17-24. A leaf node is partitioned into sections (S_1, S_2, …), their number depending on the number of dimensions indexed. Thus, the first section $L_4.S_1$ has a random sample of records in the range 0-64; $L_4.S_2$ has them in the range 0-32; $L_4.S_3$ in the range 17-32 and $L_4.S_4$ in the range 25-32. The size of each leaf is chosen as the number of records that can be stored in a disk block and so the number of leaf nodes depends on the size of the database which also determines the height of the index tree itself.

2.4 Sampling for a Query using the ACE Tree

Referring to Figure 1, consider a query Q with a range of [28-38]. The query execution algorithm proceeds by traversing down $I_{1,1}$, the root node. Both $I_{2,1}.R$ and $I_{2,2}.R$ overlaps with Q.

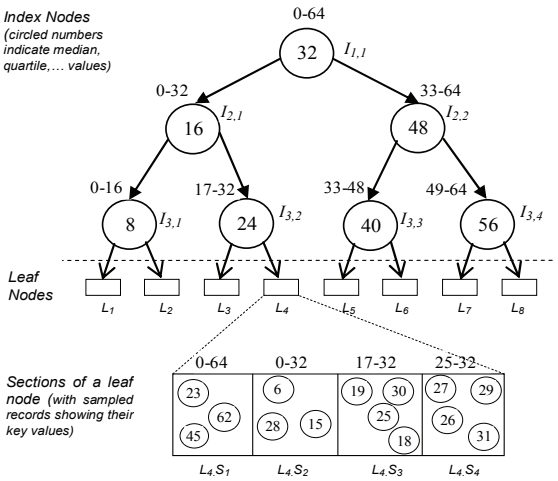


Figure 1: Structure of the ACE Tree.

A level down from $I_{2,1}$, only $I_{3,2}.R$, overlaps with Q. Traversing down to the leaf nodes, the algorithm finds the right leaf node's range [25-32] overlaps with Q and so retrieves records from L_4. The relevant records in the query's range are returned for the sample which includes record 28 from $L_4.S_2$, record 30 from $L_4.S_3$ and records 29 and 31 from $L_4.S_4$. Next, the algorithm traverses down the right node $I_{2,2}$ below the root to the leaf node L_5 and retrieves all relevant records from all sections of L_5 to the pool of sample records.

2.5 Extended ACE Tree for Multiple Dimensions

Joshi and Jermaine (2008) proposed extending the ACE Tree from a single dimension to multiple dimensions as follows: Given key attributes (dimensions), a_1, ..., a_k, split the range of values for a_1 into two sub trees of approximately equal number of keys below the root (level 1); for each node at level 2, similarly perform a binary split of the range of key values for a_2 and so on up to level k for attribute a_k. Then at level k+1, split the attribute values of a_1 again followed by a_2, etc. at further lower levels.

In real life data, a dimension's values may not split evenly into 2^n distinct ranges. For example, if a dimension has an odd number of key values, say - k_1, k_2 and k_3, with cardinalities of 30000 each; then, we cannot split them evenly into 2 distinct ranges but we can do so into 3. The height of the tree will be very large even for a moderate sized data warehouse with a relatively small number of dimensions.

3 MULTIDIMENSIONAL INDEXING

We propose the *k*-MDI tree which extends the ACE Tree index for multiple dimensions while overcoming the limitations of the ACE Tree discussed in Section 2.5. The height of the *k*-MDI tree is limited to the number of key attributes. As a multi-way tree index, it is relatively shallow even for a large number of key value ranges and so requires only a small number of disk accesses to traverse from the root to the leaf nodes.

3.1 *k*-ary Multidimensional Index (*k*-MDI)

The *k*-ary multi-dimensional index tree (*k*-MDI tree) is a *k*-ary balanced tree as described below:

1. The root node of a *k*-MDI tree corresponds to the first attribute (dimension) in the index.
2. The root points to k_1 ($k_1 \leq k$) index nodes at level 2, with each node corresponding to one of the k_1 splits of the ranges for attribute a_1.
3. Each of the nodes at level 2, in turn, points to up to k_2 ($k_2 \leq k$) index nodes at level 3 corresponding to k_2 splits of the ranges of values of attribute a_2; similarly for nodes at levels 3 to h, corresponding to attributes a_3,..., a_h.
4. At level h, each of up to k^{h-1} nodes points to up to k_h ($k_h \leq k$) leaf nodes that store data records.
5. Each leaf node has h+1 sections; for sections 1 to h, each section i contains random subset of records in the key range of the node i in the path from the root to the level h above the leaf; section h+1 contains a random subset of records with keys in the specific range for the given leaf.

Thus, the dataset is divided into a maximum of k^h leaf nodes with each leaf node, in turn, consisting of h+1 sections and each section containing a random subset of records. The total number of leaf nodes depends on the total number of records in the dataset and the size of a leaf node (which may be chosen as equal to the disk block size or another suitable size). More details on leaf nodes and sections are given in Section 3.3. In real data sets, the number of range splits at different nodes of a given level i need not be the same. For convenience, the number of splits at all levels are kept as k in Figure 2 that shows the structure of the general scheme for *k*-MDI multilevel index tree of attributes A_1, A_2, …, A_h with k ranges (R_{11}, R_{12}, ..., R_{1k}), (R_{21}, R_{22}, ..., R_{2k}), … (R_{h1}, R_{h2}, ..., R_{hk}) respectively at levels (1, …,h).

An example of the *k*-MDI tree is shown in Figure 3 from a store chain dataset with three dimensions – store, date sold and item number. The number of range splits and hence branches from non-leaf nodes vary between 2 and 4 in this example.

3.2 Leaf Nodes

Similar to the ACE tree structure, the lowest level nodes of a *k*-MDI tree point to leaf nodes containing data records. The data records are stored in h+1 sections, where h is the height of the tree. Section S_1 of every leaf node is drawn from the entire database with no range restriction on the attribute values. Each section S_i ($2 \leq i \leq h$+1) in a leaf node L is restricted on the range of key values by the same restrictions that apply to the corresponding sub-path along the path from the root to L. Thus for section S_2, the restrictions are the same as on the branch to the node at level 2 along the path from the root to L and so on.

Figure 3 shows an example leaf node projected from the sample *k*-MDI tree. The sections are indicated above the node with attribute ranges for each section below the node. The circled numbers in each section indicate record numbers that are randomly placed in the section. The range restrictions on the records are indicated below each section, where the first section S_1 has records drawn from the entire range of the database. Thus, it can contain records uniformly sampled from the whole dataset. The next section S_2 has restriction on the first dimension viz. store (for leaf node L_7 this range is store numbers 1-16). The third section S_3 has restrictions on both first and second dimensions viz. store and date. While the last section S_4 has restrictions on all the three dimensions – store, date and item.

The scheme for selection of records into various leaf nodes and sections is explained in detail in the following section.

3.3 Building the *k*-MDI Tree

The purpose of the *k*-MDI tree is to quickly retrieve relevant random samples of records for processing data warehouse queries. The records in the sample are obtained from leaf nodes by traversing the index from the root. The *k*-MDI Tree is built in the following three steps:

1. First, the dataset records are sorted by the first key attribute a_1 as the major field, followed by the second attribute a_2 and so on until the last attribute a_h.

2. The next step is to find the split points of key attribute values in the index tree at the levels 1 to h so that the number of records of the dataset that fall under each sub-tree rooted at levels 2 to h is approximately equal. The k_1-1 split points at level 1 are chosen such that the total number of records in the dataset are split into k_1 approximately equal parts; the records falling under each of the nodes at level 2 are split into k_2 approximately equal parts, and so on until the records falling under each of the nodes at level h split into k_h approximately equal parts. The number of splits at all the levels in the index should be such that the number of leaf nodes are equal to a pre-computed number based on the total number of records in the dataset and the size of each leaf node (which could be chosen as the disk block size as in the case of the ACE Tree or some other suitable size).

3. Next, a random number between 1 and h+1 is assigned to each data record as its section number. Depending on the section number and its composite key value, the record is assigned to a leaf node as follows: If the section number is 1, the record is assigned randomly to any one of the leaf nodes in the tree; if the section number is i ($2 \le i \le h$), starting from the root of the index tree, we locate the root of a sub-tree at level i in which the key of the record falls and assign the record randomly to section i of any of the leaf nodes in that sub-tree;

if the record's section number is h+1, it is assigned to the specific leaf node where the record's key value belongs. When all the records have been thus assigned section and leaf node numbers, the dataset is re-organised with records sorted according to their leaf node and section numbers.

3.4 Using the *k*-MDI Tree for Data Warehouse Queries

By using a k-MDI tree index, we can draw stratified samples for data warehousing queries from restricted ranges of key values. In this section, we first introduce two measures that are useful for the estimation of query results using such samples. The *database relevancy ratio* (DRR) of a query Q, denoted by ρ(Q) is the ratio of the number of records in a dataset D that satisfies the query conditions to the total number of records in D. For a query with no condition, ρ(Q) is 1. Similarly, the *sample relevancy ratio* (SRR) of a query Q for a sample set S, denoted by ρ(Q, S) is defined as the ratio of the number of records in S that satisfy a given query Q to the total number of records in S.

In a true random sample of records, the SRR for a query Q is expected to be equal to its DRR, i.e., E(ρ(Q, S)) = ρ(Q). A sample with ρ(Q, S) > ρ(Q) is likely to give a better estimate of the mean than a true random sample. However, for the sum of a column, the sample needs to be representative of the population, i.e., ρ(Q, S) should be close to ρ(Q).

Consider the following formula for estimating the sum (Berenson and Levine, 1992): $\hat{T} = N\hat{p}\bar{X}$, where N is the cardinality of the population, $\hat{p}$ the estimated proportion of records satisfying the query conditions and $\bar{X}$ the mean of records in the sample satisfying the query condition. In order to estimate the mean we can use all relevant sampled records from all sections of the retrieved leaf nodes, but to estimate the sum we can use sampled records

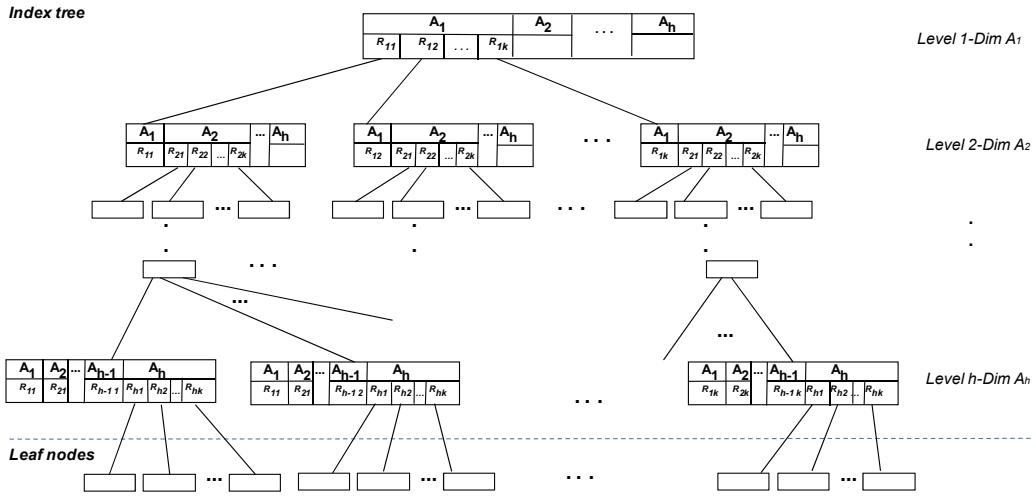


Figure 2: General structure of the k-MDI tree – A_1, A_2, …, A_h are h attributes and R_{ij} the i-th attribute's j-th range high water mark (HWM).

only from section S_1, which is the only section with records drawn randomly from the entire dataset. For estimating the sum for a query with conditions on some of the indexed dimensions we use appropriate sections of the retrieved leaf nodes to get a better estimate of the mean; the records from section S_1 are also used to get a fair estimation of the proportion records that satisfy the query conditions.

3.5 Effect of Sectioning on Relevancy Ratio

As discussed earlier, sections S_1 to S_{h+1} of each leaf node contain random collections of records with the difference that S_1 contains records from the entire dataset while other sections contain random records from restricted ranges of the key attributes. Consider a query with the same range restrictions on all three dimensions (store, date and item) as section $L_7.S_4$ in Figure 3. We are then likely to get more relevant records in the sample from the second section $L_7.S_2$ than from S_1 since records of S_2 have restrictions on the first dimension of store that matches the query condition. Records in S_3 will have restrictions on both store and date dimensions that match that of the query and so are likely to contain more relevant records than in S_2. All records in section $L_7.S_4$ will satisfy the query since the range restrictions on S_4 exactly match the query. Mathematically, for a query Q having restrictions as mentioned above:

$$\rho(Q) = E(\rho(Q, L_7.S_1)) \leq E(\rho(Q, L_7.S_2))$$
$$\leq E(\rho(Q, L_7.S_3)) \leq E(\rho(Q, L_7.S_4))$$

Using this property of the *k*-MDI tree, it is possible to quickly increase the size of a sample that is too small, by including more records from other sections of the retrieved leaf nodes.

3.6 Record Retrieval to Process a Query

The objective of using the *k*-MDI tree is to retrieve a significant number of relevant records (i.e. records that satisfy the query conditions) in the sample drawn for processing a given query. The query conditions may span sections of one or more leaf nodes which can be reached from index nodes that straddle more than one range of attribute values. These leaf nodes can be accessed by traversing the tree from the root using the attribute value ranges in the query conditions and sections from multiple leaf nodes can be combined to form the sample.

We describe the retrieval process using an example query on the sample database of Figure 4. Consider a query Q_0 about sales in store 12 for date range 1-13 and item range 12M-20M. The retrieval algorithm finds the sections of leaf nodes for this query as follows:

1. Search index level 1 to locate the relevant store range. Store 12 is in the left most range of 1-16.
2. Traverse down to index level 2 (date), indicated by a dashed arrow in Figure 4, along the first store range. Since there is a condition on date (1-13), compare the HWMs (high water marks) of the three ranges and find that it fits into two date ranges viz. the first and the second. Make a note of these date ranges.
3. Traverse down using the first date range to the next index level which has item ranges. Since there is a condition on item numbers (12M-20M), compare this range with HWMs and find that it fits into two ranges viz. the third and the fourth. Make a note of these item ranges.
4. Traverse down using the third item range to relevant leaf pages and make a note of them.
5. Iterate step 4, except this time using the fourth item range.
6. Next, repeat the above three steps i.e. steps 3 through 5; but this time using the second date range instead.
7. Now retrieve records from the relevant sections in the four leaf nodes (viz. L_3, L_4, L_7 and L_8) to form a sample for the given query.

3.7 Estimating Query Results from Samples

In decision support queries on large databases, the most common estimation performed is either of the mean or the sum of a column measure (Jin et al., 2006). We maintain a table representing a histogram of record counts for each leaf node and its sections. It is used to estimate the number the leaf nodes required to have adequate number of samples. The following steps outline our method of estimating the mean, sum, standard deviation and the confidence intervals:

1. Draw a sample set L of leaf nodes as described in Section 3.6 for the given sampling rate.
2. The following parameters are computed:
   a. Sample size – n
   b. Count of sampled records satisfying the query condition – m
   c. Count of records in all sampled S_1 sections of L – n'
   d. Count of records in all sampled $_{S1}$ sections of

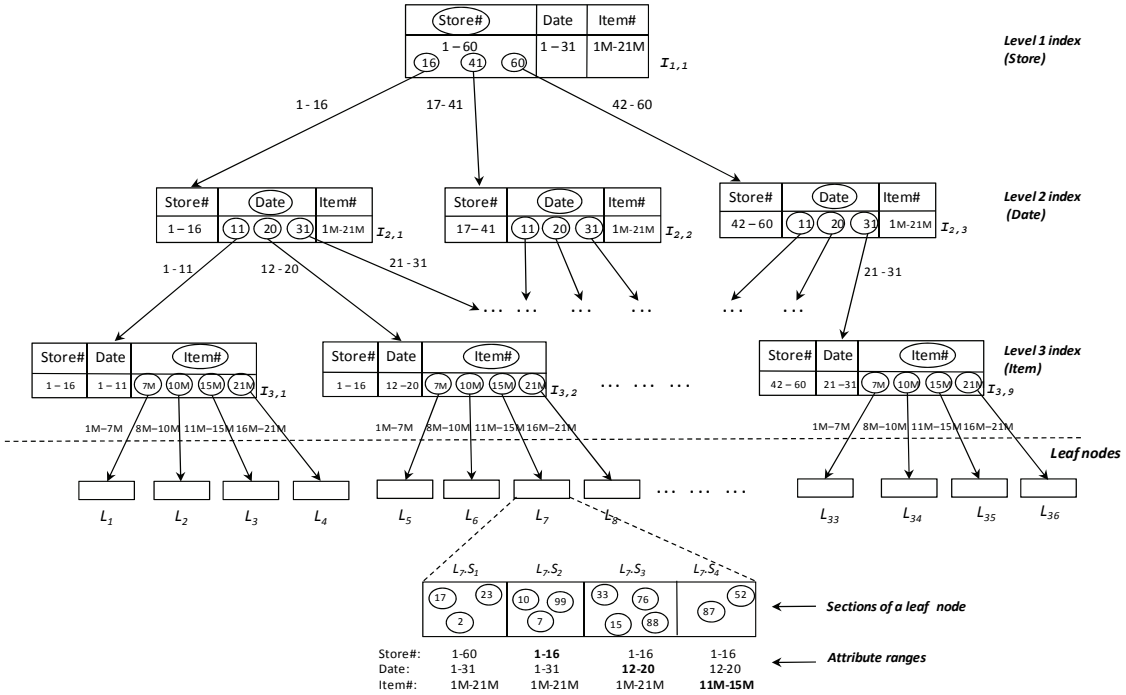


Figure 3: A leaf node (changes in range values for attributes are indicated in bold).

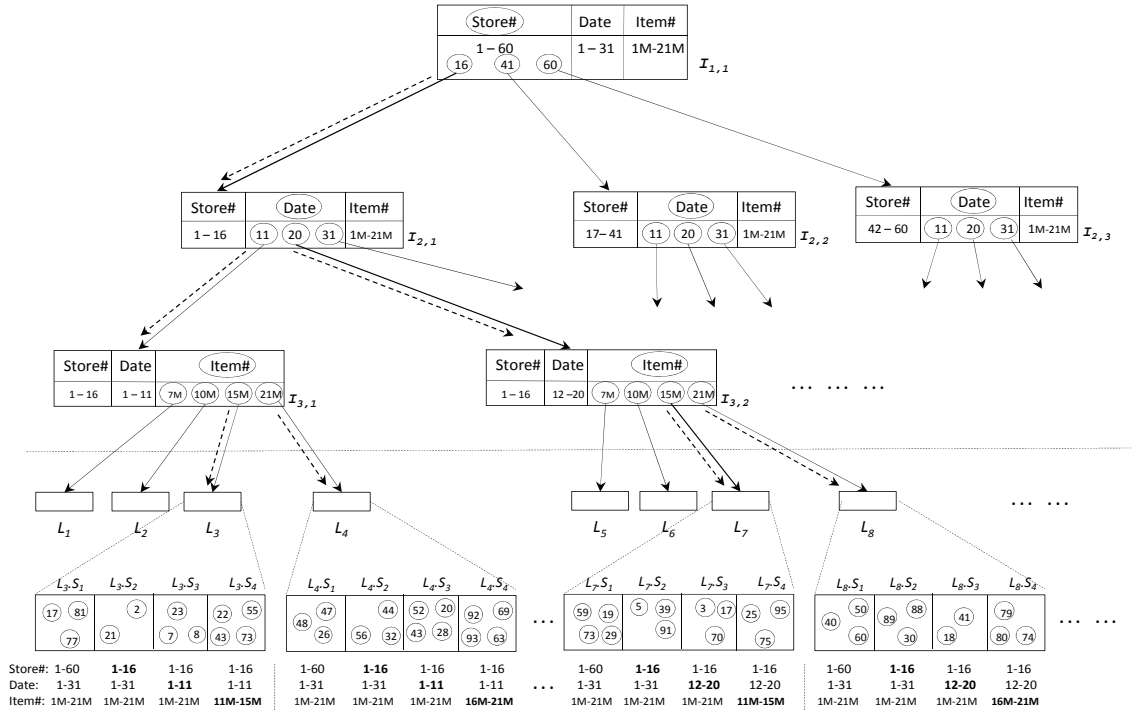


Figure 4: Navigation down index tree nodes for conditions on three dimensions.

L satisfying the query condition – m'

e.Sum of attribute (variable) value of all m records – *sum*

f. Sum of squares of attribute value of the all abovementioned m' records – $sumSq'$

g. The average sum of squares - $Z = \frac{sumSq'}{m'}$

3. Estimating the sum, average, variance and C.I. limits (Chaudhuri and Mukherjee, 1985):

a.Estimate of the number of records M that satisfy the query condition in the population (given the cardinality of the dataset *N*)

$$\widehat{M} = N\frac{m'-1}{n'-1}$$

b. Estimate of Average $\bar{x} = \frac{sum}{m}$

c.Estimate of Sum $\hat{T} = \hat{M}\bar{x}$

d. Estimate of variance of Average v($\bar{x}$)

$$= \frac{\hat{M}-m}{\hat{M}(m-1)}(Z-(\bar{x}^2))$$

e.Confidence interval lower limit

$$= \frac{m'-1}{n'-1} - Z\sqrt{\frac{(m'-1)}{n(n'-1)}\frac{(n'-m')}{(n'-1)}}$$

f. Confidence interval upper limit

$$= \frac{m'-1}{n'-1} + Z\sqrt{\frac{(m'-1)}{n(n'-1)}\frac{(n'-m')}{(n'-1)}}$$

4 EXPERIMENTAL RESULTS

To evaluate the effectiveness of our sampling technique based on the *k*-MDI tree, we performed experiments on real life supermarket retail sales data (TUN, 2011) for a month from 150 outlets. The data warehouse is structured as a star schema shown in Figure 5, with the fact table (itemscan) consisting of over 21 million rows and three dimension tables viz. storeInfo, itemDesc and storeMemberVisits. TPC-H queries (TPC Benchmarks, 2011) with suitable modifications for this sample data warehouse were used in the experiments. Most of the TPC-H queries involve SQL aggregate functions of sum(), avg() or count(). A few include min() and max() which are not easily calculated by sampling (Joshi, 2008; Hellerstein et al., 1997). So, we investigated only sum, avg and count in our experiments. The index tree structure was simulated using the Oracle DBMS. We clustered the fact table on leaf and section number to maintain the records in that sequence. This organisation supported the simulation of both the storage and retrieval of records for the experiments.

A set of three queries were used containing the SQL functions – avg(), sum(), count() with varying database relevancy ratios (DRR). The queries were of the form:

```
Select  Avg(totscanAmt), Sum(totscanAmt),
          Count(*)
  From  itemscan, storeinfo, itemdesc
  Where  storeno between s1 and s2
   And itemscan.storeno=storeinfo.storeno
   And itemscan.itemno=itemdesc.itemno
     And datesold between d1 and d2
     And itemno between i1 and i2;
```

The DDR value was set high or low for the queries by choosing a given proportion of the dimension range for the query. For example, assuming a uniform distribution of values for a dimension in the database, we can get a DRR of approximately 0.33 on a single dimension query, by picking a third of the dimension range. However, in practice we empirically varied the dimension ranges in the queries to get the desired DRR values.

The first test query had a condition on a single dimension and a high DRR value of 0.37; the second query had a lower DRR (0.05) with conditions on two dimensions; and the third query had a very low DRR (0.002) with conditions on all three dimensions. The relevance of DRR in estimating the query results may be seen from the query result estimation process of Section 3.7. In step 3a, the count $\widehat{M}$ directly depends upon the DRR, which is the statistical proportion p whose estimate is given by $\frac{m'-1}{n'-1}$. In step 3c, the estimation of sum depends on the count $\widehat{M}$ and thereby on p.

We conducted the experiments using several random samples at sampling rates of (1% - 12%) and the results were averaged for each sampling rate. The error for the three aggregate functions viz. avg, sum and count were computed as the absolute value of the difference between estimated and actual values for the whole database. Figure 6 shows the results with error rates for both average (mean) of totscanamt column, sum of totscanamt and count for the different database relevancy ratios mentioned above.

There are two graphs for each level of DRR. Figure 6a shows the error rates for the average, the sum of scan amount and count for high DRR. Figure 6b shows the confidence intervals (lower and upper limits) for the average amount for high DRR. Figure 6b also shows the estimated and actual values of the average scan amount. Figures 6c and 6d show similar information as above for the low value of DRR; Figures 6e and 6f show similar information for very low DRR.

It is seen that for the high DRR query with $\rho(Q_1)$ = 0.37, the error rates for the count, average and the sum of the total scan amount stabilize as the sampling rate is increased. The estimates are close

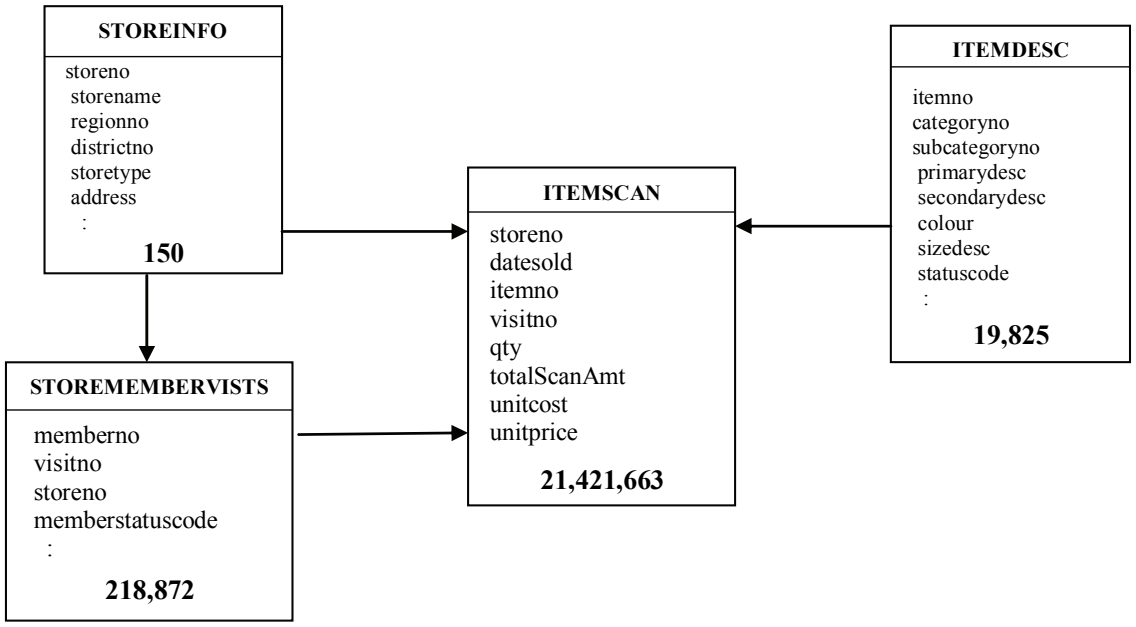


Figure 5: The schema for experimental retail sales data warehouse.

to the actual values for the lowest to the highest sampling rates used and the true value of average is always within the estimated confidence interval. For the medium DRR query with $\rho(Q_1) = 0.05$, we still get error rates below the normally acceptable rate of 5%. For query with very low DRR of $\rho(Q_1) = 0.002$, the error rates for the average scan amount, for all but 1% sampling rate, are below 5% and the true values within the C.I. limits. But the error rates for the estimated sum of scan amount and count of records are not below the acceptable limit at any sampling rate used for the very low DRR query. Thus, we cannot satisfactorily estimate the sum and the count for low values of DRR, while for medium to high values of DRR the estimations of both the sum and average are within acceptable error limits. Also, it's observed from the graphs that there is an apparent close correlation between the estimates of sum and count.

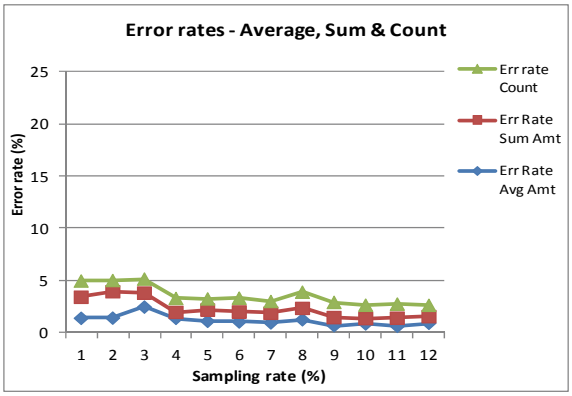
Time Improvement – Figure 7 shows the average time for processing the queries at various sampling rates and also the average time for processing these queries on the full database. It is seen that there is a significant time improvement from using the sampling scheme.

5 CONCLUSIONS

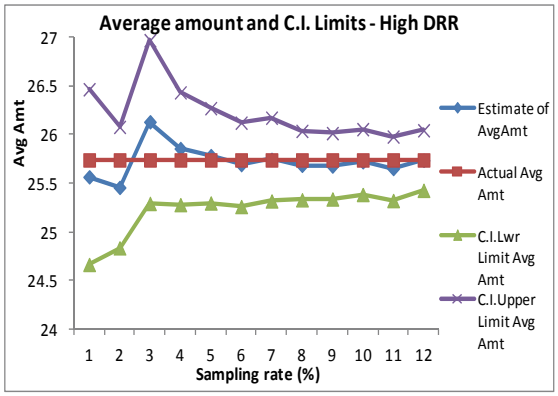
In this paper, we proposed the *k*-MDI tree indexwhich can be used to draw samples quickly for answering multi-dimensional aggregate queries from a data warehouse. The *k*-MDI tree extends the ACE binary tree as a multi-way tree index. The maximum number of levels of the *k*-MDI index is limited to the number of key attributes and so makes the access to the leaf nodes much quicker compared to a binary tree index on external storage.

We also proposed the concepts of database relevancy ratio (DRR) and sample relevancy ratio (SRR) for queries. We investigated the effect of the DRR on the accuracy of query results estimated from samples drawn using the *k*-MDI index. From the experimental evaluation of the sampling scheme on a large real dataset, it is found that even at relatively low sampling rates of 1% to 12 %, query results can be estimated accurately with a minimum of 95% confidence for queries with medium to high DRR. At a very low DRR of 0.002, the estimated values of sum and count fell outside the acceptable confidence level of 95%, but the estimated mean was within the 95% confidence interval even at very low DRR. Depending on the sampling rate, the sampling based query processing was on average 9 to 30 times faster than processing the same queries against the whole dataset.

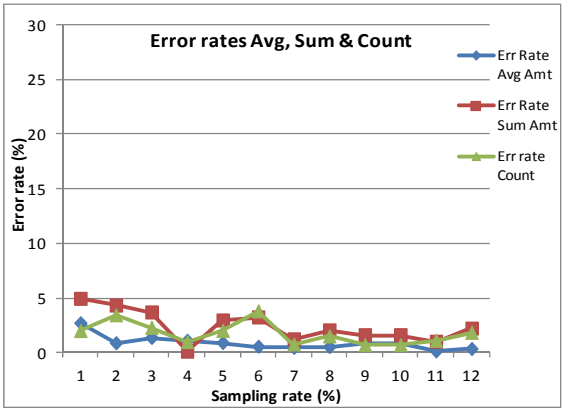
As future work, it is proposed to develop a generic tool that can be used with some parameter inputs to set up the *k*-MDI tree index for any data warehouse schema. We also plan to further evaluate the sampling based estimation scheme on data warehouses with larger dimensions.



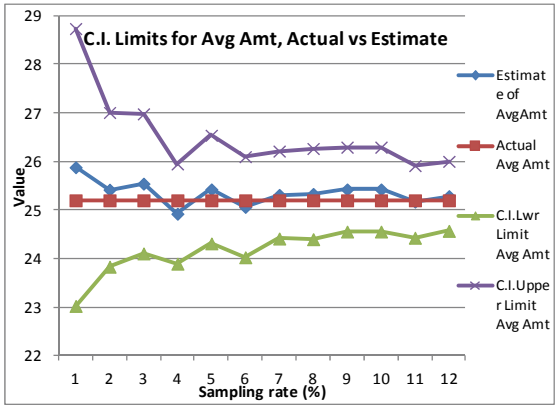
(a) Error rates for query with condition on one dimension (high relevancy ratio).



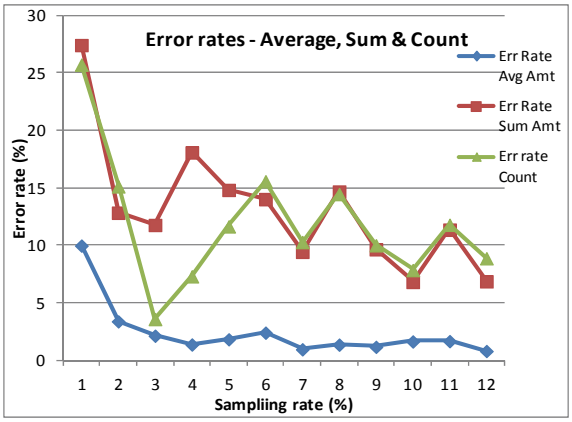
(b) Confidence interval (AVG amt) for query with condition on one dimension (high DRR).



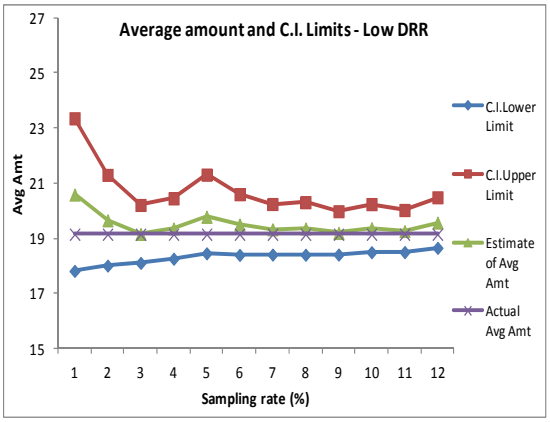
(c) Error rates for query with condition on two dimensions (medium relevancy ratio).



(d) Confidence interval (AVG amt) for query with condition on two dimensions (medium DRR).



(e) Error rates for query with condition on three dimensions (low relevancy ratio).



(f) Confidence interval for query with condition on three dimensions (low DRR).

Figure 6: Error rates of average scan amount and sum of scan amount and confidence interval of average scan amount at various sampling rates for high, medium and low relevancy ratios.

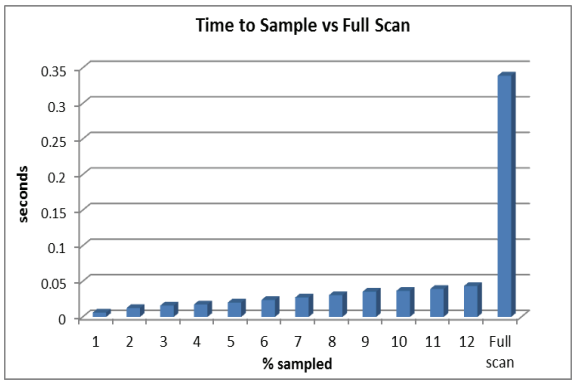


Figure 7: Query times at different sampling rates as compared to full database scan.

REFERENCES

Berenson, M. L., Levine, D. M., 1992. Basic Business Statistics - Concepts and Applications. *Prentice Hall*, Upper Saddle River, New Jersey, USA.

Bernardino, J., Furtado, P., Madeira, H., 2002. Approximate Query Answering Using Data Warehouse Striping. *Journal of Intelligent Information Systems*. 19:2, pp.145-167.

Chaudhuri, A., Mukherjee, R., 1985. Domain Estimation in Finite Populations. *Australian Journal of Statistics*. Vol. 27:2, pp. 135-137.

Hellerstein, H, Haas, P, Wang, J., 1997. Online Aggregation. *SIGMOD 1997*, pp. 171-182.

Hobbs, L., Hillson, S., Lawande, S., 2003. *Oracle9iR2 Data Warehousing*. Elsevier Science, MA, USA.

Jermaine, C., 2007. Random Shuffling of Large Database Tables. *IEEE Transactions on Knowledge and Data Engineering*. 18:1, pp.73-84.

Jermaine, C., 2003. Robust Estimation with Sampling and Approximate Pre-Aggregation. *VLDB Conference Proceedings 2003*, pp. 886-897.

Jermaine, C., Pol, A., Arumugam, S., 2004. Online Maintenance of Very Large Random Samples. *SIGMOD Conference Proceedings 2004*.

Jin, R., Glimcher, L, Jermaine, C, Agrawal, G., 2006. New Sampling-Based Estimators for OLAP Queries. *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06),* Atlanta, GA, USA.

Joshi, S., Jermaine, C., 2008. Materialized Sample Views for Database Approximation, *IEEE Transactions on Knowledge and Data Engineering*, 20:3 pp. 337-351.

Keller, G., 2009. Statistics for Management and Economics. *Cengage Learning*, Mason, OH, USA.

Kimball, R., Ross, M., 2002. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Ed. *John Wiley & Sons*, Indianapolis, USA.

Li, X., Han, J., Yin, Z., Lee, J-G., Sun, Y., 2008. Sampling Cube: A Framework for Statistical OLAP over Sampling Data. *Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD'08)*, Vancouver, BC, Canada, June.

Olken, F., Rotem, D., 1990, Random Sampling from Database File. In: *A Survey. International Conference on Scientific and Statistical Database Management, 1990*. pp. 92-111.

Spiegel, J., N. Polyzotis, 2009. TuG Synopses for Approximate Query Answering. *ACM Transactions on Database Systems*. (TODS) 34(1).

TPC Benchmarks, 2011. Transaction Processing Performance Council - TPC-H: Decision Support Benchmark. http://www.tpc.org [Accessed 20 November 2011].

TUN - Teradata University Network, 2011. http://www.teradata.com/TUN_databases. [Accessed: 13 April 2007].

Using Formal Concept Analysis to Extract a Greatest Common Model

Bastien Amar[1], Abdoulkader Osman Guédi[1,2,3], André Miralles[1,3],
Marianne Huchard[3], Thérèse Libourel[4] and Clémentine Nebut[3]
[1]*Tetis/IRSTEA, Maison de la Télédétection, 500 Rue J.-F. Breton 34093, Montpellier Cdx 5, France*
[2]*Université de Djibouti, Avenue Georges Clemenceau, BP 1904, Djibouti, Republic of Djibouti*
[3]*LIRMM, Univ. Montpellier 2 et CNRS, 161, rue Ada, F-34392, Montpellier Cdx 5, France*
[4]*Espace Dev, Maison de la Télédétection, 500 Rue J.-F. Breton 34093, Montpellier Cdx 5, France*
{bastien.amar, abdoulkader.osman-guedi, andre.miralles, therese.libourel}@teledetection.fr,
{marianne.huchard, clementine.nebut}@lirmm.fr

Keywords: Formal Concept Analysis, FCA, Greatest Common Model, GCM, Pesticide, Environmental Information System, Model Factorization, Core-concept, Domain-concept.

Abstract: Data integration and knowledge capitalization combine data and information coming from different data sources designed by different experts having different purposes. In this paper, we propose to assist the underlying model merging activity. For close models made by experts of various specialities, we partially automate the identification of a Greatest Common Model (GCM) which is composed of the common concepts (core-concepts) of the different models. Our methodology is based on Formal Concept Analysis which is a method of data analysis based on lattice theory. A decision tree allows to semi-automatically classify concepts from the concept lattices and assist the GCM extraction. We apply our approach on the EIS-Pesticide project, an environmental information system which aims at centralizing knowledge and information produced by different specialized teams.

1 INTRODUCTION AND PROBLEMATICS

Elaborating data models is a recurrent activity in many projects in different domains, for various objectives: building dictionaries of the domain, designing databases, developing software for this domain, etc. Usually, such models of the domain are required by several teams, dealing with different facets of the domain, and potentially stemming from different scientific domains. For example, in the IRSTEA institute (in which three of the authors work), the study of pesticide impact on environment involves specialists from different scientific domains: hydrology, agronomy, chemistry, etc.

Each specialist is able to model the part of the domain model it is familiar with, and finally, a consolidated domain model must be built gathering all the specialized models. This gathering activity is complex and generally carried out manually. Indeed, it requires to detect the common domain-concepts modeled in the various specialized models, so as to integrate them without redundancy in the consolidated model named greatest common model (GCM). This GCM is particularly useful to perform schema integration and knowledge capitalization.

In this paper, we address the issue of assisting this gathering activity, in the context of domain data models designed with UML class diagrams through the automated detection of common domain-concepts (with two levels of confidence) possibly enriched with new domain-concepts automatically extracted from the previous ones. This approach is based on Formal Concept Analysis (FCA), which is an exact and robust data analysis method based on lattice theory. We use FCA to detect commonalities, redundancies and introduce new abstractions, both inside the models taken individually (intra-model factorization), and inside two distinct data models taken jointly (inter-model factorization). The approach defined in this paper deals with two models, but more generally, it is able to identify the common domain-concepts of several models in order to help the designer to centralize these common concepts into a unique consolidated model (the GCM). This approach is under evaluation on a large project from the IRSTEA institute called Environmental Information System for Pesticides (EIS-Pesticides), in which two teams cooperate

to build a domain data model. The transfer team is specialized in the study of the pesticides transfer to the rivers and the practice team, mainly works on the agricultural practices of farmers.

The rest of the paper is structured as follows. In Section 2 we introduce example models taken from the EIS-Pesticides project. In Section 3, we draw the main lines of our approach, and in Section 4, we provide a short introduction to Formal Concept Analysis (FCA). In Section 5 we explain how FCA is used on input models and how the resulting lattices are analyzed so as to provide the final user clear recommendations to build the greatest common model. In Section 6, we present our produced greatest common model of our example models and we apply our approach on a larger model to evaluate its scalability. Section 7 presents the related work and Section 8 concludes the paper.

2 RUNNING EXAMPLE: THE TWO MODELS OF MEASURING STATION

The Environmental Information System for Pesticides (EIS-Pesticides) is a project (Pinet et al., 2010; Miralles et al., 2011) that has the objective to set up an information system allowing to centralize knowledge and information produced by Transfer and Practice teams (see Section 1). We illustrate our approach on a small subsystem representing part of the *measuring activity* on the catchment area (drainage basin): measuring stations monitor the major parameters involved in the transfer of the pesticides to the rivers.

Figure 1 shows the two data models of the measuring stations used in this study. They are produced by the two teams involved in the project. As these two models are very close, we have organized them by grouping at the r.h.s of measuring station (*cl_MeasuringStation*), the identical domain-concepts (that also have the same relationships). In this part of the model, the measured data are associated to the corresponding measuring device: the rainfall (*cl_Rainfall*) and the hydraulic head (*cl_HydraulicHead*) of the groundwater table are continuously recorded respectively by the rain gauge (*cl_RainGauge*) and by the piezometer (*cl_Piezometer*). Each of these measures is dated (see property *att_MeasuringDate*). On the l.h.s. of *cl_MeasuringStation*, the model *M1_MeasuringStation* allows to record the data measured by a weather station of Météo-France (a french meteorological institute): temperature (*cl_Temperature*), hygrometry (*cl_Hygrometry*) and potential evapo-transpiration (*cl_PET*) of the short green crops. These last domain-concepts are not in the model *M2_MeasuringStation* which has on the other hand a limnimeter (*cl_Limnimeter*) to measure continuously the flow rate (*cl_FlowRate*) of rivers. A technician is in charge to take samples in order to determine in laboratory the amount of pesticides in the water (*cl_PesticideMeasurement*). Finally, the wind velocity (*cl_WindMeasurement*) is a parameter coming from a weather station of Météo-France.

3 OVERVIEW OF THE PROPOSED APPROACH

The main objective of our approach is to assist the task of gathering two or more models independently defined and thus potentially involving common concepts. For that we extract from initial models their Greatest Common Model (GCM). The term "greatest common model" is chosen by analogy to the "greatest common divisor (GCD)" in arithmetic; it is more precisely defined in the following. Roughly, it contains all the common domain-concepts that are introduced in all the studied models, in a normal[1] (factorized) form.

The proposed approach is illustrated in Figure 2. The input is two (or more) models for a domain, named M_1 and M_2. In a first time, the classes of the input models are described by their owned characteristics. Formal Concept Analysis (FCA) allows entities sharing characteristics to be grouped into formal-concepts, and results in lattices providing a hierarchical view of those formal-concepts. We apply FCA on several class descriptions, resulting in several lattices. These lattices allow the identification of common concepts, specific concepts and eventually new abstractions extracted from intra- or inter- model factorization. For instance, if we describe classes by their owned attributes, the resulting lattice (cf Figure 5) extracts the r.h.s. common domain concepts of Figure 1. It also extracts new abstractions. Some new abstractions are present both in M_1 and M_2 (*e.g.* a *device* concept factorizes commonalities of rain gauge, and piezometer: inter-model factorization). Some other extracted abstractions are present only in a same model (*e.g.* a *dated measurement* concept factorizes pesticide and wind measurements in M_2: intra-model factorization). For each lattice, we have two levels

[1]Here, we refer to the relational normal form used in database schema normalization, which has the same objective: eliminate redundancies.

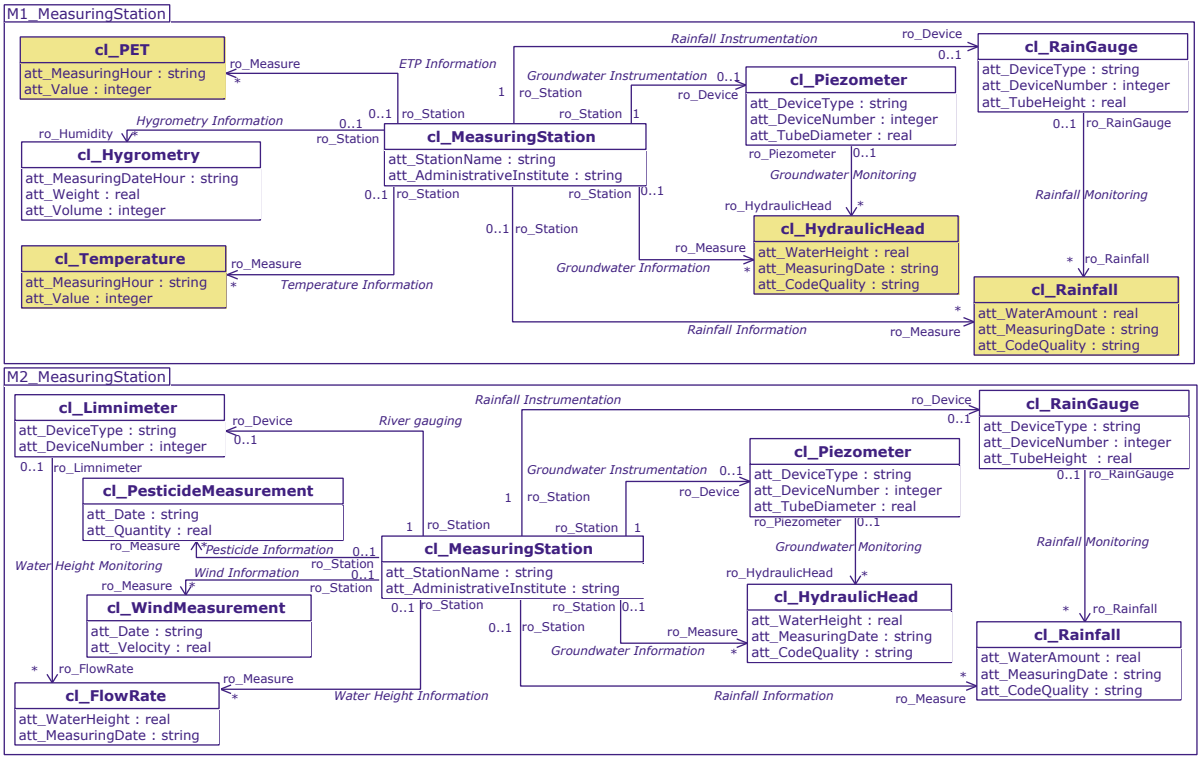


Figure 1: The two data models of measuring station produced by the two teams.

of confidence for those domain-concepts: domain-concepts which are very likely to be in the GCM, and others that have to be precisely analyzed, validated and named by the final expert. As we generate several lattices, the expert in charge of integration needs to follow a strategy for analyzing them. We propose to order the obtained lattices following the semantic hierarchy of the different factorization criteria. The lattices are then analyzed, so as to categorize formal-concepts and interpret them, if applicable, to form domain-concepts.

The domain-concepts recognized by the experts as being in the GCM are called the *core domain-concepts*. In Figure 1, the domain-concepts to the right of *cl_MeasuringStation* are certainly core domain-concepts. The *greatest common model* (GCM) is defined as the largest model factorizing the core domain-concepts of several models.

4 A SHORT INTRODUCTION TO FORMAL CONCEPT ANALYSIS

Formal Concept Analysis (FCA) (Ganter and Wille, 1999) is a method of data analysis based on lattice theory (Birkhoff, 1940). It is used in many applications relative to classification including knowledge structuring, information retrieval, association rule extraction in the data mining domain, class model refactoring, or software analysis. FCA studies entities described by their characteristics to discover formal-concepts which are maximal groups of entities sharing maximal groups of characteristics. A partial specialization order based on the entity set inclusion provides a lattice structure (the concept lattice).

A *formal context* K is a triple[2] $K = (E, C, R)$, where E is the set of entities and C the set of characteristics that describe these entities. $R \subseteq E \times C$ associates an entity with its characteristics: $(e,c) \in R$ when entity e owns characteristic c. For example, Table 1 shows the formal context of the sub-model highlighted in Figure 1 (limited to the four classes *cl_PET, cl_Temperature, cl_HydraulicHead* and *cl_Rainfall*). Classes (the entities) are described by the name of their owned attributes (characteristics).

A *formal-concept* is a pair $(Extent, Intent)$ where $Extent = \{e \in E | \forall c \in Intent, (e,c) \in R\}$ and $Intent = \{c \in C | \forall e \in Extent, (e,c) \in R\}$. These two sets represent the entities that own all the characteristics (extent) and the characteristics shared

[2] In the literature, standard notation is $K = (G, M, I)$. We use $K = (E, C, R)$ for readability reasons and to get a better understanding toward our thematic partners.

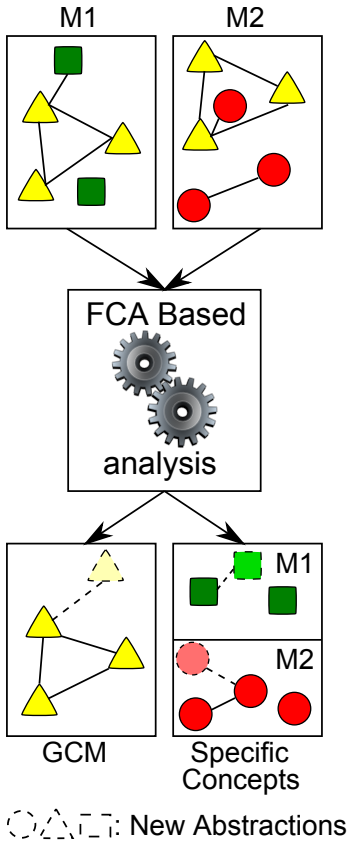


Figure 2: A schematic overview of our approach (applied on one formal context).

Table 1: The formal context of the reduced model.

	att_MeasuringHour	att_Value	att_WaterAmount	att_MeasuringDate	att_CodeQuality	att_WaterHeight
cl_PET	×	×				
cl_Temperature	×	×				
cl_Rainfall			×	×	×	
cl_HydraulicHead				×	×	×

by all entities (intent). The specialization order between two formal concepts is given by the following equivalence: $(Extent\_1, Intent\_1) < (Extent\_2, Intent\_2) \Leftrightarrow Extent\_1 \subset Extent\_2$ (equivalently $Intent\_2 \subset Intent\_1$).

In a lattice, there is an ascending inheritance of entities and a descending inheritance of characteristics. The simplified intent of a formal concept is its intent without the characteristics inherited from its super-concept intents. The simplified extent is defined in a similar way.

Nota: in this article, we distinguish simplified extent from extent. When it is not specified, we are talking about (complete) extent.

For readability reasons, all lattices presented in this paper show simplified extents and intents.

Figure 3 shows the concept lattice built from the formal context presented Table 1. Each formal-concept is represented by a box in three parts: the first contains the generated name of the formal-concept, the second part contains its simplified intent, and the last one contains its simplified extent. Let us consider `Concept_17`: it represents entities (classes) described by the characteristic *att_WaterHeight* and by the characteristics inherited from its super-concepts: *att_MeasuringDate* and *att_CodeQuality* (from `Concept_16`).

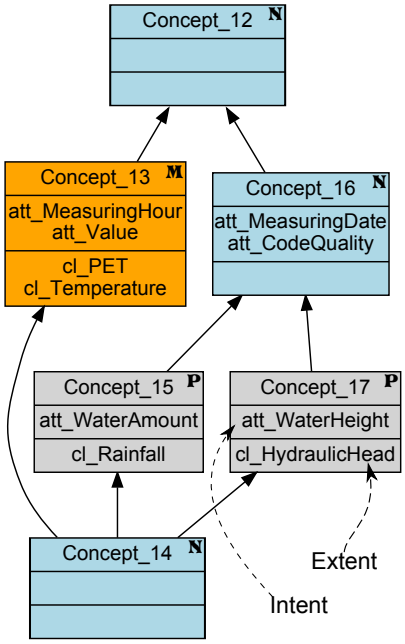


Figure 3: Class/attribute name lattice: result of FCA on Table 1.

In this work, we are interested in three categories of formal-concepts that form a partition of the set of formal-concepts:

Definition 1. Merged formal concepts *have more than one entity in their simplified extent. This means that all entities in the extent are described by exactly the same set of characteristics.*

In Figure 3, `Concept_13` is a merged formal concept: *cl_PET* and *cl_Temperature* are (exactly) described by both characteristics *att_MeasuringHour* and *att_Value*.

Definition 2. New formal concepts *have an empty simplified extent. These are new, more abstract, concepts, factoring out characteristics common to several formal-concepts.*

In Figure 3, `Concept_16` is a new formal concept, factoring out characteristics of both `Concept_15` and `Concept_17`.

Definition 3. Perennial formal concepts *have one and only one entity in their simplified extent.*

In Figure 3, both `Concept_15` and `Concept_17` are perennial. In this article, merged, new and perennial formal concepts are respectively annotated, in the figures, **M**, **N** and **P** at the right-top corner.

5 APPLYING FORMAL CONCEPT ANALYSIS TO EXTRACT CANDIDATES FOR THE GREATEST COMMON MODEL

In this section, we propose a methodology based on two automatic steps that uses Formal Concept Analysis (FCA) and an interactive step to extract the greatest common model of two input models. Given two models M_1 and M_2:

- We compute the lattices resulting from FCA applied to several formal contexts extracted from the disjoint union of the two input models $M = M_1 \oplus M_2$.
- The concepts of these lattices are analyzed thanks to a decision tree based on the analysis of the concept extent, and we obtain six concept lists (categories).

In the interactive step, these six lists are exploited to assist the expert to build the greatest common model. The next subsections precisely describe two automatic steps.

5.1 Apply FCA on the Two Models

As explained in Section 4, formal contexts describe entities by characteristics. Many different formal contexts can be extracted from a class model: it has to be defined which model elements are chosen to be the studied entities, and which features of those model elements are chosen to be their studied characteristics. Here we focus on three formal contexts extracted from the disjoint union of input models $M = M_1 \oplus M_2$:

1. the formal context of classes described by their name,
2. the formal context of classes described by their attributes,
3. the formal context of classes described by their attributes and by their roles.

Figure 4 presents the lattice obtained with the formal context of classes described by their name (*class/class name* lattice). This lattice groups in a concept the set of classes sharing the same name. For example, the merged concept `Concept_1` represents the set of classes (in extent) sharing the name (in intent) *cl_Piezometer*. In other words, FCA merged in a single concept classes that have a same name. Classes that are not duplicated in the models M_1 and M_2 remain in a perennial concept, like the *cl_PET* class in `Concept_7`. In inter-model factorization, the three categories of concepts described in Section 4 exist: the merged concept `Concept_1` has more than one entity in its simplified extent. In a similar way, the perennial concept `Concept_7` (*cl_PET*) has exactly one element in its extent. Later we will see the case where new formal concepts appear.

Figure 5 presents the lattice obtained with the formal context of classes described by the names of their owned attributes (*class/attribute name* lattice). In this lattice, a formal concept thus is a group of classes (extent) sharing a group of attribute names (intent). The lattice contains new formal concepts ($simplified\ extent = \emptyset$), e.g. `Concept_47`, that represents a new abstraction: things that are dated.

Figure 6 presents the lattice obtained with the formal context of classes described by the names of their owned attributes and roles (*class/attribute-role name* lattice). UML associations are taken into account in this lattice through those roles. For example, class *cl_FlowRate* has attribute *att_WaterHeight* and role *ro_Station* in association *Water Height Information*. The new formal concept `Concept_30` represents the classes that are linked with a Station *via* the role *ro_Station*. Class *cl_FlowRate* belongs to the extent of this concept.

5.2 Analysis of the Lattices

In this section, we present the analysis of the lattices using a decision tree to classify each concept. First, the *class/class name* lattice must be analyzed. This lattice allows the designer to group classes that have a same name. Then, we analyze the *class/attribute name* lattice that allows us to find attribute-based factorizations. As we will see, the *class/attribute-role name* lattice can be a considerable help to refine the decisions about factorization.

For each formal concept $Co_k = (E_k, I_k)$, the *complete* extent E_k has to be analyzed and the concept has

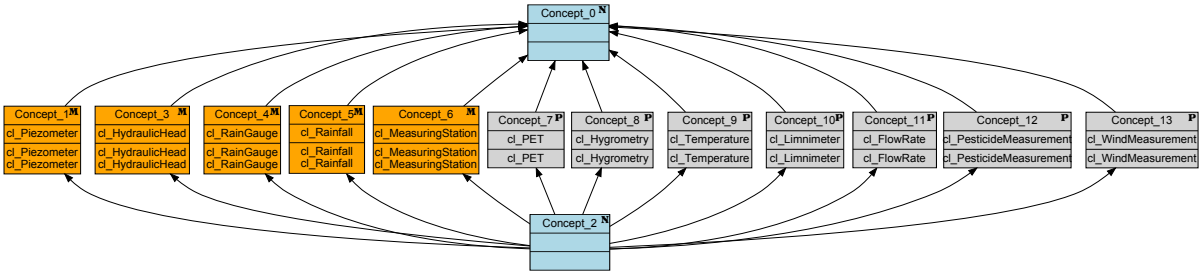


Figure 4: The class/class name concept lattice.

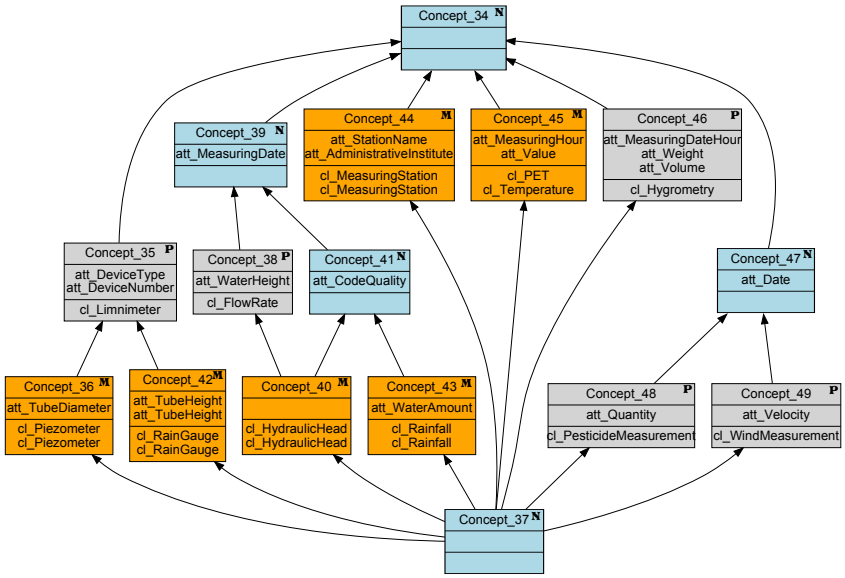


Figure 5: The class/attribute name concept lattice.

to be included in one of these lists:

- L_{GCM} is the list of core-concepts that will be included in the greatest common model.
- L_{pGCM} is the list of potential (candidate) core-concepts to be validated by an expert to be in the greatest common model.
- L_{M_1} and respectively L_{M_2} are the lists of domain concepts specific to M_1 (resp. M_2).
- L_{nM_1} and respectively L_{nM_2} are new domain concepts specific to M_1 (resp. M_2), factorizing existing domain concepts. These domain concepts are not intended to be in the greatest common model, but they can be presented to experts to improve the factorization of M_1 (resp. M_2).

Figure 7 presents the decision tree: we define $\mathcal{C}_{M_i}$ (resp. $\mathcal{C}_{M_j}$) as the set of classes in the model M_i (resp. M_j), and the decision tree is designed for two models M_i and M_j where $i \neq j$. As we apply FCA with classes as entities (characteristics being class name, attributes, and/or roles), the extent of a concept contains only classes. For each concept, we first check if the concept is a *merged concept*, a *new concept* or a *perennial concept* (nodes 1, 8 and 12 in the decision tree of Figure 7) as defined in Section 4.

Analysis of Merged Concepts: If the concept is a merged concept, then three cases are possible: its extent contains elements from both models M_i and M_j (node 2), its extent contains only elements from M_i (node 6), or its extent is empty (node 7).

If the concept extent contains elements from both models, the cardinality of the intersection between the extent and the set of model classes has to be checked. In the first case, the extent contains only one class from M_i and only one class from M_j (node 3) like `Concept_1` in the *class/class name* lattice, Figure 4. Then a corresponding domain concept should be added in L_{GCM}: it can be considered as a core-concept – a domain concept common to both models. If the extent contains only one class from M_i and several classes from M_j (node 4), or several el-

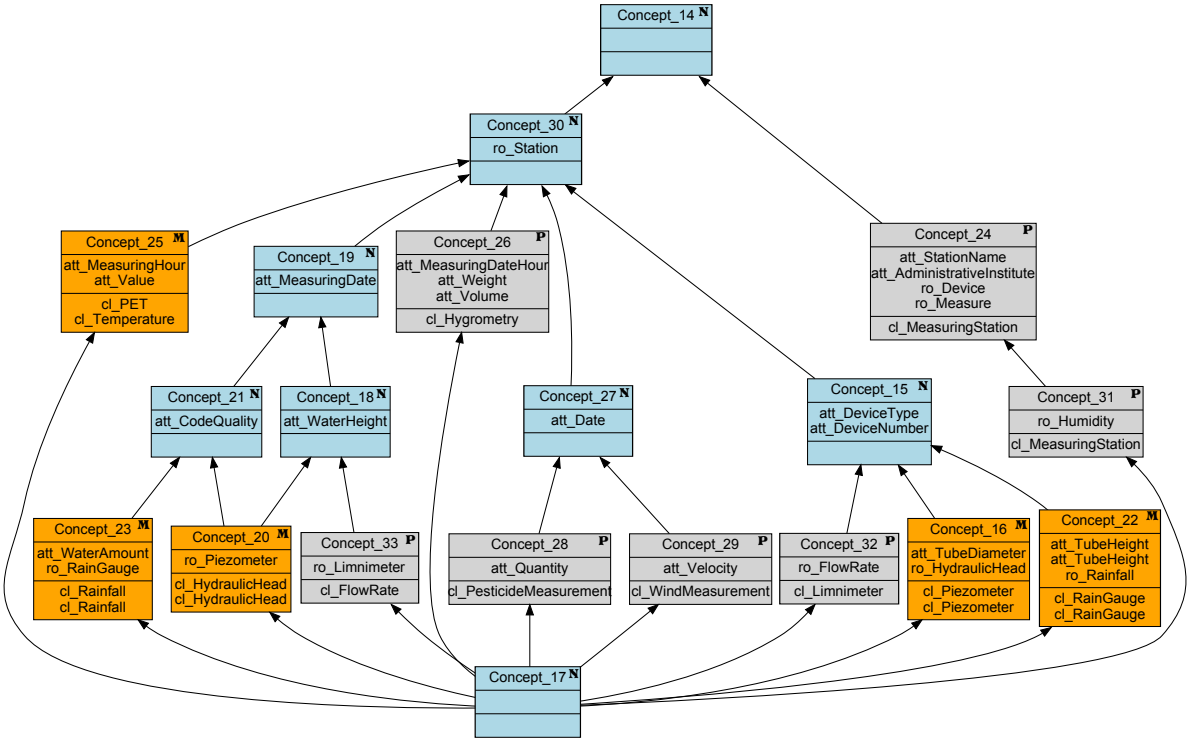


Figure 6: The class/attribute-role name concept lattice.

ements from both models (node 5), then it should be put in the L_{pGCM} list: it is a potential core-concept, but an expert intervention is necessary. He or she can choose to merge or factorize duplicated classes if they are semantically closed, in a same model (intra-model factorization), and relaunch the process to extract the greatest common model. He or she can also consider these classes as specific domain concepts and keep them in the specific model.

If the merged concept contains only classes from M_i (node 6), like the `Concept_45` in the Figure 5, it should be added to the L_{nM_i} list. Its extent contains a group of elements coming from a same model and that are described exactly by the same characteristics. It can be presented to an expert to improve the model M_i, but it is not a core-concept (they are in one model only). In the case of `Concept_45`, FCA suggests to merge the classes *cl_PET* (representing the Potential Evapo-Transpiration) and *cl_Temperature*. In this special case, these two classes are semantically different, and the expert do not want to factorize them, but in other situations he could consider this factorization to be interesting.

The node 7 describes concepts wherein the extent does not contain classes from M_i and M_j. This is inconsistent: by definition, a merged concept extent contains at least two elements (*cf* definition 1).

1 → $Merged\ Concept$
2 → $E_k \cap \mathcal{C}_{M_i} \neq \emptyset \wedge E_k \cap \mathcal{C}_{M_j} \neq \emptyset$
3 → $|E_k \cap \mathcal{C}_{M_i}| = 1 \wedge |E_k \cap \mathcal{C}_{M_j}| = 1$ → L_{GCM}
4 → $|E_k \cap \mathcal{C}_{M_i}| = 1 \wedge |E_k \cap \mathcal{C}_{M_j}| > 1$ → L_{pGCM}
5 → $|E_k \cap \mathcal{C}_{M_i}| > 1 \wedge |E_k \cap \mathcal{C}_{M_j}| > 1$ → L_{pGCM}
6 → $E_k \cap \mathcal{C}_{M_i} \neq \emptyset \wedge E_k \cap \mathcal{C}_{M_j} = \emptyset$ → L_{nM_i}
7 → $E_k \cap \mathcal{C}_{M_i} = \emptyset \wedge E_k \cap \mathcal{C}_{M_j} = \emptyset$ → $Inconsistent$
8 → $New\ Concept$
9 → $E_k \cap \mathcal{C}_{M_i} \neq \emptyset \wedge E_k \cap \mathcal{C}_{M_j} \neq \emptyset$ → L_{pGCM}
10 → $E_k \cap \mathcal{C}_{M_i} \neq \emptyset \wedge E_k \cap \mathcal{C}_{M_j} = \emptyset$ → L_{nM_i}
11 → $E_k \cap \mathcal{C}_{M_i} = \emptyset \wedge E_k \cap \mathcal{C}_{M_j} = \emptyset$ → $Bottom$
12 → $Perennial\ Concept$
13 → $E_k \cap \mathcal{C}_{M_i} \neq \emptyset \wedge E_k \cap \mathcal{C}_{M_j} \neq \emptyset$ → L_{pGCM}
14 → $E_k \cap \mathcal{C}_{M_i} \neq \emptyset \wedge E_k \cap \mathcal{C}_{M_j} = \emptyset$ → L_{M_i}
15 → $E_k \cap \mathcal{C}_{M_i} = \emptyset \wedge E_k \cap \mathcal{C}_{M_j} = \emptyset$ → $Inconsistent$

Figure 7: Decision tree.

Analysis of New Concepts: If the concept is a new concept (*cf.* definition 2, node 8), and if its extent contains elements from both models M_i and M_j (node 9) then the concept has to be put in the L_{pGCM} list: it is a potential factorization of concepts defined in M_i and M_j, so it is potentially a core-concept. Experts have to decide if this factorization is valid and if this new concept has to be included in the greatest common model. `Concept_39` in Figure 5 is an example of this type of concept. In our case study, the expert validates

this concept to be a greatest common model concept.

If the new concept extent contains only classes from one model, it can be added in the L_{nM_i} list (node 10 in the decision tree). This concept corresponds to an intra-model factorization. It is the case of Concept_47, representing things that are dated in M_2. This kind of concept is not a core-concept and should not be included in the greatest common model. It can be presented to the M_i designer in order to raise the quality of its model by a new factorization.

If the new concept extent (node 11 in the decision tree) does not contain elements from M_1 nor M_2, this means that this is the concept *Bottom*. Concept *Bottom* is present in each lattice (concepts Concept_2, Concept_37 and Concept_17). It represents elements that own all attributes and should not be used in our re-engineering process. Instead, the top concept can not be inferred only by extent analysis and it may appear in each branch of the tree. Depending on the configuration of the models analyzed, this concept may be relevant and it is classified as other concepts.

Analysis of Perennial Concepts: Node 13 in the decision tree describes perennial concepts that have in their extent classes from M_i and M_j, like Concept_35 in Figure 5. This means that there is a potential factorization of Concept_36 and Concept_42, and this factorization already exists, *cl_Limnimeter* in our example. This kind of concept has to be presented to the expert, it is thus added to the L_{pGCM} list. In our example, the designer can make *cl_limnimeter* be a super-class of *cl_piezometer* and *cl_Raingauge*, but this decision is not semantically valid: a piezometer *is not* a limnimeter. An analysis of the lattice of classes described by their attributes and role names (Figure 6) shows that it is better to create a new superclass (Concept_15) of data instrumentation, factorizing the three classes *cl_limnimeter*, *cl_Piezometer* and *cl_RainGauge*. In this case, the lattice of classes described by their attributes/roles names is useful to help the designer to take a decision.

If the perennial concept extent contains only classes from M_i (node 14) then it is a M_i domain specific concept. This concept must be added to L_{M_i}. For example, concepts Concept_7, Concept_8, Concept_48, and Concept_46 are domain concepts specific to M_i.

A perennial concept cannot have an empty extent (node 15): the definition 3 specifies that a perennial concept has one (and only one) element in its extent.

From both L_{GCM} and L_{pGCM} lists, the expert has to select the core-concepts that will be included in the GCM.

Our approach has been implemented as a profile in a case tool. A component transforms the UML models into the different types of formal contexts which are entries of FCA. Another component produces the corresponding lattices. Finally, another component generates the various lists of domain-concepts in accordance with the decision tree.

6 RESULTS

Figure 8 shows the model obtained by applying our approach: the final greatest common model of the M1 and M2 models (Figure 1). This GCM reflects also the interpretation and the validation by an expert of the new concepts. We annotated classes by associated formal concepts that represent them in the lattices (Figures 4, 5 and 6).

As expected, the same domain-concepts in both models M1 and M2 are present in the *GCM*: *cl_MeasuringStation*, *cl_Piezometer*, *cl_HydraulicHead*, *cl_RainGauge* and *cl_Rainfall*. They constitute the core-concepts of the *GCM* of M1 and M2. So, they are automatically added in the L_{GCM} list.

Our approach proposes a list of possible factorizations of domain-concepts in the L_{pGCM} list. The expert must validate the relevance of these concepts. In this example, two new concepts have been considered relevant. They are colored in figure 8.

The first corresponds to formal concepts Concept_15 (Figure 6) and Concept_35 (Figure 5) in the lattices. They factorize attributes *att_DeviceType* and *att_DeviceNumber*. This concept has been validated by experts as a new *cl_Device* class.

The second new concept corresponds to formal concepts Concept_41 and Concept_21 in the lattices. It factorizes both *att_MeasuringDate* and *att_CodeQuality* attributes. Similarly to the first new concept, experts validate this concept as a new *cl_Data* class.

Table 2 quantifies for each formal context the number of concepts in each list defined in the decision tree[3].

In order to validate the scalability of our approach, tests have been done on two versions of the complete model from the EIS-pesticides project (about 125 classes). Table 3 gives the number of concepts by list of the decision tree[3].

With the class/class name and class/attribute name lattices, experts have to analyze and to validate between 34 and 39 concepts present in the L_{pGCM} list.

[3]In these tables, new and merged concepts must be still validated by an expert.

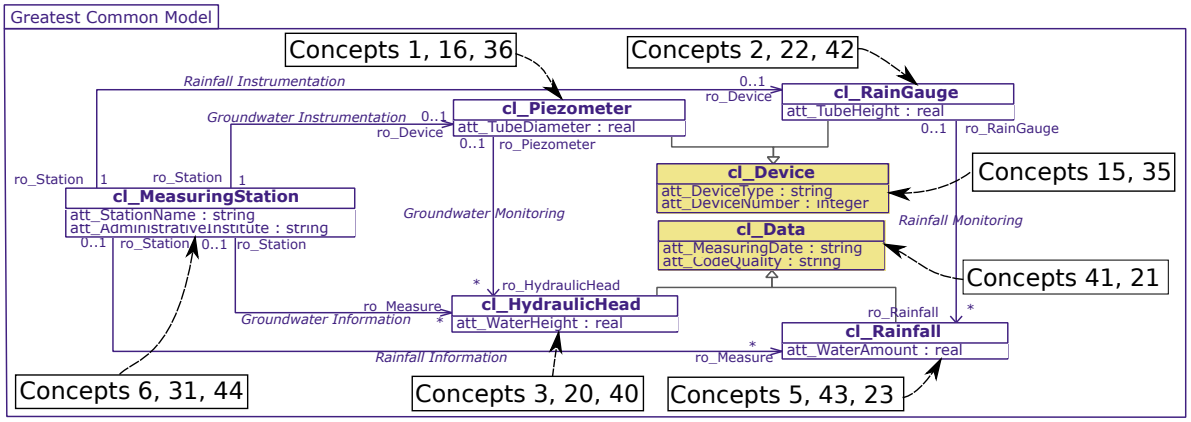


Figure 8: The greatest common model of M1 and M2 models (Figure 1).

Table 2: Result of our approach on the MeasuringStation model.

	L_{GCM}	L_{pGCM}	L_{nM1}	L_{nM2}	L_{M1}	L_{M2}
class/class name	5	1	0	0	3	4
class/attribute name	5	5	1	1	1	2
class/attribute-role name	4	7	1	1	2	4

They can obtain more precision (with also more analysis work) with the class/attribute-role name lattice, where 119 potential GCM concepts are proposed. We are currently working to assist the expert in this analysis task (Osman Guedi et al., 2011). We can also deduce from these results that the two versions of pesticide model are very close: there are only few specific concepts.

7 RELATED WORK

FCA is used to improve the abstraction quality and the duplication elimination in class models in various domains (software engineering, ontology mapping or merging). This feature led us to propose the construction of a GCM to capitalize the knowledge of various domains.

Many variants have been studied, which take into account different characteristics for classes (the entities or domain-concepts in this framework): attribute names, attribute types, operation names, operation signatures, type specialization... The relevance of this approach is related to the properties satisfied by the class model after refactoring: all duplications are eliminated and the specialization relation between formal concepts meets the inclusion of features in the class model. These previous approaches only focus on intra-model factorization. In this paper, we use FCA for *inter-model* factorization, and we need to analyze differently the lattices, to identify categories of formal-concepts useful to build the greatest common model of several input class models. We define a guide for the expert to assist the building of the GCM. Indeed, in this work, we assume that if two characteristics have the same name, then these two characteristics are identical. Some work includes semantic analysis (Falleri, 2009; Rouane et al., 2007).

In software engineering, FCA has been used to build and maintain class hierarchies (Godin and Mili, 1993; Dao et al., 2006; Arévalo et al., 2006). In this paper, our objective is different, we want to find common and specific parts between several models. The management of similarities and differences between models has been studied in the domain of model versioning (Altmanninger et al., 2009). The Smover tool uses direct comparison between a model and its previous version to detect syntactic and semantic conflict (Altmanninger et al., 2010). In order to manage model conflicts in a distributed development context, the work presented in (Cicchetti et al., 2008) proposes the use of a difference model to store differences between two versions of a same model (Cicchetti et al., 2007). These methods allow to show differences between models, but they don't aim to propose automatic core-concept detection. In the approach described in (Ohst et al., 2003), models and diagrams are considered as syntax trees, which allows the authors to design a difference operation between models. Compared to the domain of model versioning, we aim to present the GCM in a normal (factorized) form. This is why FCA is more suitable for our problem.

Table 3: Result of our approach on the complete EIS-Pesticides model.

	L_{GCM}	L_{pGCM}	L_{nM1}	L_{nM2}	L_{M1}	L_{M2}
class/class name	111	34	0	0	1	1
class/attribute name	43	39	0	0	1	2
class/attribute-role name	68	119	0	0	8	9

Formal concept analysis has been used to perform ontology mapping or merging, which is an issue close to ours (Kalfoglou and Schorlemmer, 2005; Bendaoud et al., 2008). The approach proposed by (Stumme and Maedche, 2001) uses FCA and linguistic analysis to merge ontologies in a semantic web context. In order to align ontologies, there are approaches that use a similarity measure, based on FCA (Formica, 2006) or on ontologies internal structure and association rule mining (Tatsiopoulos and Boutsinas, 2009). All these works aim to perform ontology mapping, while we work to extract the mapping result and to abstract new domain-concepts.

Since the early 80s, the database domain has studied the problem of schema integration and data matching, particularly in the database integration context. The aim of database integration context is to produce the global schema of a collection of databases (Batini et al., 1986; Rahm and Bernstein, 2001; Shvaiko and Euzenat, 2005). Producing such a global database schema is an issue close to the extraction of a greatest common model in the sense that the search for identical concepts in different schemas is a necessary step. There are a lot of work dealing with this problematic in the literature. Generally, integration is composed of different steps: schema transformation, correspondence investigation and schema integration. Our work focuses on correspondence investigation and schema integration (Parent and Spaccapietra, 1998). The integrated schema includes the GCM and the specific part of the initial schemas. There are two groups of solutions to semi-automatically find matches : rule-based solutions and learning-based solutions. Our approach is similar to rule-based solutions: we search similarity between several model elements based on their characteristics (Doan and Halevy, 2005). Unlike these approaches, the use of FCA allows to choose with finesse the way to describe the characteristics that we consider. In this article, we focus on the description of classes by their name, attribute name or role name, but FCA opens many other possibilities.

8 CONCLUSIONS

During domain modeling activity, several teams with different scientific skills usually make different models of a same domain. Each specialized team models the part of the domain model it is familiar with, and finally, a unique, consolidated domain model has to be built. This model integration requires the identification of the common domain-concepts that are present in the various specialized models.

Our contribution in this paper is an approach to assist the gathering task for several given class diagrams describing the domain. The proposed methodology is based on *Formal Concept Analysis* and the analysis of the formal-concepts using a decision tree. It allows the production of a Greatest Common Model in a normal (factorized) form. Our approach proposes two levels of confidence for candidate GCM concepts: domain-concepts which certainly will be in the GCM, and domain-concepts that have to be precisely analyzed, validated and named by experts. Moreover, the approach identifies specific-concepts and proposes possible new concepts that factorize the original models. We have validated the scalability of our approach by applying it on two versions of the EIS-Pesticides model, versions containing about 125 classes. The results of our approach were analyzed, validated and used by A. Miralles, co-author of this paper, who has a dual expertise: computer science and spraying application techniques of pesticides (Miralles et al., 1994; Miralles and Polvêche, 1998; Miralles et al., 2011).

One of the major perspective to our work is to improve the GCM through the use of Relational Concept Analysis (RCA), which is an FCA extension that will allow us to work more precisely on the relationships (UML associations) between domain-concepts. In our running example, the use of RCA would enable factorizing the *Rainfall Instrumentation* and the *Groundwater Instrumentation* associations with a new association connecting the new domain-concept *cl_Device* with the *cl_MeasuringStation* class. Similarly, RCA would extract a new association between the new *cl_Data* class and *cl_MeasuringStation*, factorizing both *RainFall Information* and *GroundWater Information* associations.

Another perspective is the use of natural language processing techniques to improve the name-based description of elements (classes, attributes, roles, etc). The knowledge of semantic relations like hyperonymy, synonymy, or homonymy between terms will refine the analysis of domain-concepts.

REFERENCES

Altmanninger, K., Schwinger, W., and Kotsis, G. (2010). Semantics for accurate conflict detection in smover: Specification, detection and presentation by example. *International Journal of Enterprise Information Systems*, 6(1):68–84.

Altmanninger, K., Seidl, M., and Wimmer, M. (2009). A survey on model versioning approaches. *International Journal of Web Information Systems*, 5(3):271–304.

Arévalo, G., Falleri, J.-R., Huchard, M., and Nebut, C. (2006). Building abstractions in class models: Formal concept analysis in a model-driven approach. In *Model Driven Engineering Languages and Systems (MoDELS)*, pages 513–527.

Batini, C., Lenzerini, M., and Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM Computer Survey*, 18:323–364.

Bendaoud, R., Napoli, A., and Toussaint, Y. (2008). Formal Concept Analysis: A unified framework for building and refining ontologies. In *International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 156–171.

Birkhoff, G. (1940). *Lattice theory*. American Mathematical Society.

Cicchetti, A., Ruscio, D., and Pierantonio, A. (2008). Managing model conflicts in distributed development. In *Model Driven Engineering Languages and Systems (MoDELS)*, pages 311–325.

Cicchetti, A., Ruscio, D. D., and Pierantonio, A. (2007). A metamodel independent approach to difference representation. *Journal of Object Technology*, 6(9):165–185.

Dao, M., Huchard, M., Hacene, M. R., Roume, C., and Valtchev, P. (2006). Towards practical tools for mining abstractions in uml models. In *International Conference on Enterprise Information Systems: Databases and Information Systems Integration (ICEIS 2006)*, pages 276–283.

Doan, A. and Halevy, A. Y. (2005). Semantic integration research in the database community: A brief survey. *AI Magazine*, 26:83–94.

Falleri, J.-R. (2009). *Contributions à l'IDM : reconstruction et alignement de modèles de classes*. PhD thesis, Université Montpellier 2.

Formica, A. (2006). Ontology-based concept similarity in Formal Concept Analysis. *Information Sciences*, 176:2624–2641.

Ganter, B. and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundation*. Springer-Verlag Berlin.

Godin, R. and Mili, H. (1993). Building and maintaining analysis-level class hierarchies using galois lattices. In *Eighth annual conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA)*, pages 394–410.

Kalfoglou, Y. and Schorlemmer, M. (2005). Ontology mapping: The state of the art. In *Semantic Interoperability and Integration*.

Miralles, A., Gorretta, N., Miller, P. C., Walklate, P., Van Zuydam, R. P., Porskamp, H. A., Ganzelmeier, H., Rietz, S., Ade, G., Balsari, P., Vannucci, D., and Planas, S. (1994). Orchard sprayers : an european program to compare testing methods. In *International symposium on fruit nut and vegetable production production engineering, Valencia Zaragoza, ESP, 22-26 mars 1993*, pages 117–122.

Miralles, A., Pinet, F., Carluer, N., Vernier, F., Bimonte, S., Lauvernet, C., and Gouy, V. (2011). EIS-Pesticide: an information system for data and knowledge capitalization and analysis. In *Euraqua-PEER Scientific Conference, 26/10/2011 - 28/10/2011*, page 1, Montpellier, FRA.

Miralles, A. and Polvêche, V. (1998). Effects of the agrochemical products and adjuvants on spray quality and drift potential. In *5th International Symposium on Adjuvants for Agrochemicals - ISAA '98*, volume 1, pages 426–432, Memphis (USA).

Ohst, D., Welle, M., and Kelter, U. (2003). Differences between versions of uml diagrams. *SIGSOFT Software Engineering Notes*, 28:227–236.

Osman Guedi, A., Miralles, A., Huchard, M., and Nebut, C. (2011). Analyse de l'évolution d'un modèle : vers une méthode basée sur l'analyse formelle de concepts. In *XXIXème Congrès INFORSID*.

Parent, C. and Spaccapietra, S. (1998). Issues and approaches of database integration. *Communication of the ACM*, 41:166–178.

Pinet, F., Miralles, A., Bimonte, S., Vernier, F., Carluer, N., Gouy, V., and Bernard, S. (2010). The use of uml to design agricultural data warehouses. In *International Conference on Agricultural Engineering (AgEng 2010)*, pages 1–10.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 10:334–350.

Rouane, M. H., Dao, M., Huchard, M., and Valtchev, P. (2007). Aspects de la réinginierie des modèles uml par analyse de données relationnelles. *Ingénierie des Systèmes d'information (RSTI série)*, 12:39–68.

Shvaiko, P. and Euzenat, J. (2005). A Survey of Schema-Based Matching Approaches Journal on Data Semantics IV. In Spaccapietra, S. and Spaccapietra, S., editors, *Journal on Data Semantics IV*, volume 3730 of *Lecture Notes in Computer Science*, chapter 5, pages 146–171. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Stumme, G. and Maedche, A. (2001). Ontology merging for federated ontologies on the semantic web. In *International Workshop for Foundations of Models for Information Integration (FMII-2001)*, pages 413–418.

Tatsiopoulos, C. and Boutsinas, B. (2009). Ontology mapping based on association rule mining. In *International Conference on Enterprise Information Systems: Databases and Information Systems Integration (ICEIS 2009)*, pages 33–40.

A Service-based Integration for an improved Product Lifecycle Management

Stefan Silcher[1], Max Dinkelmann[2], Jorge Minguez[1] and Bernhard Mitschang[1]
[1]*Graduate School advanced Manufacturing Engineering (GSaME), University of Stuttgart, Stuttgart, Germany*
[2]*Institute of Industrial Manufacturing and Management (IFF), University of Stuttgart, Stuttgart, Germany*
{*stefan.silcher, jorge.minguez, bernhard.mitschang*}*@ipvs.uni-stuttgart.de, mmd@iff.uni-stuttgart.de*

Keywords: Product Lifecycle Management, Service-oriented Architecture, Modular IT Integration.

Abstract: The continuously changing environment is nowadays a major challenge for companies. The tough competition, growing customization of products and environmental regulations forces companies to continuously adapt their business processes. In order to manage the complexity and reduce the effort for developing products and production, many IT systems are indispensable. Despite Product Lifecycle Management Technology (PLM) the growing heterogeneous IT landscapes lack of a continuous support for business processes and get quickly unmanageable. In this paper PLM technology is extended by a service-based integration approach. Therefore, a modular service-based architecture was developed which will be presented in detail. The architecture describes how the whole product life cycle can be integrated more efficiently. The characteristics and findings of our approach are presented as well as a first prototype covering the production planning.

1 INTRODUCTION

Manufacturing companies face several challenges nowadays. Firstly, in some industry sectors there exists an overcapacity of produced goods, like in the automobile industry. There is a tough competition between companies for market share, price and quality of products. Secondly, the trend for customization is growing. Companies deliver highly individualized products to customers in order to increase their competitiveness. This results in an increasing complexity when developing the products and planning their production. Thirdly, environmental regulations, like reducing energy consumption and greenhouse gas or replacement of harmful materials, force producing companies to be highly innovative in developing new technologies, materials and processes.

These challenges lead together with a highly volatile market to new requirements for manufacturing companies. To consolidate or increase the market share, they have to enhance the adaptability of their company in many ways. Their production has to be as flexible as possible to produce highly customizable products in the same production line. Changes of the factory layout have to be performed in a very short time and often with increasing frequency in order to get shorter production times. Therefore, the exchange of data between the digital and physical factory has to be improved. This means, the actual state of the physical factory should be reflected in the model of the digital factory at any time. This enables a faster reaction on appearing problems or failures in the physical factory. Additionally, the organization has to be adaptive as well. Hence, changes in the business processes have to be smoothly performed (Jovane et al., 2009).

The first approaches to organize and provide the product data over their whole life cycle came with the Product Data Management (PDM). Introducing PDM systems enormously reduces the effort to handle product data, but the systems do not consider the processes in the product life cycle. Therefore, the concept of PDM was extended to monitor and manage these processes and Product Lifecycle Management was introduced (PLM). The goal is to optimize and standardize the processes to execute them efficiently and therefore save time and money (Saaksvuori and Immonen, 2008), (Stark, 2004).

Additionally, the tasks within a process are supported by software applications. This leads to a faster execution of single tasks and consequently reduces the execution time of the process. However, several problems arise when a high number of applications are installed. One of these problems is the emergence of information silos. The characteristic of such an IT landscape is often distributed, heterogeneous and proprietary. Coupling applications by implementing

point-to-point interfaces reduces the problem of information silos, but such solutions get quickly very complex and hard to manage and maintain.

The motivation of our work is to build an architecture, which provides the needed flexibility in the IT landscape of manufacturing companies. The processes within PLM have to be continuously supported by IT systems, these systems have to be loosely coupled to provide the needed flexibility. Existing applications should be integrated into the new architecture. Additionally, the management and maintainance effort of the infrastructure has to be reduced and the availability of data within the product life cycle improved.

To realize such an architecture, a flexible solution is needed to flexibly integrate the applications of the product life cycle. A commonly used paradigm today is the Service-oriented Architecture (SOA). SOA provides a flexible integration of applications by loose coupling and reusing services, which are self-contained, platform-independent and discoverable (Erl, 2005). The use of standardized technology like Web services to implement services allows easily maintaining, extending or exchanging of an interface. Moreover, Web services can be composed to workflows, which are modeled in standardized languages like Business Process Execution Language (BPEL) (OASIS, 2007) or Business Process Model and Notation (BPMN) (OMG, 2011). These workflows can be executed by corresponding workflow engines to support the business processes.

The flexible composition of Web services within workflows enables a flexible IT support of business processes, which is necessary to quickly adapt the business processes in a highly volatile environment. The platform-independence of Web services allows exchanging data and information between heterogeneous applications. Relying on standards and loosely-coupled applications simplify maintaining, changing and extending the IT infrastructure (Weerawarana et al., 2005).

A common middleware solution to integrate Web services is the Enterprise Service Bus (ESB) (Chappell, 2004). The ESB handles the routing of messages to enable a loose coupling of applications. Most ESB solutions possess a BPEL engine, which is able to execute BPEL workflows. Additionally, the ESB can offer functionality such as message queuing capabilities or monitoring services.

This paper presents a service-based integration approach for PLM. The developed architecture is based on an ESB hierarchy to integrate the product life cycle in a modular way. The results of the implemented prototype for the digital factory demonstrate the benefits of the service-based solution. The benefits include the flexible composition of IT tasks, implemented as Web services, in workflows. The workflows support the planner of the factory by automating recurring tasks and saves therefore time and money. Additionally, the management and extensibility effort of the implemented prototype is enormously enhanced compared to the previous solution with customized scripts.

The remainder of the paper is structured as follows: In Section 2, the challenges in PLM and the problems of inadequate integration of applications are discussed. Furthermore, the service-based approach for PLM integration is presented and the current solutions in application integration are described. The implemented prototype for the integration of production planning tools is depicted in Section 3 and benefits and problems as well as further extensions are discussed. In Section 4, related work is presented before summary and outlook are given in Section 5.

2 SERVICE-ORIENTED PLM ARCHITECTURE

An efficient and effective IT support of PLM processes is one of the major challenges in today's manufacturing companies. Using IT systems, especially in the product development and production planning phase, can reduce the time-to-market of new products tremendously. At the same time failures and ramp-up-time of production are reduced as well as the quality of products is improved. Today, the time between deciding to develop a new product and its production start can be further reduced by a seamless integration of the various tools in the product life cycle. These IT systems are often legacy applications or information silos difficult to integrate. In order to enhance the efficiency of production, the challenge is to enable an efficient exchange of information within the plethora of heterogeneous and distributed IT systems used during the product life cycle.

The adoption of the SOA paradigm eases the needed information exchange between heterogeneous applications by using standardized interfaces like Web services and decoupling service functionality from its implementation. Hence, a modular integration of the various product life cycle phases is presented in Subsection 2.2. Modularity reduces the implementation and maintenance effort of the IT infrastructure. Data exchange between the Web service as well as their compositions is accomplished by an ESB. Workflows are capable of automatically executing recurring tasks within PLM and therefore save valuable time, as well as reducing errors.

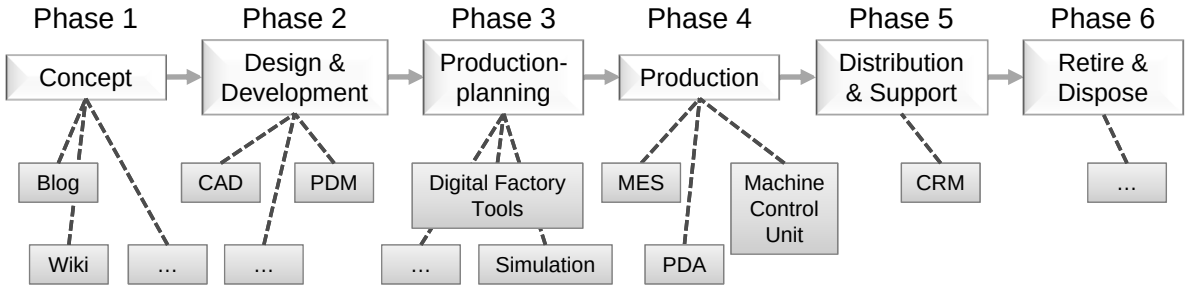


Figure 1: Phases and Tools of the Product Life Cycle.

In Subsection 2.1, the separation of the product life cycle in six phases, which is used in this paper, is presented. The service-based PLM architecture is described in Subsection 2.2. The ideas of a service-based PLM integration to production networks are extended in Subsection 2.3. Subsequently, the motivation for an improved data exchange between the digital and physical factory and the current state of the production planning are introduced in Subsection 2.4 and 2.5.

The first implementations within the Learning Factory's digital learning shell of the Institute of Industrial Manufacturing and Management (IFF) are presented in Section 3. This prototype realizes a seamless service-based integration of systems in the production planning and eases the data management thanks to automating the data exchange between the applications by implementing the process in a BPEL workflow.

2.1 Product Life Cycle

The product life cycle is heterogeneous in many ways. Therefore, it is split in different phases, each of them represent a characteristic activity. In Figure 1, the different phases of the product life cycle can be seen. The separation of the product life cycle considers not only the different activities in each phase, but also the support with tools and the management of data.

In the first phase, the 'concept' phase (C), ideas for new products are generated and market analyses are prepared. Therefore, wikis, blogs or social networks are used and often unstructured data has to be handled and interpreted.

The 'design & development' phase (DD) is determined by developing and designing the various parts of the products. Tools like Computer-aided Design (CAD) and Product Data Management systems (PDM) manage the high volume of data within this phase.

In the next phase, the 'production planning' (PP) is carried out, where process and resource data are added to the product data and linked among each other. This data is linked into Digital Factory Tools, which generate a plan for the factory. The planned production processes are simulated with simulation tools to verify the sequence. The data volume, which has to be handled, increases enormously in this phase compared to the design and development phase.

Having the product development and production planning completed, the production can start in phase four, the 'production' phase (P). Here, Manufacturing Execution Systems (MES) are responsible for applying the production orders in the production. The control unit provides a real-time control of the manufacturing facilities. Feedback of the manufacturing facilities is gathered and stored by the production data acquisition unit (PDA) for later analysis.

Selling the products, their maintenance and the customers comprise are the main tasks of phase five, the 'distribution & support' phase (DS). Customer Relationship Management Systems (CRM) are used for the linking of sold products and customers for any kind of complaints or warranty issues.

The 'retire & dispose' phase (RD) deals with the disposal of the product. Information about assembly and material composition of the product can help regain valuable material or simplify the separation of material for recycling.

2.2 Service-oriented Integration

PLM extends the concept of PDM by adding management and control of business processes for the whole life cycle. The problem today is that IT systems poorly support business processes. Especially the flexible IT support of business processes is of great importance (Papazoglou et al., 2007). Therefore, a service-based architecture is needed, which allows to implement a continuous IT support along the whole product life cycle. Additionally, business processes must be adapted in an easy manner as part of the IT infrastructure.

Therefore, a modular service-based architecture

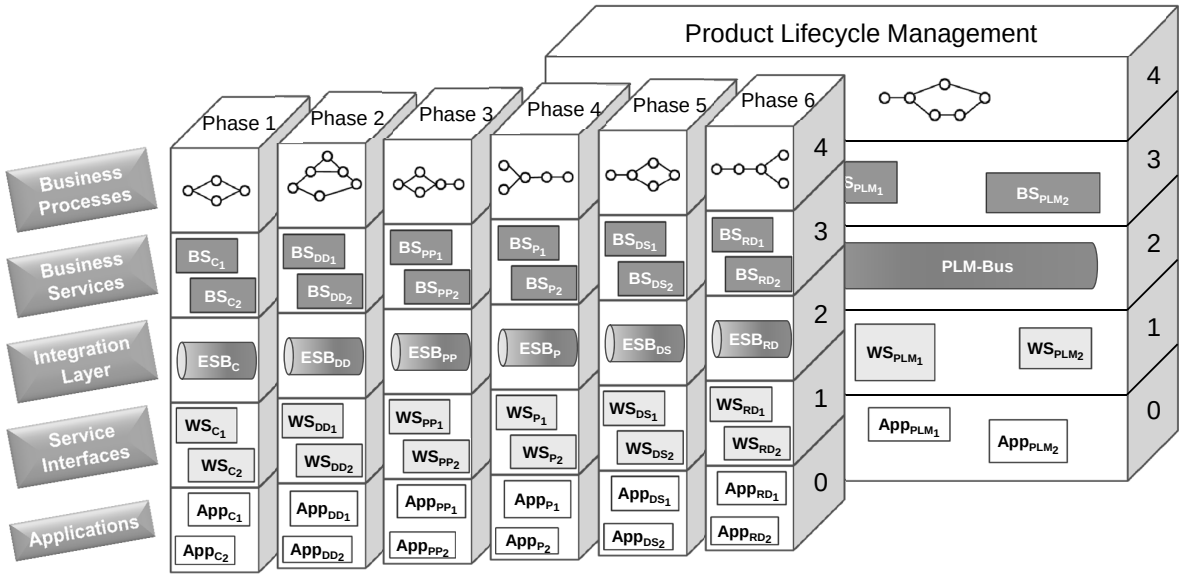


Figure 2: Service-based Integration Architecture for Product Lifecycle Management.

was developed with different integration layers, to efficiently integrate the plethora of applications, used in the product life cycle, as well as to flexibly support the business processes. The developed architecture, which is presented in Figure 2, uses a pillar for every phase of the product life cycle. Furthermore, the six phases are integrated with an additional pillar for the overall PLM (Silcher et al., 2011).

Each pillar is represented by a phase-specific ESB that integrates all the applications of the corresponding phase. These phase-specific characteristics can be supported directly. A distinction is made between phase-specific ESBs and the phase-overlapping PLM-Bus. This solution with multiple ESBs builds a hierarchy to efficiently manage all the processes in the whole product life cycle.

The vertical integration of each product life cycle phase is clearly structured by distinguishing different levels of abstraction (Minguez et al., 2010). Hence, the functionality and data, provided by each application, is exposed as Web service in small, functional units. These units can be composed to workflows, which support the execution of business processes. The five layers are explained in Subsection 2.5 in more details.

The horizontal, phase-overlapping integration is performed by the PLM-Bus, which connects the phase-specific ESB's as a central backbone. The separation of phase-specific integration and holistic PLM integration contains several advantages (Silcher et al., 2011). Particularly, the possibility of adapting the ESB to the requirements of each phase like availability, data throughput and time requirements are of great importance.

The PLM-Bus is responsible to manage the phase-overlapping data exchange, tasks and processes. The PLM-Bus should provide information about available Web services of all phases as well as authorization and authentication as a single sign-on service. Processes like change and failure management have to be coordinated by the PLM-Bus. Additionally, Enterprise Resource Planning systems and other applications, which cannot be assigned to a specific phase, should be directly integrated by the PLM-Bus.

2.3 SOA for Production Networks

Previously not considered was the fact, that most companies are not involved in the whole product life cycle. Only the Original Equipment Manufacturers (OEM's) treat the product life cycle from the development to the recycling of the product. Small and Medium-sized Enterprises (SMEs), in particular suppliers of OEMs, focus their work on one or two phases of the presented product life cycle. Therefore, a phase-specific integration is often sufficient to cover their whole business area. This is not only profitable for the SMEs, which benefit from the advantages of a service-based IT infrastructure. Especially for the OEM's, the change of the SMEs IT landscape to an SOA can be valuable.

A service-based communication improves the quality of the data exchange, due to standardized Web service interfaces, workflow definition or data exchange formats. Additional, asynchronous communication eases the loose coupling systems of the com-

panies.

Therefore, the integration of SMEs in the business process of an OEM is useful to have a defined way of communication between the companies. E.g. changes in the design of a product, made by an SME, can be necessary due to inconsistencies in assembly tests of the OEM. The OEM can describe the problems and send them to the SME, where automatically an exposed process is triggered to solve the resulting problem.

Therefore, the presented architecture can be easily extended to manage data and processes in production networks covering all involved companies, e.g., from OEM to all suppliers. The challenge would be to convince all suppliers and customers of a company to migrate their IT infrastructure to SOA. Beside the technical challenge it poses an organizational challenge.

2.4 Digital and Physical Factory

Today, one key to a more efficient factory is the coupling of the digital and physical factory. The vision is to have available an up-to-date digital copy of the physical factory at any time. Based on the current state, different simulations could be executed to forecast the short and medium term development of the factory and its production.

The problem is to provide status information of the production environment in real-time to the digital factory in order to automatically run a simulation model out of that data (Kádár et al., 2010). The presented architecture improves the availability of data in the production environment by exposing Web service interfaces and loosely coupling of applications in the infrastructure, which allows a faster data exchange between the digital and physical factory.

The IFF at the University of Stuttgart has built a Learning Factory, which contains a digital learning shell and a physical factory. The digital learning shell contains many tools for planning and optimizing the digital model of the factory. The central application is a database to store product, process and resource (PPR) data, also called PPR-Hub. Around this PPR-Hub, there are several tools for the factory optimization. Amongst others, there is a layout planning table, where a group of people can cooperatively optimize the factory layout (Kapp et al., 2005). The most important tool in a digital factory is a simulation application. Additionally, a logistic simulation tool is used to verify the planned production processes and various simulated scenarios (Kapp et al., 2003).

The presented tools and others can exchange data with the PPR-Hub by executing (Visual Basic) scripts. These scripts are customized point-to-point connections. Therefore, the inclusion of new tools or substitution of an existent tool is associated with a great effort. Substituting a tool leads to a complete reimplementation of the script.

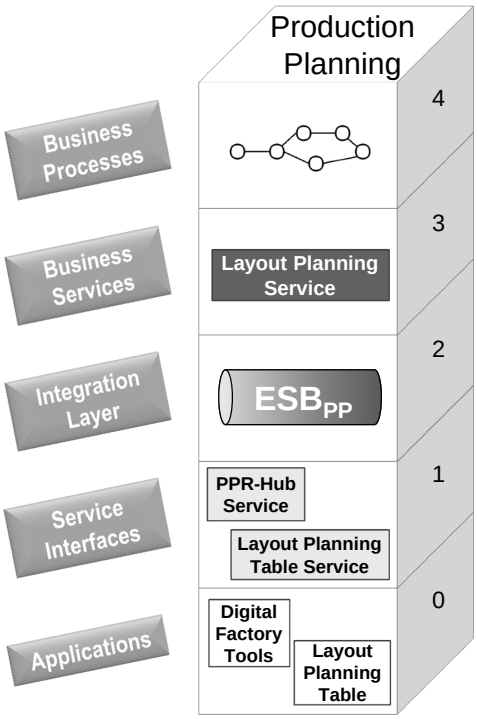


Figure 3: Integration Pillar for the Production Planning Phase.

The physical factory of the IFF, called iTRAME, consists of modular manufacturing units, which are connected with a universal plug-and-play mechanism and can be easily exchanged (Dinkelmann et al., 2011).

Currently, the factory layout of the physical factory can be automatically detected and copied with a script to the digital planning environment. Due to the high effort for implementing these scripts, no further information is used in the planning environment, e.g. process data and time information of the production. This information would help to understand the current situation in the production, when planning and real data are compared, and thus would enhance the planning accuracy, but also increases the planning effort.

The goal of implementing an SOA is to replace the tight coupling of the applications with scripts by Web service interfaces, which can be flexibly composed into workflows. The workflows can support the planner by automatically or semi-automatically executing recurring tasks in the product life cycle, e.g. when new products are introduced, changes in the product mix are detected, or machine failures appear.

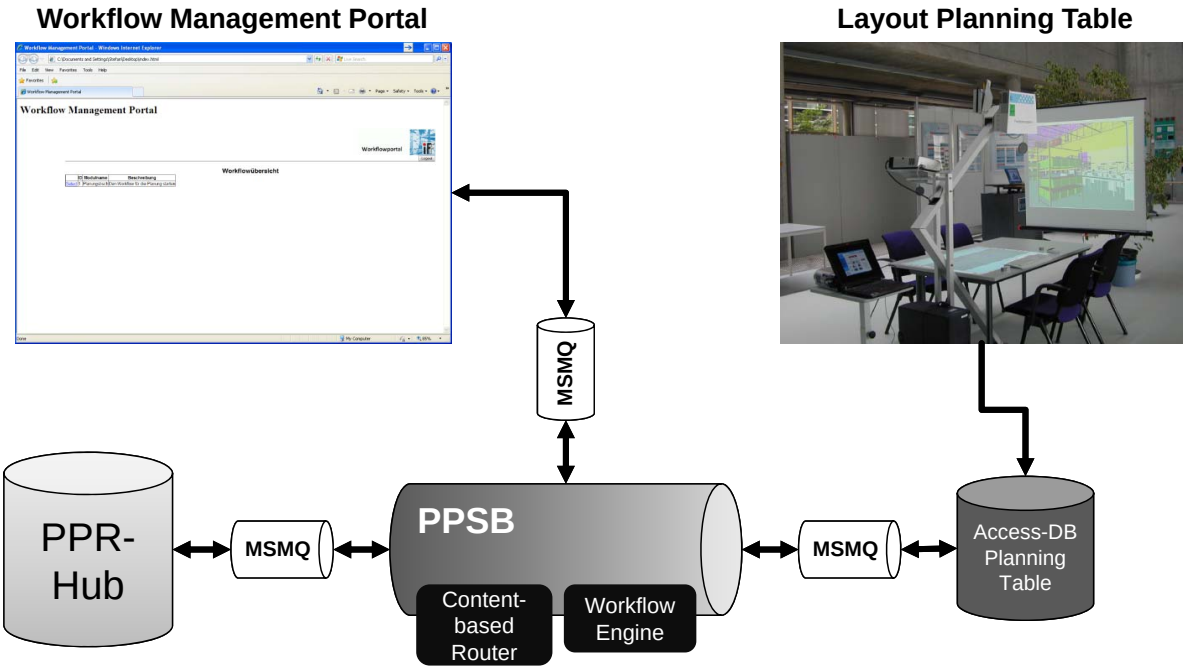


Figure 4: Architecture of Integrated IT systems.

2.5 Production Planning

Implementing the presented service-based architecture for PLM in the production planning is figured out in Figure 3.

Layer 0 contains the various applications of the production planning, e.g. digital factory tools, layout planning table, simulation tools. The data and functionality of the applications is exposed to other tools by means of Web service interfaces, which are placed in Layer 1. The ESB in the integration layer (Layer 2) routes the messages sent and received by the Web service interfaces to the desired destination. Additionally, it performs the data transformations from proprietary data formats to a canonical data format of the ESB, which is described in Subsection 3.3. The business services in Layer 3 compose the Web services to small processes, which manage the data exchange between two applications. The business process layer at the top of the pillar (Layer 4) represents the IT implementation of the business processes.

Web service technologies are platform independent, which is a great advantage when integrating proprietary heterogeneous systems like the current IT infrastructure at the IFF. In the implemented prototype, the messages are exchanged over Message Queues (MQ), which allow a reliable, asynchronous communication between the participating applications and improve the loose coupling of applications (Hohpe and Woolf, 2003). More details on our prototype implementation will be given in the subsequent chapter.

3 IMPLEMENTATION OF SERVICE-BASED INTEGRATION FOR PRODUCTION PLANNING

To show the benefits of the service-based approach, some applications of the Learning Factory's digital learning shell were integrated by the presented approach and architecture. Therefore, the databases of each application were equipped with Web service interfaces and a web-based portal was developed to control and manage workflows executed in the production planning environment.

3.1 Prototype of Service-based Production Planning

In the first phase of the implementation, the central data hub, which stores the PPR data, should be integrated with a layout planning tool (Kapp et al., 2005). The developed architecture is shown in Figure 4.

The core of this architecture is the Production Planning Service Bus (PPSB), which is based on the OpenESB, an open source implementation of an ESB (OpenESB, 2010). BPEL workflows can be executed

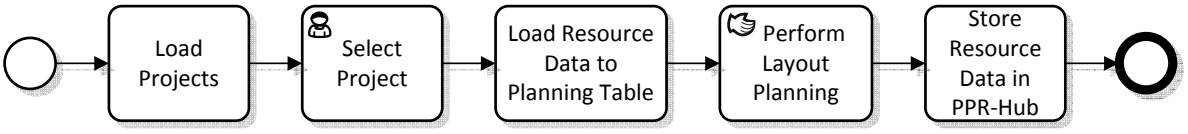


Figure 5: BPMN Workflow for Layout Planning.

in the workflow engine of the OpenESB. To administrate the workflows, the Workflow Management Portal was developed, where the production planner can start, control and if necessary restart or stop the available workflows. The usage of the implemented planning workflow is described in Section 3.2.

Using asynchronous communication improves the loose coupling between the applications. They don't communicate directly with each other; instead the messages are sent over MQs. A MQ enables a reliable communication between the participants by receiving messages and storing them persistently until the MQ has successfully delivered the message to the receiver. This means, the receiver can be temporarily unavailable without disrupting the whole system or blocking the sender of a message. Therefore, the usage of MQs makes the system more robust against network or application failures and improves the loose coupling of the applications.

In the presented prototype, Microsoft Message Queuing (MSMQ) is used to connect the systems with the OpenESB. To enable the OpenESB to communicate with the MSMQs, a MSMQ Binding Component is provided by the developers of the OpenESB.

To integrate the applications within the architecture, Web service interfaces were developed to enable the exchange of messages with other applications. For the Workflow Management Portal, PPR-Hub and Layout Planning Table a Web service interface was implemented for each. Due to the fact, that no source code was available for the Layout Planning Table, the only way to exchange data with this tool was an interface to its access database.

All the messages exchanged in the system rely on the same canonical data format. Thus, all the messages have the same structure and don't have to be transformed between the different systems. The only effort is to generate a correct message, including the data provided by the database. To reduce the effort, a common library that automatically generates the messages in the correct format is implemented to be used by all Web services.

The loose coupling of the integrated applications is ensured by using a content-based router (Hohpe and Woolf, 2003). This means, an application must not know the network address or endpoint of the destination system of a message. The message can just be sent to the ESB, where the content-based router determines the destination by inspecting the content of the message. Therefore, the endpoint can be looked up in a database, when the destination system is known. This allows transferring applications to other servers or changing their endpoints without affecting other applications. The only thing to do is to change the corresponding endpoint in the database of the content-based router and all other applications can communicate again with the changed application. The purpose of the content-based router is to decouple applications and to enhance adaptability.

3.2 Layout Planning Workflow

The workflow implemented for the prototype controls the data flow between the systems presented in the previous subsection. A BPMN model of this workflow is presented in Figure 5.

The first task "Load Projects" of the workflow sends all projects, which are stored in the PPR-Hub, in a message to the Workflow Management Portal. The message is extracted and the projects are presented in the portal.

The user is asked in the second task "Select Project", to select the project he wants to modify in the Layout Planning Table. The human icon in the top left corner of the task indicates that a human interaction is necessary in this task.

After selecting a project, the third task "Load Resource Data to Planning Table" is started, which sends a message containing the selected project to the PPR-Hub. The resource information of this project is thereupon sent over the ESB to the Layout Planning Table and stored in its database.

In the fourth task "Perform Layout Planning", the user can perform the layout planning to optimize the material flow, the logistic processes, and so forth. The hand icon in the top left corner of the task indicates that this task has to be performed manually by the user.

When the user finished the layout planning, the last task "Store Resource Data in PPR-Hub" is executed. This task generates a message to send the optimized planning data over the ESB back to the PPR-Hub, where they are stored and are now available for other systems in the production planning.

3.3 Message Format

To efficiently exchange messages between the different applications, an application-independent canonical data format was defined. Using a canonical data format is best practice in integration, because it reduces the complexity of data formats when integrating new applications into the infrastructure (Chappell, 2004). Thus, the data of each application has to be transformed to and from the canonical format, instead of one transformation from each to every application. This reduces the complexity of integration and improves the extensibility and scalability of the integration approach.

The defined message format consists of two parts and is based on XML. The first part contains information about the message type and the routing. The content-based router derives the destination of the message from the message type and writes them in the routing information part of the message.

The second part of the message contains the data of a project. Currently, this part entails information about the production resources, but can be easily extended with product, process and order information in the future, when more applications are integrated into the architecture.

```
<?xml version="1.0" encoding="utf-8"?>
<ProjectMessage xmlns:xsi="...>
  <CommonInfos xmlns="http://tempuri.org/ESB">
    <RoutingInfos>
      <OriginSystem>PPR-Hub</OriginSystem>
      <DestinationSystem>
        <SystemID>WP</SystemID>
        <SystemURI>http://localhost:9007/
                  PortalService/</SystemURI>
      </DestinationSystem>
    </RoutingInfos>
    <MessageType>0</MessageType>
       .   .   .
  </CommonInfos>
  <Projects xmlns="http://tempuri.org/ESB">
    <Resources>
      <ModuleTypes>
        <ModuleID>1</ModuleID>
        <Modulename>modulename</Modulename>
        <Type>1</Type>
        <FileName>test.cad</FileName>
        <BitmapName>test.bmp</BitmapName>
        <Module>
          <ObjectID>1</ObjectID>
          <Objectname>objectname</Objectname>
          <Position>
            <Position_X>300</Position_X>
            <Position_Y>80</Position_Y>
            <Position_Z>0</Position_Z>
            <Rotation_X>0</Rotation_X>
            <Rotation_Y>0</Rotation_Y>
            <Rotation_Z>0</Rotation_Z>
          </Position>
        </Module>
        <Module>
          .   .   .
        </Module>
      </ModuleTypes>
    </Resources>
    <ProjectID>1</ProjectID>
    <Projectname>test project</Projectname>
  </Projects>
</ProjectMessage>
```

The messages are generated by the Web services, which read the data from the database of the source system and fill them into the XML schema. After completing the message, it is sent to the router of the PPSB, where the destination is derived from the message content and forwarded to the destination Web service interface. After receiving the message, the information is extracted from the XML message and stored in the database of the destination system.

3.4 Review and Extension of the Prototype

For the manageability of our IT infrastructure approach, this prototype demonstrates great advantages compared to the previously implemented point-to-point interfaces between the applications:

- The ESB as central integration backbone eases the connection of the heterogeneous applications to this prototype.
- The use of a canonical data format reduces the number of different transformations, which leads to a better extensibility and scalability of the infrastructure.
- The content-based router enables the loose coupling of the applications to the ESB by introducing a central database for the registration of application endpoints.
- The implementation of MQs boosts the loose coupling at the connectivity layer. Additionally, the robustness of the complete infrastructure is improved due to temporarily failures of networks or applications.
- Changing from a synchronous to an asynchronous communication increases the performance by reducing unnecessary blocking of applications when waiting till a message is send or is available to receive.

The presented prototype allows the production planner to manage the data of the integrated applications at a single point, the Workflow Management

Portal. In the portal, the planner can see the available workflows and start, stop or restart them.

Proprietary applications can be integrated in various ways into an SOA. To access the functionality of a program, an available interface can be used or a new one can be implemented, provided that the source code is available. If neither is available, the functionality of a program cannot be easily integrated in a workflow, as in our case. Nevertheless, the data can be accessed by an interface over the program logic or directly over the database of a program. In the prototype the databases of the integrated programs were equipped with an Web service interface which lead to an enhances data exchange between the applications. To fully automate the planning processes, functionality has to be exposed as Web services to be able to execute them within a workflow activity.

Extending the prototype with a simulation tool like the logistic simulation tool would make sense. The optimized layout could be verified by simulating the production processes and the throughput can be measured. Therefore, the simulation tool has to be equipped with a Web service interface to receive the necessary data for the simulation.

The canonical message format has to be extended to include beside the resource data also product and process data. In the currently used data format, this extension is already provided and can be easily performed. Additionally, the Web service interface of the PPR-Hub has to be extended to read and write the product and process data. On the other hand, the Web service interface of the Layout Planning Table remains unchanged. Now, there are two alternatives to integrate the simulation tool in a workflow. The presented layout planning workflow can be extended to include the simulation tool or a new simulation workflow can be implemented to control the data exchange between the PPR-Hub and the simulation tool. In the second case, the layout planning workflow and the new simulation workflow have to be executed consecutively.

4 RELATED WORK

In the last few years, the main PLM vendors like Dassault Systèmes, PTC and Siemens PLM Software extended their PLM solutions with a service-based approach to get the desired continuous integration of the product life cycle (CIMdata, 2006). However, they adopt their proprietary integration middleware and thus resulted in restricted interoperability properties: lacking to integrate systems of other vendors, missing flexibility in business process support, and applications are not loosely coupled to the integration middleware. Furthermore, the interfaces are not open, so it is hard or impossible for other software vendors to connect their applications to these middleware systems.

Rantzau et al. implemented a Data Change Propagation System called CHAMPAGNE for heterogeneous information systems (Rantzau et al., 2002). The CHAMPAGNE platform manages dependencies between the schemas of different distributed applications. Compared to the presented prototype in this paper, CHAMPAGNE implements a tight coupling to the participating applications. Hence, changes in a system can only be made when the propagation scripts are changed accordingly, which my become quite some hassle.

5 CONCLUSIONS AND OUTLOOK

Highly volatile markets and growing competition force companies to continuously increase their effectiveness. Their flexibility has to be improved to adapt to the constantly changing environment. This can be achieved by a more flexible support of business processes and the IT infrastructure. Additionally, the applications, which support a business process task, have to be better integrated to improve the data and information flow. The vision is a continuous integration of all applications used in the product life cycle to accelerate the data exchange.

The paper presents a service-based architecture to integrate the different phases of the product life cycle. The phase-overlapping integration is performed by the PLM-Bus, which allows exchanging data and coordinating processes between the phases. The benefits of this architecture are a clear separation between different levels of abstraction as well as the possibility to adapt each ESB to the requirements of each phase like availability, data throughput and time requirements. Possible extensions to integrate customers and suppliers in this infrastructure to get a consistent information exchange in the production network were discussed.

Furthermore, the developed prototype based on the Production Planning Service Bus performs a service-based integration of the production planning environment at the IFF. The benefits and problems of the prototype and the integration of proprietary applications are discussed and an outlook on useful extensions of the implementation in the production planning is given.

The next step in the service-based integration of

PLM is the implementation of the PLM-Bus to efficiently couple the production planning and production phase. The goal is to establish a bidirectional communication between the digital and physical factory to automatically adopt the current production status for the planning and to accomplish an optimized planning in the production environment.

ACKNOWLEDGEMENTS

The authors extend their sincere thanks to Fabian Laux, who contributed in developing and implementing the prototype of the service-based production planning integration.

Furthermore, the authors would like to thank the German Research Foundation (DFG) for financial support of this project as part of the Graduate School of Excellence advanced Manufacturing Engineering (GSaME) at the University of Stuttgart.

REFERENCES

Chappell, D. A. (2004). *Enterprise Service Bus: Theory in Practice*. O'Reilly Media, 1st edition.

CIMdata (2006). Service-oriented architecture for plm - an overview of ugs soa approach. Technical report, CIMdata, Inc.

Dinkelmann, M., Riffelmacher, P., and Westkmper, E. (2011). Training concept and structure of the learning factory advanced industrial engineering. In ElMaraghy, H. A., editor, *Enabling Manufacturing Competitiveness and Economic Sustainability*, Proceedings of the 4th International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV2011), pages 624–629. Springer Berlin Heidelberg.

Erl, T. (2005). *Service-Oriented Architecture: Concepts, Technology, and Design*. Prentice Hall International, illustrated edition edition.

Hohpe, G. and Woolf, B. (2003). *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley Longman, Amsterdam.

Jovane, F., Westkämper, E., and Williams, D. (2009). *The ManuFuture Road: Towards Competitive and Sustainable High-adding-value Manufacturing*. Springer, Berlin, 1st edition.

Kádár, B., Lengyel, A., Monostori, L., Suginishi, Y., Pfeiffer, A., and Nonaka, Y. (2010). Enhanced control of complex production structures by tight coupling of the digital and the physical worlds. *CIRP Annals - Manufacturing Technology*, vol. 59(1):437–440.

Kapp, R., Le Blond, J., and Westkämper, E. (2005). Fabrikstruktur und logistik integriert planen: Erweiterung eines kommerziellen werkzeugs der digitalen fabrik fr den mittelstand. *wt Werkstattstechnik online*, vol. 95 (2005), No. 4:191–196 (german).

Kapp, R., Löffler, B., Wiendahl, H.-H., and Westkämper, E. (2003). Der logistik-prüfstand: Skalierbare logistiksimulation von der lieferkette bis zum arbeitsgang. *wt Werkstattstechnik online*, vol. 93 (2003), No. 1/2:31–38 (german).

Minguez, J., Ruthardt, F., Riffelmacher, P., Scheibler, T., and Mitschang, B. (2010). Service-based integration in event-driven manufacturing environments. In *WISE 2010 Workshops*, volume 6724 of *Lecture Notes in Computer Science*, pages 295–308. Springer.

OASIS (2007). Web services business process execution language version 2.0.

OMG (2011). Business process model and notation (BPMN) version 2.0.

OpenESB (2010). http://www.logicoy.com/esb.php (last visited: November 2011).

Papazoglou, M., Traverso, P., Dustdar, S., and Leymann, F. (2007). Service-oriented computing: State of the art and research challenges. *Computer*, vol. 40(11):38–45.

Rantzau, R., Constantinescu, C., Heinkel, U., and Meinecke, H. (2002). Champagne: data change propagation for heterogeneous information systems. In *Proceedings of the 28th international conference on Very Large Data Bases*, VLDB '02, pages 1099–1102. VLDB Endowment.

Saaksvuori, A. and Immonen, A. (2008). *Product Lifecycle Management*. Springer, Berlin, 3rd edition.

Silcher, S., Minguez, J., and Mitschang, B. (2011). Adopting the manufacturing service bus in a service-based product lifecycle management architecture. In *Proceedings of the 44th International CIRP Conference on Manufacturing Systems: ICMS '11; Madison, Wisconsin, USA*, pages 1–6. Online.

Stark, J. (2004). *Product Lifecycle Management: 21st Century Paradigm for Product Realisation (Decision Engineering)*. Springer, Berlin, 1st edition.

Weerawarana, S., Curbera, F., Leymann, F., Ferguson, D. F., and Storey, T. (2005). *Web Services Platform Architecture: Soap, WSDL, WS-Policy, WS-Addressing, WS-Bpel, WS-Reliable Messaging and More*. Prentice Hall International, USA.

Bayesian Networks for Matcher Composition in Automatic Schema Matching

Daniel Nikovski[1], Alan Esenther[1], Xiang Ye[1], Mitsuteru Shiba[2] and Shigenobu Takayama[2]
[1]*Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139, U.S.A.*
[2]*Mitsubishi Electric Corporation, 5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, Japan*
{*nikovski, esenther, ye*}*@merl.com,* {*Shiba.Mitsuteru@ap, Takayama.Shigenobu@db*}*.MitsubishiElectric.co.jp*

Keywords: Data Integration, Virtual Databases, Uncertain Schema Matching.

Abstract: We propose a method for accurate combining of evidence supplied by multiple individual matchers regarding whether two data schema elements match (refer to the same object or concept), or not, in the field of automatic schema matching. The method uses a Bayesian network to model correctly the statistical correlations between the similarity values produced by individual matchers that use the same or similar information, in order to avoid overconfidence in match probability estimates and improve the accuracy of matching. Experimental results under several testing protocols suggest that the matching accuracy of the Bayesian composite matcher can significantly exceed that of the individual component matchers.

1 INTRODUCTION

The problem of automatic schema matching (ASM) between two or more database schemas arises in many applications, such as data migration, when one database has to be incorporated into another, virtual databases, where a single interface is used to access multiple databases, and data analysis, when multiple databases are stored in a data warehouse with a single schema. When two database schemas that describe the same problem domain are given (e.g. purchase orders, real-estate listings, books, etc.), the objective of an automatic schema matching (ASM) method is to discover which pairs of elements from the two schemas are likely to match, that is, likely to refer to the same entity (e.g. shipping address, house price, book title, etc.), and possibly to also estimate the confidence of such a match.

The ASM problem is usually very difficult, because when database designers create database schemas, they rarely provide full and unambiguous information about what individual schema elements represent. Even if any such information exists, it is usually not meant for computer processing. Rather, database designers usually choose suitable words or abbreviations for the names of data elements, so as to facilitate future maintenance of the data schemas by themselves or other humans. Because of this common practice, lexical analysis of the names of data elements could be an effective approach to ASM. For example, the names "Street", "Str", and "StreetName" can be recognized to refer to a street, possibly in an address, and lexical analysis by string matching can reveal this similarity. A different type of information that might be useful for ASM is the structure of the data schemas, if present. In many cases, schemas are not represented by a flat list of element names, but the elements are organized in a hierarchy. For example, the element "CustomerName" might have three subelements, "FirstName", "MiddleInitial", and "FamilyName". Using such structural information is another approach to ASM. Many more approaches exist: for example, when the actual values of two database fields come from the same statistical distribution (e.g., over names, numbers, etc.), this can serve as evidence that the corresponding schema elements match. Dictionaries, thesauri, and other auxiliary data sources have been used for ASM purposes, too (Rahm and Bernstein, 2001).

Due to the difficulty of the problem, no single method has been shown to perform best on all ASM tasks. This has led to the idea that multiple basic matchers of the types described above can be used together in a composite matcher (Do and Rahm, 2002; Tang and Li, 2006). The purpose of the composite matcher is to combine the output of the individual matchers and arrive at a more accurate set of likely matches. In most cases, the output of an individual matcher k for a given pair of elements $S_1.E_i$ and $S_2.E_j$ is a similarity value v_k in the interval $[0,1]$, where

$v_k = 0$ means no similarity, and $v_k = 1$ means full confidence that the two elements match. When given a library of K different individual matchers, the objective, then, is to find a composite similarity measure v that is a function of the individual outputs $v_k, k = 1, K$.

Several methods for combining similarity values have been proposed. The LSD system (Doan et al., 2003) uses machine learning techniques to estimate weighting coefficients w_k such that the final similarity measure v is a weighted average of the individual similarity measures: $v = \sum_{k=1}^{K} w_k v_k$. The COMA system (Do and Rahm, 2002) extends this approach with the minimum and maximum operators: $v_{min} = \min_k w_k v_k$ and $v_{max} = \max_k w_k v_k$.

Although experimental results suggest that these methods for combining similarity values lead to matching accuracy that is higher than that of the accuracy of the individual matchers, it can be recognized that they are specific approaches to the fundamental problem of combining evidence from multiple sources (in this case, multiple individual matchers), and make very specific assumptions about the statistical structure of the evidence. These assumptions might or might not be warranted in practice. We propose a general method for correct modeling of any kind of statistical structure in the evidence, based on Bayesian networks and probabilistic reasoning, and a statistically grounded method for composing matcher evidence using these Bayesian networks.

2 BAYESIAN NETWORKS FOR COMBINING OUTPUTS OF MULTIPLE SCHEMA MATCHERS

When combining evidence from multiple sources, one of the major problems and causes for errors is the improper modeling of correlation and other forms of statistical dependence between variables in the problem domain. For example, when two very similar matchers k and l are applied to an ASM problem, their outputs v_k and v_l will be highly correlated — when v_k is high, then v_l will be high, too, and vice versa. For example, a lexical matcher based on edit (Levenshtein) distance would assign a medium-level similarity to the pair of element names “Street” and “State”; similarly, a lexical matcher based on the Jaccard distance between the sets of letters in the two elements would assign such similarity to the pair. For another pair of elements, for example “Street” and “Address1”, both lexical matchers would compute low similarity, becasue in this case similarity cannot be established on the basis of string matching. In either case, not only is the computed similarity misleading as regards to the correct match, but both matchers provide the same kind of evidence (both positive or both negative), so its (in this case, harmful) influence is reinforced. If a weighted sum of the two similarity values is used, the same evidence will be counted twice, in practice, which will result in a phenomenon known as overconfidence. One of the matchers is almost redundant, and including it in the composition process might actually decrease the accuracy of matching. This effect has been observed in other fields where evidence has to be combined, such as medical diagnosis, and one possible tool for handling it has been belief reasoning in Bayesian networks. Our method for combining matcher output is based on such a network.

2.1 Representation

A Bayesian network (BN) is a probabilistic graphical model that represents a set of random variables and their conditional dependencies by means of a directed acyclic graph (DAG). An edge in the DAG between two nodes signifies that the variable Y corresponding to the child node is statistically conditionally dependent on the variable X corresponding to the parent node. This dependence is expressed in a conditional probability table (CPT) stored in the child node for Y. If $X \in Par(Y)$, where $Par(Y)$ is the set of parent nodes of Y, this table contains probability entries $Pr(Y = y|Par(Y) = z)$ for every possible combination of values x that X can take on and configurations (sets of values) z that the variables in $Par(X)$ can take on. Likewise, when there is no direct edge between two nodes, they are assumed to be conditionally independent given their parents. In particular, when two nodes have a common parent, but no edge between them, they are assumed to be conditionally independent given the value of their parent. The presence (or absence) of edges in the DAG of a Bayesian network is a way to express the statistical dependence (correlation) between variables.

A Bayesian network to be used for combining outputs of individual matchers in an ASM task is shown in Figure 1. Its DAG is a tree of depth four, with some additional edges between some of the nodes. The meaning of the nodes is as follows:

1. At the first (top) level, the root node corresponds to a Boolean variable signifying whether two schema elements match. This is the final hypothesis that has to be evaluated.
2. The nodes at the second level of the trees represent independent ways in which the two element

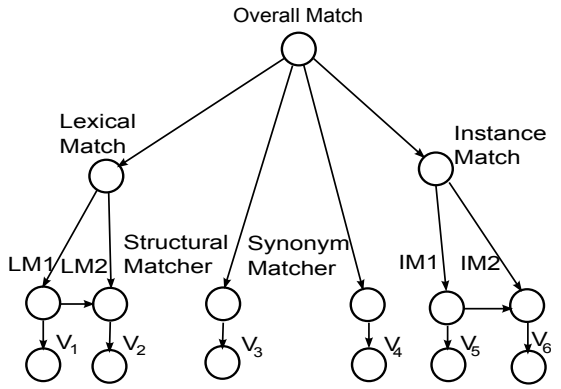


Figure 1: A Bayesian network for combining the output of multiple individual matchers.

names can match (lexical, structural, instance-based, etc.). It is expected that these variables are largely uncorrelated, because they use different information to test for possible matches. They also each correspond to clusters of individual matchers whose output is correlated. In Figure 1, one cluster represents the hypothesis that the two elements match lexically, and the other cluster represents the hypothesis that the instances (values) of the two elements in their respective databases match.

3. The nodes at the third level of the tree are also Boolean and represent the individual hypothesis that the two elements match, according to a single matcher. In Figure 1, these include two lexical matchers LM1 and LM2, one structural matcher, one synonym matcher, and two instance matchers IM1 and IM2.
4. The leaves of the tree, at the fourth level, represent the similarity values V_k, $k = 1,K$ of the individual matchers whose outputs have to be combined (in this case, for the sake of illustration, $K = 6$). These variables are continuous, and their possible values are the real numbers v_k.

The overall structure of the BN expresses the understanding that when two elements match (or don't), the outputs of the structural matcher, synonym matcher, the lexical match variable, and the instance match variable will be statistically independent. This is what is to be expected on a matching task, because these matchers all use different information from the two data schemas in order to compute an estimate about whether the elements match. However, the outputs of the two lexical matchers LM1 and LM2 would be correlated, as expected if they use the same information (the names of the two elements). That is why there exists an edge between nodes LM1 and LM2. Similarly, the output of the two instance matchers would be correlated, too, because they would both use the same information to base their estimates on (namely, the contents of the two corresponding database fields). Accordingly, an edge between nodes IM1 and IM2 reflects this dependency. This structure of the BN, then, corresponds to our understanding of which matchers produce highly correlated outputs, and which ones are statistically independent.

2.2 Parameter Estimation

In addition to the graph of the BN, if the network is to be used for inference, the parameters in its CPTs have to be specified, too. This can be done by means of labeled cases, where pairs $e_l = (S_1.E_i, S_2E_j)$ of elements $S_1.E_i$ and $S_2.E_j$, $l = 1,\ldots,N$ have been run through all K matchers, to produce the corresponding similarity values $v_{l,k}$, $l = 1,\ldots,N$, $k = 1,\ldots,K$, and the correct labeling for some or all of the remaining Boolean variables has been supplied, too.

If labels for all Boolean variables have been supplied, then the estimation of the probabilities in the CPTs of the Boolean nodes could be reduced to frequency counting. That is, the entry $Pr(Y = y|Par(Y) = z)$ is equal to the ratio of the number of cases when Y had a specific value y (either True or False) and the parents $Par(Y)$ of Y were in configuration z, and the number of times the parents of Y were in configuration z (regardless of the value of Y). For the continuous nodes V_k, a suitable parametric model for the similarity values must be chosen. One possible model is a normal (Gaussian) distribution with mean $\mu$and variance σ^2. Then, two separate normal distributions $N(\mu_{k,+}, \sigma^2_{k,+})$ and $N(\mu_{k,-}, \sigma^2_{k,-})$ are estimated for positive (matching) and negative (non-matching) cases (pairs of elements), respectively. The mean $\mu_{k,+}$ is the average of the similarity values $v_{k,i}$ of all data cases where the parent node X_k of V_k has been labeled with value True. The parameter $\sigma_{k,+}$ is the sampled standard deviation of these cases. Analogously, the parameters $\mu_{k,-}$ and $\sigma_{k,-}$ are the sample mean and standard deviation of $v_{k,i}$ over all cases when the parent node X_k has been labeled with the value False.

It is also possible to estimate the parameters in the CPTs when only some of the nodes have been labeled. A typical situation arises when a human designer has provided feedback about whether the two elements match (that is, has assigned a Boolean value to the root node of the BN), but has not explained why they match (that is, whether the match is lexical, instance-based, structural, based on a dictionary, etc.) This situation is more challenging, but as long as the graph of the network is known and fixed, it is still possible

to estimate the most likely values of the parameters in its CPT. This problem is known as parameter learning with partially observed data in Bayesian networks, and can be solved by means of gradient ascent in the likelihood function or the Expectation Maximization algorithm, among other methods (Heckerman, 2001; Thiesson, 1995).

Assuming there is a data set Σ of N independent training cases, the log-likelihood scoring function is

$$log\, L(\Theta|\Sigma) = \frac{1}{N}\sum_{i=1}^{M}\sum_{l=1}^{N} log\, P(X_{il}|Pa(X_i), \theta_i),$$

where Σ denotes the training data set, $Pa(X_i)$ denotes the parents of the node X_i, $i = 1, \ldots, M$, and Θ is the parameter vector $\Theta = \{\theta_1, \ldots, \theta_M\}$.

However, we only have partial observations, which means that there are several hidden nodes with no labels. For each training case, one pair of elements $S_1.E_i$ and $S_2.E_j$ is run through all K individual matchers to produce the corresponding similarity values $v_{i,j,k}$, and a true label of two elements matching or not for the root node $OverallMatch$ is provided by the human designer. With known structure and partial observation, we can use the EM (expectation maximization) algorithm to find a locally optimal maximum-likelihood estimate of the parameters (Murphy, 2003). After learning parameters from a training data set, each discrete node has a conditional probability table (CPT) specifying the probability of each state of the node given each possible combination of parents' states.

2.3 Inference

Given the individual similarity values $V_k = v_k$, $k = 1, K$ that have been reported by all individual matchers, and a full Bayesian network with CPTs estimated from data, we can evaluate the probability that the two elements match on the basis of all evidence, by means of a standard computational process known as belief updating. One possible method to perform belief updating is to construct the join tree of the Bayesian network, and use if for inference. This can be done by means of a number of commercial or freely available reasoning engines. The continuous variables V_k, under the chosen Gaussian parametrization, can be incorporated into the process of belief updating in the form of virtual (uncertain) evidence (Pan et al., 2006). To supply virtual evidence to a belief updating engine, all that is needed is the likelihood ratio of the observed values v_k for the similarity value variables V_k:

$$L(V_k = v_k|X_k) \doteq \frac{Pr(V_k = v_k|X_k = T)}{Pr(V_k = v_k|X_k = F)} = \frac{N(v_k|\mu_{k,+}, \sigma^2_{k,+})}{N(v_k|\mu_{k,-}, \sigma^2_{k,-})},$$

where $N(v|\mu, \sigma^2)$ is the probability that measurement v comes from normal distribution with mean μ and variance σ^2, and X_k is the parent node of V_k in the BN.

After the process of belief updating concludes, all Boolean nodes in the network will be assigned probability values according to the observed evidence (values) v_k for the similarity value variables V_k. The probability of the root node is the final estimate that the two elements match, given the combined evidence of the individual matchers.

3 EXPERIMENTAL RESULTS

In order to evaluate the match accuracy of any matcher described below, we used five XML schemas for purchase orders, CIDX, Excel, Noris, Paragon and Apertum, kindly provided to us by the University of Leipzig. The figure of merit for evaluation of the accuracy of matching was the popular f-measure, defined as the harmonic mean of precision and recall, as used in the information retrieval community. If the number of true matches identified by the matching system as such (hits) is A, the number of true matches not identified as such (misses) is B, and the number of cases when two elements do not match, but the matcher incorrectly declares a match (false positives) is C, the f-measure F can be computed as $F = 2A/(2A+B+C)$.

We developed 13 basic schema matchers and evaluated the ability of the proposed Bayesian method to combine their outputs so as to improve the accuracy of matching. Of these, 11 were lexical matchers: CosineSimilarity, HammingDistance, JaroMeasure, LevenshteinString, BigramDistance, TrigramDistance, QuadgramDistance, PrefixName, SuffixName, AffixName, SubstringDistance. One matcher, PathName, was structural, comparing the entire paths of the two elements in their respective XML schemas. The last basic matcher was neither lexical nor structural: the Synonym matcher declared a match if and only if the two tested elements were found in a list of synonyms relevant to the domain of purchase orders. Based on their method of operation, the similarity values computed by the 11 lexical matchers can be expected to be highly correlated and statistically dependent; in contrast, the synonym matcher could be expected to produce output that is largely independent of the lexical matchers. Experimental evaluation of their pairwise dependence confirms this intuition: Figure 2 shows the pairwise correlation between all 13 pairs of matchers, evaluated from all pairs of elements in all ten pairs of schemas. Clearly, all 11 lexical matchers

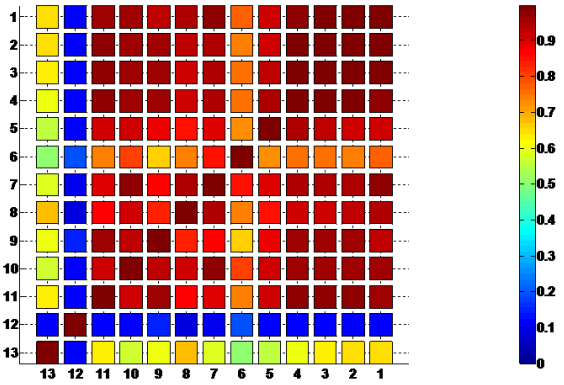


Figure 2: Pair-wise correlations between all pairs of basic matchers, numbered as follows: 1: Edit Distance; 2: Substring Distance; 3: Bi-Gram Distance; 4: Tri-Gram Distance; 5: Quad-Gram Distance; 6: Cosine Similarity; 7: Hamming Distance; 8: Jaro Measure; 9: Affix Name; 10: Prefix Name; 11: Suffix Name; 12: Path Name; 13: Synonym.

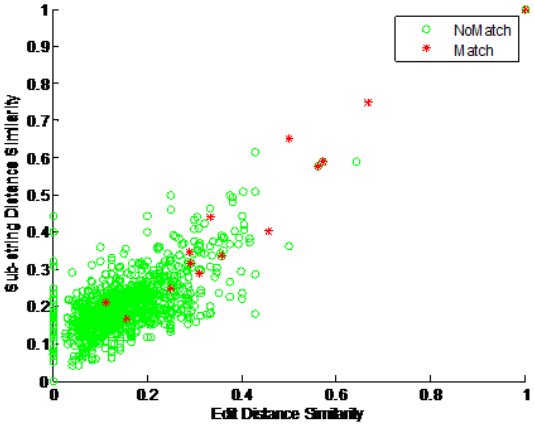


Figure 3: Scatter plot of similarity values computed by the Edit Distance (LevenshteinString) and SubstringDistance matchers. Their output is clearly correlated, resulting in a correlation coefficient of 0.9892.

are highly correlated, whereas their correlation with the Synonym matcher is minimal. Somewhat surprisingly, the structural matcher, PathName, is the least correlated with any other matcher.

The kind of major correlation that exists between lexical matchers is illustrated in Figure 3 that shows a scatter plot of the similarity values computed by the LevenshteinString (edit distance) matcher and the SubstringDistance matcher. Their high correlation (0.9892) makes one of them almost redundant, if the other one is present.

Regarding the experimental evaluation of matching accuracy, as with any machine learning method, care should be given to the training and testing evaluation protocol, that is, which data are used for training and which data are used for testing. We used three evaluation protocols, as described below.

3.1 Testing on Training Data set

This is the simplest evaluation protocol, where we use the same data set for testing and training. Its purpose is to evaluate how well we can fit the training data. Under this protocol, we define ten matching tasks that correspond to all possible pairs of the five XML schemas. For each matching task (pair of schemas), we build a dedicated Bayesian composite matcher that is specific for this task. The same data set, then, is used as evidence to predict the belief for every pair of elements. This is the most lenient evaluation protocol, since the learning algorithm has seen during training the data that will be used for testing.

After a similarity matrix is computed for all pairs of elements of two database schemas, an additional global matching step called Max1/Delta is performed to produce the final match decisions, based on the understanding that most often (but not always) mappings between database elements are one-to-one (Do and Rahm, 2002). Since this procedure is sensitive to the exact value of the Delta parameter, we present below results as a function of that paramater. After global match decisions have been obtained, they are compared with the ground truth, and the f-measure for this pair of schemas is computed. These f-measures are averaged over all pairs of tasks in the testing data set (in this case, ten pairs of tasks), in order to arrive at the final overall f-measure.

Figure 4 shows a comparison between all 13 basic matchers and the Bayesian Composite Matcher (BCM). The accuracy of the BCM reaches 0.819 and is significantly higher than that of any other matcher. It is also practically constant for a wide range of the parameter Delta. The performance of Path Name matcher is better than other individual matchers, because it is a hybrid matcher combining two basic match techniques.

3.2 Leave-One-Out Cross Validation (LOOCV)

A more realistic testing protocol is under the leave-one-out cross validation (LOOCV) method, where training and testing data are clearly separated. Each of the ten pairs of schemas is used for testing, using a BCM that was learned using the other nine pairs of schemas. The results are averaged over the ten pairs, as follows:

1. Build training and testing data sets for 10 test tasks. For instance, if the similarity matrix of

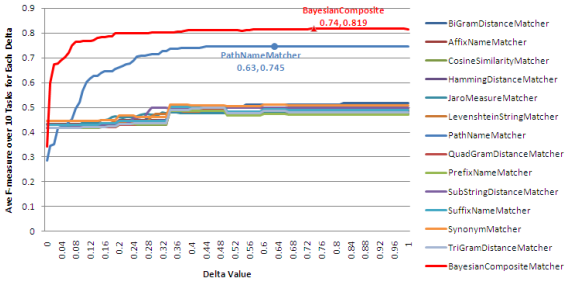


Figure 4: Comparison of average f-measure between the Bayesian Composite Matcher and all other matchers.

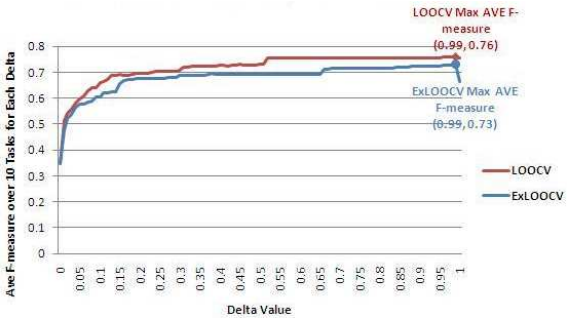


Figure 5: Comparison of Bayesian composite matcher performance under LOOCV and exclusive LOOCV testing protocols.

Excel $\leftrightarrow$ *Noris* is used as testing set, the training data set for this test task is a collection of similarity matrices of the remaining 9 schema pairs.

2. Learn one Bayesian composite matcher for each task based on its training data.
3. Implement $Max1/Delta$ selection approach on the composite similarity matrix generated by each Bayesian Composite Matcher.

3.3 Exclusive Leave-One-Out Cross Validation (ExclLOOCV)

The second protocol described above still allowed the training algorithm to see data from the pair of schemas that would be used for testing, but not the ground truth for their direct match. To eliminate any exposure of the training algorithm to data that would be used for testing, we modified the LOOCV procedure as follows. For each task, if the test pair is $A \leftrightarrow B$, the training examples only come from the three remaining schemas not involving either A nor B. For example, if one test set is *Excel* $\leftrightarrow$ *Noris*, it will be tested with the Bayesian composite matcher that has used only the following three pairs of schemas for training: *CIDX* $\leftrightarrow$ *Apertum*, *CIDX* $\leftrightarrow$ *Paragon*, and *Apertum* $\leftrightarrow$ *Paragon*. This is the maximally realistic testing protocol.

Figure 5 shows a comparison between the two variants of the LOOCV evaluation protocol for the Bayesian Composite Matcher. It can be seen that the accuracy drops to 0.76 under usual LOOCV and 0.73 under exclusive LOOCV.

4 RELATED WORK

As mentioned in the first section, many methods for creating composite matchers have been tried, and this section explains the difference between them and the proposed approach. One major distinction between these methods is whether they rely on manual tuning of the composition structure and parameters, or such parameters are estimated from a training set and verified on an independent test set. The composition methods developed in the COMA (Do and Rahm, 2002; Do and Rahm, 2007) and GLUE (Doan et al., 2003) systems are based on manual tuning of the composition parameters, so comparison with learning methods for tuning parameters is not entirely correct; a composite matcher that is manually tuned with a specific set of schemas in mind can certainly be expected to be more accurate than a learning matcher that is tested under a cross-validation protocol.

Among the learning methods for composing matchers, our approach is most similar to the one proposed by Marie and Gal (Marie and Gal, 2007), who have approached the problem from a Bayesian network perspective, too, arguing that a disciplined approach to handling match uncertainty has to be applied. However, their approach is based on Naive Bayes networks, that is, two-level Bayesian networks with one root node that corresponds to the matching event, and many leaf nodes that are directly children to the root node. It can be shown that such a Naive Bayes network has the same classification properties as a logistic regression model, and the decision surface is linear, similar to the one used in the LSD and GLUE systems (Doan et al., 2003; Doan et al., 2003). In contrast, a full (non-naive) Bayesian network like the one proposed in this paper can model arbitrary correlations and decision surfaces.

Furthermore, the Bayesian network proposed in this paper is also different from the Bayesian network classifiers used in the YAM system (Duchateau et al., 2009) in that our network includes unobservable nodes corresponding to types of matchers; in contrast, YAM employs the BayesNet classifier from the WEKA library that can learn the structure of a fully observable network by adding and removing edges, but cannot add unobservable nodes (Witten and

Frank, 2005). Unobservable nodes corresponding to a type of matcher (e.g. lexical, dictionary-based, structural, etc.) present a natural way of representing the conditional dependency between multiple matchers of the same type, because they restrict the edges of the graph only to the nodes of the same type. In contrast, a fully-connected BN without hidden nodes would require an exponential number of CPT parameters to be estimated, which would make it practically impossible to collect the data necessary for estimating them. This problem is further compounded by the continuous values of the similarity values produced by basic matchers — in fact, it is not immediately clear how YAM would have been able to learn a fully connected BN with 13 continuous nodes representing the similarity values of each basic matcher, from the few thousand examples available from the PO dataset under the two LOOCV protocols.

On the other hand, non-linear classifiers such as decision trees (Duchateau et al., 2008) can indeed represent non-linear decision surfaces from a limited number of training examples, but are not inherently probabilistic, and the binary decisions output by them are not easy to use in the global assignment process that determines the entire mapping between two schemas from the pair-wise matches between their individual elements. Other probabilistic approaches to the automatic schema matching problem include the use of an attribute dictionary in the AUTOMATCH system, where training examples of matching schemas are used to compile the dictionary, and candidate elements from new schemas are compared probabilistically to the dictionary. Although this approach does result in probabilistic estimates of matches, the compilation of the dictionary requires many training examples, and is best suited to domains where many pairs of entire schemas have to be matched repeatedly.

5 CONCLUSIONS AND FUTURE WORK

We have proposed a novel method for creating composite matchers for the purpose of automatic schema matching. Its main advantage is the explicit modeling of the conditional statistical dependence between the similarity values computed by individual basic matchers. Experiments suggest that it combines successfully the outputs of such matchers, and achieves matching accuracy significantly exceeding that of the individual matchers. Furthermore, its outputs are estimates of the genuine probabilities of match, which allows the application of decision-theoretic methods for optimal judgment whether elements match, or not. Further work will focus on leveraging the clear semantics of the computed probabilities for improving the accuracy of the global matching algorithm, as well as on improving the computational properties of the proposed Bayesian method.

REFERENCES

E. Rahm, P. A. Bernstein, A Survey of Approaches to Automatic Schema Matching, VLDB Journal, 10:4 2001.

H. H. Do, E. Rahm, COMA - A System for Flexible Combination of Schema Matching Approaches, in Proceedings of the 28th International Conference on Very Large Data Bases (VLDB), 2002.

W. Li, C. Clifton, A Tool for Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Network, Journal of Data and Knowledge Engineering 33: 1, 49-84, 2000.

A. Doan, P. Domingos, and A. Halevy., Learning to Match the Schemas of Databases: A Multistrategy Approach, Machine Learning Journal, no. 50, pp. 279–301, 2003.

S. Bergamaschi, S. Castano, M. Vincini, D. Beneventano, Semantic Integration of Heterogeneous Information Sources, Journal of Data and Knowledge Engineering 36: 3, 215-249, 2001.

H. H. Do, E. Rahm, Matching Large Schemas: Approaches and Evaluation, Journal of Information Systems, Vol. 32, Issue 6, Sep. 2007.

A. H. Doan, P. Domingos, A. Halevy, Reconciling Schemas of Disparate Data Sources: A Machine Learning Approach, SIGMOD 2001.

D. W. Embley, Multifaceted Exploitation of Metadata for Attribute Match Discovery in Information Integration. WIIW 2001.

D. Heckerman, A Tutorial on Learning Bayesian Networks, Journal of Learning in Graphical Models, pp. 301-354, 2001.

K. Murphy, An Introduction to Machine Learning and Graphical Models, the Intel Workshop on Machine Learning, Sep. 2003.

J. Tang, J. Z. Li, Using Bayesian Decision for Ontology Mapping, Journal of Web Semantics, Vol. 4, Issue 4, Dec. 2006.

Thiesson, B., Accelerated Quantification of Bayesian Networks with Incomplete Data, Proceedings of the Conference on Knowledge Discovery in Data, 1995, pp. 306-311.

Rong Pan, Yun Peng, Zhongli Ding, Belief Update in Bayesian Networks Using Uncertain Evidence, 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), 2006, pp.441-444.

A. Marie and A. Gal. Managing Uncertainty in Schema Matcher Ensembles. Proceedings of the 1st International Conference on Scalable Uncertainty Management. Washington, DC, October 2007, pp. 60-73.

A. H. Doan, J. Madhavan, R. Dhamankar, P. Domingos, A. Halevy, Learning to Match Ontologies on the Seman-

tic Web, The VLDB Journal 12 (4), 2003, pp. 303-319.

F. Duchateau, Z. Bellahsene and R. Coletta, A Flexible Approach for Planning Schema Matching Algorithms, OTM Conferences (CooPIS), 2008, pp. 249-264.

F. Duchateau, R. Coletta, Z. Bellahsene, R. J. Miller, Not Yet Another Matcher, Proceedings of CIKM'09, Hong-Kong, China, November 2009, pp. 2079-2080.

Ian H. Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann, 2005.

Berlin, J., A. Motro: Database Schema Matching Using Machine Learning with Feature Selection. CAiSE 2002, pp.452-466.

A Temporal Search Engine to Improve Geographic Data Retrieval in Spatial Data Infrastructures

Fabio Gomes de Andrade[1], Cláudio de Souza Baptista[2] and Ulrich Schiel[2]
[1]*Department of Informatic, Federal Institute of Education, Science and Technology of Paraíba, Cajazeiras, Brazil*
[2]*Department of Systems and Computing, University of Campina Grande, Campina Grande, Paraíba, Brazil*
fabio@ifpb.edu.br, {baptista, ulrich}@dsc.ufcg.edu.br

Keywords: Geographical Information Retrieval, Spatial Data Infrastructures, Temporal Search Engine.

Abstract: Recently, spatial data infrastructures have become an important solution to ease the finding of geographical data offered by different organizations. Nevertheless, the catalog services provided by these infrastructures still have some important drawbacks that limit the geographic information retrieval based on temporal constraints. Examples of these drawbacks include the lack of both a more detailed description of temporal information, and ranking. Aiming to overcome these limitations, this paper describes a new temporal search engine for solving feature type retrieval offered by catalog services. To reach this goal, our search engine is based on a model that stores temporal information about each service and its respective feature types described in the SDI catalog service. Moreover, the paper proposes a temporal ranking metric to evaluate the relevance of each feature type retrieved for the user's query.

1 INTRODUCTION

Recently, spatial data infrastructures (SDIs) have become an important solution to ease the finding of geographical data offered by different organizations (Williamson et al., 2003). SDIs usually offer catalog services, which can be used by both providers and clients. Providers use this service to announce their resources, while clients use it to find the data of their interest.

The current catalog services improve geographic data retrieval, but still have serious limitations. One of these limitations is the low support to temporal searches. The time-based searches offered by the current catalogs are performed using attributes such as temporal extension and creation/modification date of the resources. Nevertheless, the values of these attributes are often omitted; or contain imprecise information. Such characteristic considerably reduces the quality of temporal searches. Other important drawback is the lack of mechanisms to evaluate the importance of each resource retrieved from a user's query.

Aiming to overcome these limitations, in this paper we propose a new search engine for solving temporal queries in SDIs. The proposed solution offers two contributions. The first one consists of a model that improves the description of the temporal features of the services described in the SDI catalog service. The second contribution consists in the development of a ranking mechanism that is based on the temporal features of each feature type and ideas from the classical information retrieval. Such ranking is used to evaluate how relevant each feature type is for the user's query, considering only the temporal dimension.

It is important to keep in mind that geographical data are characterized by three dimensions: space, theme and time. The solution described in this paper approaches only the time dimension. However, this solution has been integrated to a broader search engine, which is able to solve queries with spatial, thematic and temporal constraints. Details about the semantic ranking implementation can be found in (Andrade and Baptista, 2011).

The remaining of this paper is organized as follows. Section 2 describes how the current SDIs offer temporal information. Section 3 focuses on the model used to represent temporal information. Section 4 presents the temporal ranking measurement. Section 5 discusses the implementation and the experimental evaluation. Section 6 summarizes the main related works. Finally, in section 7, we conclude the paper.

2 TEMPORAL INFORMATION IN SDI

In order to ease the standardization and access to their geographical data, many SDIs are being implemented as a set of services (Bernard and Craglia, 2005). In such infrastructures, the datasets offered by a provider are commonly supplied as a set of feature types. These feature types, in turn, are encapsulated and delivered for the users through services standardized by the Open Geospatial Consortium [OGC], such as Web Map Service (WMS) (OGC, 2004) and Web Feature Service (WFS) (OGC, 2005).

When a provider registers a service in the SDI catalog, it must provide metadata that describe different features of the dataset offered by the service. Part of this information is related to the temporal extension, where the provider must inform the time interval with respect to the dataset. The way as this information is described depends on the metadata standard adopted by the infrastructure. In the ISO 19115 metadata standard, the temporal reference of a resource is usually described through a temporal interval. Such interval is defined by two attributes called *beginPosition* and *endPosition*, defining, respectively, the initial and final limits. The value of these attributes is commonly described in the ISO 8601 format.

One of the causes that limit the resolution of temporal queries in the present SDIs is the fact that the values of attributes that describe the temporal extension of the data are often omitted by the provider. In this case, the only information offered that is related to time is the creation/modification date of the metadata record, which do not describe the temporal extension with precision.

Another limitation is related to the details supplied during the registration. Presently, most geographic data providers create a single metadata record to describe their dataset. Hence, only one temporal extension is defined to characterize all the feature types offered by a service. Figure 1 shows two metadata records from different providers, called M_1 and M_2. Each record describes the feature types offered by a service.

In the service described by M_1, there are feature types covering the periods of 2000, 2002 and 2005. In this record, it was defined that the temporal extension (TE) of the service is the interval between 2000 and 2005, which corresponds to the smallest interval covering all of its feature types. Such a situation leads to two kinds of disadvantages. In order to understand the first one, let us consider a query where the user looks for feature types concerning the year of 2003. In this case, as the extension of M_1 intersects the period defined in the query, the record ends up being retrieved, though it does not offer any feature type concerning the period defined in the request. The second disadvantage occurs because the catalog service always retrieves the service as a whole. So, the user is in charge of accessing the service and identifying the feature types that satisfy the criteria defined in the query. This task can be, often, tedious and time consuming, since many services offer a large number of feature types.

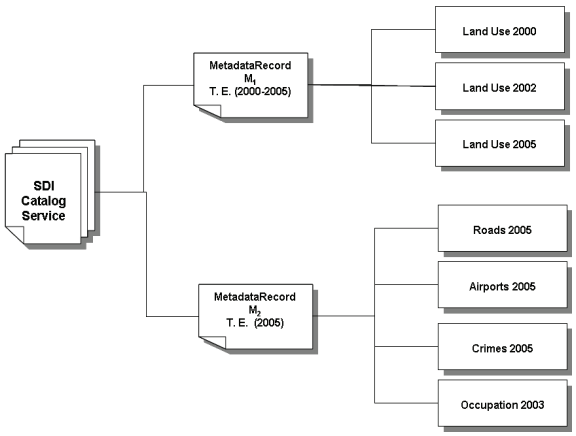


Figure 1: Example of services temporal description.

Other problems concerning the resolution of temporal queries occur due to inconsistencies between the temporal extension defined in the metadata record and the temporal extensions concerning the feature types. Observing Figure 1 again, it is possible to notice that the record M_2 defines the interval 2005 as its temporal extension, though it also has a feature type concerning another temporal interval. This kind of situation occurs because many providers use as temporal extension just the period associated to most of its feature types. So, in the case the user performs a query for feature types concerning the period of 2003, the catalog service will not retrieve the service described by M_2, though the service has data which satisfy the criteria defined in the request.

Besides the limitations previously described, another problem of the present catalog services is that they assume that all the resources that satisfy the selection criterion defined in the query have the same relevance for the user. So, a resource that covers the whole interval requested by the user is considered as relevant as one that covers only a part of the requested interval. This makes the possibly

more relevant services to be presented later to the user. This is an undesired characteristic, especially in queries which return a large number of results. The main consequence is that, in these queries, the user may lose time in trying to find the service having the most relevant information or, in worse cases, this user my end up not evaluating the resource.

3 DESCRIBING TEMPORAL INFORMATION

The first stage in the development of our search engine consists in defining a model to represent temporal information of the services offered by the SDI. In order to fulfill requirement R_1, the designed schema stores the temporal information of each service and of each feature type that it offers. Figure 2 presents the conceptual schema.

It is important to have in mind that the simplicity of the schema is due to the fact that most part of the service information keeps being stored in the metadata record. So, our schema stores just the information needed to resolve the temporal queries.

The temporal extensions of a service and of its feature types are defined through the attributes *beginTime* and *endTime*, present in their respective entities. Both attributes are represented as timestamps. Moreover, the model stores some descriptive attributes of each service and of each feature type, such as name, title and textual description. These attributes are shown to the user during the exhibition of the query result. After evaluating the information shown by these attributes, the user may request a complete view of the record that describes the service offering the feature type of interest. Such retrieval is performed by the attribute *metadataIdentifier*, which keeps a reference for the service metadata record.

After defining the model to be used for data representation, we created a methodology to extract temporal information from the service and feature types. In order to facilitate the reader's understanding, this process will be referred to as *temporal annotation* throughout this paper.

3.1 Services Annotation

The first stage of the process of extracting temporal information consists in identifying the temporal interval covered by the service. The information is obtained through the processing of information in the metadata record of the service.

The service annotation is done through the value of the attributes that define its temporal extension. When the values of these attributes are provided by the metadata record that describes the service, they are retrieved and normalized to a temporal interval. This interval is then defined as the service temporal extension. Another attribute verifies the service update frequency. The verification is intended to check whether the service is continuously updated. If so, the model considers the service persistent and, consequently, it has no final limit. In our solution, we consider persistent just the services with the following values for update frequency: *annually, continually, daily, monthly* and *weekly*.

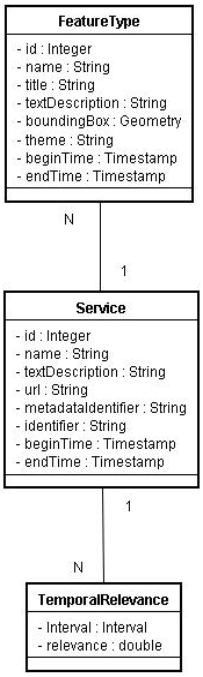


Figure 2: Conceptual schema.

When the temporal extension of the service is not present in its metadata, its temporal extension is identified through the values of other attributes. The attributes are queried in a priority order. In the case an attribute contains temporal information, its value is extracted, standardized and associated to the service, ending the annotation process. In the opposite case, the next attribute is queried and the process is repeated until all attributes are checked. Currently, the attributes queried in order to obtain temporal information of a service, in priority order, are: keywords, title and textual description. Finally, in the case temporal information is not found in any of these attributes, the creation date of the metadata record is used as temporal reference for this service.

Among the attributes queried during the annotation process, some are represented directly as dates, as the temporal extension and the inclusion date of the record. When the temporal information is obtained through these attributes, it is necessary to convert the attribute value into a temporal interval, in the format used by the data schema.

The date standardization process depends on the format of the temporal information. In the case the attribute value is a simple date (for instance, 1/10/2010), its conversion to an interval can be done in two forms, depending on its update frequency. If the service is continuously updated, the date is converted into a persistent interval, which means that the final limit corresponds to the moment at which the query is performed. Otherwise, the value is standardized to an interval containing the limits of the value found for the attribute. The granularity of this interval (day, month, year) is identical to the granularity of the attribute value. For example, if the value represents a day, the generated interval will represent a day too. The same occurs for attributes that represent a month or a year.

While attributes concerning temporal extension and inclusion date of the record already supply the values in the form of dates, some of the queried attributes, such as service name and textual description, are offered as a set of strings. When these attributes are queried, their text values must be processed in order to extract temporal information. This processing, which is occurs with use of machine learning techniques, is performed in two stages.

In the first step, the text corresponding to the attribute value is processed for analysis of the parts of the speech. This stage is intended to identify and classify the radical of the elements that appear in the text. This task is done by a framework called *TreeTagger*.

In the second stage, the result of the previous stage is processed in order to find temporal expressions. This information can be found by both numerical values, such as dates, and textual elements, such as the words *today, yesterday and tomorrow*. As the result of this stage, an XML file containing all the temporal expressions identified in the text is generated. This file is coded in the *TimeML* standard (Pustejovsky et al., 2003), an XML-based standard for specification of temporal expressions in documents. This task is carried out by a framework called *Heideltime*.

The processing made by *Heideltime* automates the recognition of temporal expressions. However, each identified expression is treated and annotated individually, with no relationship among them. So, the file generated by this framework must be processed, and the temporal information that is found must be converted into a temporal interval. The standardization of temporal annotations found in a text depends on the number of annotations found. If the file has just one temporal annotation, the feature type interval corresponds to the smallest interval that covers all the temporal information found. Finally, the absence of temporal annotations means that no temporal element was found in the analyzed text. This situation indicates that the temporal reference cannot be obtained through this attribute.

3.2 Feature Types Annotation

Differently from the annotation of services, the temporal annotation of a feature type is performed with basis on the information contained in the document describing the capabilities of the service. Such a document is obtained by invoking the operation *getCapabilities* of the service being annotated.

An important characteristic of the capabilities document is that, differently from the metadata record, it has no specific attribute to define temporal information of the feature types offered by the service. So, the temporal annotation of these elements must be done through the identification of temporal expressions present in the values of some attributes. In order to perform this task, for each feature type, their keywords, titles and textural descriptions are queried following the priority order.

Since all attributes used for feature types annotation are textual, the temporal information contained in these elements must be extracted through the processing of the text corresponding to their values. The procedure adopted to perform this task is the same used in the annotation of services. The values obtained after this process are used as the temporal extension of the feature type that is being processed. In the case the temporal extension of the feature type cannot be obtained from any of the verified attributes, we assume that its value is the same one obtained for its respective service.

4 TEMPORAL RANKING

After defining how the temporal information will be retrieved from the services and stored in the database, we developed a search engine for retrieval of feature types that meet a certain temporal

constraint. To implement it, we defined a metric that evaluates how relevant each feature type is for the query. Such a metric is computed from other two measurements: the degrees of overlap and of temporal relevance.

4.1 The Degree of Overlap

The first measurement used to evaluate the temporal ranking is the *degree of overlap*, which evaluates the similarity between the temporal interval requested in the query and the temporal interval covered by the feature type under evaluation. This metric is computed by means of the Tversky equation (Tversky, 1977). This equation was chosen because it considers, during the evaluation of similarity between objects, the characteristics that they may have or not in common. Let t_1 be the temporal interval defined in the user's query and t_2 the interval associated to the feature type under evaluation. Then, the *degree of overlap* between them is computed by equation 1.

$$od(t_1, t_2) = \frac{|t_1 \cap t_2|}{|t_1 \cap t_2| + \alpha |t_1 / t_2| + (1-\alpha)|t_2 / t_1|} \quad (1)$$

where:

- $| t_1 \cap t_2 |$ represents the extension, in milliseconds, of the intersection between the intervals t_1 and t_2;
- $| t_1 / t_2 |$ represents the extension of the interval in t_1, but not in t_2;
- $| t_2 / t_1 |$ represents the extension of the interval in t_2, but not in t_1;
- The constant α represents the weight that the complement of the interval t_1 has to the evaluation of the overlap between the intervals. Presently, the value 0.9 is used for this constant. To determinate this value, we used a technique called weighting (Fox and Shaw, 1993). To estimate the value of α we used the Pearson correlation coefficient.

4.2 The Temporal Relevance

The second metric used to evaluate the temporal ranking of a feature type is the degree of *temporal relevance*. As in the classical information retrieval, this metric evaluates how relevant a certain temporal interval is to a certain service (Baeza-Yates and Ribeiro-Neto, 1999). Hence, when two feature types offered by different services have the same degree of overlap (or close values) with respect to the user's query, the model prioritizes feature types offered by services whose temporal extension has higher relevance.

In order to compute the relevance of a temporal interval to the service, it is necessary to evaluate first the frequency in which this interval occurs in the service. This metric is called *raw frequency (temp_f)* and is computed by comparing the temporal interval under evaluation with the temporal extension covered by each feature type offered by the service.

During the development of this metric, we evaluated three forms to compute the frequency of a temporal interval t in a service S. The first solution consisted in computing the number of associated feature types whose temporal extension was identical to t. Despite being easier to compute, this kind of approach did not consider that the intervals were different from t, but that they completely contained that interval, and also did not consider that some parts of the interval were present in intervals that intersected it. In the second approach, the frequency was computed with basis on the number of intervals that were identical or totally covered the interval under evaluation. The disadvantage of this solution is that it does not consider the overlap between the evaluated interval and the other intervals that do not cover it totally.

In the third approach, which ended up being adopted for the proposed metric, the frequency is computed by summing the *degree of overlap* between t and the temporal interval associated to each feature type offered by the service. So, all the presences (total or partial) on the interval t in the temporal interval associated to each feature type are considered when evaluating the frequency of this interval. Equation 2 describes the computation of this frequency. In this equation, t represents the temporal interval under evaluation, while S is the service to which the relevance of t is being computed. Moreover, Ti_{temp} represents the temporal extension of a feature type T offered by the service, and n represents the number of feature types offered by the service.

$$temp\_f(t, S) = \sum_{i=1}^{n} overlap(Ti_{temp}, t) \quad (2)$$

After computed, the value of the *raw frequency* of the temporal interval is used to compute its *normalized frequency* (*temp_tf*). This frequency is computed by the proportion between the *raw frequency* and the number of feature types offered by the service (Equation 3). As in the previous equation, t represents the temporal interval which is under evaluation and S represents the service to which the relevance will be computed.

$$temp\_tf(t, S) = \frac{temp\_f(t, S)}{n} \quad (3)$$

Another variable used to evaluate the degree of relevance of a temporal interval is the *inverse frequency* (*temp_isf*). Its objective is to evaluate how important a temporal interval is to the whole set of services offered by the SDI. The value of the *inverse frequency* (Equation 4) is computed by the proportion between the number of services offered and the number of services in which the interval is totally covered by at least one feature type.

$$temp\_isf(t) = \log \frac{N}{ni} \quad (4)$$

After computing the *normalized frequency* and the *inverse frequency*, their values are used to determine the degree of temporal relevance of an interval to the service. This *degree of relevance* is computed through the product of both frequencies, as in Equation 5.

$$tr(t,S) = temp\_tf(t,S) * temp\_isf(t) \quad (5)$$

4.3 The Temporal Similarity

Once computed, the values of the *degrees of overlap* and *temporal relevance* are combined to determine the *temporal ranking* of a feature type. The value of this metric is used to classify and sort the feature types that are retrieved form a user's query.

Given a temporal interval t defined in the user's query and a feature type T offered by the SDI, the *temporal ranking* of T is obtained through Equation 6. In this equation, T_T represents the temporal interval covered by T, while S represents the service that offers this feature type.

$$ranking(t,T) = w_1 * od(t,T_T) + w_2 * tr(T_T,S) \quad (6)$$

where:

- *ranking* represents the temporal similarity between a temporal interval t defined in the user's query and a feature type T being evaluated;
- *od* represents the *degree of overlap* between the interval defined in the query (t) and the interval covered by the feature type under evaluation (T_T);
- *tr* represents the *degree of relevance* of the temporal interval associated to the feature type under evaluation to its respective service;
- w_1 and w_2 represent the weights that the degree of overlap and the degree of temporal relevance has for the calculation of the degree of temporal similarity. Each weight must have a value between 0 and 1, and their sum must always be equal to 1. Presently, we use values of 0.84 and 0.16, respectively, for w_1 and w_2. These values were defined using the same statistical technique used to determinate the weights in Equation 1.

5 IMPLEMENTATION AND EVALUATION

Once defined the temporal ranking metric, our temporal search engine was implemented. This section first focuses on the implementation issues. After, the results obtained from the experimental evaluation are presented.

5.1 Implementation Issues

After defining the model used to represent information and the metric used to retrieve this information, we implemented a prototype for our search engine. This engine was incorporated to a tool called SESDI (Semantic-Enabled Spatial Data Infrastructures) (Andrade and Baptista, 2011), used for discovery of geographic data in SDIs. The architecture used for its implementation is comprised of two main subsystems: the data acquisition module and the query resolution module. The first module contains the components responsible for the extraction of the temporal data from the services registered in the SDI and from their respective feature types. The query resolution module, in turn, has the components that use the temporal ranking, describe is the previous section, to resolve temporal queries.

5.1.1 The Data Acquisition Process

The data acquisition process consists in collecting temporal data that will be used during the query resolution process. The data acquisition process is performed periodically, in order to find new services included and/or updated, keeping the database updated.

The process consists in obtaining information about new services. For this, the acquisition module calls the *getRecords* operation of the registration service of the infrastructure. This call has a filter that selects only the services whose inclusion/modification dates are more recent than the last verification made by the module. Each service retrieved in the first stage is then processed in order to make its temporal annotation.

The processing of each service is subdivided into several stages. The first one consists of extracting the temporal extension covered by the service. After obtaining this information, the module calls the *getCapabilities* operation of the service to obtain a document with information about the feature types offered by this service. After this, each feature type is processed to make its temporal annotation.

After extracting the temporal information from the service and from its feature types, the next stage consists in using the obtained information to generate the temporal relevance data of the service. For this, the module computes the temporal relevance of each temporal interval associated to at least one of the feature types offered by the service.

In the last stage of the data acquisition process, the information obtained after the extraction of temporal information and generation of temporal relevance of the service are made persistent in the database of the prototype. In this process, occurs the storage of the information about geographic data services and their feature types, with their temporal information, besides temporal relevance data generated for each service. Presently, these data are made persistent in a database implemented in the DBMS PostgreSQL/PostGIS.

5.1.2 The Query Resolution Process

The query module is responsible for the resolution of temporal queries. All queries are performed through a set of dynamic web pages featured by the tool. Each requisition received by the query module has as input parameter a temporal interval of the user's interest. The processing of queries is divided in four steps: temporal filtering, matchmaking, relevance filtering and ordering.

In the temporal filtering step, the search tool selects, among all the feature types recorded in the database, those whose temporal expression intersects the interval defined in the query. This task is done through a simple query in the SQL database. The result of this step is a set containing all the feature types that meet this constraint.

The second step consists in using the temporal ranking to compute the relevance of each feature type retrieved in the first stage. During the matchmaking process, persistent intervals end up receiving as final limit the timestamp value corresponding to the time at which the query was requested. At the end of this step, for each retrieved result, the result obtained for the temporal ranking is associated.

In many situations, temporal queries may return results with very low relevance for the query. This characteristic, in some situations, may be undesired, especially during the processing of queries that return a large number of results. In order to avoid such a situation, the user may define a minimum threshold at the moment of the query. When this happens, the third step in the processing of a query consists in removing from the final result all feature types whose relevance is below this threshold.

Finally, the last step consists in organizing the remaining results according to their relevance values. For this, a sorting algorithm is executed, in order to list the retrieved feature types in descendant relevance order. Figure 3 shows the result of a requisition for historic feature types concerning the year of 1964.

Figure 3: Temporal query result.

5.2 Experimental Evaluation

After implementing our temporal search engine, an experimental evaluation was performed, in order to compare its performance to that of the present catalog service. To make this validation, the catalog service of the North-American SDI was used as case study. The information of this service was collected and processed to perform the temporal annotation of each service and their respective feature types. The results obtained after this processing were stored in the database of the search engine. Presently, this database has 103 services and 12,914 feature types.

During the validation process, several queries were made for different time intervals. For each temporal interval used, two queries were performed. The first query was performed in the catalog service, and retrieved any record whose temporal extension intersected the interval defined in the query or whose publication date intersected the query interval (in the case of records with no temporal extension value defined). The second query was performed using the SESDI tool, which retrieved all the feature types whose temporal extension was intersected by the interval defined in the requisition. The results of

both queries were used to compare the performance of these two approaches. This comparison was performed according to recall and precision metrics. Recall corresponds to the proportion between the number of relevant results retrieved and the total of existing relevant results. Precision, in turn, is obtained through the proportion between the number of relevant results retrieved and the total of retrieved results.

5.2.1 Recall Evaluation

The experiments results were analyzed in two steps. In the first one, we made the evaluation at service level. This evaluation was intended to check the impact of our solution with respect to the number of different services retrieved by a query. This verification allows us to observe the number of services that have at least one feature type which is relevant to the query, but which end up not being retrieved by the catalog service. Figure 4 shows a graphic obtained through the comparison of results with respect to coverage. This graphic shows the performance of both SESDI and catalog service along the queries that have been executed during the evaluation process. Moreover, axis x represents the queries, while axis y represents the obtained performance for each approach.

The analysis of Figure 4 allows us to see that the model used by SESDI led to a big improvement in the coverage of the queries. The model used by the tool obtained a mean recall of 88.45%, while the catalog service had a mean recall of 45.17%. This difference can be explained by the fact that our search engine stores temporal information of services and feature types, while the catalog service makes queries only on information stored at services level. This difference allows our search engine to retrieve any services that offer at least one feature type whose temporal extension matches the constraints defines in the query, even if the temporal extension of the service does not meet the criteria defined in the query. This characteristic is impossible for the catalog service. Moreover, the large number of services that do not have a defined value for the temporal extension contribute to this difference, since the creation and modification dates of the metadata do not express this information precisely.

In the second step of the validation process, the two approaches were compared through the number of retrieved feature types. This evaluation is intended to obtain an overview of the number of feature types that are not retrieved due to the limitations of the present catalog services, as well as measuring how much the model used by our search engine improves the retrieval of this kind of information. The recall comparison of the two approaches with respect to the retrieval of feature types is shown in the graphic of Figure 5.

Figure 5 shows that, when the recall of the queries is compared at feature type level, the performance difference between the two approaches is still high. In this kind of evaluation, the model used by our search engine presented a mean recall of 95.56%, while the catalog service obtained a mean recall of 43.78%. The reasons that led to this difference are the same that cause the recall difference with respect to the retrieval of services. However, the large number of feature types offered by some services makes the difference between the performances of both approaches to be still bigger than what happens in the comparison at services level.

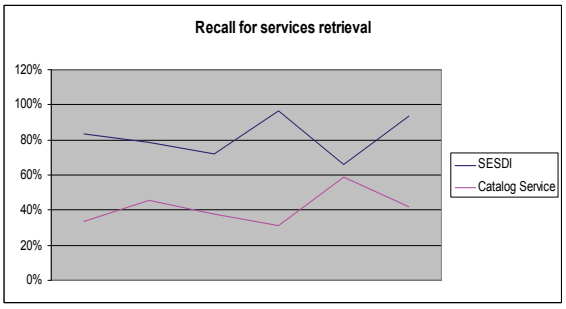


Figure 4: Graph of recall for services retrieval.

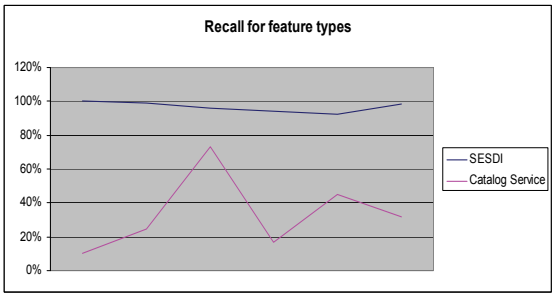


Figure 5: Graph of recall for feature types retrieval.

5.2.2 Precision Evaluation

Besides recall, the approaches were compared with respect to precision. Figure 6 shows the comparison of precision between the two approaches with respect to the number of services retrieved by each solution. The results show that the model used by SESDI had a better performance in some queries, while the catalog service achieves more precise

results in some requisitions. The mean precision of the catalog service was of 93.72%, while SESDI achieved a mean precision of 89.48%. The analysis of the results shows that the precision loss of the SESDI is not caused by the model used by its search engine, but is due to the performance of the temporal annotation process. While the catalog service performs its searches with basis on information of the catalog service, which are manually provided, the temporal search engine used by SESDI performs its queries with basis on information extracted manually during the temporal annotation process. This process is subject to errors, since some temporal expressions may be misinterpreted when extracted from textual attributes.

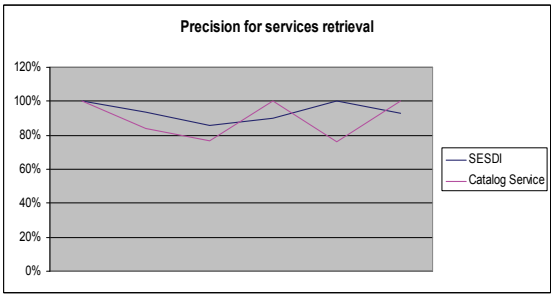


Figure 6: Graph of precision for services retrieval.

When the precision of the approaches is compared at feature type level, the performance of the two approaches is similar to the performance at services level. The graphic obtained for this comparison is shown in Figure 7. While the catalog service had a mean precision of 92.14%, SESDI achieved a mean precision of 88.95%.

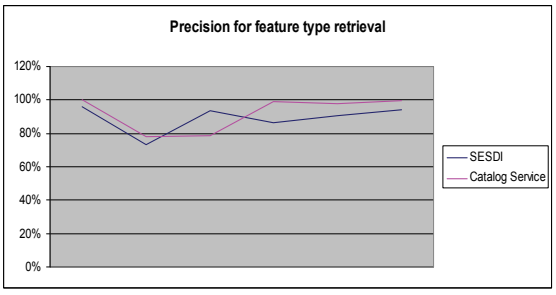


Figure 7: Graph of precision for feature types retrieval.

In general, the results obtained through the validation prove the feasibility of the model proposed in this paper. The main element that leads to this conclusion is that the model increases considerably the coverage of temporal queries, allowing the retrieval of many feature types even if the temporal description of their respective service is not supplied or not described in a consistent fashion.

6 RELATED WORK

The use of the temporal dimension to improve the information retrieval has been addressed by many studies over the years. Hübner and Visser (2003) proposed a solution during the implementation of the BUSTER project (Vögele et al., 2003). In that work, the temporal extension of each resource is represented through temporal periods. The limits of these periods can be described in precise, persistent or fuzzy forms, or can be defined with respect to other periods. During the information retrieval process, an algorithm based on Allen's temporal logic (Allen, 1998) is used to infer the semantic relationships between each period. The use of logic and inference allows the development of more powerful search tools. Nevertheless, this kind of approach does not offer a ranking. Moreover, the high computational cost of this kind of solution hinders its application to collections with a large amount of data.

A ranking-based solution was developed by Alonso, Gertz and Baeza-Yates (2006). In their work, documents are grouped in clusters into a timeline according to their temporal information. Inside each cluster, the documents are organized by a ranking, which is obtained through the metrics *tf-idf*, with respect to the matching of the text in the document and the text in the query. Manica et al., (2010) developed a search engine which enables the retrieval of temporal information in XML documents. In that work, the processing of queries is performed in two stages: a keyword matching and a temporal query, which is applied to the nodes that are closer to those retrieved in the previous stage. In both works, the ranking used to organize the documents does not consider the temporal extension of each resource.

Another ranking-based solution was developed by Jin et al., (2010). In that work, the keywords that form a document are associated to temporal intervals, which are obtained from expressions contained in the document. For each combination formed by a keyword and a temporal interval, a ranking value is computed through *tf-idf* techniques. The disadvantage of that work is that its ranking considers just the importance of the keywords, not considering the relevance of each temporal expression found in the document.

Another work that addresses temporal information retrieval in documents was developed

by Strötgen and Gertz (2010). In that work, spatiotemporal information of a document is extracted through the processing of the text and can be explored by users after a query. However, the means to evaluate the temporal ranking of each document are not supplied.

The analysis of the above studies shows that the temporal information retrieval is still an open problem. This analysis also shows that many studies explore the temporal dimension during the resolution of queries, but do not use (or use superficially) this information to establish the ranking of the retrieved results. This highlights the need for a more specific ranking, generated from a deeper analysis of this kind of information. Moreover, we can notice the lack of effective solutions to retrieve temporal data in the geospatial domain. This limitation, allied to the key importance that the time represents for this domain, highlights the importance of the work presented in this paper.

7 CONCLUSIONS

The temporal dimension has great importance for the retrieval of geographic data. However, the retrieval of geographic data with basis on temporal criteria is still a hard task for the present SDIs. The absence of a detailed description of the temporal extension of the services and the lack of a temporal ranking are some of the characteristics that cause this limitation.

Aiming to overcome those limitations, this paper described a new temporal search engine. The main contributions consist in the development of a new model that improves the description of the temporal extension at service and feature type levels, and the development of a ranking for the feature types retrieved during a query.

Some future works still should be undertaken to improve our research. An important task to be developed consists of extending our approach to handle others types of temporal information, such as imprecise temporal information. Besides, we should improve the integration of our temporal search engine with the other similarity metrics. This task will enable us to evaluate the performance of our tool when solving queries concerning more then one dimension. Finally, other important improvement to be undertaken consists of integrating our solution to the current catalog service interface.

REFERENCES

Allen, J. F., (1998). Maintaining Knowledge about Temporal Intervals. *Communications of the ACM,* 26(11), 832-843.

Alonso, O., Gertz, M. and Baeza-Yates, R. A., (2006). Clustering of search results using temporal attributes. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press.

Andrade, F.G. and Baptista, C. S., (2011). Using Semantic Similarity to Improve Information Discovery in Spatial Data Infrastructures. *Journal of Information and Data Management,* 2(2), 181-194.

Baeza-Yates, R. and Ribeiro-Neto, B., (1999). *Modern information Retrieval*. Wokingham: Addison-Wesley.

Bernard, L., and Craglia, M. (2005). SDI - From Spatial Data Infrastructure to Service Driven Infrastructure. In *Workshop on Cross-learning between Spatial Data Infrastructures, and Information Infrastructures.*

Fox, E. A. and Shaw, J. A., (1993). Combination of Multiple Searches. In *Second Text Retrieval Conference*. NIST Special Publications.

Hübner, S. and Visser, U., (2003). Temporal Representation and Reasoning for the Semantic Web. In *Twenty-First International Florida Artificial Intelligence Research Society Conference.* AAAI Press.

Jin, P., Li, X., Chen, H., and Yue, L., (2010). CT-Rank: A Time-aware Ranking Algorithm for Web Search. *Journal of Convergence Information Technology*, 5(6), 99-111.

Manica, E., Dorneles, C. F., and Galante, R. M., (2010). Supporting Temporal Queries on XML Keyword Search Engines. *Journal of Information and Data Management*, 1(3), 471-486.

Open Geospatial Consortium. (2004). *OGC Web Map Service Interface*. Retrieved February 26, 2012, from: http://portal.opengeospatial.org/files/?artifact_id=4756

Open Geospatial Consortium, (2005). *Web Feature Service implementation specification*. Retrieved February 26, 2012, from: https://portal.opengeospatial.org/files/?artifact_id=8339.

Pustejovsky J., Castaño J., Ingria R., Saurí R., Gaizauskas R., Setzer A. and Katz G., (2006). TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Fifth International Workshop on Computational Semantics.*

Strötgen, J. and Gertz, M., (2010). A System for Exploring Spatio-Temporal Information in Documents. *Proceedings of the VLDB Endowment*, 3(1), 1569-1572.

Tversky A., (1977). Features of similarity. *Psychological Review*, 84 (4), 327-352.

Vögele, T., Hübner, S. and Shuster, G., (2003). BUSTER - An Information Broker for the Semantic Web. *Künstliche Intelligenz*, 17(3), 31- 34.

Williamson I., Rajabifard, A. and Feeney, M. F., (2003). *Developing Spatial Data Infrastructures: From Concept to Reality*. London: Taylor & Francis.

SNMPFS
An SNMP Filesystem

Rui Pedro Lopes[1,2], Tiago Pedrosa[1,2] and Luis Pires[1]
[1]*Polytechnic Institute of Bragança, Bragança, Portugal*
[2]*IEETA, University of Aveiro, Aveiro, Portugal*
{*rlopes, luica, pedrosa*}*@ipb.pt*

Keywords: Network Management, SNMP, File system.

Abstract: The intensive operation of constantly growing, heterogeneous networks is a challenging task. Besides the increase in size and complexity, data networks have become a critical factor for the success of many organizations. During the last decade, many network management architectures where developed and standardized, aiming at the definition of an open environment to control and optimize its operation. Many of these architectures use specific protocols and specialized tools, to allow remote monitoring and configuration of network equipment and computational platforms.
One of the best well known paradigms of human-computer interaction is the file concept and the associated file system. Files have been around since early days in the personal computer industry and users are perfectly comfortable associating the work produced in applications to be stored in medium.
In this paper we propose a file system interface to network management information, allowing users to open, edit and visualize network and systems operation information.

1 INTRODUCTION

A file system is a database typically storing large blocks of information. The information is stored in the form of files, structured as a hierarchy of directories. Each entry in the file system, including directories and files, is characterized by a limited group of attributes (instant of creation, instant of the last access, instant of the last change, permissions, owner, group). This paradigm is, perhaps, one of the best well known mechanism for storing information, available in several operating systems as well as in some embedded devices, such as PDAs, mobile phones and even digital cameras.

Usually, files are stored locally in persistent memory, such as hard-disks or memory cards. It is also common, particularly in enterprises, to use network file systems to store files in a remote server, accessed through a special protocol like NFS (Shepler et al., 2000) or SMB. In this situation, the network server exports part of the local file system to a set of selected clients, allowing them to remotely access files and directories. However, this centralized, single server approach, suffers some scalability issues related to throughput, capacity and fault tolerance.

Distributed file systems are designed to improve scalability and fault tolerance by transparently balancing the access between servers. It provides the same view to every client and is responsible for maintaining coherence of data through distributed locking and caching (Ghemawat et al., 2003; Howard et al., 1988; Anderson et al., 1996).

Yet another paradigm, cloud storage systems, such as Dropbox[1], transparently synchronize the local copy of data with a remote datacenter, allowing user access to personal and shared files anytime, anywhere.

Based on this paradigm, we considered the possibility of accessing network management information as a set of virtual file systems. Network resources, usually accessed through SNMP (Harrington et al., 2002), COPS (Durham et al., 2000), or other network management protocol, are seen as remote shares, to be mounted in a regular workstation file system and the instrumentation and configuration information accessed as regular files.

The nature of the information as well as the purpose of this *SNMP File System* (`SNMPFS`) is radically different from traditional distributed file systems (DFS). This makes the current requirements different

[1]http://www.dropbox.com

of traditional DFS. First, distributed file systems are used to store files belonging to one or more users. In a network management scenario, the information is generated both by network resources and users. The former is used for instrumentation and the later is used for configuration.

Second, the content of the majority of network management backed files is constantly changing, because of the dynamic nature of network parameters. As an example, consider a value representing the number of transmitted packets or a value representing CPU load. In regular file systems, used to store personal and application files, entries seldom change. This allows better cache hit rates than in the former situation.

Third, network management files are very small, resulting from the parameters they represent. Frequently, a single `int` is used and, some times, a small table of values is sufficient.

Fourth, the whole `SNMPFS` is composed of several network services and resources, such as routers, firewalls, web servers, and so on. Each resource exports the information resulting from the instrumentation of working parameters thus playing a part in a potentially huge cluster of distributed file systems.

Fifth, faults are usually frequent, resulting from connectivity problems, hardware failures, human action and others. The system should cope with this issues, by recovering when possible and replicating when necessary.

2 FILE SYSTEM DESIGN

The goal of the `SNMPFS` is to unify around a unique name space all of the enterprise network management agents. For concept proving, we are using SNMP agents, since they are common in organizations and widely implemented by network devices. This approach allows integrating the tree of management objects of distinct SNMP agents in a single file system. Resuming, the goals of this approach are the following:

- allow the use of simple files and directories handling tools (`cd`, `cp`, `cat`, ...);
- allow to consult and change the values through simple tools (file redirection, text editors);
- allow the creation of pipes:
  `cat sysUpTime | toxml | mail --s ''sysUpTime'' admin@host.com`;
- allow using office tools, such as Calc or Excel, to view and alter tabular information;

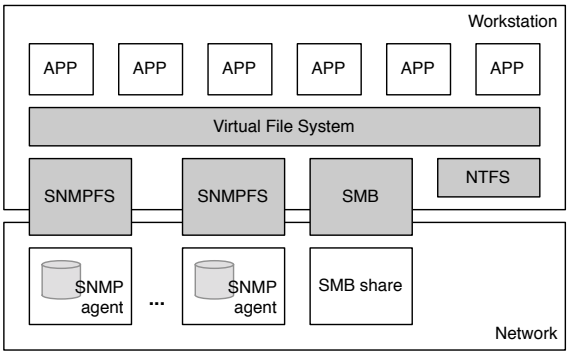


Figure 1: SNMPFS global architecture.

- allow reducing the complexity of the management system.

The information generated by each SNMP agent is accessed through a specific file system (`SNMPFS`), working as a gateway between the regular file operations (`open`, `read`, `write`, `append`, `close`, ...) and SNMP commands (Figure 1).

As other file systems, like Server Message Block (SMB), also known as Common Internet File System (CIFS), to access SMB (Windows) shares across a network or NTFS for local storage, applications access data through a uniform layer (VFS - Virtual File System). Specific file system details are of the responsibility of each file system technology (SNMPFS, SMB, NTFS and so on). As a result, all the information is stored under the same naming tree.

2.1 Topology

The system topology is straightforward: on one side, several SNMP agents, associated with diverse enterprise resources; on the other side, one or more workstations mount the agents' information through the `SNMPFS` (Figure 2).

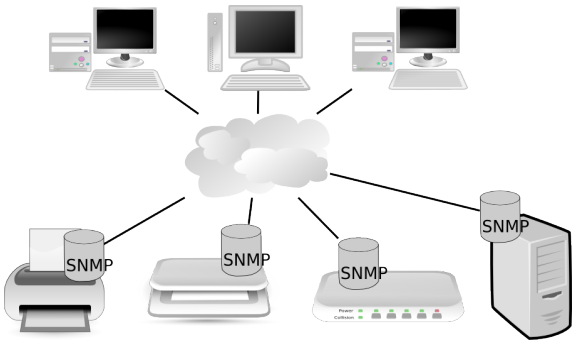


Figure 2: Simplified topology.

It is possible that two or more workstations mount the same agents in its local file system. For read operations this does not present any problem. However,

for update operations it is possible that potentially many processes try to change the same value.

2.2 Locking

In conventional distributed file systems, file locking is essential for coordinating access to shared information among cooperating processes. If multiple processes are writing to the same file it is necessary to regulate the access through some kind of locking mechanism. In SNMPFS, locking is performed by the agent, in accordance with the managed object definition, since SNMP agents may also be updated by network management applications concurrently. Some MIBs provide a mechanism to regulate concurrent access. The Expression MIB (Kavasseri, 2000), for example, has tables with a special column used to instantiate the row – RowStatus (McCloghrie et al., 1999b).

2.3 Security

SNMP is inherently insecure. Although true for the versions 1 and 2c, SNMPv3 present a modular security architecture based on cryptographic protocols and algorithms. It is mandatory that SNMPv3 implementations support the HMAC-MD5-96 protocol for authentication. They can also support the HMAC-SHA-96 for authentication and the CBC-DES for privacy (Blumenthal and Wijnen, 2002). More recently, a new privacy protocol was added. (Blumenthal et al., 2004) describes the Advanced Encryption Standard (AES) for SNMPv3 in the SNMP User-based Security Model which can be used as an alternative to the CBC-DES.

The SNMPv3 security service provides data integrity, data origin authentication, data confidentiality and message timeliness as well as limited replay protection. It is based on the concept of a user, identified by a *userName*, with which security information is associated. In addition to the user name, an authentication key (*authKey*) is shared between the communicating SNMP engines, ensuring authentication and integrity. A privacy key (*privKey*), also symmetric, ensures confidentiality.

Complementing the communication security, the SNMPv3 model also provides access control through a view-based access control model (Wijnen et al., 2002). This model grants or denies access to MIB portions (view subtrees) according to the predefined configuration and the current user permissions.

The security details for SNMPv3, either for authentication, integrity, confidentiality and access control, dictates the security functions for the SNMPFS. We have to pass to the file system the authentication and the access control required by the SNMP model. The authentication problem is performed by the system when mounting the file system. A similar approach is followed for NFS or SMB shares:

```
mount -t smb //server/share /mnt -o
username=aUser,password=xxx.
```

If the server recognizes the username and password, the host is allowed to access the file system and a user ID (uid) is associated with it.

Access control is enforced by file permissions. In Unix, each file has a set of permissions (read, write, execute) for the file owner, group and others. For example, the permissions

```
-rwxr-x---
```

gives the owner the possibility to read, write and execute the file, the group to read and execute and no other user can read, write or execute.

SNMPFS translates each file permission to the View-based Access Control mechanism of the SNMPv3.

2.4 Attributes

File system entries, such as files or directories, are characterized by a set of attributes which describes their fundamental aspects, such as size, date, permissions, name and others. The name and number of attributes is typically static, meaning that it is not possible to add or remove further information to each file system entry latter on.

An attribute which is necessary to better describe the data types and the structure of an SNMP agent is the MIB tree it implements. The MIB tree is described in a set of MIB files which contain each node name, data type, restrictions and role. With this information, the SNMPFS can present to the user a more meaningful set of file names as well as file types (a table, a string, an int, etc). In particular, this information is valuable for tables, which the SNMPFS exports as Coma Separated Values (CSV) format and can be opened and edited by a spreadsheet, such as Microsoft Excel or OpenOffice Calc.

To be able to access the meta-information about management data, the SNMPFS has the possibility to load MIB files from a specific directory. This information will allow the files to have a more meaningful name as well as adapting the content to the nature of the information it stores.

3 IMPLEMENTATION DETAILS

The `SNMPFS` implements a gateway between regular file system access primitives and SNMP commands. Generally, file systems are implemented at kernel space, however, we chose to implement the file system in userspace to facilitate the development and test process. We used FUSE (FUSE, 2011), a stable and well known API for file system development.

The SNMP information is somehow austere, mainly because of the Object IDs (OIDs) that it uses to identify each managed object. The OID is a sequence of integer values, separated by a '.' (dot): 1.3.6.1.2.1.1.1.0. This sequence defines a path in the agent's tree of objects, referring to a specific value (Figure 3). This path, for example, points to a specific object, which refers to a value resulting from device instrumentation.

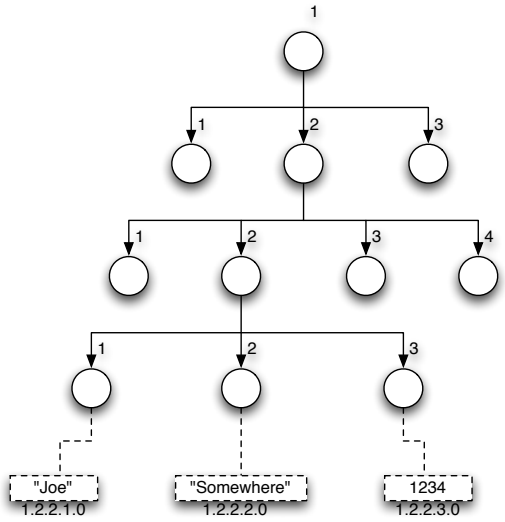


Figure 3: Scalar values organization in the agent.

In Figure 3, for example, the OID 1.2.2.1 refers to the lower left node in the tree. This node is associated with a specific value, a scalar, in this case, which contains the string "Joe". The scalar is viewed as an additional node, a leaf, and is referred by adding a '.0' to the OID.

Each object, the circles in the figure, has a set of attributes or meta-information, which allows the user to get the semantics of the value (what does "Joe" stands for). The attributes, as well as the overall structure, is described in a file, called a Management Information Base (MIB), which associates a descriptive, meaningful, name to each object and further describes the data type, access restrictions, OID structure and others. From the user perspective, this also allows mapping the sequence of integers to a short name: it is easier to refer to each object by the short name, instead of the OID (`get sysDescr`, instead of `get 1.3.6.1.2.1.1.1.0`).

After parsing the MIB files, the `SNMPFS` performs this mapping, storing the meta-information to improve the information provided to the user by the file system. The OID in the sequence of integers format will only be used when the MIB is not available.

3.1 The MIB Parser

As mentioned above, the MIB structure is important to `SNMPFS` to provide a more useful and meaningful view of the file system. MIB files are written in a subset of the Abstract Syntax Notation One (ASN.1), called "Structure of Management Information" (McCloghrie et al., 1999a). The MIB must be parser so that the tree, nodes, types and objects extracted. We are using Marser (Marser, 2007), an API to parse SMI (v1 and v2).

Moreover, the information from the MIB allows to identify tabular information, which further helps the file system to present the information to the user in a more manageable way. In this case, tables will be available is CSV format, allowing the user to read and modify it using a spreadsheet application.

Tables are represented as a further extension to the OID tree (Figure 4). As in the previous case, where each scalar is retrieved from a leaf, tabular values are retrieved from several leafs, hanging on the OIDs that represent the columns (in the figure, the table is referred with the OID 1.2.3, which has the columns 1.2.3.1 and 1.2.3.2).

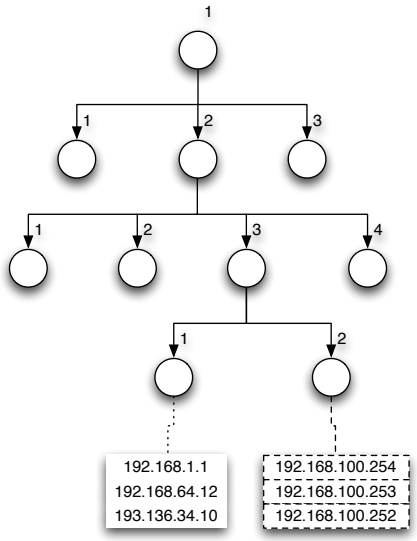


Figure 4: Tabular values organization in the agent.

One or more of the columns are the index, which

allows to retrieve the rows in the table. So, to get the first value of the table, it is necessary to issue: `get 1.2.3.2.192.168.1.1`, which yields `192.168.100.254`.

3.2 Files

When dealing with agents with an unknown structure, the user usually has to explore the management information tree, retrieving all the information that the agent stores. This operation is called 'walk', because it allows to visit all the places "hidden" in the agent. The `SNMPFS` has a special file which allows doing precisely this. The file, called `_walk.txt`, lists all the managed objects retrieved as the result of a 'walk' operation. By simply opening this file in a text editor, the user will be able to immediately see the objects the SNMP agent implements:

```
sysDescr;1.3.6.1.2.1.1.1.0
sysObjectID;1.3.6.1.2.1.1.2.0
sysUpTime;1.3.6.1.2.1.1.3.0
sysContact;1.3.6.1.2.1.1.4.0
sysName;1.3.6.1.2.1.1.5.0
sysLocation;1.3.6.1.2.1.1.6.0
sysServices;1.3.6.1.2.1.1.7.0
sysORLastChange;1.3.6.1.2.1.1.8.0
sysORID;1.3.6.1.2.1.1.9.1.2.1
sysORID;1.3.6.1.2.1.1.9.1.2.2
sysORID;1.3.6.1.2.1.1.9.1.2.3
sysORID;1.3.6.1.2.1.1.9.1.2.4
sysORID;1.3.6.1.2.1.1.9.1.2.5
sysORID;1.3.6.1.2.1.1.9.1.2.6
...
```

With the information obtained in the file, the user can configure the file system, describing which files should be available and what is the name they should have. The configuration is written is XML and define all the aspects of the file-system: the agent's address, access credentials, MIBs to load, which nodes to show and where to mount:

```
<device name="device">
  <mount dir="tmp" />

  <mibs dir="../mibs/">
    <mib file="SNMPv2-MIB"/>
    <mib file="RFC1213-MIB"/>
    <mib file="IF-MIB"/>
  </mibs>

  <snmp address="192.168.1.1" port="161"
      version="v2c" community="public" />

  <entries>
    <scalar label="sysUpTime" />
    <scalar label="sysDescr" />
    <table label="ifTable" />
    <table oid=".1.3.6.1.4.1.63.501.3.2.2"
        file="myTable">
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.1" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.2" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.3" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.4" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.5" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.6" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.7" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.8" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.9" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.10" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.11" />
      <col oid=".1.3.6.1.4.1.63.501.3.2.2.1.12" />
    </table>
  </entries>
</device>
```

The previous configuration file will result in the appearance of 5 files in the 'tmp' directory – two tables, two scalars and the `_walk.txt` (Figure 5).

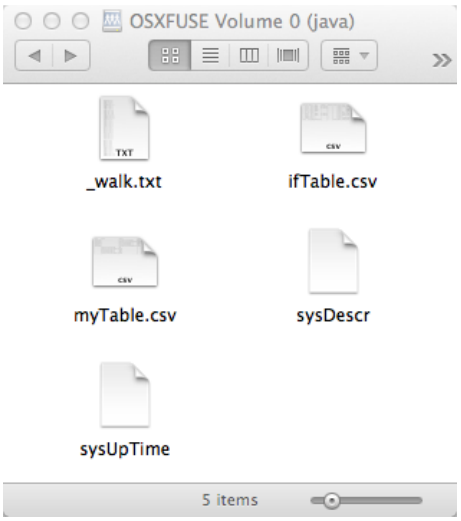


Figure 5: Screenshot of an SNMPFS directory.

3.3 Values and Tables

Each file representing a scalar simply has to get the value from the agent each time it is read. In the previous configuration file, the scalar 'sysUpTime' and 'sysDescr' show as files, containing the information from the agent located in 192.168.1.1.

Tables, because of the tree like structure, require more processing. The algorithm we follow is:

```
read configuratiom file;
for each entry
  if is table
    read columns;
    if columns empty
      read columns from MIB;
  for each column
    while has more leafs
```

```
        get leaf;
        store in row,column;
end;
```

The fetch of values and store in row and column format allows to build the content around the CSV format, that can be manipulated by a spreadsheet application (Figure 6).

4 USE CASES

In system and network management, the administrators are used to create scripts to automate some of the typically repetitive and/or boring maintenance activities. Many of this scripts work by reading, writing and updating configuration files, altering the way daemons and services work (such as e-mail, HTTP, SSH, ...). The `SNMPFS` enables mapping SNMP information to a file system structure, thus contributing to the integration of monitoring, configuration and accounting processes. In other words, the `SNMPFS` will foster administrators to develop new tools easier and to use the same tools in different scenarios because of the transparency and integration of the file system paradigm.

The extension of the file system with files related to instrumentation and configuration information from network devices will further alleviate the burden. By mounting several devices in the same file system, where each directory represents a different agent, enables the administrators to be able to make queries or change values on all the equipment, just by browsing the file system.

4.1 Monitoring

Monitoring operations require the user to retrieve, process, analyse and visualize instrumentation information. For example, several parameters can be queried to get the status of remote hosts (Table 1).

Table 1: Monitoring examples.

Query	Object	Type
the number of users on a system	hrSystemNumUsers	scalar
the bytes transmitted	ifInOctets;ifOutOctets	table
the storage areas	hrStorageTable	table
the processor load	hrProcessorLoad	table

Other common operation is to build the topology map of the network, representing the connections and hosts structure. This is done with the help of the IP forwarding table, maintained in the switches and routers. Each table gathers the MAC addresses in each port, allowing the correlation of addresses into building a visual representation of the network topology. By replicating this information (a simple copy will do), will enable to create a view of the network a specific times.

4.2 Configuration

One challenge on integrated management of networks is how to apply policies that are transversal to more that one equipment. The integration of several SNMP agents in the same file system can enable the administrator to create sound scripts to apply the policy.

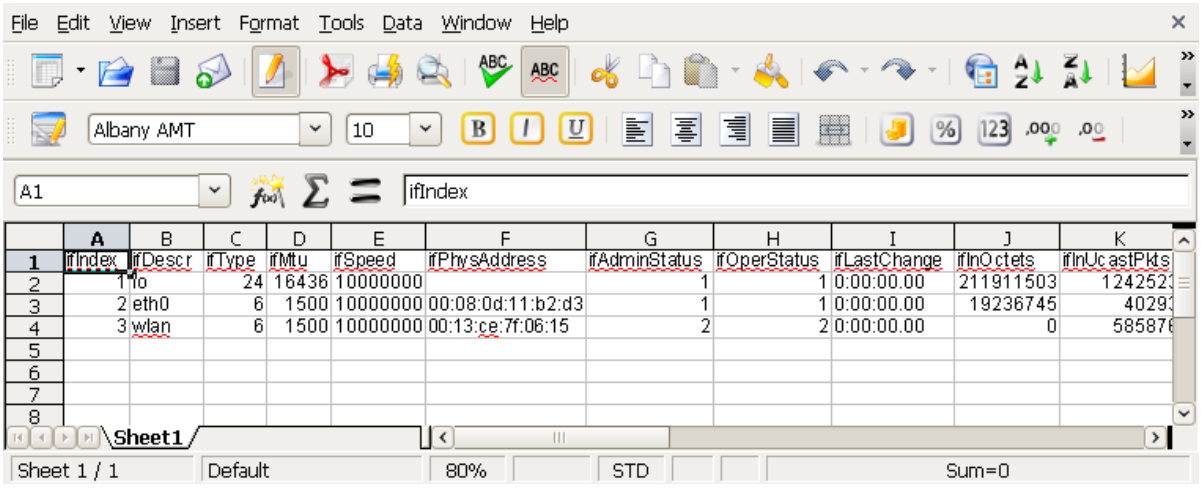
In complement to these use cases, one that is being currently addressed is the use of version control systems for maintaining snapshots of SNMP agents' configuration. Each type and version of equipment has different ways for manipulating their configurations, because of the MIBs they implement. With the `SNMPFS`, administrators can make use of already accepted solutions for version control. A Distributed Version Control, such as GIT[2] enables to have a master repository for each agent as well as a local repository in the management stations.

The first mount of the agent, a new repository is created and the files are added, tagged as the initial commit. This local repository is pushed to the master repository, which will work as another, more general, versioning peer. The master is configured for post commit routines, that will update the directory where the agent is mounted. The master will apply the changes in the file system, configuring the equipment accordingly and changing the properties that enables the updated of the configurations. This enables that different administrators can work on their stations using the repositories for changing configurations and the push the configurations to the master that will change the agents' status. Moreover, using tags to identify specific configuration versions allows to easily change form one configuration to another.

4.3 Scheduling Operations

Modern operating systems have tools that enables users to schedule jobs (commands or shell scripts) to run periodically at certain times or dates. One such tool, popular un Unix-like operating systems, is `cron`, used to automate system maintenance or administration. Proper configured cron jobs allows to activate some options at some time on agents, for example, to shutdown several devices at a specific time. Moreover, it also allows to watch files for specific values for, for example, triggering some configuration change or event (send emails, executing commands).

[2] http://git-scm.com/



	A	B	C	D	E	F	G	H	I	J	K
1	ifIndex	ifDescr	ifType	ifMtu	ifSpeed	ifPhysAddress	ifAdminStatus	ifOperStatus	ifLastChange	ifInOctets	ifInUcastPkts
2	1	lo	24	16436	10000000		1	1	0:00:00.00	211911503	1242521
3	2	eth0	6	1500	10000000	00:08:0d:11:b2:d3	1	1	0:00:00.00	19236745	40291
4	3	wlan	6	1500	10000000	00:13:ce:7f:06:15	2	2	0:00:00.00	0	585878
5											
6											
7											
8											

Figure 6: Editing the ifTable with a spreadsheet.

5 CONCLUSIONS

Accessing and updating information is a frequent operation in virtually any activity. Because of the evolution of computing platforms, the electronic information is associated with the concept of files, residing in a generic storage mechanism. Usually, the files are updated by general use applications, such as office suites or drawing editors. Often, the content is in plain text, allowing standard editors to retrieve and update the information.

Because of the ubiquity of files, modern operating system have an extensive set of tools to deal with the maintenance of files, such as renaming, creating, copying, backing up and restore, and so on. In enterprises, work files are typically stored in network storage, made available through a virtual file system in local desktop computers.

Network and system management (NSM) is a major concern for maintaining the system in good working conditions. Many of the tasks involved in NSM require monitoring and updating information resulting from instrumentation procedures in applications, services and equipment. The paradigm in traditional NSM models rely on client-server protocols, through special purpose applications.

The `SNMPFS`, proposed in this paper, integrates network devices, applications and service management in a common platform and paradigm – the file system. In this way, network management operations can benefit from the existing powerful operating system tools to monitor, update instrumentation, configuration and monitoring information.

REFERENCES

Anderson, T. E., Dahlin, M. D., Neefe, J. M., Patterson, D. A., Roselli, D. S., and Wang, R. Y. (1996). Servless network file systems. *ACM Transactions on Computer Systems*, 14(1):41–79.

Blumenthal, U., Maino, F., and McCloghrie, K. (2004). The Advanced Encryption Standard (AES) Cipher Algorithm in the SNMP User-based Security Model. RFC 3826 (Proposed Standard).

Blumenthal, U. and Wijnen, B. (2002). User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3). RFC 3414 (Standard).

Durham, D., Boyle, J., Cohen, R., Herzog, S., Rajan, R., and Sastry, A. (2000). The COPS (Common Open Policy Service) Protocol. RFC 2748 (Proposed Standard). Updated by RFC 4261.

FUSE (2011). Filesystem in userspace. http://fuse.sourceforge.net.

Ghemawat, S., Gobioff, H., and Leung, S.-T. (2003). The google file system. Technical report, Google.

Harrington, D., Presuhn, R., and Wijnen, B. (2002). An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks. RFC 3411 (Standard).

Howard, J. H., Kazar, M. L., Menees, S. G., Nichols, D. A., Satyanarayanan, M., Sidebotham, R. N., and West, M. J. (1988). Scale and performance in a distributed file system. *ACM Transactions on Computer Systems*, 6(1):51–81.

Kavasseri, R. (2000). Distributed Management Expression MIB. RFC 2982 (Proposed Standard).

Marser (2007). Marser - a mib parser. http://marser.sourceforge.net.

McCloghrie, K., Perkins, D., and Schoenwaelder, J. (1999a). Structure of Management Information Version 2 (SMIv2). RFC 2578 (Standard).

McCloghrie, K., Perkins, D., and Schoenwaelder, J. (1999b). Textual Conventions for SMIv2. *RFC 2579 (Standard)*.

Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C., Eisler, M., and Noveck, D. (2000). NFS version 4 Protocol. *RFC 3010 (Proposed Standard)*. Obsoleted by RFC 3530.

Wijnen, B., Presuhn, R., and McCloghrie, K. (2002). View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP). *RFC 3415 (Standard)*.

Mining Generalized Association Rules using Fuzzy Ontologies with Context-based Similarity

Rodrigo Moura Juvenil Ayres and Marilde Terezinha Prado Santos
Department of Computer Science, Federal University of São Carlos, Rod. Washington Luis, São Carlos, Brazil
{rodrigo_ayres, marilde}@dc.ufscar.br

Keywords: Generalized Association Rules, Fuzzy Ontologies, Post-processing, Context-based Similarity.

Abstract: In crisp contexts taxonomies are used in different steps of the mining process. When the objective is the generalization they are used, manly, in the pre-processing or post-processing stages. On the other hand, in fuzzy contexts, fuzzy taxonomies are used, mainly, in the pre-processing step, during the generation of extended transactions. A great problem of such transactions is related to the generation of huge amount of candidates and rules. Beyond that, the inclusion of ancestors in the same ends up generating problems of redundancy. Besides, it is possible to see that many works have directed efforts for the question of mining fuzzy rules, exploring linguistic terms, but few approaches have proposed new steps of the mining process. In this sense, this paper propose the *Context* FOntGAR algorithm, a new algorithm for mining generalized association rules under all levels of fuzzy ontologies composed by specialization/generalization degrees varying in the interval [0,1]. In order to obtain more semantic enrichment, the rules may be composed by similarity relations, which are represented at the fuzzy ontologies in different contexts. In this work the generalization is done during the post-processing step. Other relevant points are the specification of a generalization approach; including a grouping rules treatment, and an efficient way of calculating both support and confidence of generalized rules during this step.

1 INTRODUCTION

An important task in data mining is the mining association rules, introduced in (Agrawal et al., 1993). In traditional algorithms of association, like Apriori, the rules are generated based only on existing items in the database. This characteristic makes an excessive amount of rules be produced. In this sense, the domain knowledge, represented via taxonomies, can be used in order to obtain more general patterns, facilitating the user's comprehension. The association task using taxonomic structures is called mining generalized association rules, and was introduced by (Srikant and Agrawal, 1995) and (Jiawei Han and Fu, 1995).

According to the authors, ancestors of taxonomy are inserted into database transactions, which are called extended transactions. Then, from these extended transactions, it is applied an algorithm for extract the final set of rules, which can be composed by traditional rules and generalized ones. However, the inclusion of ancestors in the database transactions results the generation of many candidate itemsets, in addition, algorithms using such transactions ends up generating redundant patterns, making it extremely necessary the use of interest measures for eliminate redundancies. On the other hand, some works, like (Carvalho et al., 2007) for example, show that the post-processing stage can be more advantageous, because few candidates and rules are generated. Moreover, it is eliminated the need of measures used for prune redundant rules, since the process is made based on the traditional patterns generated.

However, in many applications of the real world ontologies and taxonomies may not be crisp, but fuzzy (Wei and Chen, 1999), because some applications do not have classes of objects with pertinence criteria precisely defined (Zadeh, 1965). In this context, Wei and Chen (Wei and Chen, 1999) introduced the use of fuzzy taxonomies. They considered the partial relationships possibly existing in taxonomies, where an item may partially belong to more than one parent. For instance, tomato may partially belong to both fruit and vegetable with different degrees. Wei and Chen thus defined a fuzzy taxonomic structure and considered the extended degrees of support, confidence and interest

measures for mining generalized association rules. However, most of the works are focused in to improve methods of to obtain generalized fuzzy association rules, which are the ones composed by linguistic terms, but few works have directed efforts for improve the exploring of generalized rules under fuzzy concept hierarchies, mainly in relation to the stage that they are used.

Besides, some works, like (Miani et al., 2009) and (Escovar et al., 2006), explore the semantic enrichment through similarity relations. However, these works do not consider that the degree of a similarity relation, between two or more elements, it is also related to the point of view or to the context analysed. For example, consider the problem of compare two vegetables, tomato and khaki, in relation to two different points of view (contexts), appearance and flavour. In respect to the appearance context, would be possible to check that tomato is very similar to khaki, with a very high degree of similarity; but in relation to the flavour, would be possible to check that both are bit similar, with a minor degree of similarity.

Thus, this paper presents the *Context* FOntGAR algorithm for mining generalized association rules, using fuzzy ontologies composed by relationships of specialization/generalization varying in the interval [0,1], and similarity relations with different degrees according to the context. The generalization can to occur in all levels of fuzzy ontologies. The paper is organized as follow: Section two shows some related works. Section three presents the *Context* FOntGAR algorithm. The section four presents the experiments, and the section five shows the conclusions.

2 BACKGROUND

Aiming to obtain general knowledge, the generalized association rules, which are rules composed by items contained in any level of a given taxonomy, were introduced by (Srikant and Agrawal,1995). There are many works using crisp taxonomic structures. These works are distinguished, mainly, in function of the stage (of the algorithm processing) in which these structures are used.

In the pre-processing, the generalized rules are obtained through extended databases, and these bases are generated before the pattern generation. Extended databases are the ones composed by transactions containing items of the original database and ancestors of the taxonomy. In the post-processing the generalized rules are obtained after the generation of the traditional rules, through a sub-algorithm that uses some generalization methodology based on the patterns generated.

In (Wu and Huang, 2011), the mining is made using an efficient data structure. The goal is to use the structure for find rules between items in different levels of a taxonomy tree, under the assumption that the original frequent itemsets and association rules were generated in advance. Thus, the generalization occurs during the post-processing step. In relation to the post-processing, (Carvalho et al., 2007) proposed the GARPA algorithm. The algorithm, unlike what was proposed by (Srikant and Agrawal, 1995), do not insert ancestor items in the database transactions. The generalization was done using a method of replacing rule items into taxonomy ancestors. From the quantitative point of view, this process is more advantageous than proposed by (Srikant and Agrawal, 1995), because implies a smaller amount of candidates, and consequently of rules generated, dispensing the use of measures for pruning redundant rules.

In mining generalized rules, most of the works using fuzzy logic are mainly focused in to obtain generalized fuzzy association rules, which are the ones composed by fuzzy linguistic terms, such as young, tall, and others. In such approaches are used crisp taxonomies and the linguistic terms are generated based on fuzzy intervals, normally generated through clustering. Besides, these works are directed to explore quantitative or categorical attributes. In this context we can to point, for example, the works (Hung-Pin et al., 2006), (Mahmoudi et al., 2011), (Cai et al., 1998), (Hong et al., 2003) and (Lee et al., 2008). On the other hand, few works use fuzzy taxonomies in order to obtain their rules. In this case, the focus is not the exploring of patterns composed by linguistic terms, but it is how to explore taxonomic structures composed by different specialization/generalization degrees.

The problem of mining generalized rules using fuzzy taxonomies was proposed by (Wei and Chen, 1999). They included the possibility of partial relationship in taxonomies, i.e., while in crisp taxonomies the specialization/generalization degrees are 1, in fuzzy structures such degrees vary in the interval [0,1]. So, the degree μ_{xy} which any node y belongs to its ancestor x can be derived based upon the notions of subclass, superclass and inheritance, and may be calculated using the max-min product combination. Specifically,

$$\mu_{xy} = \max_{\forall l:\, x \to y} \left(\min_{\forall e \text{ on } l} \mu_{le} \right) \quad (1)$$

Where l: $x \to y$ is one of the paths of attributes x and y, e on l is one of the edges on access l, μ_{le} is

the degree on the edge *e* on *l*. If there is no access between *x* and *y*, $\mu_{xy} = 0$ (Wei and Chen, 1999).

In addition to defining such structures, they also consider extended degrees of support and confidence. The degree of the extended support (*Dsupport*) is calculated based on this μ_{xy}. If *a* is an attribute value in a certain transaction $t \in T$, *T* is the transaction set, and *x* is an attribute in certain itemset *X*, then, the degree μ_{xa} can be viewed as the one that the transaction *{a}* supports *x*. Thus, the degree that *t* supports *X* may be obtained as follows:

$$\mu_{tX} = support_{tX} = \min_{\forall x \in X} (\max_{\forall a \in t} \mu_{xa})) \quad (2)$$

Furthermore, an $\sum count$ operator is used to sum up all degrees that are associated with the transactions in T, in terms of how many transactions in T support X:

$$\sum_{\forall t \in T} count\,(support_{tX}) = \sum_{\forall t \in T} count\,(\mu_{tX}) \quad (3)$$

Thus, the support of a generalized association rule $X \rightarrow Y$, let $X \cup Y = Z \subseteq I$, can be obtained as follows, where $|T|$ is the total of transactions in the database:

$$\sum_{\forall t \in T} count\,(\mu_{tZ}) / |T| \quad (4)$$

Similarly, the confidence ($X \rightarrow Y$), called *Dconfidence*, can be obtained as follows:

$$\sum_{\forall t \in T} count\,(\mu_{tZ}) / \sum_{\forall t \in T} count\,(\mu_{tX}) \quad (5)$$

It is important to say in (Wei and Chen, 1999) only the concepts are defined and in (Chen and Wei, 2002) the authors proposed two algorithms to realize the mining, one working with the mentioned taxonomies, and other working with these taxonomies and linguistic terms. The first was called FGAR, and the second was called HFGAR, both algorithms use the same concept of extended transactions.

A similar work can be found in (Keon-Myung, 2001), however, it is related to the mining generalized quantitative association rules. The authors use two different structures: fuzzy concept hierarchies and generalization hierarchies of fuzzy linguistic terms. In the first, a concept may have partial relationship with several generalized concepts, and the second is a structure in which upper level nodes represent more general fuzzy linguistic terms.

As well as Wei and Chen (Wei and Chen, 1999), (Keon-Myung, 2001) also use the technique of extended transactions. Besides, it is considered the use of interest measures for prune redundant rules. According to (Wen-Yang et al., 2010), the works using fuzzy taxonomies, like proposed by (Wei and Chen 1999), require the same be static, ignoring the fact they cannot necessarily be kept unchanged. For example, some items may be reclassified from one hierarchy tree to another for more suitable classification.

In this sense, the work (Wen-Yang et al., 2010) introduces an algorithm where the final set of rules generated can be updated according to the evolution of the structures. The evolution can to occur due four basic causes: insertion, deletion, renaming and reclassification of items. Fuzzy taxonomies are used and, as well as (Wei and Chen, 1999), (Keon-Myung, 2001), and (Wen-Yang et al., 2010), the generalized rules are obtained using extended transactions.

Thus, in respect to the use of fuzzy taxonomies, composed by degrees of specialization/ generalization varying in the interval [0,1], the works (Wei and Chen, 1999) , (Keon-Myung, 2001), and (Wen-Yang et al., 2010), are the most relevant found in the literature.

On the other hand, some works, like (Escovar et al., 2006) and (Miani et al., 2009) are directed to the semantic of the data mined. They use ontologies for extract associations of similarity existing between items of the database. These relations are represented in the leaves of ontology, but the specialization/generalization degrees are constant 1, like crisp ontologies. The work (Miani et al., 2009) is an extension of (Escovar et al., 2006), and the main differences are the introduction of a redundancy treatment and a step of generalizing non-frequent itemsets. However, both algorithms are limited, since generalizes at only one level of ontology (leaf nodes to parents).

As said, these works do not consider the question of context in the similarities represented at the leaves. In this line, the work (Cerri et al., 2010) propose an Upper Fuzzy Ontology With Context Representation (UFOCoRe), an approach that represent multiple relationship strengths in a single ontology, so that it is possible to express different relationship semantics depending on the context chosen. The approach does not define context ontology like the ones used in context-aware systems, but it allows organizing the context information of multiple perspectives in single domain ontology. As described, there are few works dealing with mining generalized association rules under fuzzy taxonomies. Besides, most of the works are inserted in the line of mining generalized fuzzy association rules, which is a concept smoothly different, since for it are used crisp taxonomies and the fuzzy generalized rules are obtained, most of the

time, with the utilization of linguistic terms. Besides, it is possible to see a bias, which is the realization of the generalization process exploring fuzzy taxonomies during the pre-processing stage, through extended transactions. In this sense, considering the concept of fuzzy taxonomies, presented in (Wei and Chen, 1999), no work to date was proposed for obtain generalized rules during the post-processing stage including the questions of similarity relations considering context.

3 THE PROPOSED ALGORITHM

The aim of the *Context* FOntGAR is post-process a set of specialized association rules (AR) using fuzzy ontologies, in order to obtain a reduced non-redundant and more expressive set of generalized rules, facilitating the user's comprehension. Figure 1 illustrates all steps of the *Context* FOntGAR algorithm. The steps colored in grey are the main points of our algorithm.

3.1 Main Ideas

The process of generating traditional association rules is based on Apriori (Agrawal and Srikant, 1994), and as an mining association rule algorithm, it needs of an user-provided minimum support and minimum confidence parameters to run. Moreover, it needs of a minGen, a side and a context parameters:

- ***minsup***, which indicates the minimum support;
- ***minconf***, represents the mininum confidence;
- ***minGen***, which represents the minimum quantity of descendants in different specialized rules;
- ***minSim,*** which is the minimum similarity used in the reasoner inferences (Miani et al., 2009);
- ***side***, which represents the side of generalization;
- ***context,*** which represents the context used in the similarity;

The minsup, minconf, minGen and minSim parameters are expressed by a real value in the interval [0,1]. The side parameter is expressed by a string left, right or lr, indicating the generalization side. The generalization can be done on one side of the rule (antecedent or consequent) or both sides (lr: left and right side). While the left side indicates relations between classes of items and specialized items, the side *right* indicates relations between the specialized items and classes of items. The side *lr* indicates relations between classes. The similarities are represented in the leaves of ontology. Relations with similarity degree value greater than or equal to the user-provide minSim (Miani et al., 2009) can be show in the rules generated, increasing the semantic enrichment of the same. The generalization is made through a sub-algorithm that uses a methodology of grouping and replacement in the rules. In this methodology, two or more rules are grouped in order to be replaced by a unique generalized rule. Several groups can be generated, and the grouping is done based on the parameter side and on the fuzzy ontology. In this case, two or more rules having identical parents in the side of generalization are grouped in a same group.

It is important to say that a group is generated only if two or more rules can be grouped, because is not reasonable generalize a unique rule. As several groups may be generated, various generalized rules may be obtained. During the grouping, the ancestors analyzed are the immediate ones of items present on rules in question, which are the ancestor presents in the current level of generalization. The parameter *side* indicates the generalization side. Thus, when this parameter is set with *left* or *right*, if two or more rules have the same elements in the opposite of *side*, and have identical parents in relation to the items present in the *side*, then these rules are placed in a same group. For example, supposing ontology of bread and milk, where bread is a *breadA, breadB, breadC, breadD, breadE*, and milk is a *milkA, milkB milkC*. Suppose the algorithm generates, during the extracting patterns stage, a set of traditional rules *milkA* → *breadA, milkA* → *breadB, milkA* → *breadC*, which are the ones composed only by leaf nodes.

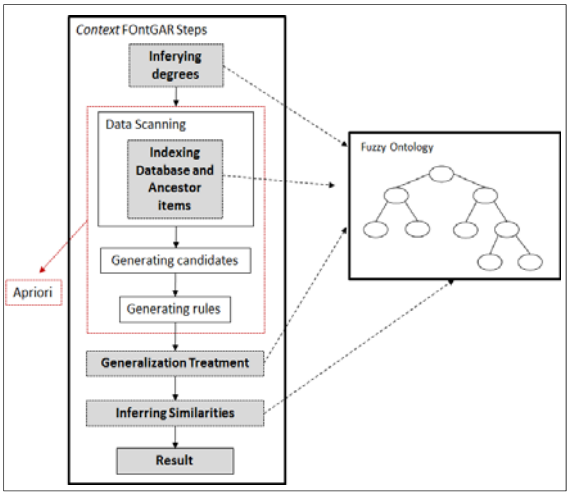


Figure 1: Steps of the *Context* FOntGAR.

When the parameter *side* is *lr,* if two or more rules have the same parents in relation to the

Pseudo-Code of the Generalization Treatment

```
1  if (side = left) or (side = right) or (side = lr) then;
2       level:= 1;
3       nonGeneralizedRules:= all traditional rules;
4       while (level < total of levels) do
5         ontologyVerification(nonGeneralizedRules);
6         aux:= result of ontologyVerification;
7         groupingRules(result of nonGeneralizedRules, aux and side);
8         groupedrules:= result of groupingRules;
9         for (all groups in groupedRules) do
10            all rules in a group are represented by a general rule;
11            verify if the minGen is satisfied;
12            verify other generalization criteria;
13            if (replacement can occur) then
14              do the calculus of support of the general rule;
15            end if
16            if (the general rule is frequent)
17              all rules of the group are replaced by the general rule;
18              generalizedRules:= the general rule;
19            else if (level = 1) then
20                rules of the group will be show in the result;
21        end for
22        if (generalizedRules contains generalized rule) then
23           level: level + 1;
24           for (all rules of generalizedRules) do
25             add the rule generalized into nonGeneralizedRules;
26           end for;
27        end if;
28        if (generalizedRules is empty) then
29          break;
30      end while;
31  end if;
```

Figure 2: Pseudo-code of generalization.

antecedent items, and, respectively, have the same parents in relation to the consequent items, then these rules will be grouped together. For example, considering that traditional rules *milkA* → *breadA, milkB* → *breadB, milkC* → *breadC* have been generated. Comparing these rules, we can see that they have the same parent in relation to the antecedent, and respectively, they have the same parent in relation to the consequent. Thus, these rules will be grouped together.

It is important to say the rules used in the grouping can be composed by any quantity of items. At first, the patterns used during the generalization are the traditional ones generated by the extracting patterns stage. Posteriorly, the obtained generalized rules are treated in the same way, in order to obtain a new set of generalized rules. Thus, it is a recursive process. An important point is that generalized rules can be generated without the use of all descendants of an ancestor. In this sense, to avoid an over-generalization, a set of specialized rules contained in a group can be substituted by a more general rule only if a minGen parameter (Miani et al., 2009) was satisfied. Consider that the minGen value is 0.6 (60%), and the side is lr, the rule milk → bread will be generated even if there is no rule for each kind of bread and milk in the current group, but only if 60% of descendants of bread and milk are present in this set of rules. Thus, the use of minGen could produce a semantic loss. In this sense, in order to guide the user's comprehension, the algorithm show the items which have not participate in the generalization process. For example, suppose the item breadE is not present in the specialized AR set, the generalized rule are shown as milk → bread (-breadE), indicating that the item breadE did not compose the generalization.

In this research, for represent a fuzzy ontology with specialization/generalization degrees varying in [0,1] and context in similarity relations, we follow the ideas described in two meta-ontologies, proposed in (Agrawal and Srikant, 1994), and (Cerri et al., 2010) respectively. Both are upper ontologies as it represent fuzzy constructs to be inherited and/or instantiated by specific domain ontologies. Such ontologies are based on OWL DL (Smith et al., 2004), a W3C recommendation supported by several reasoners and application programming interfaces used to develop ontology-based applications.

3.2 The Algorithm Step by Step

First, the ontology reasoner is used to infer the membership degrees of the leaves in relation to the ancestors, through the equation 1 of the section two. These degrees are stored in a data structure. The steps of data scanning, generating candidates and generating rules are done similarly to the Apriori.
At end of generating rules we have a set of specialized rules, which will be used on the generalization treatment. Then, the generated rules and the side of generalization are passed to the groupingRules function (line 7), which is responsible by the grouping treatment mentioned

above. Posteriorly, for each group generated, all rules in a group are represented by a more general rule (line 10). So, the minGen parameter (line 11) is checked, besides, it is verified if antecedent ∩ consequent = 0 and if no consequent item is ancestor of any antecedent item (line 12). If such verifications are satisfied (line 13), the calculus of support is done. If the general rule is not frequent then the generalization is not made. In this case, if the level is 1 (line 19), the rules of the corresponding group are inserted in the result. But if the general rule is frequent, the rules of the corresponding group are replaced by the same, and it is inserted in the result.

After that, if there are generalized rules, the same are used in the next level of generalization. If this situation is true for all next levels, the generalization process will be done until a level below the ontology root. However, if there is no generalized rule at a certain level, then will be impossible generalize in the next levels. When this happens, the generalization process is concluded. After the generalization treatment, the algorithm uses the ontology reasoner to obtain the similarity relations. So, these relations are used in the non-generalized rules. Finally, after that, the algorithm enters its final stage, which is the results generation.

3.3 Calculating the Support and Confidence Degrees

Considering the fuzzy taxonomy of Figure 3, *Fruit* → *Meat* is a generalized rule and {Fruit, Meat} is their itemset format. The support is calculated based on the sum of all degrees of transactions that support simultaneous occurrences of {Fruit, Meat}. However, {Fruit, Meat} is obtained and known only during the post-processing. Then, for obtain the degree of each transaction, it would be necessary a new scanning in the database. As many generalized rules may be generated, the quantity of new scanning also may be huge, and depending on the quantity of rows of the database, the performance of the algorithm would be affected.

In *Context* FOntGAR we use two data structures (Figure 3 and Figure 4) to allow the calculating of support avoiding additional scan. Such structures are composed by keys and values. In Figure 3, a key is an item of the database or an ontology ancestor. Each key points a value, which is a vector storing the transaction identifiers where the key appear. The vector is an object of the class Vector in Java, dynamically created. The equation used in the calculus of support is derived of the Equation 2 (section two). So, if we partitioned the same in two subparts (Part 1 and Part 2), we have:

- **Part 1** = $\max_{\forall\, a \in t}(\mu_{xa})$
- Part 2 = $\min_{\forall\, x \in X}$ ($\boldsymbol{Part\ 1}$).

As said, we can have many generalized rules, but we don't know what will be generated. So, the itemset format of each may be any $X = \{x_1, \ldots, x_n\}$, where X is the generalized rule, and $x_1, \ldots, x_n$ are items of the rule. That way, during the first scan, we do the computation of Part 1, which is the degree that each transaction *t* supports an ancestor *x*. Based on the results of Equation 1, found at beginning of the algorithm, these degrees are calculated and stored in a data structure (Figure 4), where a key is the ancestor *x* (which will be present in generalized rules), and each key points a value, which is a vector storing the degrees mentioned. Thus, since the result of Part 2 correspond to min operator for the degrees related to any rule $\{x_1, \ldots, x_n\}$, we use the stored degrees of $x_1, \ldots, x_n$ for calculating the Part 2, obtaining the support of any generalized rule.

An important point is that if $\mu_{tX} = 0$ the transaction does not supports x_n, then the degree μ_{tx} is not stored in the vector. Thus, each vector linked in a key of the Figure 3 has the same quantity of positions of the vector pointed out by the same key of the Figure 4. Besides, in such vectors, the values of correspondent positions are related. For example, through Figure 3 we can see that the key Fruit is present in three transactions, T1, T2 and T4. Then, from the Figure 4.5 we can infer that the degree which T1, T2 and T3 support Fruit is 1, 0.7 and 0.7, in the same order.

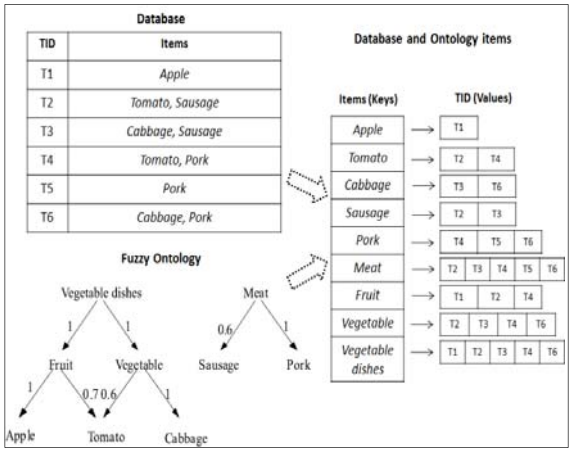


Figure 3: Indexing items and ancestors.

Now, consider an example about how calculate the support of the rule *Fruit* → *Meat*: First, the algorithm uses the structure shown in the Figure 3 for verify the quantity of transactions in the

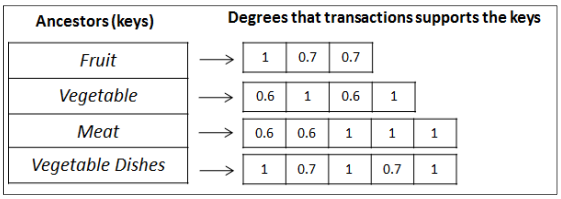


Figure 4: Storing the transaction support degrees.

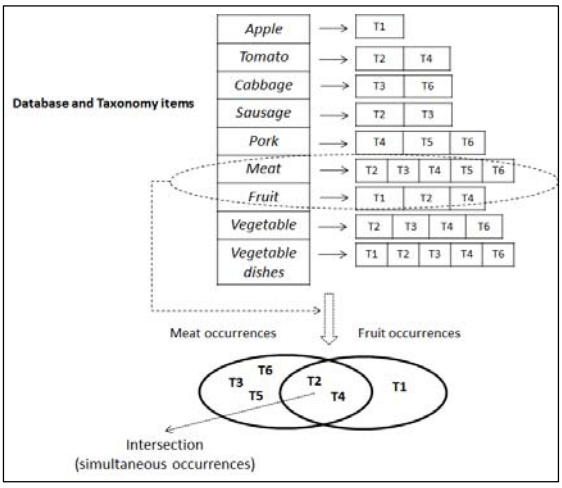


Figure 5: Idea used in the calculating of support.

intersection of values stored in vectors of these keys, since it represents all simultaneous occurrences of Fruit and Meat on the dataset transactions. Figure 5 illustrates this idea. In this case we have two occurrences of {*Fruit, Meat*}.

Then, in relation to each key, the algorithm uses the positions of these transactions in Figure 3 to found the degree which each transaction supports these ancestors. Such degrees are present in the same positions of the vectors linked at Fruit and Meat on the Figure 4. In this case we have: Fruit: 0.7/T2, 0.7/T4; Meat: 0.6/T2, 1/T4, which are results of Part 1. Based on these degrees, we use Part 2 to calculate the μ_{tX}, where *X* is {*Fruit, Meat*}.

For T2 we have:

$$\mu_{tX} = \min_{\forall\, x \in X} (Part\ 1) = \min(0.7, 0.6) = 0.6$$

For T4 we have:

$$\mu_{tX} = \min_{\forall\, x \in X} (Part\ 1) = \min(0.7, 1) = 0.7$$

So, according to Equation 3, we have 0.6 + 0.7 = 1.3. Furthermore, the Equation 4 is used to calculate the support, which is 0.21. Although we presented a specific example, the process applies to any rule.

3.4 Inferring Similarity Relations According to the Context

As said before, for represent our fuzzy ontology, we follow the ideas described in two meta-ontologies, proposed in (Agrawal and Srikant, 1994), and (Cerri et al., 2010). The approach proposed in (Cerri et al., 2010) allows to represent, in a single ontology, distinct relationships according to different contexts.

In relation to fuzzy relationships, they introduce the *ctx:ContextFuzzyRelationMembership* class, responsible for associating fuzzy relationships to several contexts.

Ctx:ContextFuzzyRelationMembership is subclass of the *fuzz:FuzzyRelationMembership* class from the fuzzy ontology, thus it inherits *fuz:fuzzyRelationDomain*, *fuz:fuzzyRelationRange*, *fuz:fuzzyRelationProp* and *fuz:membershipDegree* properties. The context association is represented by *ctx:hasContext* and *ctx:context* properties, which link contexts to fuzzy relationships (*fuz:FuzzyRelation*) and fuzzy degrees respectively. By using such constructs, a domain expert can model fuzzy relationships from different perspectives, with specific fuzzy degrees according to each context.

In our algorithm, the similarity degree values between items are represented in the fuzzy ontology leaves, which specify the semantics of the database contents. This step navigates through the fuzzy ontology structure to identify semantic similarity between items, according to the pre-defined *context* parameter. If according to a user-provide *context* the similarity degree between items is greater than or equal to the minSim parameter cited in section 3.1, a semantic similarity association is found and this association is considered similar enough. A fuzzy association of size 2 is made by these pair of items found and are expressed by the symbol ~ indicates the similarity relation between items, for example, $item_a$~$item_b$.

After that, this step verifies the presence of similarity cycles as proposed in (Escovar et al., 2005). These are fuzzy associations of size greater than 2 that only exists if the items are, in pairs, sufficiently similar. The minimum size of a cycle is 3, and the maximum is the number of sibling leaf nodes, for example, $item_a$~$item_b$~$item_c$. According to (Escovar et al., 2005), based on the concept of fuzzy intersection, the similarity degree value of a cycle is the minimum value found among the pairs. For example, if in a context $item_a$~$item_b$ are 0.8 similar; $item_b$~$item_c$ are 0.7 similar; $item_a$~$item_c$ are 0.5 similar, then

$item_a$~$item_b$~$item_c$ are 0.5 similar. Similarity cycles are obtained through the transitive property (Zadeh, 1965). All similarity relations and similarity cycles with degree values greater than or equal to the *minsim* are stored (as strings) by the algorithm. After that, this step does a search in the rules generated checking if the same have items that are included in some relation or cycle stored. In positive cases, these items are replaced by the correspondent string stored. We can say the positive cases are related to the traditional rules which have not been generalized, since the similarity relations are associated only to the leaf nodes. For example, suppose the rule: $item_a, item_d \quad item_b \rightarrow item_f$, $item_h$. Considering that there is a similarity relation $item_a$~$item_b$, then the stored correspondent string, $item_a$~$item_b$, it is inserted in the rule, replacing the single items $item_a$ and $item_b$. So only the new rule, $item_a \sim item_b, \quad item_d \rightarrow item_f, item_h$, it is show by the algorithm.

We can say that our approach is totally different than (Miani et al., 2009) and (Escovar et al., 2006). In these works, the inclusion of similarities in the rules is done through a concept of fuzzy item, which are a type of similarity representation. Such items are inserted in the set of candidates, during the candidate generation, and are used to generate the rules. Besides, a calculus of fuzzy occurrences also is done. Another different point is that (Miani et al. 2009) and (Escovar et al., 2006) do not consider the inclusion of context in the similarity relation.

4 EXPERIMENTS

This section shows some experiments performed to validate the *Context* FOntGAR algorithm. Two real datasets were used. The first dataset (DB-1) contains information about Years of study, Race or ethnicity and Sex, and was provided by Brazilian Institute of Geography and Statistics (IBGE). DB-1 contains 10000 transactions with 12 distinct items. The second data set (DB-2) contains a one day sale of a supermarket located in São Carlos city. DB-2 contains 1716 transaction with 1936 distinct items.

Two fuzzy ontologies were created, one for the DB-1, called Ont-1 ontology, and other for the DB-2, called Ont-2 ontology. The Ont-1 was constructed contained one level of abstraction, except by the root, and Ont-2 was constructed with four levels of abstraction, except by the root. In both ontologies the average value of specialization/generalization degrees was 0.8. Both ontologies were modeled in OWL (Web Ontology Language) and the Jena Framework was used to allow navigation through ontology concepts and relations.

In order to compare and illustrate the performance of *Context* FOntGAR, the experiments were carried out with respect to two major aspects. First, with the DB-1, the GARPA algorithm (Carvalho, Rezende et al. 2007) under a corresponding crisp taxonomy, NARFO (Miani et al. 2009) under a corresponding crisp ontology and *Context* FOntGAR algorithm under the Ont-1 were run. The purpose was to show what the effect of fuzzy extensions could be. In this comparison, 2 experiments have been conducted. Second, with the DB-2 and Ont-2, the *Context* FOntGAR was executed. The purpose was to show how the generalization treatment could improve the reduction in the rules amount. This experiment checks the compaction rate, which represents the percentage of reduction in the volume of rules.

4.1 Performance Comparisons

We performed 2 experiments with real data and taxonomic structures mentioned above, changing a different parameter in each experiment. The experiments were done with default values of parameter, except for the one being varied. By default, minsup = 0.02, minconf = 0.4 and mingen = 0.2. The side of generalization was set to lr in all algorithms.

Number of Transactions

In Figure 6, the vertical axis is the average of reading time per transaction (in milliseconds) in relation to the first scanning in the database. Here was compared the first scan on NARFO and the first scan on *Context* FOntGAR. We varied the number of transactions from 2000 to 10000. From Figure 6, it is possible see that the gap between *Context* FOntGAR, and NARFO show that the scanning with fuzzy ontologies is more time consuming than scanning with crisp ontologies. There are two reasons. First, the membership degree calculation demands more time. Second, the data structures generation contributes for increase the runtime. However, we can see that the gap tends keep stable with the increase of the number of transactions. This shows that the computational complexity is linear with the number of transactions, which is the same as the crisp algorithm. The difference between the two curves turns to be constant.

In Figure 7 we changed the minimum degree of support from 0.05% to 0.2%. The vertical axis is the

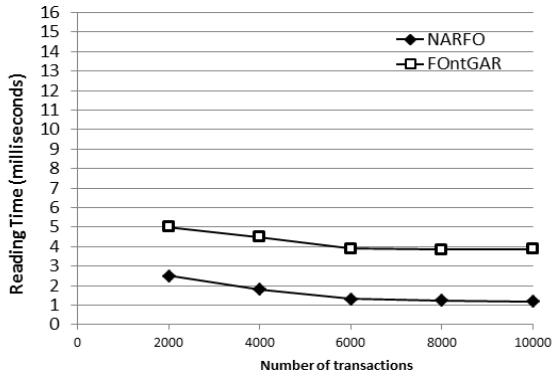


Figure 6: Scanning time (per transaction).

Minimum Degree of Support

total execution time in seconds. Notably, with the increase of minsup, the runtime of both *Context* FOntGAR and GARPA decreases. The reason is that when the minsup increases the amount of traditional rules decrease, and consequently a minor quantity of rules are post-processed. However, we can see that GARPA consumes more time than *Context* FOntGAR. The reason is that GARPA demands more time during the calculating of support, because a new scan is done in the database for each generalized rule obtained. So, depending on the quantity of rules and rows of the dataset, the runtime can be very high. On the other hand, apart from provide an indexed access to data, in *Context* FOntGAR, the data structures avoid the necessity of new scans in the database, decreasing the runtime.

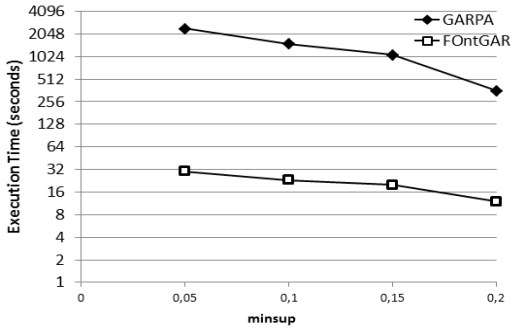


Figure 7: Comparison in relation to the runtime.

Compaction Rate in Context FOntGAR

The Figure 8 shows that the compaction rate is high, especially when values of minGen are low. This means that for high values of minGen the number of generalized rules decreases and consequently the number of traditional rules increases, reflecting in the amount generated.

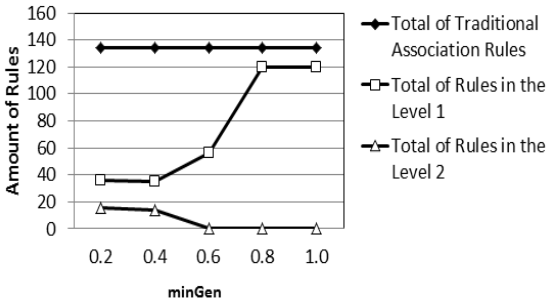


Figure 8: Compaction rate in *Context* FOntGAR.

4.2 Exploring Rules with Similarity Relations

In order to explore rules with similarity relations the DB-2 and Ont-2 were used. For explore different contexts Ont-2 was extended through the meta-ontology mentioned above. Two contexts were inserted, flavour and appearance. The Table 1 shows some leaf items and their respective similarity degree values, in relation to the two contexts. The part shown represents the similarTo relationship between the spinach and mustard according to context appearance. The similarity degree is set to 0.7.

Table 1: Similarity Degree Values.

Similarity Contexts			
items		Appearance	Flavour
Coca-Cola	Pepsi	0.8	0.6
Pepsi	Brazilian Coke	0.8	0.5
Tomato	Khaki	0.7	0.3
European Chocolate	Brazilian Chocolate	0.8	0.6
spinach	lettuce	0.7	0.4
spinach	mustard	0.7	0.4

In Table 1 the similarity degree values are given in pairs of items. For example, spinach and mustard have similarity 0.7 in context of appearance. Besides, based on the table 1 two similarity cycles can be found in the ontology. Depending on the similarity value, the selection of context may cause change in the similarities represented in the rules. Our experiment was carried out employing the parameters values: minimum support (minsup) =0.2, minimum confidence (minconf)=0.2, and minimum similarity (minsim)=0.3. Some examples of rules generated are:

Appearance Context:

- spinach~lettuce~mustard, coffee → onion, potato
- tomato~khaki, bread → soap, detergent
- milk → EuropeanChocolate~BrazilianChocolate

5 CONCLUSIONS

This paper proposes the *Context* FOntGAR algorithm, a new algorithm for mining generalized association rules under all levels of fuzzy ontologies, including similarity relations in the rules. The experiments show that *Context* FOntGAR makes an efficient generalization treatment, reducing the amount of rules. This work presents several contributions. First, it is introduced an algorithm which uses fuzzy ontologies with context-based similarity relations during the post-processing stage. Considering the bias found in the literature, our algorithm makes an important improvement on the state of the art. Another important contribution is that *Context* FOntGAR improves the semantic in the rules and generates non-redundant patterns without use pruning measures, since the generalized ones are obtained based on the traditional rules. For future works we are doing some improvements in the *Context* FOntGAR algorithm. We are improving the use of *mingen*, based on the user's preferences.

ACKNOWLEDGEMENTS

We wish to thank the Determinants of Educational Performance Project (CAPES/INEP).

REFERENCES

Agrawal, R., T. Imielinski, et al. (1993). M*ining association rules between sets of items in large databases*, Washington, DC, USA, ACM.

Agrawal, R. and R. Srikant (1994). Fast algorithms for mining association rules. *Conference on Very Large Databases (VLDB). Santiago, Chile, Morgan Kaufmann Publischers Inc.: 487-499.*

Cai, C. H., Ada, et al. (1998). Mining Association Rules with Weighted Items. *International Database Engineering and Application Symposium.*

Carvalho, V. O. D., S. O. Rezende, et al. (2007). Obtaining and evaluating generalized association rules. *9th International Conference on Enterprise Information Systems, ICEIS 2007, Funchal, Madeira; 12 June 2007 through 16 June 2007.*

Cerri, M. J., C. Yaguinuma, et al. (2010). UFOCoRe: Exploring Fuzzy Relations According to Specifics Contexts. *International Conference on Software Engineering & Knowledge Engineering (SEKE 2010). San Francisco Bay, USA: 529-534.*

Chen, G. and Q. Wei (2002). "Fuzzy association rules and the extended mining algorithms." *Information Sciences - Informatics and Computer Science: An International Journal* **147**(1-4): 201-228.

Escovar, E. L. G., M. Biajiz, et al. (2005). "SSDM: A Semantically Similar Data Mining Algorithm." *20 Brazilian Symposium of Databases.*

Escovar, E. L. G., C. A. Yaguinuma, et al. (2006). Using Fuzzy Ontologies to Extend Semantically Similar Data Mining. *21 Brazilian Symposium on Databases. Florianópolis, Brazil: 16-30.*

Hong, T. P., K. Y. Lin, et al. (2003). "Fuzzy data mining for interesting generalized association rules." *Fuzzy Sets and Systems 138(2): 255-269.*

Hung-Pin, C., T. Yi-Tsung, et al. (2006). A Cluster-Based Method for Mining Generalized Fuzzy Association Rules. *Innovative Computing, Information and Control, 2006. ICICIC '06. First International Conference on.*

Jiawei Han and Y. Fu (1995). Discovery of Multiple-Level Association Rules from Large Databases. *21º VLDB Conference. Zurich, Switzerland: 420-431.*

Keon-Myung, L. (2001). Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies. *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th.*

Lee, Y.-C., T.-P. Hong, et al. (2008). "Multi-level fuzzy mining with multiple minimum supports." *Expert Systems with Applications: An International Journal* **34**(1): 459-468.

Mahmoudi, E. V., E. Sabetnia, et al. (2011). Multi-level Fuzzy Association Rules Mining via Determining Minimum Supports and Membership Functions. *Intelligent Systems, Second International Conference on Modelling and Simulation (ISMS), 2011.*

Miani, R. G., C. A. Yaguinuma, et al. (2009). *NARFO Algorithm: Mining Non-redundant and Generalized Association Rules Based on Fuzzy Ontologies. Enterprise Information Systems.* J. Filipe and J. Cordeiro, Springer Berlin Heidelberg. **24:** 415-426.

Smith, M. K., C. Welt, et al. (2004). "*W3C Proposed Recomendation: OWL Web Ontology Language Guide*." Retrieved 2 dezembro, 2010, from

Srikant, R. and R. Agrawal (1995). Mining Generalized Association Rules. *Proceedings of the 21th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc.*

Vo, B. and B. Le (2009). "Fast Algorithm for Mining Generalized Association Rules." *International Journal of Database Theory and Application* **2**(3): 1-12.

Wei, Q. and G. Chen (1999). Mining generalized association rules with fuzzy taxonomic structures. *Fuzzy Information Processing Society, 1999. NAFIPS. 18th International Conference of the North American.*

Wen-Yang, L., T. Ming-Cheng, et al. (2010). Updating generalized association rules with evolving fuzzy taxonomies. *IEEE International Conference on Fuzzy Systems (FUZZ), 2010.*

Wu, C.-M. and Y.-F. Huang (2011). "Generalized association rule mining using an efficient data structure." *Expert Systems with Applications* 38(6): 7277-7290.

Zadeh, L. A. (1965). "Fuzzy sets." *Information and Control* **8**(3): 338-353.

SHORT PAPERS

Product Quantization for Vector Retrieval with No Error

Andrzej Wichert
Department of Informatics, INESC-ID/IST, Technical University of Lisboa, Lisboa, Portugal
andreas.wichert@ist.utl.pt

Keywords: High Dimensional Indexing, Multi-resolution, Quantization, Vector Data Bases.

Abstract: We propose a coding mechanism for less costly exact vector retrieval for data bases representing vectors. The search starts at the subspace with the lowest dimension. In this subspace, the set of all possible similar vectors is determined. In the next subspace, additional metric information corresponding to a higher dimension is used to reduce this set. We demonstrate the method performing experiments on image retrieval on one thousand gray images of the size 128×96. Our model is twelve times less complex than a list matching.

1 INTRODUCTION

In this paper we propose a hierarchical product quantization for less costly vector retrieval. The space in which a vector is represented is decomposed into low dimensional subspaces and quantize each subspace separately. Each subspace corresponds to a subvector described by a mask. A global feature corresponds to a subvector of a higher dimension, a local feature to a subvector of a lower dimension. During similarity-based vector retrieval, the search starts from the images represented by global features. In this representation, the set of all possible similar images is determined. In the next stage, additional information corresponding to the representation of more local feature is used to reduce this set. This procedure is repeated until the similar vectors can be determined. The described method represents a vector indexing method that speeds up the search considerably and does not suffer from the curse of dimensionality. The method is related to the subspace trees (Wichert et al., 2010).

We describe a mathematical model of the hierarchical product quantization. The paper is organized as follows:

- We show why the search with the product quantization is an improvement to simple quantization due to the curse of dimensionality.
- We introduce a new indexing method, called hierarchical product quantization and preform experiments on image retrieval on one thousand gray images of the size 128×96 resulting in vectors of dimension 12288.

2 HIERARCHICAL PRODUCT QUANTIZATION

2.1 Euclidean ε-similarity

Two vectors $\vec{x}$ and $\vec{y}$ are similar if their Euclidian distance is smaller or equal to ε, $d(\vec{x},\vec{y}) \leq \varepsilon$. Let DB be a database of s vectors $\vec{x}^{(i)}$ dimension m in which the index i is an explicit key identifying each vector,

$$\{\vec{x}^{(i)} \in DB | i \in \{1..s\}\}. \tag{1}$$

The set DB can be ordered according to a given query vector $\vec{y}$ using an Euclidian distance function d. This is done by a monotone increasing sequence corresponding to the increasing distance of $\vec{y}$ to $\vec{x}^{(i)}$ with an explicit key that identifies each vector indicated by the index i,

$$d[y]_t := \{d(x^{(i_n)},y)_t \mid \forall t \in \{1..s\}, \forall i_t \in \{1..s\} :$$
$$d(x^{(i_1)},y)_1 \leq d(x^{(i_2)},y)_2$$
$$\leq ... \leq d(x^{(i_n)},y)_t .... \leq d(x^{(i_s)},y)_s\} \tag{2}$$

if $\vec{y} \in DB$, then $d[y]_1 := 0$. The set of similar vectors in correspondence to $\vec{y}$, $DB[y]_\varepsilon$, is the subset of DB, $DB[y]_\varepsilon \subseteq DB$ with size $\sigma = |DB[y]_\varepsilon|$, $\sigma \leq s$:

$$DB[y]_\varepsilon := \{x^{(i)} \in DB \mid d[y]_t = d(x^{(i)},y) \leq \varepsilon\}. \tag{3}$$

2.2 Related Work

Could we take advantage of the grouping of the vectors into clusters?

The idea would be to determine the most similar cluster center which represents the most similar category. In the next step we would search for the most similar vectors $DB[y]_\varepsilon$ only in this cluster. By doing so we could save some considerable computation. Such a structure can be simply modeled by a clustering algorithm, as for example k-means. We group the images into clusters represented by the cluster centers c_j. After the clustering cluster centers $c_1, c_2, c_3, ..., c_k$ with clusters $C_1, C_2, C_3, ..., C_k$ are present with:

$$C_j = \{x | d(x, c_j) = \min_i d(x, c_i)\} \quad (4)$$

$$c_j = \{\frac{1}{|C_j|} \sum_{x \in C_j} x\}. \quad (5)$$

Suppose $s = \min_i d(y, c_i)$ is the distance to the closest cluster center and r_{max} the maximal radius of all the clusters. Only if $s \geq \varepsilon \geq r_{max}$ we are guaranteed to determine $DB[y]_\varepsilon$. Otherwise we have to analyze other clusters as well. When a cluster with a minimum distance s was determined, we know that the images in this cluster have the distance between $s + r_{max}$ and $s - r_{max}$. Because of that we have to analyze additionally all the clusters with $\{\forall i | d(y, c_i) < (s + r_{max})\}$. It means that in the worst case we have to analyze all the clusters. The worst case is present when the dimension of the images is high. High dimensional spaces (like for example dimensions > 100) have negative implications on the number of clusters we have to analyze. These negative effects are named as the "curse of dimensionality." Most problems arise from the fact that the volume of a sphere with the constant radius grows exponentially with increasing dimension.

Could hierarchical clustering overcome those problems? Indeed, traditional indexing methods are based on the principle of hierarchical clustering of the data space, in which metric properties are used to build a tree that then can be used to prune branches while processing the queries. Traditional indexing trees can be described by two classes, trees derived from the kd-tree and the trees composed by derivatives of the R-tree. Trees in the first class divides the data space along predefined hyper-planes regardless of data distribution. The resulting regions are mutually disjoint and most of them do not represent any objects. In fact with the growing dimension of space we would require exponential many objects to fill the space. The second class tries to overcome this problem by dividing the data space according to the data distribution into overlapping regions, as described in the second section. An example of the second class is the M-tree (Paolo Ciaccia, 1997). It performs exact retrieval with 10 dimensions. However its performance deteriorates in high dimensional spaces. Most indexing methods operate efficiently only when the number of dimensions is small (< 10). The growth in the number of dimensions has negative implications in the performance; these negative effects are also known as the "curse of dimensionality."

A solution to this problem consists of approximate queries which allow a relative error during retrieval. M-tree (Ciaccia and Patella, 2002) and A-tree (Sakurai et al., 2002) with approximate queries perform retrieval in dimensions of several hundreds. A-tree uses approximated MBR instead of a the MBR of the R-tree. Approximate metric trees like NV-trees (Olafsson et al., 2008), locality sensitive hashing (LSH) (Andoni et al., 2006) or product quantization for nearest neighbor search (Jegou et al., 2011) work with an acceptable error up to dimension 1000.

We introduce hierarchical product quantizer, who preforms exact vector queries in high dimensions. The hierarchical product quantizer model is related to the subspace-tree. In a subspace-tree instead of quamtizing the subvectors defined by masks the mean value is computed (Wichert, 2008), (Wichert et al., 2010). Mathematical methods and tools that were developed for the analysis of the subspace tree, like the correct estimation of ε (Wichert, 2008) and the algorithmic complexity (Wichert, 2008) can be as well applied for the hierarchical quantization method. Hierarchical quantization for image retrieval was first proposed by (Wichert, 2009).

2.3 Searching with Product Quantizers

The vector $\vec{x}$ of dimension m is split into f distinct subvectors of dimension $p = dim(m/f)$. The subvectors are quantizied using f quantiziers:

$$\vec{x} = \underbrace{x_1, , x_2, \cdots, x_p}_{u_1(\vec{x})}, \cdots, \underbrace{x_{m-p+1}, \cdots, x_m}_{u_f(\vec{x})} \quad (6)$$

$$u_t(x) \in x | t \in \{1..f\}$$

We group the subvectors of dimension $p = dim(m/f)$ into clusters represented by the cluster centers c_j of dimension p. After the clustering cluster centers $c_1, c_2, c_3, ..., c_k$ with clusters $C_1, C_2, C_3, ..., C_k$ are present with

$$C_j = \{u_t(x) | d(u_t(x), c_j) = \min_i d(u_t(x), c_i)\} \quad (7)$$

$$c_j = \{\frac{1}{|C_j|} \sum_{u_t(x) \in C_j} u_t(x)\}. \quad (8)$$

We assume that all subquantuzers have the same number k of clusters. To a query vector y we determine the most similar vector x of the database using the qunatizied codes and the Euclidean distance function d.

$$d(U(\vec{x}),U(\vec{y})) = \sqrt{\sum_{t=1}^{f} d(u_t(\vec{x}),(u_t(\vec{y}))^2}$$

$$d(U(\vec{x}),U(\vec{y})) = \sqrt{\sum_{t=1}^{f} d(c_{t(x)},c_{t(y)})^2} \quad (9)$$

We represent vectors by the corresponding cluster centers:

$$U(\vec{x}) = \underbrace{c_{i1},c_{i2},\cdots,c_{ip}}_{u_1(\vec{x})=c_{1(x)}=c_i},\cdots,\underbrace{c_{j1},\cdots,c_{jp}}_{u_f(\vec{x})=c_{f(x)}=c_j} \quad (10)$$

By using $d(U(\vec{x}),U(\vec{y}))$ instead of $d(\vec{x},\vec{y})$ an estimation *error* is produced:

$$d(U(\vec{x}),U(\vec{y})) + error = d(\vec{x},\vec{y}) \quad (11)$$

To speed up the computation of $d(U(\vec{x}),U(\vec{y}))$ all the possible $d(c_{t(x)},c_{t(y)})^2$ are pre-computed and stored in a look-up table. The size of the look-up table depends on the number k, it is k^2. The bigger the value of k, the slower the computations due to the size of the look-up table. However the bigger the value of k the smaller is the estimation error. To determine ε similar vectors according to the Euclidean distance to a given query y, we have to compute $d(\vec{x},\vec{y})$ for all vectors x out of the database. If the distances computed by the quantizied product $d(U(\vec{x}),U(\vec{y}))$ are smaller or equal than the distances in the original space $d(\vec{x},\vec{y})$, a lower bound which is valid in both spaces can be determined. The distance of similar objects is smaller or equal to ε in the original space and, consequently, it is smaller or equal to ε in the quantizied product as well. The use of a lower bound between different spaces was first suggested by
(Faloutsos et al., 1994), (Faloutsos, 1999). Because of the estimation error the lower bound is only valid for a certain ω value:

$$d(U(\vec{x}),U(\vec{y})) - \omega \leq d(\vec{x},\vec{y}). \quad (12)$$

How can we estimate the ω value for all $\{\vec{x}^{(i)} \in DB | i \in \{1..s\}\}$? Suppose $s = \min_i d(u_t(y),c_i)$ is the distance to the closest cluster center and r_{max} the maximal radius of all the clusters. That means that in the worst case we have d to subtract r_{max} for each subvector before computing the Euclidian distance function, $\omega = k \times r_{max}$.

If we compute the Euclidian distance between all vectors and $\{\vec{x}^{(i)} \in DB | i \in \{1..s\}\}$ compare it to the Euclidian distance between$\{U(\vec{x})^{(i)}$, we find that $\omega << k \times r_{max}$.

We can estimate the ω value by computing the Euclidian distance between a a random sample of vectors and their product quantizier representation. If the lower bound is satisfied with the correct value ω, all vectors at a distance lower than ε in the original space are also at a lower distance in the product quantizier representation. The distance of some that are above ε in the original space may be below ε in the product quantizier representation. This vectors are called false hits. The false hits are separated from the selected objects through comparison in the original space.

The set of ε similar vectors in correspondence to a query vector $\vec{y}$ is computed in two steps. In the first step the set of possible candidates is determined using product quantizier representation. The speed results from the usage of the look-up table. In the second step the false hits are separated from the selected objects through comparison in the original space. A saving compared to a simple list matching is achieved if the set of possible candidates is sufficiently small in comparison to the size of database. An even greater saving can be achieved, if one applies this method hierarchically.

2.4 Searching with Hierarchical Product Quantizers

We apply the product quantizier recursively. The vector $\vec{x}$ of dimension m is split into f distinct subvectors of dimension $p = dim(m/f)$. The subvectors are quantizied using f quantiziers, the resulting quantizied vectror are quantizied using e quantiziers with $g = dim(m/e)$ and $f > e$

$$\vec{x} = \underbrace{\underbrace{x_1,x_2,\cdots,x_p}_{u1_1(\vec{x})},\cdots,\cdots,\cdots,\underbrace{x_{m-p+1},\cdots,x_m}_{u1_f(\vec{x})}}_{u2_1(U(\vec{x}))\qquad\qquad u2_e(U(\vec{x}))} \quad (13)$$

with following hierarchical representation,

$$\begin{aligned} &U1(\vec{x}) = U(\vec{x}) = c_{i1},c_{i2},\cdots,c_{ip},\cdots,c_{j1},\cdots,c_{jp} \\ &U2(\vec{x}) = U(U1(\vec{x})) = c_{i1},c_{i2},\cdots,c_{ig},\cdots,c_{j1},\cdots,c_{jg} \\ &\cdots \\ &Un(\vec{x}) = U(U(n-1)(\vec{x})) = c_{i1},c_{i2},\cdots,c_{il},\cdots \\ &\cdots,c_{j1},\cdots,c_{jl} \end{aligned} \quad (14)$$

and

$$d^*(Uk(\vec{x}),Uk(\vec{y}) = d(Uk(\vec{x}),Uk(\vec{y})) - \omega_k \leq d(\vec{x},\vec{y}), \quad (15)$$
$$k \in \{1..n\}$$

The DB is mapped by the first product quantizers $U1$ into $U1(DB)$, by the k th quantizers Uk into $Uk(DB)$.

The set $Uk(DB)$ can be ordered according to a given query vector $Uk(\vec{y})$ using an Euclidian distance function with ω_k as explained before

$$d[Uk(y)]_t := \{d^*(Uk(x^{(i)}),Uk(y)) \mid \forall t \in \{1..s\} : d^*[Uk(y)]_t \leq d^*[Uk(y)]_{t+1}\}$$

for a certain ε value,

$$Uk(DB[y])_\varepsilon := \{Uk(x)_t^{(i)} \in Uk(DB) \mid d[Uk(y)]_t = d^*(Uk(x)^{(i)},Uk(y)) \leq \varepsilon\}$$

with the size $Uk(\sigma) = |Uk(DB[y]_\varepsilon)|$ and $\sigma < U1(\sigma) < U2(\sigma) < \cdots < s$.

To speed up the computation of $d(U(\vec{x}),U(\vec{y}))$ all the possible $d(c_{j(x)},c_{j(y)})^2$ are pre-computed and stored in a look-up table For simplicity we assume that the the cost for a look-up operation is a constant $c = 1$. This is the case, given the size of the look-up tables for each hierarchy is constant. Consequently the computational dimensions of the quantized vector $\vec{x}$ is $dim(Uk) = m/f$, where $dim(Uk)$ is the number of distinct subvectors of dimension f of the vector $\vec{x}$. It follows, that $dim(Uk)$ is the number of quantiziers. The higher the hierarchy, the lower the number of the used quantiziers. Given that $dim(U0) =: m$ (the dimension of the vector $\vec{x}$), the computational cost of a hierarchy on n level is:

$$cost_n = \sum_{i=1}^{n} Ui(\sigma) \cdot dim(U(i-1) + s \cdot dim(Un) + n \tag{16}$$

where the last summand n represents the cost of the look-up operation. The cost $cost_n$ of retrieving a dozen most similar vectors out of the database DB to a query vector $\vec{y}$, is significantly lower as the cost of simple list matching $s \cdot m$.

To estimate ε we define a mean sequence $d[Uk(DB)]_n$ which describes the characteristics of an vector database of size s:

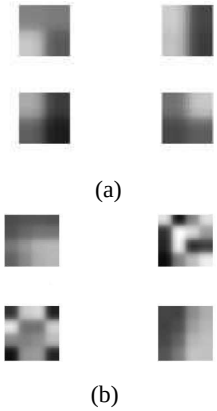
$$d_s[Uk(DB)]_n := \sum_{i=1}^{s} \frac{d[Uk(x^{(i)})]_n}{s}. \tag{17}$$

We will demonstrate this principle on an example of high dimensional vectors representing gray images.

2.5 Example: Image Retrieval

The high dimensional vectors correspond to the scaled gray images, representing the gray level distribution and the layout information. Two images $\vec{x}$ and $\vec{y}$ are similar if their distance is smaller or equal to ε, $d(\vec{x},\vec{y}) \leq \varepsilon$. The result of a range query computed by this method is a set of images that have gray level distribution that are similar to the query image.

We preform experiments on image retrieval on one thousand ($s = 1000$) gray images of the size 128×96 resulting in vectors of dimension 12288. Each gray level is represented by 8 bits, leading to 256 different gray values. The image database consists images with photos of landscapes and people, with several outliers consisting of drawings of dinosaurs or photos of flowers (Wang et al., 2001). We use a hierarchy of four $n = 4$.



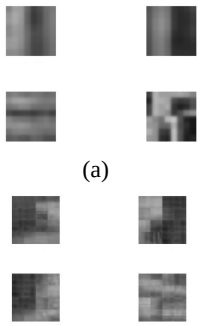
(a)

(b)

Figure 1: (a) Two examples of of squared masks M of a size 2×2. (b) Two examples of of squared masks M of a size 4×4.

- In the first level of hierarchy each image corresponds to a of dimension 12288. It is is split into 3072 distinct subvectors of dimension $4 = dim(12288/3072)$. Each subvector corresponds to a squared mask M of a size 2×2. A natural grouping of the components into subvectors is achieved by the coverage of the image with 3072 masks M (see, Figure 1 (a)). The subvectors of dimension four are grouped into clusters represented by 256 cluster centers.
- In the second level of the hierarchy the resulting quantizied images are quantizied using 768 quantiziers with $16 = dim(12288/768)$. Each subvector corresponds to a squared mask M of a size 4×4 (see, Figure 1 (b)). The subvectors of dimension sxten are grouped into clusters represented by 256 cluster centers. We follow the procedure recursively additionally two times.
- The resulting quantizied images are quantizied using 192 quantiziers with $64 = dim(12288/192)$. Each subvector corresponds to a squared mask M of a size 8×8 (see, Figure 2 (a)). The subvectors of dimension 64 are grouped into clusters represented by 256 cluster centers.
- The resulting quantizied images are quantizied using 48 quantiziers with $256 = dim(12288/48)$.

Each subvector corresponds to a squared mask M of a size 16×16 (see, Figure 2 (b)). The subvectors of dimension256 are grouped into clusters represented by 256 cluster centers. (Note: The number of cluster centers remains constant through the hierarchy.)



(a)

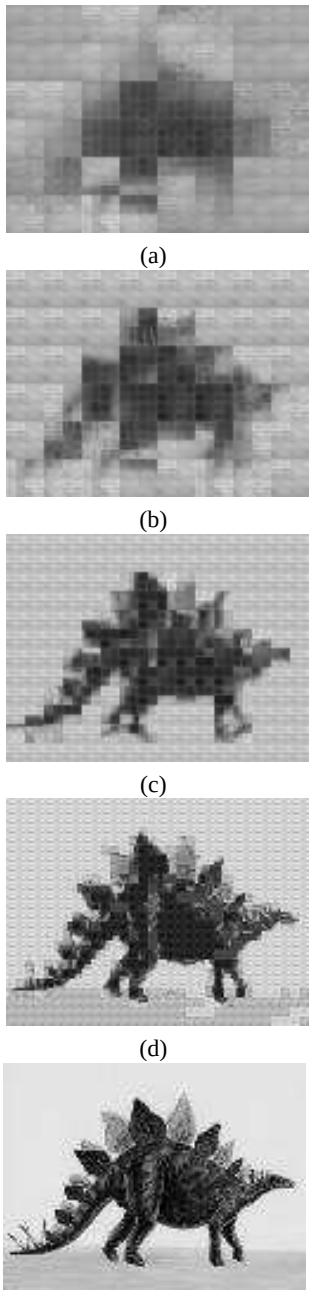
(b)

Figure 2: (a) Two examples of of squared masks M of a size 8×8. (b) Two examples of of squared masks M of a size 16×16.

The computational dimensions of the quantized vectors are $dim(U1) = 3072$, $dim(U2) = 768$, $dim(U3) = 192$ and $dim(U4) = 48$. An example of the quantizied representation of an image is indicated in the Figure 3. In each layer the image is described with less accuracy, so that the following layers represent less information. The set of ε similar vectors in correspondence to a query vector $\vec{y}$ is computed in 5 steps. For the estimation of the value of ε we use the characteristics, see Equation 17 and Figure 4

- In the first step the set of possible candidates is determined using product quantizier representation in the $U4(DB)$ of the computational dimension 48 and determine the subset $U4(DB[y])_\varepsilon$.
- Recursively out of the set $U4(DB[y])_\varepsilon$ we determine $U3(DB[y])_\varepsilon =$ $U3(U4(DB[y])_\varepsilon)_\varepsilon$,
- $U2(DB[y])_\varepsilon = U2(U3(DB[y])_\varepsilon)_\varepsilon$,
- $U1(DB[y])_\varepsilon = U2(U2(DB[y])_\varepsilon)_\varepsilon$ and
- finally the set $DB[y]_\varepsilon$ in which the false hits are separated from the selected objects through comparison in the original space by $DB[y]_\varepsilon = U0(U1(DB[y])_\varepsilon)_\varepsilon$.

To retrieve in the mean 5 most similar images the estimated ε value is 6036. We estimate the values ω_k and ε by a random sample of hundred vectors and their product quantizier representation. We computed the mean value $Uk(\overline{\sigma})$ over all possible queries (one thousand queries, each time we take an element out of the database out and and preform a query). For $\varepsilon = 6036$ the values are $U0(\overline{\sigma}) = 5$, $U1(\overline{\sigma}) = 20$, $U2(\overline{\sigma}) = 95$, $U_3(\overline{\sigma}) = 315$ and $U4(\overline{\sigma}) = 835$. To retrieve the 5 most similar images to a given query image of the image test database, the mean computation costs are according to Equation 16:



(a)

(b)

(c)

(d)

(e)

Figure 3: Gray image and its quantized representation, (e) is equal to the original space, (d) corresponds to U1(DB), (c) to U2(DB), (b) to U3(DB) and (a) to U4(DB).

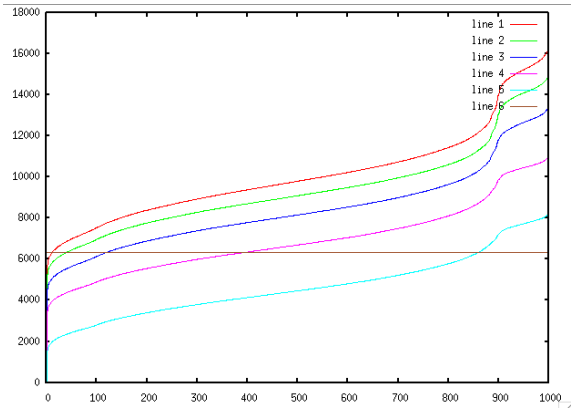


Figure 4: Characteristics of $s = 1000$, $d_s[U0(DB)]_n =$ line 1, $d_s[U1(DB)]_n =$ line 2, $d_s[U2(DB)]_n =$ line 3, $d_s[U3(DB)]_n =$ line 4 and $d_s[U4(DB)]_n =$ line 5. Line 5 represents ε.

$$(12288 \cdot 20 + 3072 \cdot 95 + 768 \cdot 315 + 192 \cdot 835 + \\ + 48 \cdot 1000 + 4 = 987844$$

which is 12.4 times less complex than a list matching which requires $12288 \cdot 1000$ operations. Further optimization of our results could be achieved by better quantization training (clustering algorithms).

3 CONCLUSIONS

We propose hierarchical product quantization for vector retrieval with no error for vector based databases. Through quantization by hierarchical clustering the distribution of the points in the high dimensional vector space can be estimated. Our method is exact and not approximative. It means we are guaranted to find the most similar vector according to a distance or similarity function. We demonstrated the working principles of our model by empirical experiment on one thousand gray images which correspond to 12288 dimensional vectors.

ACKNOWLEDGEMENTS

This work was supported by Fundação para a Ciência e Tecnologia (FCT): PTDC/EIA-CCO/119722/2010.

REFERENCES

Andoni, A., Dater, M., Indyk, P., Immorlica, N., and Mirrokni, V. (2006). Locality-sensitive hashing using stable distributions. In MIT-Press, editor, *Nearest Neighbor Methods in Learning and Vision: Theory and Practice*, chapter 4. T. Darrell and P. Indyk and G. Shakhnarovich.

Ciaccia, P. and Patella, M. (2002). Searching in metric spaces with user-defined and approximate distances. *ACM Transactions on Database Systems*, 27(4).

Faloutsos, C. (1999). Modern information retrieval. In Baeza-Yates, R. and Ribeiro-Neto, B., editors, *Modern Information Retrieval*, chapter 12, pages 345–365. Addison-Wesley.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., and Equitz, W. (1994). Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3(3/4):231–262.

Jegou, H., Douze, M., and Schmid, S. (2011). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128.

Olafsson, A., Jonsson, B., and Amsaleg, L. (2008). Dynamic behavior of balanced nv-trees. In *International Workshop on Content-Based Multimedia Indexing Conference Proceedings, IEEE*, pages 174–183.

Paolo Ciaccia, Marco Patella, P. Z. (1997). M-tree: An efficient access method for similarity search in metric spaces. In *VLDB*, pages 426–435.

Sakurai, Y., Yoshikawa, M., Uemura, S., and Kojima, H. (2002). Spatial indexing of high-dimensional data based on relative approximation. *VLDB Journal*, 11(2):93–108.

Wang, J., Li, J., and Wiederhold, G. (2001). Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963.

Wichert, A. (2008). Content-based image retrieval by hierarchical linear subspace method. *Journal of Intelligent Information Systems*, 31(1):85–107.

Wichert, A. (2009). Image categorization and retrieval. In *Proceedings of the 11th Neural Computation and Psychology Workshop*. World Scientific.

Wichert, A., Teixeira, P., Santos, P., and Galhardas, H. (2010). Subspace tree: High dimensional multimedia indexing with logarithmic temporal complexity. *Journal of Intelligent Information Systems*, 35(3):495–516.

A Constraint-based Mining Approach for Multi-attribute Index Selection

B. Ziani[1], F. Rioult[2] and Y. Ouinten[1]
[1]*LIM, Computer Science Department, University of Laghouat, Laghouat, Algeria*
[2]*GREYC (CNRS - UMR 6072), Université de Caen, Caen, France*
{*bziani, ouinteny*}*@mail.lagh-univ.dz, Francois.Rioult@unicaen.fr*

Keywords: Data Warehouse Physical Design, Bitmap Join Index Selection, Data Mining, Constraint Mining.

Abstract: The index selection problem (ISP) concerns the selection of an appropriate indexes set to minimize the total cost for a given workload under storage constraint. Since the ISP has been proven to be an NP-hard problem, most studies focus on heuristic algorithms to obtain approximate solutions. The problem becomes more difficult for indexes defined on multiple tables such as bitmap join indexes, since it requires the exploration of a large search space. Studies dealing with the problem of selecting bitmap join indexes mainly focused on proposing pruning solutions of the search space by the means of data mining techniques or heuristic strategies. The main shortcoming of these approaches is that the indexes selection process is performed in two steps. The generation of a large number of indexes is followed by a pruning phase. An alternative is to constrain the input data earlier in the selection process thereby reducing the output size to directly discover indexes that are of interest for the administrator. For example, to select a set of indexes, the administrator may put limits on the number of attributes or the cardinality of the attributes to be included in the indexes configuration he is seeking. In this paper we addressed the bitmap join indexes selection problem using a constraint-based approach. Unlike previous approaches, the selection is performed in one step by introducing constraints in the selection process. The proposed approach is evaluated using APB-1 benchmark.

1 INTRODUCTION

Data Warehousing and On-line Analytical Processing (OLAP) are becoming critical components of decision support. They are especially designed to enable executives, managers, and analysts to take better and faster decisions. Data warehouses are generally modelled according to a star schema that contains a central, large fact table, and several dimension tables that describe the facts (Inmon, 2002), (Kimball and Ross, 2007).

Queries defined on a star schema are called *star join queries*. They are complex and use several join operations that are very costly. Such queries will be performed on tables having potentially billions of records. As a result, it becomes crucial to accelerate query evaluation. Among the techniques adopted in relational data warehouses to improve query performance, materialized views and indexes are presumably the most effective ones (Chaudhuri and Narasayya, 2007). Data warehouses administrators then handle the fastidious task of choosing an advantageous configuration of indexes to enhance the system performance.

For a given data warehouse, the total number of distinct indexes can be extremely large; hence it is not always practicable to create all the indexes due to the limited amount of storage space that we can physically maintain. The approaches dealing with the index selection problem are composed of two steps:

1. Generation of candidate indexes for a given workload;
2. Selection of a final configuration that minimizes the cost of the workload, while observing the storage space limit.

The first step reduces the space of potential indexes by eliminating non relevant attributes. The final configuration (step 2) is mostly selected using greedy algorithms (Agrawal et al., 2000). The proposed approaches prune the set of generated indexes so that the constraint space is satisfied. However, this pruning process is performed **after** the generation of a large number of candidate indexes.

An alternative is to constrain the generation of indexes in order to produce fewer and more relevant outputs. In this paper we propose a constraint-based mining approach to solve the index selection problem. We believe that constraint-based mining will enable

administrators to focus on a subset of most advantageous indexes and that it avoids the generation of unwanted indexes.

The remainder of this paper is organized as follows: in Section 2 we present existing works related to bitmap join indexes selection problem and constraint-based mining. Section 3 describes the proposed approach for the bitmap join indexes selection. We experimentally study the efficiency of our approach in Section 4. We conclude the paper and present future directions in Section 5.

2 RELATED WORK

2.1 Bitmap Join Index Selection

The index selection problem has been studied first in traditional databases context (Chaudhuri and Narasayya, 1997), (Agrawal et al., 2000), (Chaudhuri et al., 2004), (Feldman and Reouven, 2003), (Frank et al., 1992), (Valentin et al., 2000). With the advent of data warehouse, indexation has become an important option in physical design and its importance is well recognized (Golfarelli et al., 2002). The index selection problem has been proven to be NP-hard (Chaudhuri et al., 2004). Thus, most studies in the literature have focused on finding approximate solutions using greedy strategies or heuristics-based approaches.

The aim of the proposed approaches is to determine a set of candidate indexes from a given workload of queries, then to propose a final indexes configuration providing the best profit, under storage space constraint. However, considered indexes usually concern one table. Bitmap join indexes are multi-attribute indexes involving several tables. Selecting a suitable configuration of Bitmap join indexes is more complicated than the classical mono-table indexes, since it requires the exploration of a large search space. To the best of our knowledge, only few studies dealing with the problem of selecting bitmap join indexes are carried out (Aouiche et al., 2005), (Bellatreche et al., 2007), (Bellatreche and Boukhalfa, 2010), (Ziani and Ouinten, 2011). Due to the large number of candidate indexes, the proposed approaches mainly focused on pruning the search space of potential indexes. They have used frequent itemsets (Aouiche et al., 2005), (Bellatreche et al., 2007), (Ziani and Ouinten, 2011) or heuristic strategies (Bellatreche and Boukhalfa, 2010) to perform the pruning process. In (Aouiche et al., 2005), (Bellatreche et al., 2007) the *Close* algorithm (Pasquier et al., 1999) for mining closed frequent itemsets is used to prune the search space of candidate indexes. Due to the large number of indexes generated as closed frequent itemsets, the authors in (Ziani and Ouinten, 2011) propose a maximal frequent itemsets based approach to perform the selection.

In (Bellatreche and Boukhalfa, 2010), the authors propose an intuitive algorithm for bitmap join indexes selection. As an initial configuration, the algorithm selects an index for each query having indexable attributes. When the size of the configuration exceeds the storage capacity $\mathcal{S}$, some selected indexes should be reduced until the satisfaction of $\mathcal{S}$.

The principal weakness of the proposed approaches is the large number of generated indexes, that is very difficult to manage, according to the system limitations (number of indexes per table and storage space constraint). Indeed, the pruning is done after the generation of the indexes configuration.

An alternative is to constrain the input data earlier in the selection process, thereby reducing the output size to directly discover indexes that are of interest for the administrator. We believe that a *constraint-based approach* will help to mine a reduced and more relevant indexes configuration.

2.2 Constraint-based Pattern Mining

Mining frequent itemsets (FI) in datasets is a demanding task common to several important data mining applications, that look for interesting patterns within databases (*e.g.*, association rules, correlations, sequences, episodes, classifiers, clusters). It was originally proposed in (Agrawal and Srikant, 1994), (Agrawal et al., 1993) with the Apriori algorithm.

The drawback of mining frequent itemsets is that, if there is a large frequent itemset with size s, then almost all 2^s candidate subsets of the itemset might be generated and tested. Furthermore, the number of frequent itemsets grows very quickly as the minimum support threshold decreases.

Moreover, the huge size of the output complicates the task of the analyst, who has to extract useful knowledge from a very large amount of frequent patterns. To overcome this problem, the paradigm of pattern discovery based on constraints was introduced with the aim at providing a tool for driving the discovery process towards potentially interesting patterns. Using constraints can be of a great help to purge a lot of patterns that are irrelevant for the user.

Constraint-based mining has then been widely addressed, with really different approaches. The mostly used constraints are the minimum or maximum support threshold, including (or being included in) some specific itemset, aggregated computation (sum, aver-

age, min, max, when items are associated to a measure). As this paper does not specifically contribute to the field of constraint-based mining, we just briefly recall below the main contributions.

Most of them combine anti-monotone constraints and monotone one (Pei and Han, 2000; Bucila et al., 2003). A constraint is monotone (resp. anti-) if it preserved while itemset specialization (resp. generalization). Many useful constraints fall within the anti-monotone category, such as the minimum support threshold or uper-bounding the aggregated sum. This allows for powerful pruning of the search space, because this space is built through specialization. The maximum support threshold is a typical monotone constraint.

Other approaches directly prune the dataset (Bonchi et al., 2003) or consider the problem as an inductive database issue and formalize the constraints as queries, in a dedicated constraint-base mining environment (Boulicaut et al., 2005),(Jeudy and Boulicaut, 2002).

3 CONSTRAINT-BASED INDEX SELECTION

To illustrate the motivation of our approach, let us see an example. Suppose that a given approach recommends a set $C_{idx} = \{I_1, I_2, \ldots, I_k\}$ of k indexes. The administrator may keep an index I_j knowing that it needs acceptable storage space, or reject it because it has previously shown negligible improvement for the system performance.

Indeed, depending on the cardinality of the attributes, the indexing process may be more or less efficient. If the cardinality is very large or very small, an index might not bring a very significant improvement (Vanichayobon and Gruenwald, 1999). On the other hand, it is not beneficial to create an index on a small table. Hence, table size is another parameter which can be taken into account. The administrator decides whether a table is large or not, and only the indexes on attributes belonging to large tables are selected.

More formally, let $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ be the set of indexable attributes and $\mathcal{D}$ the extraction context (Query/Attributes) for a given workload $\mathcal{W}$. If $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ is a set of k functions, denoting the properties of interest (constraints) for each index $I \subseteq \mathcal{A}$, our approach to solve the index selection problem requires to compute all the itemsets (indexes) occurring in the extraction context $\mathcal{D}$ and satisfying the set of constraints $\mathcal{C}$, i.e:

$$\{I \subseteq \mathcal{A} | C_1(I) \wedge C_2(I) \wedge \cdots \wedge C_k(I)\}$$

The architecture of our approach is illustrated in Figure 1. As data mining based approaches, it constructs an extraction context by identifying the indexable attributes from a given workload. Then, it performs a constraint-based extraction (involving administrator expertise) to generate the desired indexes. Unlike the classical frequent itemsets mining based approaches, we do not built an initial indexes configuration and we do not need to use a greedy algorithm to recommend a final configuration.

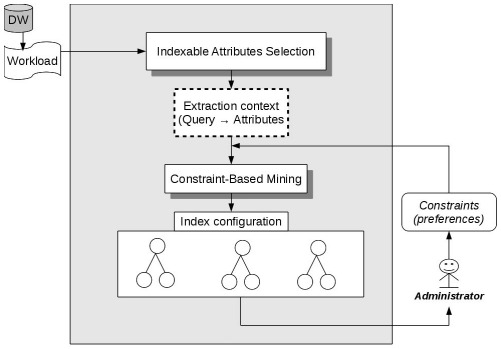


Figure 1: Constraint-based indexes selection.

4 EXPERIMENTAL STUDY

4.1 Description of the Experiment

The aim of our experiments is to evaluate our approach by observing the impact of using various constraints on the selected indexes. In the first experiment we study the impact of including constraints in the mining process on the number of generated indexes and the corresponding storage space. In the second expriments, we compare the performance of our approach with the baseline case where no indexes are created as well as the approaches using classical frequent itemsets mining, where the attributes frequency is the unique parameter used to generate a configuration of indexes. To evaluate the interestingness of the indexes generated by our approach, we use the same cost models proposed in similar works (Aouiche et al., 2005).

To perform a constraint-based selection, we have used the `MUSIC-dfs` tool (Mining with a User-SpecifIed Constraint, Depth-First Search approach) (Soulet et al., 2006). This tool provides a flexible and rich constraint query language. The user can iteratively develop complex constraints integrating various knowledge types.

In this study we are more interested in the quality of the generated indexes. Thus, our comparisons

are performed with no restrictions on available disk space. We have used the APB-1 Benchmark of the OLAP Council (OLAP-Council, 1998). The APB-1 Benchmark simulates a star schema data warehouse. It consists of one fact table *Actvars* and four dimension tables *ProLevel, TimeLevel, CustLevel, and ChanLevel*. We have considered 12 indexable attributes (Table 1).

Table 1: Characteristics of the indexable attributes.

Code	Attribute	Cardinality	Size of the dimension table
A	Class_Level	605	9
B	Quarter_Level	4	900
C	Group_Level	300	9
D	Family_Level	75	9
E	Line_Level	15	9
F	Division_Level	4	9
G	Year_Level	2	900
H	Month_Level	12	900
I	Retailer_Level	99	9000
J	Gender_Level	2	9000
K	All_Level	5	24
L	City_Level	4	9000

We have also used the same workload used in (Bellatreche and Boukhalfa, 2010). It consists of 60 star join queries involving aggregation operations and multiple joins between the fact table and dimension tables. We considered the following constraints:

- the support (frequency) of the generated indexes. This support shows the representativity of the index in the workload of queries;
- the length (number of attributes) of the generated indexes. It directly impacts on the width of the space needed for storing the index;
- the cardinality of the attributes in the generated indexes. This factor also impacts on the storage space width;
- the size of the dimension table to which an attribute belongs, that impacts on the height to the index.

4.2 Experimental Results

Experiment 1: Number and Size of Indexes vs Constraints. We begin with a baseline experiment where the frequency is the unique parameter taken into account (as classical approaches). Then, we conducted several experiments using the MUSIC-dfs to compute the generated configurations, for different combinations of constraints. The experiments depicted here are performed with a minimum support of 5% (3 queries). Using this threshold, the workload basically generates 56 indexes. This represents a very high value when compared to the size of the workload (60 queries).

Consequently, constraints are added to improve the number of generated indexes. Figures 2 and 3 show respectively the number and the total size occupied by the generated indexes for different constraint combinations. We applied different constraints to examine their impact on the generated configurations. It is interesting to observe that the characteristics of a generated configuration (i.e, number of indexes and total indexes size) depends on the complexity of the constraint (i.e., the number of combinations). This behavior allows the administrator to experiment with a broad set of configurations to select the most interesting one.

Experiment 2: Workload Cost vs Constraints. We compare the performance of our approach with the baseline case where no indexes are created as well as the approaches using classical frequent itemsets technique. For each constraint, we evaluate the workload cost using the generated indexes. The results we obtained are ploted in Figure 4. They show that we achieve a better performance using constraints on both the support (frequency) of indexes and the cardinality of the attributes. For complex constraints, there are very few or no generated indexes and thus the cost of the workload increases.

5 CONCLUSIONS

In this paper we have proposed a constraint-based framework for the index selection problem. Our approach leverages and extends principled methods of mining frequent itemsets for the index selection problem. The key contribution is that we show how constraint-based mining can be adapted in a flexible way that balances the characteristics of the workload and the administrator preferences for the index selection problem.

The existing approaches consist of a fully automatic procedure. Like any conventional process of data mining, this can lead to obvious, unhelpful, or undesirable knowledge (indexes). Our approach associates, on the one hand the high capacity of automatic selection mining frequent itemsets, and in the other hand, the necessary expertise of the administrator. Experimental results show that our approach is effective because it allows for more directly computing the useful indexes with precisely describing the requirements.

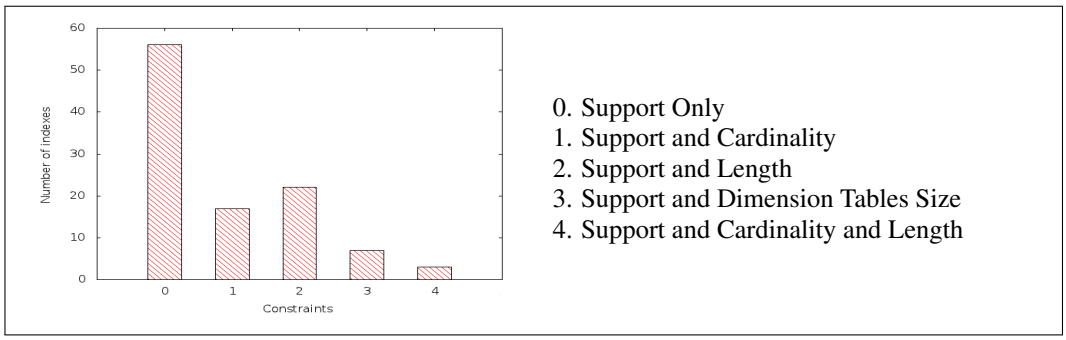


Figure 2: Number of generated indexes vs constraints.

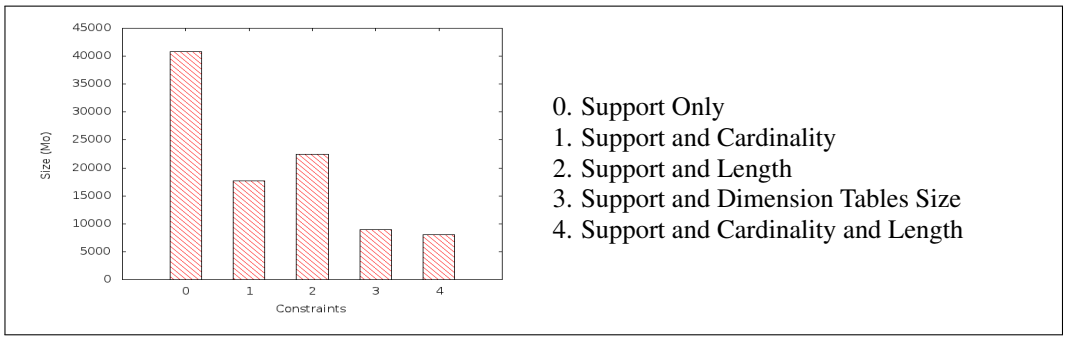


Figure 3: Size of generated indexes vs constraints.

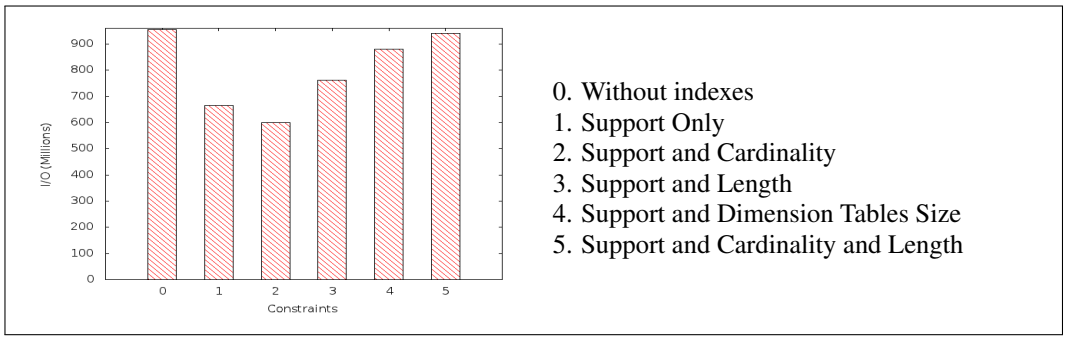


Figure 4: Workload cost vs constraints.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD International Conference on Management of Data, Washington, D.C*, pages 207–216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *International Conference on Very Large Data Bases, Santiago de Chile, Chile*, pages 487–499.

Agrawal, S., Chaudhuri, S., and Narasayya, V. (2000). Automated selection of materialized views and indexes in sql databases. In *VLDB*, pages 496–505.

Aouiche, K., Darmont, J., Boussaid, O., and Bentayeb, F. (2005). Automatic selection of bitmap join index in data warehouses. In *7th International Conference, DaWaK, Copenhagen, Denmark*, pages 64–73.

Bellatreche, L. and Boukhalfa, K. (2010). Yet another algorithms for selecting bitmap join index. In *12th International Conference, DAWAK , Bilbao, Spain*, pages 105–116.

Bellatreche, L., Missaoui, R., Necir, H., and Drias, H. (2007). Selection and pruning algorithms for bitmap index selection problem using data mining. In *9th International Conference, DaWaK, Regensburg, Germany*, pages 221–230.

Bonchi, F., Giannotti, F., Mazzanti, A., and Pedreschi, D.

(2003). Exante: Anticipated data reduction in constrained pattern mining. In *Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'03), Cavtat-Dubrovnik, Croatia*, pages 47–58.

Boulicaut, J.-F., Raedt, L. D., and Mannila, H., editors (2005). *Constraint-Based Mining and Inductive Databases, European Workshop on Inductive Databases and Constraint Based Mining, Hinterzarten, Germany, March 11-13, 2004*, volume 3848 of *Lecture Notes in Computer Science*. Springer.

Bucila, C., Gehrke, J. E., Kifer, D., and White, W. (2003). Dualminer: A dual-pruning algorithm for itemsets with constraints. *Data Mining and Knowledge Discovery*, 7(4):241–272.

Chaudhuri, S., Datar, M., and Narasayya, V. (2004). Index selection for databases: A hardness study and a principled heuristic solution. *IEEE Trans. Knowl. Data Eng*, 16:1313–1323.

Chaudhuri, S. and Narasayya, V. (1997). An efficient cost-driven index selection tool for microsoft sql server. In *23rd International Conference on Very Large Data Bases*, pages 146–155.

Chaudhuri, S. and Narasayya, V. (2007). Self-tuning database systems: a decade of progress. In *33rd international conference on Very large data bases*, pages 3–14.

Feldman, Y. A. and Reouven, J. (2003). A knowledge-based approach for index selection in relational databases. *Expert Syst. Appl.*, 25:15–37.

Frank, M., Omiecinski, E., and Navathe, S. (1992). Adaptive and automated index selection in rdbms. In *3rd International Conference on Extending Database Technology, Vienna, Austria*, pages 277–292.

Golfarelli, M., Rizzi, S., and Saltarelli, E. (2002). Index selection for data warehousing. In *4th Intl. Workshop DMDW, Toronto, Canada*.

Inmon, W. (2002.). *Building the Data Warehouse*. John Wiley & Sons, Inc., New York, NY, USA, 2nd edition.

Jeudy, B. and Boulicaut, J.-F. (2002). Constraint-based discovery and inductive queries: Application to association rule mining. In Hand, D. J., Adams, N. M., and Bolton, R. J., editors, *Pattern Detection and Discovery*, volume 2447 of *LNCS*, pages 110–124. Springer.

Kimball, R. and Ross, M. (2007). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, Inc., New York, NY, USA, 2nd edition.

OLAP-Council (1998). Apb-1 olap benchmark, release ii. *http://www.olapcouncil.org/*.

Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999). Discovering frequent closed itemsets for association rules. In *7th International Conference on Database Theory*, pages 398–416.

Pei, J. and Han, J. (2000). Can we push more constraints into frequent pattern mining? In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD'00)*, pages 350–354, Boston, USA. New York : ACM Press.

Soulet, A., Klema, J., and Crmilleux, B. (2006). Efficient mining under flexible constraints through several datasets. In *Workshop on Knowledge Discovery in Inductive Databases co-located with PKDD'06*.

Valentin, G., Zuliani, M., Zilio, D., Lohman, G., and Skelley, A. (2000). Db2 advisor: An optimizer smart enough to recommend its own index. In *ICDE*, pages 101–110.

Vanichayobon, S. and Gruenwald, L. (1999). Indexing techniques for data warhouses queries. Technical report, University of Oklahoma, School of computer science.

Ziani, B. and Ouinten, Y. (2011). Enhancing multi-attribute indexes selection using maximal frequent itemsets. In *EGCM, Tanger, Morocco*, pages 65–77.

A UML & Spatial OCL based Approach for Handling Quality Issues in SOLAP Systems

Kamal Boulil, Sandro Bimonte and Francois Pinet
Irstea, UR TSCF, 24 avenue des Landais, 63172 Aubière, France
{kamal.boulil, sandro.bimonte, francois.pinet}@irstea.fr

Keywords: Spatial DataWarehouse, Spatial OLAP, Quality, UML, OCL, Integrity Constraints.

Abstract: Spatial Data warehouses and Spatial OLAP systems are Business Intelligence technologies allowing efficient and interactive analysis of large geo-referenced datasets. In such a kind of systems the goodness of analysis depends on: the warehoused data quality, how aggregations are performed, and how warehoused data are explored. In this paper, we propose a framework based on a UML profile and OCL-defined integrity constraints to grant quality in the whole SOLAP system. We also propose an automatic implementation in a classical ROLAP architecture to validate our proposal.

1 INTRODUCTION AND MOTIVATION

Spatial Data Warehouse (SDW) and Spatial OLAP (SOLAP) systems are Business Intelligence (BI) technologies allowing effective storage and on-line spatio-multidimensional analysis of huge volumes of geo-referenced data which can be collected from multiple heterogeneous data sources (Malinowsky et al., 2008). These systems are based on the spatio-multidimensional model, which extends the conventional OLAP model with spatial concepts such as spatial measures and spatial dimensions which provide support for the representation and storage of spatial data, and spatial operators allowing users to interactively explore and aggregate warehoused data. A typical Spatial Relational OLAP (Spatial ROLAP) architecture is composed of three tiers: (i) the SDW tier historizes and manages integrated (spatial) data using a spatial Relational DBMS; (ii) the SOLAP server implements SOLAP operators that compute and handle spatial data cubes; (iii) the SOLAP client tier provides decision-makers with interactive visual displays that trigger SOLAP operators.

The heterogeneity of data sources in these systems may lead to several data quality problems (Boulil et al., 2011). In order to grant data quality in SDW, some approaches have been proposed to "repair" data by means of statistical techniques, data mining techniques, etc. (Ribeiro et al., 2011). At the same time, Integrity Constraints (IC) have been recognized as effective methods to express rules that control the consistency and completeness of warehoused spatial data (Salehi, 2009). Moreover, the goodness of spatio-multidimensional analysis also depends on the correct aggregation of measures in respect to summarizability conditions (or aggregation constraints), which check for example that the measure and aggregate function types are compatible (Lenz et al., 1997). However, in SOLAP systems the goodness of the analysis also requires another control when exploring (aggregated) data in order to avoid misinterpretation of meaningless SOLAP query results (Levesque et al., 2007), e.g., the query "Sales per country after December 26, 1991" returns empty results for USSR that could be interpreted by users as an absence of sales instead of realizing that a result is impossible for this period. On the other hand, conceptual design of complex systems such as data warehouses has been widely recognized as being necessary for successful BI projects (Malinowski and Zimányi, 2008) since it allows designers defining schemas that are easy to understand by decision makers. In this context, UML (Unified Modeling Language) is widely accepted as the Object-Oriented standard for modelling various aspects of software systems, and also SDW systems (Pinet and Schneider, 2009). Indeed, any approach using UML minimizes the efforts of designers and decision-makers in developing and implementing the data schema. It can be also interpreted by CASE tools. In the same

way, defining IC at a conceptual level allows handling quality issues at the early stages of development (Boulil et al., 2011), minimizing implementation efforts. In this context, (Ghozzi et al., 2003) propose ad-hoc conceptual multidimensional models allowing the expression of some data IC by means of logical predicates. (Malinowskiand Zimányi, 2008) propose an extension of the ER model for the design of spatio-temporal data warehouses. They define a set of ad-hoc pictograms to express spatial data IC (i.e., spatial topological relationships between spatial members). (Glorioand Trujillo, 2008) propose a UML profile for SDW, but they consider a very small number of data IC. A survey on aggregation issues is presented in (Mazón et al., 2009). They express simple structural constraints (e.g., facts should be linked to dimensions with one-to-many associations) with UML multiplicities. In (Pinetand Schneider, 2009), complex structural aggregation constraints are expressed with Object Constraint Language (OCL). OCL represents an effective solution to define data IC at the conceptual level in a clear, non-ambiguous and platform-independent way. Indeed, (Boulil et al.2011) present the definition, on the top of a UML-based SDW conceptual model, of a large number of data IC on warehoused spatial data by means of Spatial OCL, which is an extension of OCL for spatial data (Pinet et al., 2007). They also propose an automatic implementation in the Spatial DBMS Oracle Spatial 11g. (Lavesque et al., 2007) propose a framework for identifying quality risks in ETL, and SOLAP systems. They define 3 types of quality problems (data sources, OLAP data cubes and GIS functionalities) and define them by means of paper forms. They also propose an implementation in the JMAP SOLAP system.

Finally, to best of our knowledge, no work proposes a unique framework to express at the conceptual abstraction level IC on spatial warehoused data, aggregation, and spatio-multidimensional queries, and their automatic implementation in a classical ROLAP architecture.

Thus, in this paper we present three main contributions.

For first, we extend/reformulate the definition of (S)DW IC for handling quality issues in SOLAP systems; we use IC to perform three quality control types:

(a) Data quality control ensures that warehoused spatial data are valid (e.g., geometries of cities must be topologically included in the geometries of their states);

(b) Aggregation quality control ensures that aggregations of measures are correct and meaningful (e.g., the sum of the unit prices does not make sense) (Lenz et al., 1997);

(c) SOLAP exploration control avoids problems of interpretation induced by meaningless SOLAP query results (e.g., sales in USSR after 26 December 1991) (misuse data cube risks as defined by (Levesque et al., 2007)).

Secondly, motivated by a lack of a unique conceptual framework to define SOLAP IC, we propose a UML-OCL based conceptual framework. Finally, we propose an automatic implementation of such framework in a classical Relational SOLAP architecture.

2 SOLAP IC CLASSIFICATION

In this section, we present an extension of our previous SDW IC classification (Boulil et al., 2011) by introducing a new class, Query IC class. This classification (Figure 1) serves as a reference guide for the process of handling the three types of quality issues in a SOLAP system.

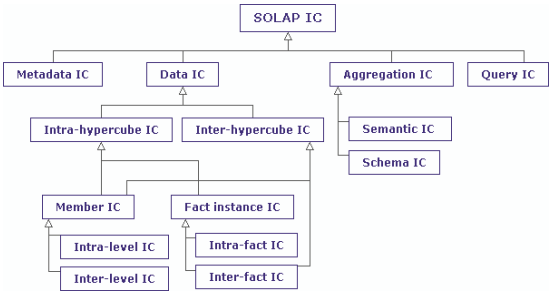


Figure 1: SOLAP IC classification.

Before detailing the classification, we present the case study which will be used all along the paper to describe our proposal. It concerns an environmental SDW, with a *temporal dimension* that groups days into months and months into years, and a *spatial dimension* representing cities with their regions and countries. The measure is the *temperature value*. Using this SDW, decision-makers can answer to SOLAP queries like these:"*What is the minimal temperature per year and country?*" or "*What is average temperature per month and country?*". In order to answer these queries, decision-makers use the min and average aggregate functions to aggregate the temperature values.

Let us now provide explanations and some examples of these IC classes using the previously described case study.

As shown in (Boulil et al., 2011), *Metadata IC* verify the consistency of metadata of different integrated data sources (e.g., spatial members and measures must be defined with the same geographic scale).

Data IC ensure the logical consistency and completeness of warehoused spatial data, for example:

Example 1: "*the geometry of each city must be topologically included in the geometry of its region*" or

Example 2: *"no facts (e.g., temperature values) should exist for USSR after 26 December 1991"*.

These constraints can be defined on all elements of the SDW such as facts, members, etc.

Aggregation IC guarantee correct and meaningful aggregations of measures. In particular, *semantic constraints* address the problem of the applicability of aggregate functions to measures according to the semantic nature and the type of measures, aggregate functions and dimensions. For example:

Example 3: "*Sum of temperature values does not make sense*"

Schema constraints are conditions that must be satisfied by dimension hierarchies and dimension-fact relationships to avoid double counting and incomplete aggregates. For example, dimensions and facts should be linked by one-to-many relationships (Mazón et al., 2009).

Query IC refer to conditions that guarantee that SOLAP queries are valid in the sense that their results are not always empty in order to avoid problems of misinterpretation. For example, the SOLAP query "*What are the average temperatures in USSR in 2010?*" returns an empty result since no temperature value is stored for USSR after 26 December 1991 (the previous data IC of Example 2). Even if this IC is implemented as data IC, classical SOLAP tools allow decision-makers to formulate this query by combining these two members (USSR and 2010) returning an empty value. This leads to a problem of interpretation: this empty value may be perceived as if there were no temperature values registered for USSR during 2010, instead of realizing that this combination of members (USSR and 2010) is invalid. Consequently, to avoid this misinterpretation we define the following query constraint:

Example 4: *"It is incorrect to combine USSR with days after 26 December 1991 in a SOLAP query"*.

Although, this query example could be resolved by using particular spatio-multidimensional data structures such as DW versioning structures, Query IC allow designers to model any invalid query which can be independent of time-versioning aspects (for example, some products cannot be sold in certain stores).

3 THE FRAMEWORK

Before describing our conceptual framework for defining SOLAP IC, we present main concepts of a UML profile and Spatial OCL.

The UML profiles are a way to customize UML for particular domains or platforms by extending its metaclasses (class, property, etc.). A profile is defined using three extension mechanisms: stereotypes, tagged values and constraints. A stereotype is an extension of a UML metaclass. Tagged values are properties of stereotypes. Finally, a set of OCL constraints precise each stereotype's application semantics. OCL provides a platform-independent method to model constraints. It can be interpreted by code generators to generate code automatically. OCL constraints can be defined at the meta-model level (e.g., UML profile) and also at the model level (the profile instance). Spatial OCL is an extension of OCL that supports spatial topological relationships (inside, intersect, etc.) (Pinet et al., 2007).

In order to define SOLAP data, aggregation and query IC at a conceptual level, we propose a framework based on a UML profile and Spatial OCL (Figure 2).

The main idea is to a have a unique UML profile that defines 3 interconnected models to conceptually represent:

a) SDW data structures (*SDW model*),

b) how measures are aggregated to meet the analysis requirements (*Aggregation model*), and

c) *Query IC model*

and then define IC with Spatial OCL using these models. In particular Data IC are defined by designers using Spatial OCL on the top of the instance of SDW model, Aggregation IC are defined as Aggregation model's stereotypes constraints using OCL, and Query IC are defined using the Query IC model and Spatial OCL. Due to space reasons we do not detail the proposed profile, but we provide some examples. Details on the SDW and aggregation models can be found in (Boulil et al.,

2011). It is important to note that we have chosen to define a UML profile instead of a metamodel since the UML metamodel's elements are sufficient to capture all the SOLAP applications' semantics including all the multidimensional data structures (Glorioand Trujillo, 2008) and all the identified IC types.

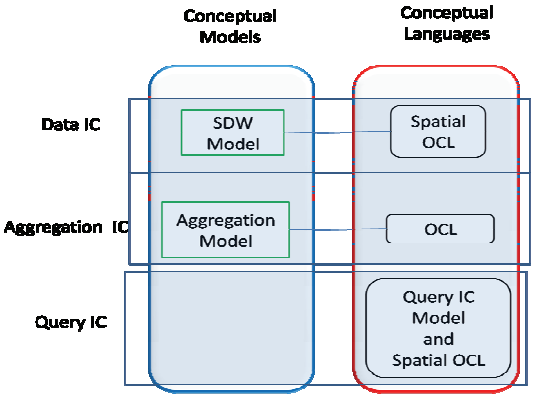


Figure 2: UML-OCL based conceptual framework.

The *SDW model* allows the definition of SDW data structures and the expression of Data IC on the top of these structures using Spatial OCL (Boulil et al., 2011).

The SDW case study represented using the *SDW model* is shown on Figure 3. This *SDW model* instance contains two dimensions: (i) a spatial dimension composed of 3 spatial levels (stereotyped as <<SpatialAggLevel>>), *City*, *Region* and *Country*; and (ii) a temporal dimension composed of three temporal levels *Day*, *Month* and *Year*. The numerical measure *temperature* (<<NumericalMeasure>> stereotype) is defined as an attributed of the fact class *Temperature* (<<Fact>> stereotype).

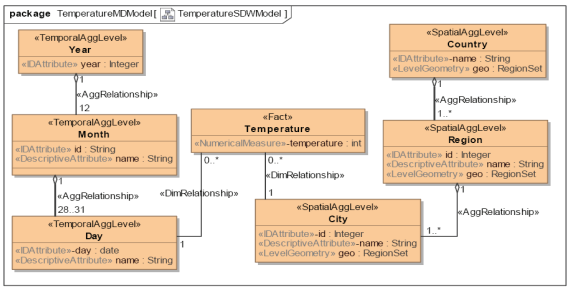


Figure 3: A *SDW model* instance.

Once the SDW model instance has been defined, data integrity constraints can be expressed using Spatial OCL. For example, the Data IC of Example 1 is expressed as follows:

```
context Region inv DataIC1:
self.geo.isInside(country.geo) or
self.geo.coveredBy(country.geo)
```

The Data IC of Example 2 is expressed using OCL in the following way:

```
context Temperature inv DataIC2:
not (
 self.day.day >= '1991-12-26' and
 self.city.region.country.name = 'URSS'
)
```

The *Aggregation model* represents how measures are aggregated along dimensions according to decision-makers' analysis needs. The instance of Aggregation model for our case study, which represents that the *temperature* measure (*aggregatedAttribute* tagged value) is aggregated along all the dimensions using the average aggregate function (*aggregator=Avg* tagged value), is depicted in Figure 4.

In (Boulil et al., 2011) we have identified a set of aggregation constraints that grant meaningful aggregations of measures. These constraints are valid for all SOLAP applications. Thus, we have implemented them as OCL constraints in the Aggregation Model package of the profile. They are checked by the CASE tool at the design stage when validating the conceptual model.

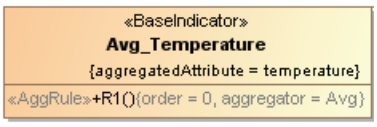


Figure 4: Aggregation model instance.

For example, in order to force the user to not aggregate non-additive (or value per unit) measures (for example the temperature; Example 3) using the sum aggregate function, the following OCL statement is defined in the profile:

```
context AggRule inv notSumValuePerUnitMeasure:
if(
 baseIndicator.aggregatedAttribute.OclIsKindOf(Measure)
 and baseIndicator.aggregatedAttribute.addType =
  'ValuePerUnit'
) then aggregator.name <> 'Sum'
```

Finally, designers can express IC on SOLAP queries using the *Query IC model*. Typically, a SOLAP query is a combination of measures and members from different dimensions. Thus, the Query IC model can be used for example to define invalid combinations of member sets. These member sets are specified as attributes with the <<MemberSet>> stereotype. The value domain of a <<MemberSet>> attribute is a subset of members of a dimension level, whose definition is precised with

the *condition* tagged value, which is an OCL statement defined on the context of the dimension level to select a subset of its members.

An example of an instance of the Query IC model is depicted on Figure 5, where the user states that combining days (<<MemberSet>> *day*) after 26 December 1991 (*condition= After1991-12-26*, whose OCL expression is shown in Figure 6) with the USSR (<<MemberSet>> *country*) is meaningless in any SOLAP query.

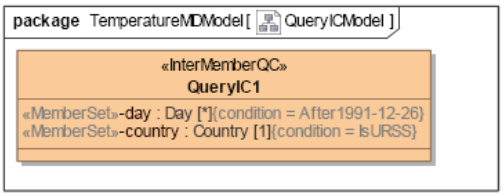


Figure 5: Query IC model instance.

```
context Day inv After1991-12-26:
self.day >= '1991-12-26'
```

Figure 6: OCL used by the Query IC of Figure 5.

4 IMPLEMENTATION

In this section, we present our architecture to automatically implement SOLAP IC (Figure 7). The main idea is to automatically implement each kind of IC in a different tier of the SOLAP architecture. The conceptual definition of each IC is automatically translated into the implementation language used by each tier. In particular, Data IC are translated using SpatialOCL2SQL and implemented in the SDW tier; Query IC are translated by our automatic code generator (called UML2MDX) and implemented in the OLAP server and the SOLAP client, and finally Aggregation IC are implemented in our UML profile using OCL and controlled during the design stage by the MagicDraw CASE tool.

Our SOLAP architecture (Figure 7) is based on: the Spatial DBMS Oracle Spatial 11g, the ROLAP Server Mondrian and a SOLAP client JRubik. Mondrian connects to a relational database and enables the execution of OLAP queries expressed using MDX (MultiDimensional eXpressions) that is a standard language for querying multidimensional databases. JRubik provides a graphical presentation layer on top of Mondrian and allows cartographic representations of OLAP queries using the SVG format.

In order to automatically implement data IC in the Oracle Spatial 11g, we have used the code generator Spatial OCL2SQL. Spatial OCL2SQL is a Java open source tool which integrates the spatial extensions of OCL called OCL 9IM and OCL ADV (Pinet et al., 2007). It automatically generates SQL scripts for Oracle Spatial from Spatial OCL conceptual constraints.

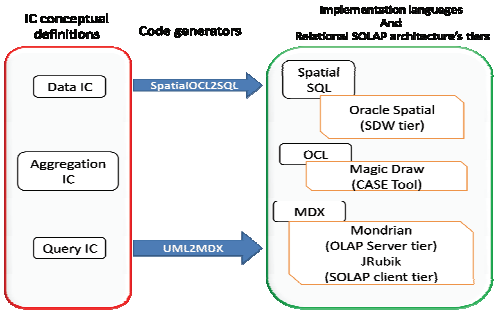


Figure 7: Automatic implementation of SOLAP IC.

In our case study, the previously defined OCL data IC of Example 2 is transformed in the following SQL query:

```
select * from TEMPERATURE SELF where not (
not (
 (select DAY from DAYS where DAY_PK in
  (select DAY_FK from TEMPERATURE where
   TEMPERATURE_PK = SELF.TEMPERATURE_PK)
 ) >= 19911226 and
 (select COUNTRY_NAME from CITIES where CITY_PK in
  (select CITY_FK from TEMPERATURE where
   TEMPERATURE_PK = SELF.TEMPERATURE_PK)
 ) = 'URSS'
));
```

This query selects the facts (TEMPERATURE table's tuples) that do not satisfy the constraint of Example 2.

The Aggregation IC are implemented as OCL profile inherent constraints in the MagicDraw CASE tool. MagicDraw supports OCL at the meta-model level (UML profile). In other terms, MagicDraw is able to check OCL constraints defined on UML stereotypes. This allows checking Aggregation IC at design stage independently of the specific SOLAP architecture used and without providing any implementation efforts. For example, if the designer defines an instance of the Aggregation model by using the Sum for the temperature measure, MagicDraw checks the OCL Aggregation IC of Example 3 and informs him that the constraint is violated.

In order to implement Query IC, we use MDX, which is the defacto standard of OLAP Servers and Clients. Thus, the choice of Mondrian as OLAP server is not a limitation for our generic architecture. The main idea is to translate the Query IC into MDX

formula, which are stored in the OLAP Server and then visualized in the SOLAP client. These formulas, when executed, inform user about the quality of query results. For each Query IC type we have defined an MDX template. The templates are fulfilled using a Java method (*UML2MDX*) that parses the XMI files associated to the Query IC. Different visual policies are associated with different combinations of members from these sets to be displayed in the SOLAP client tier: green colour for valid cells, yellow colour for aggregated cells that include valid and invalid cells and red colour for invalid cells. Figure 8 shows an example of OLAP query where these visual policies are applied according the MDX formula implementing the Query IC of Figure 6: valid cells such as those combining USSR with dates before 1991-12-26 (e.g. 1991-12-01) are displayed with green colour; invalid cells that involve for example USSR and dates after 1991-12-26 (e.g. 1991-12-27, 2010-1, 2010) are displayed with red colour, other cells are displayed with yellow colour, such as 1991-12 with USSR because it is the aggregation of valid (e.g. 1991-12-01 with USSR) and invalid cells (e.g. 1991-12-27 with USSR).

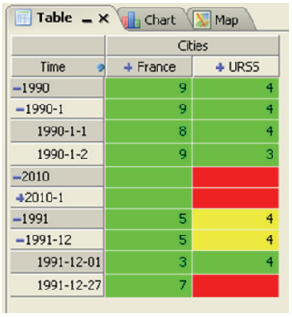


Figure 8: Query IC visualization of Example 4.

5 CONCLUSIONS

In this paper, we first show that the SOLAP analysis goodness depends on 3 quality types: data, aggregation and query qualities. Thus, we (i) extend the concept of integrity constraints to consider all these quality types; (ii) propose a framework based on a UML profile and Spatial OCL to express these SOLAP IC at the conceptual level; and (iii) show their automated implementations in a typical ROLAP architecture. Our current work is on improving the UML2MDX tool by integrating Spatial MDX expressions and defining cartographic-related visualization policies in order to implement spatial query IC.

As in our current automatic implementation only considers the snowflake schema SDW implementations, we are working on the consideration of the star-schema implementations. Finally, we will work on the formal validation of the completeness of our classification, and the expressiveness of our conceptual framework.

REFERENCES

Boulil, K., Bimonte, S., Pinet, F. (2011). Un modèle UML et des contraintes OCL pour les entrepôts de données spatiales : De la représentation conceptuelle à l'implémentation. *Ingénierie des Systèmes d'Information*, 16(6) 11-39

Ghozzi, F., Ravat, F., Teste, O., Zurfluh, G. (2003). Constraints and Multidimensional Databases. In *5th International Conference on Enterprise Information Systems*, 104-111

Glorio, O. and Trujillo, J. 2008. An MDA Approach for the Development of Spatial Data Warehouses. *In 10th International Conference on Data Warehousing and Knowledge Discovery*, Berlin-Heidelberg: Springer, 23-32

Lenz, H.-J. and Shoshani, A. 1997. Summarizability in OLAP and statistical data bases. In *International Conference on Scientific and Statistical Database Management*, IEEE, 132-143

Levesque, M.-A., Y. Bédard, M. Gervais, R. Devillers, (2007). Towards managing the risks of data misuse for spatial datacubes. In *5th International Symposium on Spatial Data Quality*, June 13-15, Enschede, Netherlands

Malinowski, E. and Zimányi, E. (2008). *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*. Berlin: Springer-Verlag.

Mazón, J.-N., J. Lechtenbörger, et al. (2009). A survey on summarizability issues in multidimensional modeling. *Dataand Knowledge Engineering* 68(12): 1452-1469.

Pinet, F., Duboisset, M. and Soulignac, V. (2007). Using UML and OCL to maintain the consistency of spatial data in environmental information systems. *Environmental modellingand software*, 22(8) 1217-1220

Pinet, F., Schneider, M. (2009) A Unified Object Constraint Model for Designing and Implementing Multidimensional Systems. *Journal of Data Semantics* 13, 37-71

Ribeiro, L., Goldschmidt, R., Cavalcanti, M. 2011. Complementing Data in the ETL Process. In *13th International Conference Data Warehousing and Knowledge Discovery*, Berlin-Heidelberg: Springer, 112-123

Salehi, M. (2009). *Developing a Model and a Language to Identify and Specify the Integrity Constraints in Spatial Datacubes*. Doctoral thesis. Faculté des études supérieures de l'Université Laval, Canada.

Labeling Methods for Association Rule Clustering

Veronica Oliveira de Carvalho[1], Daniel Savoia Biondi[1],
Fabiano Fernandes dos Santos[2] and Solange Oliveira Rezende[2]
[1]*Instituto de Geociências e Ciências Exatas, UNESP - Univ Estadual Paulista, São Paulo, Brazil*
[2]*Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, Brazil*
veronica@rc.unesp.br, danielsavoia@gmail.com, {fabianof, solange}@icmc.usp.br

Keywords: Association Rules, Post-processing, Clustering, Labeling Methods.

Abstract: Although association mining has been highlighted in the last years, the huge number of rules that are generated hamper its use. To overcome this problem, many post-processing approaches were suggested, such as clustering, which organizes the rules in groups that contain, somehow, similar knowledge. Nevertheless, clustering can aid the user only if good descriptors be associated with each group. This is a relevant issue, since the labels will provide to the user a view of the topics to be explored, helping to guide its search. This is interesting, for example, when the user doesn't have, a priori, an idea where to start. Thus, the analysis of different labeling methods for association rule clustering is important. Considering the exposed arguments, this paper analyzes some labeling methods through two measures that are proposed. One of them, Precision, measures how much the methods can find labels that represent as accurately as possible the rules contained in its group and Repetition Frequency determines how the labels are distributed along the clusters. As a result, it was possible to identify the methods and the domain organizations with the best performances that can be applied in clusters of association rules.

1 INTRODUCTION

Association rules are widely used in many distinct domain problems due to their ability to discover the frequent relationships that occur among sets of items stored in databases. Although this characteristic motivates its use, the main weakness of the association technique occurs when it is necessary to analyze the mining results. The huge number of rules that are generated makes the user's exploration a difficult task. Many approaches have been developed to overcome this post-processing problem, such as *Querying*, *Evaluation Measures*, *Pruning*, *Summarizing* and *Grouping* (Zhao et al., 2009; Natarajan and Shekar, 2005; Jorge, 2004). There are other ways to reduce the number of rules before post-processing be done, using, for example, extraction algorithms that are not exhaustive as *Apriori* (Agrawal and Srikant, 1994). However, the focus of this work is the post-processing phase. Thus, it is considered, in this work, that it is better not to eliminate rules (knowledge) during the extraction process, but to work with all of them later.

Grouping is a relevant approach related to the structure of the domain, since it organizes the association rules, previously obtained by algorithms like *Apriori* (Agrawal and Srikant, 1994), in groups that contain, somehow, similar knowledge. These groups can improve the presentation of the mined patterns, providing the user a view of the domain to be explored (Reynolds et al., 2006; Sahar, 2002). The papers that use clustering for post-processing association rules, as seen in (Reynolds et al., 2006; Jorge, 2004; Sahar, 2002; Toivonen et al., 1995), are only concerned with the domain organization. However, it is essential that the organizations be used to aid the user during the exploration process, minimizing its effort. Aiding can be obtained from a structured domain by: (i) highlighting the groups (clusters[1]) that are interesting to be explored; (ii) generating good labels for the groups that allow an easier browsing in the domain.

Regarding (i), (Carvalho et al., 2011), for example, proposed the PAR-COM methodology that, by combining clustering and objective measures, reduces the association rule exploration space by directing the user to what is potentially interesting. Thus, the user only explores a small subset of the groups that contain the potentially interesting knowledge. Regarding (ii), it is essential that groups be represented by

[1]The words groups and clusters are used in this paper as synonymous.

labels that can provide the user a view of the topics contained in the exploration space, helping to guide its search. Finding good labels is a relevant issue in many tasks as in Text Mining (TM) and Information Retrieval (IR) (see some applications in (Manning et al., 2009)). It is necessary, for example, that good descriptors be presented to the user to facilitate exploratory analyses, interesting when the user doesn't have, a priori, an idea where to start. Furthermore, although many methods have been proposed to label document clusters in TM and IR, the papers related to association rule clustering have not explored this issue. Thus, as in other tasks, the analysis of different labeling methods for association rule clustering is also relevant, since it is necessary to identify the methods that present good results. Besides, the integration of good labeling methods with other methodologies can allow association rule clustering to become a powerful post-processing tool. The integration with PARCOM (Carvalho et al., 2011), for example, can enable the identification of the potentially interesting topics in the domain.

Considering the exposed arguments, this paper aims to analyze some labeling methods in order to identify: **(a)** the methods that are more adequate for association rule clustering; **(b)** the domain organizations that provide the best results, since the performance of the methods are affected by them, i.e., by a clustering algorithm combined with a similarity measure; **(c)** a consequence of **(b)**, the domain organizations that best structure the knowledge. Two measures are proposed and used to evaluate the methods. The ideal is that the labels of each cluster represent as accurately as possible the knowledge of its group (Precision (*P*) measure) and be as different as possible of the labels of the other groups (Repetition Frequency (*RF*) measure). It is important to mention that this paper doesn't fit in the post-processing approaches itself. The labeling methods here presented have to be applied to clustering of association rules, i.e., along with a post-processing methodology.

The paper is structured as follows: Section 2 presents some related works; Section 3 and Section 4 the labeling methods that were selected and the measures that were proposed to evaluate the experiments results, respectively; Section 5 the configurations used in experiments; Section 6 the results and discussion; Section 7 the conclusions and future works.

2 RELATED WORKS

Since this paper aims to analyze some labeling methods for association rule clustering, this section presents some papers related to the clustering approach and the labeling methods they use.

In order to structure the extracted knowledge, different clustering strategies have been used for post-processing association rules. In (Reynolds et al., 2006) clustering is demonstrated through partitional (K-means, PAM, CLARANS) and hierarchical (AGNES) algorithms using Jaccard as the similarity measure. In this case, the Jaccard between two rules r and s, expressed by $\text{J-RT(r,s)}=\frac{\#\{t \text{ matched by } r\}\cap \#\{t \text{ matched by } s\}}{\#\{t \text{ matched by } r\}\cup \#\{t \text{ matched by } s\}}$, is calculated considering the common transactions (t) the rules match – we refer to this similarity measure as Jaccard with Rules by Transactions (J-RT). A rule matches a transaction t if all the rule items are contained in t. (Jorge, 2004) demonstrates the use of clustering through hierarchical algorithms (Single Linkage, Complete Linkage, Average Linkage) also using Jaccard as the similarity measure. However, the Jaccard between two rules r and s, expressed by $\text{J-RI(r,s)}=\frac{\#\{\text{items in } r\}\cap \#\{\text{items in } s\}}{\#\{\text{items in } r\}\cup \#\{\text{items in } s\}}$, is calculated considering the items the rules share – we refer to this measure as Jaccard with Rules by Items (J-RI). (Toivonen et al., 1995) proposes a similarity measure based on transactions and uses a density algorithm to carry out the clustering of the rules. (Sahar, 2002) also proposes a similarity measure based on transactions considering (Toivonen et al., 1995)'s work, although using a hierarchical algorithm to carry out the clustering.

All the above papers, related to the structure of the domain, are only concerned with the domain organization. In general, each paper only uses one family of clustering algorithms along with one similarity measure to cluster the association rules and a unique labeling method to present the mined results to the user. (Reynolds et al., 2006) and (Jorge, 2004) select as labels of each group the items that appear in the rule which is more similar to all the other rules in the group (the medoid of the group). (Toivonen et al., 1995) doesn't mention how the labels are found, but provides some traces that the labels represent the more frequent and distinct items in the group. On the other hand, (Sahar, 2002) proposes an approach to summarize each cluster by finding the patterns $a \Rightarrow c$ that cover all the rules in the cluster; a and c are items in the domain and a pattern $a \Rightarrow c$ covers a rule $A \Rightarrow C$ if $a \in A$ and $c \in C$. As observed, although the proposed approach is used to summarize the clusters and not, in fact, to define the cluster's labels, the idea can be used for this purpose.

Although many methods have been proposed to label document clusters in tasks of Text Mining (TM)

and Information Retrieval (IR), as in (Moura and Rezende, 2010; Lopes et al., 2007; Kashyap et al., 2005; Fung et al., 2003; Glover et al., 2002; Popescul and Ungar, 2000; Larsen and Aone, 1999; Cutting et al., 1992), the papers related to association rule clustering have not explored this issue. However, as presented in next section, many of these methods used to label document clusters are similar to the ones used to label association rule clusters, i.e, they are, somehow, related. Thus, some methods, apart from the ones presented in the next section, could be adapted from TM and IR for association rule clustering.

3 LABELING METHODS

Aiming to analyze some labeling methods (LM) for association rule clustering regarding their behavior in relation to precision and distinctiveness, four methods were selected and implemented. These methods represent the ideas of many of the methods previously described and cited in Section 2 (both for association rules (AR) as for documents (TM and IR)). In order to understand the methods, consider a clustering composed of three clusters of association rules: C_1={r_1: coffee $\Rightarrow$ butter; r_2: milk $\Rightarrow$ coffee; r_3: milk & butter $\Rightarrow$ coffee}; C_2={r_1: butter $\Rightarrow$ coffee; r_2: milk $\Rightarrow$ butter}; C_3={r_1: butter $\Rightarrow$ milk; r_2: coffee $\Rightarrow$ milk}. The example is merely illustrative. The four methods described below are **LM-M**, **LM-T**, **LM-S** and **LM-PU**.

In **LM-M** (*L*abeling *M*ethod *M*edoid) the labels of each cluster are built by the items that appear in the rule of the group which is more similar to all the other rules in the cluster (the medoid of the group). So, is computed the accumulated similarity (a_s) of each rule considering its similarity with respect to the other rules and the one with the highest value is selected. Considering C_1 of the above example and that r_1 covers {t_1, t_3, t_5, t_7}, r_2 {t_1, t_3, t_5, t_7, t_9}, r_3 {t_3, t_5, t_7}, the similarities s(r_1,r_2)=s(r_2,r_1)=$\frac{4}{5} = 0.8$, s(r_1,r_3)=s(r_3,r_1)=$\frac{3}{4} = 0.75$, s(r_2,r_3)=s(r_3,r_2)=$\frac{3}{5} = 0.6$, considering J-RT (Section 2), are obtained and the following a_s are found: a_s(r_1)=s(r_1,r_2)+s(r_1,r_3)=1.55; a_s(r_2)=s(r_2,r_1)+s(r_2,r_3)=1.40; a_s(r_3)=s(r_3,r_1)+s(r_3,r_2)=1.35. Thus, r_1 is selected and C_1's labels are defined to be {coffee, butter}. These similarities among rules can be obtained through any similarity measure, as the ones presented in Section 2. In this paper we used J-RT as in the most of the literature works. The papers related with this idea are (Reynolds et al., 2006; Jorge, 2004) from AR and (Kashyap et al., 2005; Larsen and Aone, 1999; Cutting et al., 1992) from TM and IR. In this case, the user can also know the existing relationship among the labels through the rule.

In **LM-T** (*L*abeling *M*ethod *T*ransaction) the labels of each cluster are built by the items that appear in the rule of the group that covers the largest number of transactions. A rule covers a transaction t if all the rule items are contained in t. Considering C_1 of the above example and that r_1 covers {t_1, t_3, t_5, t_7}, r_2 {t_3, t_5, t_7}, r_3 {t_1, t_3, t_5, t_7, t_9}, r_3 is selected and C_1 labels are defined to be {milk, butter, coffee}. The paper related to this idea is (Fung et al., 2003) from TM and IR. In this case, the user can also know the existing relationship among the labels through the rule.

In **LM-S** (*L*abeling *M*ethod *S*ahar due to its reference to (Sahar, 2002)), a simplified version of the process described in (Sahar, 2002) from AR and explained in Section 2, the labels of each cluster are built as follows: (i) considering a set $I = \{i_1, ..., i_m\}$ containing all the distinct cluster items, a set $R = \{r_1, ..., r_n\}$ containing all the possible relationships $a \Rightarrow c$, where $a, c \in I$ – each one of these relationships represents a rule pattern; (ii) the number of rules that each pattern $r_i \in R$ covers is computed (N_c); a pattern $a \Rightarrow c$ covers a rule $A \Rightarrow C$ if $a \in A$ and $c \in C$; (iii) the pattern with the highest cover is selected; in the event of a tie all tied pattern are selected; (iv) all the selected patterns compose a set $P \subseteq R$; (v) at the end, all the distinct items in P compose the labels. Considering C_1 of the above example we have: I={coffee, butter, milk}, R={$r_1 : coffee \Rightarrow butter$, $r_2 : butter \Rightarrow coffee$, $r_3 : coffee \Rightarrow milk$, $r_4 : milk \Rightarrow coffee$, $r_5 : butter \Rightarrow milk$, $r_6 : milk \Rightarrow butter$}, N_c={$r_1 : 1$, $r_2 : 1$, $r_3 : 0$, $r_4 : 2$, $r_5 : 0$, $r_6 : 0$} and P={r_4}. Thus, C_1's labels are defined to be {milk, coffee}. In this case, the user can also know the existing relationship among the labels through the rule(s).

In **LM-PU** (*L*abeling *M*ethod *P*opescul and *U*ngar due to its reference to (Popescul and Ungar, 2000)) the labels of each cluster are built by the N items in the cluster that present the best tradeoff between frequency and predictiveness; formally we have: $f(i_n|C_n) * \frac{f(i_n|C_n)}{f(i_n)}$. The $f(i_n|C_n)$ measure computes the frequency f of each item i_n in its cluster C_n. The $\frac{f(i_n|C_n)}{f(i_n)}$ measure computes the frequency f of each item i_n in its cluster C_n divided by the item frequency in all the clusters. The i_n items are all the distinct items that are present in the rules of the cluster. Each time an item i_n occurs in a rule its frequency is incremented by one. Therefore, the labels are built by the N items that are more frequent in their own cluster and infrequent in the other clusters. Considering C_1 of the above example, its distinct items {coffee,

butter, milk} and $N = 1$ we have: coffee=$3 * \frac{3}{5}$=1.8; butter=$2 * \frac{2}{5}$=0.8; milk=$2 * \frac{2}{5}$=0.8. Thus, C_1's labels are defined to be {coffee}. The papers related to this idea are (Toivonen et al., 1995) from AR and (Lopes et al., 2007; Glover et al., 2002; Popescul and Ungar, 2000) from TM and IR. In this case, the user doesn't know the existing relationship among the labels.

4 EVALUATION METHODOLOGY

In order to evaluate the precision and distinctiveness of the four labeling methods, two measures, presented in Equations 1 and 2, were proposed, where N refers to the number of clusters. Both measures range from 0 to 1. To understand the measures, consider a clustering composed of three clusters of association rules: C_1={coffee $\Rightarrow$ butter; milk $\Rightarrow$ butter} with the labels {butter, coffee, milk}; C_2={butter $\Rightarrow$ coffee; milk $\Rightarrow$ coffee} with the label {milk}; C_3={butter $\Rightarrow$ milk; coffee $\Rightarrow$ milk} with the labels {butter, milk}. The example is merely illustrative.

Precision (P), in Equation 1, measures how much the labeling method can generate labels that really represent the rules contained in the clusters. This measure is an adaptation of Recall used in Information Retrieval (see (Manning et al., 2009)); however, in this case, the relevant items to be retrieved are all the rules in a cluster. Considering the above example, the illustrative method has a P of 0.83 ($P(C) = \frac{\frac{2}{2}+\frac{1}{2}+\frac{2}{2}}{3}$), since the labels of C_2 represent only one rule of a total of two. It is considered that a rule is represented (covered) by a set of labels if the rule contains at least one of the labels. Thus, it is expected that a good method must have a high precision. However, it is not enough to be precise if the labels appear repeatedly among the clusters. Therefore, Repetition Frequency (RF), in Equation 2, measures how much the distinct labels that are present in all the clusters don't repeat. Considering the above example, the illustrative method has a RF of 0.33 ($RF(C) = 1 - \frac{2}{3}$): one of the three distinct labels (butter, coffee, milk) that are present in clusters doesn't repeat. The higher the RF value, the better the method, i.e., less repetitions implies in better performance. Observe that RF can be used to compute the repetition frequency if we omit "1-" of Equation 2; however, in this case, the lower the RF value, the better the method. Thereby, the choice of not computing the repetition was to standardize the interpretation of the measures.

$$P(C) = \frac{\sum_{i=1}^{N} P(C_i)}{N}, \text{ where} \quad (1)$$

$$P(C_i) = \frac{\#\{\text{rules covered in } C_i \text{ by } C_i \text{ labels}\}}{\#\{\text{rules in } C_i\}}$$

$$RF(C) = 1 - \frac{\#\{\text{distinct labels that repeat in the clusters}\}}{\#\{\text{distinct labels in the clusters}\}} \quad (2)$$

Considering the labeling methods and the above measures, some experiments were realized, which are next described.

5 EXPERIMENTS

Some experiments were carried out to evaluate the labeling methods regarding precision and distinctiveness through P and RF. The four data sets used in experiments are presented in Table 1. The first three are available in *R Project for Statistical Computing* through "arules" package[2]. The last one was donated by a supermarket located in São Carlos city, Brazil[3]. All the transactions of the Adult and Income contain the same number of items (referred here as standardized-transaction data sets), different from Groceries and Sup (referred here as non-standardized-transaction data sets). Thus, the labeling methods were evaluated on different types of data. The rules were mined using an *Apriori* implementation developed by Christian Borgelt[4] with a maximum number of 5 items per rule and excluding the rules of type $\emptyset \Rightarrow X$, where X is an item contained in data. With the Adult set 6508 rules were generated using a minimum support (min-sup) of 10% and a minimum confidence (min-conf) of 50%; with Income 3714 rules considering a min-sup of 17% and a min-conf of 50%; with Groceries 2050 rules considering a min-sup of 0.5% and a min-conf of 0.5%; with Sup 7588 rules considering a min-sup of 0.7% and a min-conf of 0.5%. These parameter values were chosen experimentally considering the exposed arguments in Section 1 and (Carvalho et al., 2011)'s work.

Table 1: Details of the data sets used in experiments.

Data set	# of transactions	# of distinct items
Adult	48842	115
Income	6876	50
Groceries	9835	169
Sup	1716	1939

Since the papers described in Section 2 only use one family of clustering algorithms and one similar-

[2]http://cran.r-project.org/web/packages/arules/index.html.
[3]http://sites.labic.icmc.usp.br/research/Cjto-Sup.data.
[4]http://www.borgelt.net/apriori.html.

ity measure to cluster the association rules, it was decided to use one algorithm of each family and the two most used similarity measures (J-RI and J-RT (Section 2)). The Partitioning Around Medoids (PAM) was chosen within the partitional family and the Average Linkage within the hierarchical family. PAM was executed with k ranging from 5 to 50 considering a step of 5. The dendrograms generated by Average Linkage were cut in the same ranges (5 to 50 considering a step of 5). All the choices were made considering an analysis of many clustering configurations presented in (Carvalho et al., 2012). Table 2 summarizes the configurations used in the experiments.

Table 2: Configurations used in the experiments.

Data sets	Adult; Income; Groceries; Sup
Algorithms	PAM; Average Linkage
Similarity measures	J-RI; J-RT
k	5 to 50, step of 5

Table 3: Results for *P* and *RF* considering the ADULT and INCOME data sets.

Labeling method	Mean of *P*	Mean of *RF*
LM-M	0.995310	0.321458
LM-T	0.923752	0.340560
LM-S	0.965381	0.416278*
LM-PU	0.997238*	0.305087
Clustering algorithm	**Mean of *P***	**Mean of *RF***
PAM	0.969465	0.285709
Average	0.971375*	0.405983*
Similarity measure	**Mean of *P***	**Mean of *RF***
J-RI	0.970287	0.269874
J-RT	0.970553*	0.421818*

Table 4: Results for *P* and *RF* considering the GROCERIES and SUP data sets.

Labeling method	Mean of *P*	Mean of *RF*
LM-M	0.924978	0.700539*
LM-T	0.771151	0.696544
LM-S	0.899201	0.641688
LM-PU	0.971076*	0.662681
Clustering algorithm	**Mean of *P***	**Mean of *RF***
PAM	0.873818	0.564347
Average	0.909385*	0.786379*
Similarity measure	**Mean of *P***	**Mean of *RF***
J-RI	0.930973*	0.616215
J-RT	0.852230	0.734511*

Considering the configurations in Table 2, the four labeling methods (LM-M; LM-T; LM-S; LM-PU) were applied in the different domain organizations. In relation to the labeling methods, LM-M and LM-T select only one rule as label, LM-S one or more rules, in case of tie, and LM-PU the 5 items that present the best tradeoff between frequency and predictiveness. Thus, in average, all the labeling methods generate the same amount of labels per cluster. In the end, the performance of each labeling method was evaluated through *RF* and *P*, whose results are presented in the next section. It is important to remember that the aim of the measures is to evaluate, respectively, how much the method can find labels that represent as accurately as possible the knowledge contained in their own groups and how the labels are distributed along the clusters. The ideal is to identify methods that have high values for both measures.

6 RESULTS AND DISCUSSION

As mentioned before, the performance of the labeling methods were evaluated through *P* and *RF*. Thus, in order to identify the methods that are more adequate for association rule clustering and the domain organizations that provide the best results, an analysis based on the mean of each measure was done. Tables 3 and 4 present the results – the best values are marked with "*". Each mean was obtained considering all the results of the experiments[5], which were grouped according to the criteria shown (labeling method, clustering algorithm, similarity measure) and according to the different types of data (standardized-transaction (Table 3) and non-standardized-transaction (Table 4)). It is important to mention that since the results are deterministic no statistical test was done. It can be observed that:

- in the standardized-transaction data sets (Table 3) the method that presents the best result regarding *P* is LM-PU and considering *RF* LM-S. Thereby, the user can choose one of them based on his interests: accurate or distinctiveness. However, it is possible to note that in all the methods *RF* presents low values while *P* presents high values. Thus, it is better to use LM-S when the user wants a tradeoff between *P* and *RF*, since it improves *RF* (difference above 0.1) while maintaining a good *P* (difference of 0.03).
- in the non-standardized-transaction data sets (Table 4) the method that presents the best result regarding *P* is LM-PU and considering *RF* LM-M. Thereby, the user can choose one of them based on his interests: accurate or distinctiveness. On the other hand, it is possible to note that both methods have similar values (difference of 0.05

[5]All the results of the experiments are available in http://veronica1.rc.unesp.br/public/ICEIS-2012-R.pdf.

Table 5: Examples of labels obtained in some of the experiments using Average+J-RT and $k = 5$.

Experiment	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Income+ LM-S	age=14-34 dual_incomes=not_married language_in_home=english	age=35+ language_in_home=english	language_in_home=english number_in_household=1	language_in_home=english occupation=professional/managerial	language_in_home=english sex=female years_in_bay_area=10+
SUP+ LM-M	agua_tonica_antartica coca_cola	coca_cola gatorade	deterglimpol oleo_girassol_salada_bunge	fartrigo_renata gelatina_royal leite_moca	coca_cola leite_salute

in P and of 0.04 in RF). Thus, both of them could be used when the user wants a tradeoff between P and RF. However, it seems more adequate to use LM-M in spite of LM-PU since LM-M (i) can be more easily computed with partitional algorithms, (ii) can allow the user to know the existing relationship among the labels and (iii) presents a better value for RF (above 0.7) while maintaining a good P (above 0.9). Finally, it is possible to note that these types of data sets present better RF values in relation to the RF values in Table 3.

- the algorithm that presents the best performance in all the tests is Average (Tables 3 and 4).
- the similarity measure that presents the best performance in almost all the tests is J-RT (Tables 3 and 4). The only exception is P in Table 4, where J-RI presents a better performance.

Considering the exposed arguments, it can be observed that: (i) for standardized-transaction data sets the method that seems to be more adequate for association rule clustering is LM-S; (ii) for non-standardized-transaction data sets the method that seems to be more adequate for association rule clustering is LM-M; (iii) the methods present better results when the clustering is obtained through Average; (iv) J-RT seems to be a good similarity measure to be used along with Average; (v) as a consequence of (iii), it is possible to verify that Average represents the domain organization which best separates the domain knowledge, independently of the similarity measure used – it can be inferred that a domain is well separated if a domain organization, along with an adequate labeling method, provides good labels. These conclusions cover the three objectives stated in Section 1 (letters **(a)** to **(c)**). Besides, these results can be used with other methodologies, as the methodology described in (Carvalho et al., 2011), to make the association rule clustering a powerful post-processing tool.

Finally, Table 5 presents examples of labels obtained in some of the experiments using Average+J-RT and $k = 5$. One data set of each type of data (standardized or non-standardized) is shown along with its labeling method, according to the results above discussed, that had the best performance. The items that occur more than once are underlined. It can be observed that: (i) the labels of Income describe, with good precision and distinctiveness ($P = 0.835$; $RF = 0.875$), some specificities well defined of the domain – cluster 2, for example, is related to people above 35 years and cluster 5 to people who are female and live for more than 10 years in the San Francisco Bay area; (ii) on the other hand; the labels of SUP describe, also with good precision and distinctiveness ($P = 0.788$; $RF = 0.889$), some types of beverages that can be purchased, as clusters 1, 2 and 5, which are related with distinct shop styles: cluster 1 with water, cluster 2 with soft drink and cluster 5 with milk; (iii) the items that occur in many clusters labels are very frequent in their data sets (language_in_home=english: 91%; coca_cola: 22%), which means that they can be used as complementary information of the clusters. Thus, as observed, it is essential that good labels be found, since they can aid the users in exploratory analyses by guiding their search.

7 CONCLUSIONS

Due to the huge amount of association rules that are obtained, considering the exposed arguments in Section 1, many approaches were suggested, as clustering. However, for clustering to be useful to users it is essential that good descriptors be associated with each cluster to help, for example, in guiding their search. Thus, the analysis of different labeling methods for association rule clustering is a relevant issue. Considering the exposed arguments, this paper analyzed some labeling methods. Two measures were proposed and used to evaluate the methods. Precision, P, measures how much the methods can find labels that represent as accurately as possible the rules contained in their own groups. Repetition Frequency, RF, measures how the labels are distributed along the clusters. As a result, it was possible to identify the methods and the domain organizations with the best performances that can be applied in clusters of association rules.

As future work we will explore some approaches that aim to improve the labels through a generalization process. We want to explore the impact of generic labels on *P* and *RF* to analyze if the results of the labeling methods can be improved. From this generalization process we intend to discover a topic for each cluster considering the context given by the user through ontology. Given, for example, the labels "rice", "bean" and "salad", the topic could be food or lunch, depending on the knowledge codified in the ontology.

ACKNOWLEDGEMENTS

We wish to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (processes numbers: 2010/07879-0 and 2011/19850-9) and Fundação para o Desenvolvimento da Unesp (FUNDUNESP) for the financial support.

REFERENCES

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In Bocca, J. B., Jarke, M., and Zaniolo, C., editors, *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB 1994*, pages 487–499.

Carvalho, V. O., Santos, F. F., and Rezende, S. O. (2011). Post-processing association rules with clustering and objective measures. In *Proceedings of the 13th International Conference on Enterprise Information Systems*, volume 1, pages 54–63.

Carvalho, V. O., Santos, F. F., Rezende, S. O., and Padua, R. (2012). PAR-COM: A new methodology for post-processing association rules. *Lecture Notes in Business Information Processing*, 102. In press. Available due May 19.

Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329.

Fung, B. C. M., Wang, K., and Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In *Proceedings of the 3rd SIAM International Conference on Data Mining*, pages 59–70.

Glover, E. J., Pennock, D. M., Lawrence, S., and Krovetz, R. (2002). Inferring hierarchical descriptions. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 507–514.

Jorge, A. (2004). Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In *Proceedings of the 4th SIAM International Conference on Data Mining*. 10p.

Kashyap, V., Ramakrishnan, C., Thomas, C., and Sheth, A. (2005). Taxaminer: An experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, 1(2):240–266.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 16–22.

Lopes, A. A., Pinho, R., Paulovich, F. V., and Minghim, R. (2007). Visual text mining using association rules. *Computers & Graphics*, 31(3):316–326.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press. 544p.

Moura, M. F. and Rezende, S. O. (2010). A simple method for labeling hierarchical document clusters. In *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications*, pages 336–371.

Natarajan, R. and Shekar, B. (2005). Interestingness of association rules in data mining: Issues relevant to e-commerce. *SADHANA – Academy Proceedings in Engineering Sciences (The Indian Academy of Sciences)*, 30(Parts 2&3):291–310.

Popescul, A. and Ungar, L. (2000). Automatic labeling of document clusters. Unpublished manuscript. http://www.cis.upenn.edu/∼popescul/Publications/popescul00labeling.pdf.

Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504.

Sahar, S. (2002). Exploring interestingness through clustering: A framework. In *Proceedings of the IEEE International Conference on Data Mining*, pages 677–680.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., and Mannila, H. (1995). Pruning and grouping discovered association rules. In *Workshop Notes of the ECML Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pages 47–52.

Zhao, Y., Zhang, C., and Cao, L. (2009). *Post-mining of association rules: Techniques for effective knowledge extraction*. Information Science Reference. 372p.

Modeling the Performance and Scalability of a SAP ERP System using an Evolutionary Algorithm

Daniel Tertilt[1], André Bögelsack[1] and Helmut Krcmar[2]
[1] *fortiss GmbH, An-Institut der Technischen Universität München, Guerickestr. 25, 80805 München, Germany*
[2]*Technische Universitaet Muenchen, Boltzmannstraße 3, 85748 Garching, Germany*
{tertilt, boegelsack}@fortiss.org, krcmar@in.tum.de

Keywords: Performance, Modeling, ERP, Synthetic Benchmark, Evolutionary Algorithm.

Abstract: Simulating the performance behavior of complex software systems, like Enterprise Resource Planning (ERP) systems, is a hard task due to the high number of system components when using a white box simulation approach. This paper utilizes a black box approach for establishing a simulation model for SAP ERP systems on the basis of real world performance data, which is gathered by using a synthetic benchmark. In this paper we introduce the benchmark, called Zachmanntest, and demonstrate that by using an evolutionary algorithm basing on the results of the Zachmanntest, the exact performance behavior of the ERP system can be modeled. Our work provides insights on how the algorithm is parameterized e.g. for the mutation and crossover probability, to receive optimal results. Furthermore we show that the evolutionary algorithm models the performance and scalability of an ERP system with an error less than 3.2%. With this approach we are able to build simulation models representing the exact performance behavior of a SAP ERP system with much less effort than required when using a white box simulation approach.

1 INTRODUCTION

The performance of an enterprise SAP ERP system is a business critical factor, as the ERP system often builds the basis for many semi-automated business processes. The throughput and response time of the ERP system determines how fast the business operations can be performed. Any change on the ERP system in hardware, software or user behavior is a business critical action.

Software performance prediction is an approach to reduce the risk of bad system performance after such a change. Simulation approaches like layered queuing networks (Franks et al., 2009) are conventionally used to predict the performance behavior of SAP ERP systems (Sithole et al., 2010). Simulation approaches though require an insight into the system (white box approach), which is not always given. The white box approach becomes more hardly to handle, when more than 60,000 SAP ERP system's components have to be represented in an appropriate simulation model. Besides, existing models often only assume the performance behavior of the components (for example an exponential function), which leads to incorrect simulation results. In order to avoid incorrect simulation models and results, a black box approach might be used first.

This paper uses a black box approach for creating a simulation model based on performance data from a real-world SAP ERP system. We strictly follow the proposed simulation way of Jain (Jain, 1991), where any simulation should be based on reliable performance data. The appropriate performance data set is gathered by executing a synthetic benchmark, called Zachmanntest (Bögelsack et al., 2010). We follow the black box approach by executing the test from outside of the SAP ERP system and only record the performance results. Applying a white box approach to the analyzed SAP ERP system would be possible too, but very costly due to the system's complexity. The results of the black box approach are then used to build up the simulation model. Whenever though this data is used as input to a subsequent system like a simulation engine, it has to be transformed into a mathematical model. An algorithm that exactly solves the mathematical modeling for any given set of data is highly complex, resulting in an inacceptable execution time. To avoid this execution time we use a model approximation using an evolutionary algorithm. As long as the

approximation error is less than the estimated measurement error, the approximation does not negatively affect the exactness of the subsequent system. This paper proves that evolutionary algorithms can be used to establish a model representing the performance behavior of a SAP ERP system under different configurations and workloads. We describe the model approximation we performed on the results of the Zachmanntest, using our evolutionary algorithm implementation Mendel. Our research shows what affects the efficiency of the algorithm, and how it has to be configured to obtain best results. In addition we depict the modeling results and interpret the usability of evolutionary algorithms for black box ERP performance prediction.

The rest of the paper is organized as follows: section 2 provides an overview about related work in the field of performance simulation of ERP systems and the usage of evolutionary algorithms for the purpose of performance modeling. Section 3 describes the background and functionality of the Zachmanntest in detail. The explanation of the evolutionary algorithm and the establishment of it are explained in section 4. Section 5 summarizes the paper and provides an outlook.

2 RELATED WORK

Exploring related work in the area of modeling and simulating SAP ERP systems should be divided into two subareas: 1) the application of any modeling and simulation approach to SAP ERP systems and 2) the application of evolutionary algorithms to the IS field for any modeling or simulation purpose.

Regarding the first subarea there are several papers available, all dealing with the common problem of how to simulate a SAP ERP system, which consists of more than 60,000 programs. Modeling the performance of SAP ERP system is firstly mentioned in (Bögelsack et al., 2008), whereas the authors state out how they would tackle the modeling problem of a complex software product like SAP ERP system. The approach is afterwards extended in (Gradl et al., 2009). Here a concrete modeling approach called Layered Queuing Network (LQN) is used and a first model is populated manually with performance measurement data and simulated afterwards. The same approach of utilizing LQN is used in (Rolia et al., 2009) to show the appropriateness of the LQN approach. Further research of the authors lead to (Rolia et al. 2010), where a resource demand estimation approach is presented.

In the area of applying evolutionary algorithms to a IS-related problems, first papers are published in the area of logistic problems, e.g. for the pallet loading problem as in (Herbert/Dowsland 1996). However, applying evolutionary algorithms in the area of simulation and especially performance simulation is very common. (Tikir et al., 2007) shows the application of evolutionary algorithms in the field of High Performance Computing. In (Justesen 2009) a simulation model combined with an evolutionary algorithm to find optimal processing sequences for two cluster tools from the wafer manufacturing.

3 PERFORMANCE MEASUREMENT AND WORKLOAD CREATION

Following the ideas of (Jain, 1991) and (Law, 2008) every simulation must be either based on or validated by performance measurement results and obtaining data from real-world applications is the best case for this. Generally spoken, application and synthetic benchmarks can be used to obtain valuable performance results. In this chapter we explain a synthetic benchmark, called Zachmanntest, which is used to gather performance results. Those results form the basis of our simulation model and algorithm.

3.1 Application and Synthetic Benchmark

Measuring the performance of SAP ERP systems is a hard task as there are two different perspectives of how to measure the performance and how to implement a measurement process. First, the usage of so called application benchmarks is proposed. Application benchmarks contain a sequence of typical application usage steps. An exemplary step would be the creation of a customer order or a production order. The set of typical application usage steps form the application benchmark, which is then somehow instrumented with a performance metric, e.g. the number of created production orders. The most commonly known application benchmark in the sector of SAP ERP systems is the sales and distribution benchmark (SD benchmark). Application benchmarks are used very often, which can be proven by the large number of available SD-

benchmark results (see (SAP, 2010)). One drawback of application benchmark is that they are hard to implement and need a huge testing environment.

Second, the usage of synthetic benchmarks is proposed for measuring the performance of a SAP ERP system. The synthetic approach derives from the need of testing the performance of a very specific element in the SAP ERP system. A synthetic benchmark is a set of elementary operations in the SAP ERP system (Curnow/Wichmann, 1976). For example, applying a TPC-benchmark for measuring the performance of the underlying database system, is a popular approach to get an understanding of the system's performance (Doppelhammer et al., 1997). One drawback of any synthetic benchmark is that the benchmark is very focused. However, the major advantage is, that a synthetic benchmark can be easily applied to the system and performance results can be gained quickly.

In this paper we utilize a synthetic benchmark to measure the performance and scalability of a SAP ERP system. We chose the synthetic benchmark, because it is easy to apply to the SAP ERP system, it gains the necessary results for our simulation approach and the benchmark steps are transparent to us.

3.2 Zachmanntest – A Synthetic Main Memory Benchmark

3.2.1 Zachmanntest: Architecture

The Zachmanntest consists of two Advanced Business Application Programming (ABAP) programs. The first program is an easy to use entry mask to specify the test execution parameters. The second one is the ultimate test executable, which produces a lot of main memory operations in the application server. In fact, those main memory operations are operations on so called internal tables. Each program of the SAP ERP system, which is somehow interacting with the database management system and stores/reads data from it, uses this concept. From our point of view, this operation is a universal one and therefore a suitable example for a synthetic benchmark. A synthetic benchmark requires a specific sequence of operations/programs to be executed during runtime (Curnow and Wichmann, 1976). This is achieved by specifying the following steps during the execution. Please note that we used pseudo-code instead of ABAP statements:

```
 1:While time < max_run
 2:  Create internal table
 3:  Fill internal table with data
 4:  While iteration <loop_cnt
 5:  Randomly select data set
 6:  Read selected data set
 7:  Increase throughput counter
 8:  Endwhile
 9:  Delete internal table
10:Endwhile
11:Print throughput counter
```

The value max_run defines the runtime (default: 900 seconds) after which the execution of the Zachmanntest is aborted. The value loop_cnt (default: 1,000,000) defines a numerical value for how often the internal table should be cycled. By executing the entire Zachmanntest, one instance of the test executable is instantiated. The Zachmanntest produces a heavy main memory load on the application server.

3.2.2 Performance Metric

The Zachmanntest is meant to quantify the performance of the underlying main memory system from a SAP perspective. Generally, there are several performance metrics available, e.g. response time metrics or throughput metrics. The performance metric of the Zachmanntest is throughput, measured in rows per seconds. For example, after finishing one run of one Zachmanntest, the throughput of the SAP ERP system results in about 9,000 rows per second. This metric is to be interpreted as follows: in the case of one instantiated benchmark in the SAP ERP system, approx. 9,000 rows per second can be accessed for this benchmark instance. When handling two benchmark instances at the same time (we refer to them as two Zachmanntests) the throughput might be less or equal. This is because the maximum available throughput will be shared between both Zachmanntests.

The throughput metric is the best metric for the purpose of our simulation, as it can be easily applied to the simulation model. The throughput is expressed in a very simple numerical only way. Thus it can be applied to our simulation without the need of any transformation or mathematical operation.

4 MODELING THE PERFORMANCE USING EVOLUTIONARY ALGORITHMS

The next step after measuring the performance of the ERP system using the Zachmanntest is to make the measured data usable for performance and scalability prediction. For this, the measured data has to be transformed into a mathematical model. For multi-dimensional data, an exact solution becomes very complex in terms of the model size and solution determination, making it unusable for simulation approaches. Furthermore there is no guarantee that exactly one optimal model exists for the measured performance data – several Pareto-optimal solutions might be possible (Zitzler and Thiele, 1999) when factors like the model length and evaluation time are considered. To limit the maximum model size, as well as to reduce the time for solution determination, an evolutionary algorithm is used to approximately model any given set of performance data.

4.1 Description of the Evolutionary Algorithm Approach

The basic idea behind any evolutionary algorithm is the imitation of Darwin's idea of natural evolution. The best individuals or genomes of a generation survive and reproduce. Hence, an evolutionary algorithm is a random search method performing multi-criteria optimization on an n-dimensional search area. The algorithm consists of multiple individuals, competing on a limited resource. The algorithm performs several iterations, each resulting in a new generation of individuals. A fitness function is used to determine every individual's fitness, resulting in the decision if an individual is allowed to pass its genome to the next generation or not. Mutation and crossover is performed whenever a genome is passed to a new generation's individual, allowing moving or jumping in the search area.

In our actual prototype Mendel (named after the researcher Gregor Johann Mendel), the limited resource is the fixed size of individuals and the rule that 50% of the individuals are passed to the next generation, while new individuals replace the other 50%. The fitness of an individual is defined by the negative geometrical distance of the generated model from the underlying measured performance data. An error value s_{Err} is calculated as defined in formula 1, with $r_{measured_i}$ being the i[th] measured value, $r_{modeled_i}$ the ith modeled value, and n the number of measured performance values. Simply saying, an individual is fitter than another if its model fits closer to the measured data (i.e. the model has a smaller error value s_{Err}).

$$s_{Err} = \frac{\sum_{i=0}^{n} \frac{\left|r_{measured_i} - r_{modeled_i}\right|}{r_{measured_i}}}{n}$$

4.2 Model Representation and Mathematical Operators

The model of an individual is stored in a genome structure. Every odd element in the genome is either a fixed number, or a parameter, and every even element is an operator. The genome is interpreted from right to left, assuming a right bracketing.

Figure 1 is a visualization of the exemplary model $a + \frac{x}{b^{y-\sin(c*z)}}$ with *a*, *b*, *c* being fixed numbers and *x, y, z* parameters.

A	+	*x*	/	*b*	^	*y*	- *sin*	*c*	*	*z*

Figure 1: Genome coding of an exemplary model.

4.3 Configuration of the Evolutionary Algorithm

The performance and efficiency of the evolutionary algorithm is strongly dependent on its configuration (Zitzler/Thiele, 1999). Commonly used configuration parameters are shown in Table 1.

Table 1: Configuration parameters of the evolutionary algorithm.

Parameter	Description
Population Size	The number of individuals. Larger population size results in higher model variance, but also increases the resource usage per iteration.
Genome Length	The length of the genome. Longer genomes result in more complex models.
Mutation Probability	The probability for mutation when a model is passed to the next generation.
Crossover Probability	The probability for crossover when a model is passed to the next generation.

For identifying the optimal configuration for modeling the given performance data we carried out five calculations for every combination of configuration parameters, interrupted the evolutionary algorithm after five minutes, and compared the resulting models. As the evolutionary algorithm is a non-deterministic algorithm, we

compared the median value of the five calculations per configuration.

4.3.1 Population Size and Genome Length

Population size defines the number of parallel threads that are used for modeling, while the genome length defines the length of the model. Both parameters are correlated, as they both affect the resource usage of the evolutionary algorithm. A bigger population requires to evaluate and pass more models per iteration, while the genome length determines the required CPU cycles to evaluate and the memory to store the model.

To get an indication for an appropriate population size range we performed the modeling with 100, 1,000, 5,000, 10,000 and 20,000 individuals. The results of this first iteration showed that a population size bigger than 5,000 does not provide usable results on the given hardware configuration.

The same ranging was done for the genome length. Modeling was performed for 11, 21, 41 and 201 genome length, showing that a genome longer than 41 elements is not performing in the given context. Figure 2 depicts the average modeling error for all combinations of population size and genome length.

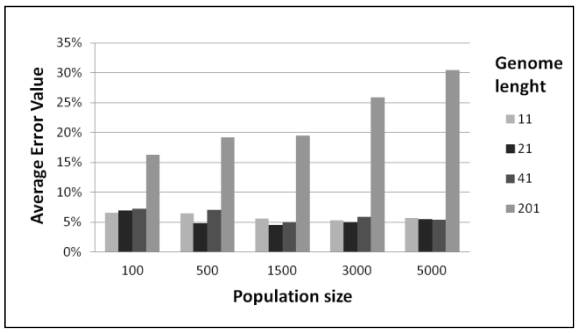


Figure 2: Effect of population size and genome length on modeling accuracy.

Higher population size results in more modeling variation, which again results in a higher chance of the model converging to the measured data. The optimal population size though is determined by the number of available CPUs. Too big populations (in our case > 3,000 individuals) result in increased wait times, reducing the efficiency of the algorithm.

From the data presented in the diagram it is obvious that a too long genome also reduces the modeling accuracy. On the one hand this inaccuracy is caused by a reduced number of iterations performed in the given timeframe due to an increased resource need for the model evaluation. On the other hand an analysis of intermediate result revealed that with a long genome mutation becomes inefficient. In every iteration mutation changes one genome element. However, the longer a genome is the higher is the chance that it contains elements with small effects. Hence the possibility of mutations advancing the model noticeably is decreased. Short genomes though reduce the model flexibility, inhibiting the approximation of complex measured data. For the given ERP data a population size of 1,500 or 3,000, and a genome length of 21 proved to return the best results.

4.3.2 Mutation and Crossover Probability

Mutation and crossover, as defined by Goldberg (1989), build the random searching operations of the evolutionary algorithm. Both operations are performed with a given probability when a model is passed to a new generation. To determine the effect of the mutation and crossover probability the average error value is compared for each combination of mutation and crossover probability. Figure 3 shows all the combinations resulting in an average modeling error value of less than five percent.

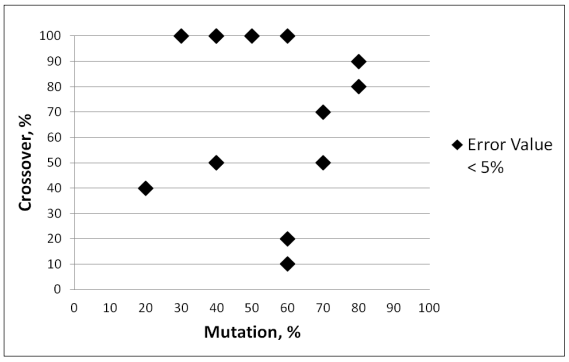


Figure 3: Effect of mutation and crossover probability on modeling accuracy.

It is obvious that high crossover or mutation probability leads to accurate models. Zero or small mutation probability (< 20%) avoids convergence towards an optimum, while zero or small crossover probability restricts the jumping in the search area, forces the algorithm to getting caught in a local optimum.

4.4 Modeling Results

Given the correct configuration, the evolutionary algorithm results in models approximating very

close the given scalability data. In our case study the model fits to the given data with an error smaller than four percent.

Figure 4 visualizes the modeled scalability data compared to the measured data for an ERP system configured with 12 work processes. It is visible that the model comes very close to the measured data. Providing the presented model the evolutionary algorithm achieved an error of less than 0.7 percent, in a modeling time of five minutes.

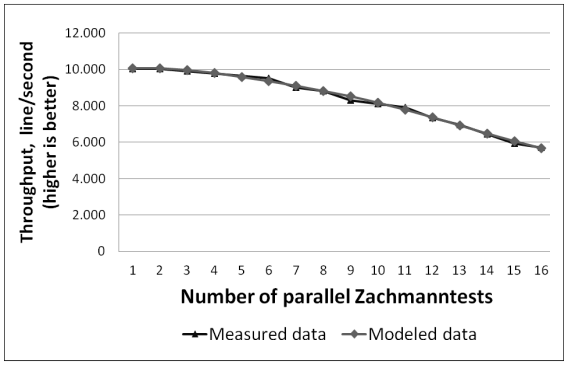


Figure 4: Comparison of measured and modeled data for 12 workflow processes.

Table 2 shows the error values (EV, in percent) of all work process (WP) configurations. For each configuration, modeling was performed for exactly five minutes.

Table 2: Modeling error values for all measured work process configurations.

WP	6	7	8	9	10	11	12	13	14
EV	2.2	3.2	2.8	0.9	1.7	1.2	0.7	2.0	1.4

Compared to other works in the field of evolutionary algorithms (see (Tikir et al. 2007) for example), our reached error values are very low. Hence, we rate our gained error values as very good ones.

5 CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

This paper presents our black box approach to create a simulation model, which is based on an evolutionary algorithm and real world performance data from a SAP ERP system. The presented results show that, given the correct configuration, evolutionary algorithms perform well in modeling scalability data of ERP systems with an error value under 3.2%. The modeling error of approximately two percent is less than the assumed measurement error, and thus acceptable. A negative side of the non-determinism of the evolutionary algorithm is that an acceptable model is only found in approximately ninety percent of all modeling runs in an acceptable time, while in the other cases the algorithm takes hours to result in a usable model. This effect is independent on the given performance data but results from the random model generation and mutation. We neglected the effect by setting a timeout, after which the algorithm was restarted.

One of the biggest benefits of using the evolutionary algorithm proved to be its ability to model any kind of data without being adopted. This characteristic allows the modeling of multiple sets of data automatically without any manual effort, and allows the integration of the algorithm into an automatic scalability and performance prediction framework, bridging for example from the measured scalability and performance data to the simulation engine.

5.2 Future Work

This paper shows how to use the black box approach for modeling a very complex SAP ERP system in a first step. However, such a software system must be modeled in a more detailed way. Thus our goal is to extend the simulation model with more components and to use real life monitoring data to establish an evolutionary algorithm, which is able to reproduce the exact performance behavior of the entire system.

Evolutionary algorithms as implemented in our prototype Mendel, suite well in modeling the performance and scalability data when the data is equally distributed. When an equal distribution is not given, the used fitness function might result in a model not representing properly the scalability of the ERP system. This might be the case if, for example, a big data set is available for low load, but only few data for high load. Then a well matching model for all the low load data, not matching the high load data, might result in a good fitness value. This effect will be neglected by implementing clustering of the scalability data and solving each cluster on its own.

Future work will also be to identify the optimal configurations of the evolutionary algorithm for different usage scenarios. As presented in this paper the configuration strongly affects the modeling error and the time the algorithm needs to finish.

REFERENCES

Bögelsack, A., Jehle, H., Wittges, H., Schmidl, J., Krcmar, H., 2008. An Approach to Simulate Enterprise Resource Plannung Systems. *6th International Workshop on Modelling, Simulation, Verification and Validation of Enterprise Information Systems.* Barcelona, Spain.

Curnow, H., Wichmann, B., 1976. A synthetic benchmark. In: *The Computer Journal*, Vol. 19 No. 1, pp. 43.

Doppelhammer, J., Höppler, T., Kemper, A., Kossmann, D., 1997. Database performance in the real world: TPC-D and SAP R/3. *Proceedings of the 1997 ACM SIGMOD international conference on Management of data.* Tucson, Arizona, United States: ACM.

Goldberg, D. E., 1989. *Genetic algorithms in search, optimization, and machine learning,* Addison-Wesley Professional, Upper Saddle River, NJ, USA.

Gradl, S., Bögelsack, A., Wittges, H., Krcmar, H., 2009. Layered Queuing Networks for Simulating Enterprise Resource Planning Systems. *6th International Workshop on Modelling, Simulation, Verification and Validation of Enterprise Information Systems*. Milano, Italy.

Herbert, E., Dowsland, K., 1996. A family of genetic algorithms for the pallet loading problem. In: *Annals of Operations Research*, Vol. 63 No. 3, pp. 415-436.

Jain, R., 1991. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling: Techniques for Experimental Design, Measurement, Simulation and Modelling,* John Wiley & Sons, Inc.

Justesen, P. D., 2009. *Multi-objective Optimization using Evolutionary Algorithms.* Department of Computer Science, University of Aarhus.

Law, A. M., 2008. How to build valid and credible simulation models. *Proceedings of the 40th Conference on Winter Simulation (pp. 39-47)*. Miami, Florida: Winter Simulation Conference.

Rolia, J., Casale, G., Krishnamurthy, D., Dawson, S., Kraft, S., 2009. Predictive modelling of SAP ERP applications: challenges and solutions. *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools (pp. 1-9)*. Pisa, Italy: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Rolia, J., Kalbasi, A., Krishnamurthy, D., Dawson, S., 2010. Resource demand modeling for multi-tier services. *Proceedings of the first joint WOSP/SIPEW international conference on Performance engineering (pp. 207-216)*. San Jose, California, USA: ACM.

SAP, 2010. *SAP Standard Application Benchmarks*. http://www.sap.com/solutions/benchmark/index.epx, accessed at 12.3.2010.

Tikir, M. M., Carrington, L., Strohmaier, E., Snavely, A., 2007. A genetic algorithms approach to modeling the performance of memory-bound computations. *Proceedings of the 2007 ACM/IEEE conference on Supercomputing (pp. 1-12)*. Reno, Nevada: ACM.

Zitzler, E., Thiele, L., 1999. Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach. In: *IEEE Transactions on Evolutionary Computation*, Vol. 3 No. 4, pp. 257 - 271.

Modeling Structural, Temporal and Behavioral Features of a Real-Time Database

Nada Louati[1], Rafik Bouaziz[1], Claude Duvallet[2] and Bruno Sadeg[2]
[1]*MIRACL-ISIMS, Sfax University, BP 1088, 3018, Sfax, Tunisia*
[2]*LITIS, UFR des Sciences et Techniques, BP 540, 76 058, Le Havre Cedex, France*
{*nada.louati, raf.bouaziz*}*@fsegs.rnu.tn*, {*claude.duvallet, bruno.sadeg*}*@univ-lehavre.fr*

Keywords: Real-time, Database, MARTE, UML-RTDB, Profile.

Abstract: Real-time databases are different from traditional databases in that they have timing constraints on data and on transactions upon the data. The design of this kind of databases must consider both temporal aspects of data and timing constraint of transactions in addition to the logical constraints of the database. This paper proposes a new UML profile that extends UML with concepts related to real-time databases design. These extensions aim to accomplish the conceptual modeling of a real-time database according to three aspects: structural, temporal and behavioral. Our propositions is based on MARTE (Modeling and Analysis of Real-Time and Embedded systems) profile which provides capabilities of modeling concepts to deal with real-time and embedded system features. The proposed profile is illustrated by a case study in the context of Ait Traffic Control System.

1 INTRODUCTION

Real-Time DataBases (RTDBs) are typically used to manage environmental data in computer control applications, such as air traffic control, automated manufacturing, and military command and control (Ramamritham, 1993). The design of such RTDBs differs from both that of real-time systems and that of conventional database management systems. The designers of RTDBs much consider both temporal aspects of data and timing constraints of transactions in addition to logical constraints of the database

The design of RTDBs is performance-and semantic-dependent. It must consider factors such as sensor data, derived data and quality of data management, temporal semantics in transactions scheduling algorithms and concurrency control protocols, to meet the timing constraints defined by the real-time applications (Idoudi et al., 2010). This paper proposes a new UML profile that incorporates these concepts. It is based on a subset of concepts inspired from the HLAM, NFP, and TIME packages of MARTE. The motivations behind these extensions are three-folds: (i) to have new notations distinguishing quantitative and qualitative features of RTDBs, (ii) to facilitate the modeling of timing aspects of data and transactions, (iii) to accomplish the conceptual modeling of RTDBs. This profile not only captures the structural aspects of a RTDB features, but also the temporal and behavioral aspects.

The remainder of this paper is structured as follows: in Section 2, we present the related works. In Section 3, 4, and 5 we respectively address the modeling of structural, temporal, and behavioral aspects of RTDBs. In Section 6, we present a case study to more illustrate our proposals. In Section 7 we conclude the paper and we give some perspectives of our work.

2 RELATED WORK

In recent years, several UML profiles have been proposed for modeling real-time systems to depict its real-time constraints. Only a few of these UML profiles address RTDBs modeling.

In (DiPippo and Ma, 2000), DiPippo and Ma describe an UML package, called RT-Object, for specifying RTSORAC object. This UML package contains real-time attributes, real-time methods, compatibility functions and constraints. The RT-Object package is based on a past standard of UML which is UML 1.3 Extension Mechanisms package. Furthermore, imprecise computation encapsulated within the RTSORAC object model is described in the context of Epsilon Serializability (Ramamritham and Pu, 1995),

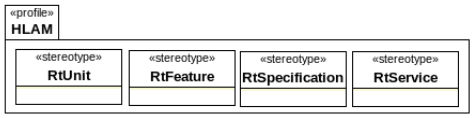


Figure 1: HLAM package specification in MARTE.

and does not support the notion of quality of data introduced in (Amirijoo et al., 2006).

Another work proposed the UML-RTDB profile (Idoudi et al., 2008a) (Idoudi et al., 2008b) to express RTDB features in a structural model. It supplies concepts for RTDBs modeling such as real-time attributes, real-time methods and real-time classes. However, the proposed extensions are based on the Evolutionary Stereotype concept (Debnath et al., 2003) which is not a standard way of extending UML.

MARTE (OMG, 2009) is an industry standard of the OMG for model-driven development of embedded systems (OMG, 2009). It aims at providing support for specification, design, validation, and verification stages in real-time and embedded system development. The richness of MARTE profile in terms of concepts offers an interesting common modeling basis to adequately specify many design features of RTDBs. However, the notions of real-time data and real-time transactions, which are the basis features of RTDBs are missing in MARTE.

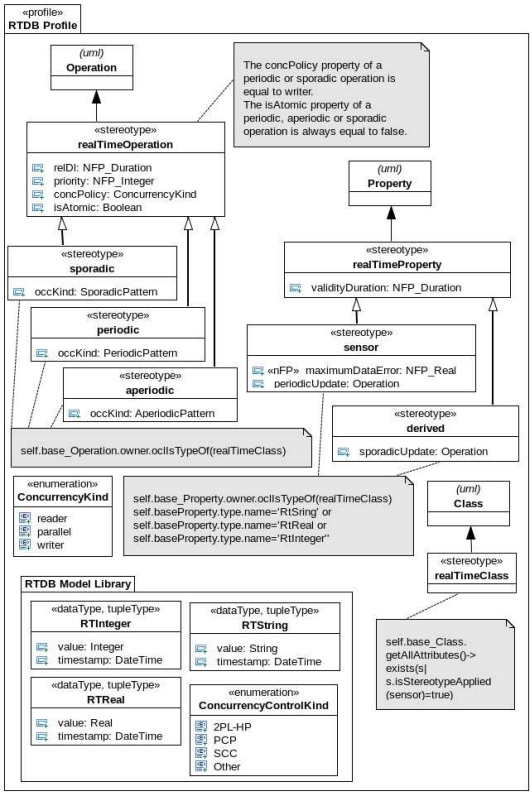


Figure 2: RTDB Profile Metamodel.

3 MODELING OF STRUCTURAL ASPECTS

3.1 High Level Application Modeling in MARTE

The HLAM package of MARTE proposes concepts that enable to describe both quantitative and qualitative features of real-time applications at a high abstraction level. Figure 1 depicts the basic stereotypes associated with the HLAM package. An *RtUnit* may be seen as an autonomous execution resource that owns one or more schedulable resources to handle incoming messages. *RtUnits* may provide real-time services. The *RtService* stereotype has been introduced for that purpose. The *RtFeature* stereotype is used to annotate model elements with real-time features according to set of *RtSpecification* associated with this stereotype. It can be attached to multiple kinds of modeling elements (i.e., behavioral features, actions, messages, signals, and ports). For a full description of all these stereotypes, the reader may refer to the specification document of MARTE (OMG, 2009).

3.2 Modeling of RTDB Structural Features

We propose new stereotypes expressing RTDB features in a structural model on the one hand, and according to the MARTE profile on the other hand. Figure 2 shows the extensions proposed to some metaclasses belonging to the class diagram. In order to take advantages of some MARTE concepts, the proposed profile imports stereotypes from HLAM, NFP, and VSL sub-profiles.

3.2.1 Stereotypes for Real-Time Data

Real-time attributes are divided into two types: *sensor attributes* and *derived attributes* (Ramamritham et al., 2004). Sensor attributes are used to store sensor data which are issued from sensors. Derived attributes are used to store derived data which are calculated from sensor data. Thus, we define two stereotypes, *sensor* and *derived*, in order to declare respectively sensor and derived attributes in the class diagram. As illustrated in Figure 2, we define an abstract stereotype, called *realTimeProperty*, that factorizes *validity duration* property which indicates the amount of time

during which the attribute value is considered valid. We use the *MARTE NFP_Duration* datatype as a type for the *validity duration* feature.

We characterize each *sensor* stereotype by two properties : *maximum data error* and *periodicUpdate*. *Maximum data error* indicates the maximum deviation tolerated between the current attribute value and the updated value. It is of type *NFP_Real*. *Maximum data error* represents a non-functional property specifying the upper bound of the error. We propose to associate the *Nfp* stereotype of MARTE profile to the *maximum data error* attribute. *PeriodicUpdate* refers to a periodic operation of the class that owns the sensor attribute. The role of this periodic operation is to update the current *value* and the *timeStamp* fields of the *sensor* attribute. It is of type *Operation*.

As shown in Figure 2, we characterize *derived* stereotype by *sporadicUpdate* property, that refers to a sporadic operation of the class that owns the derived attribute. This sporadic operation is used to update the *current value* and the *timeStamp* fields of the *derived* attribute.

Each real-time attribute value is characterized by a *timestamp*, which indicates the time at which it was last updated. So, for each real-time attribute value corresponds a *timestamp*, which distinguishes it from other attribute's values. Whereas, the values of the *validity duration* and *maximum data error* fields are the same for the same real-time attribute.

3.2.2 Stereotypes for Real-Time Transactions

A real-time transactions may be aperiodic, periodic, or sporadic (Ramamritham et al., 2004). It has timing constraints such as deadline and period. We define three stereotypes, aperiodic, periodic, and sporadic (cf. Figure 2) in order to declare respectively aperiodic, periodic, and sporadic operations in the class diagram. Each of these stereotypes is characterized by a *deadline*, which indicates the last time by which the method execution must be completed. Thereby, we define an abstract stereotype, called *realTimeOperation*, which is used to annotate model elements (i.e. Operations) with real-time features according to set of RtSpecification associated with this stereotype. RtSpecification possesses nine tagged values among which: relDl, occKind, and priority. The relDl attribute specifies the deadline of a method execution. The occKind attribute indicates the arrival transaction specification (i.e. periodic, aperiodic, or sporadic). For a periodic (respectively aperiodic and sporadic) transaction, the PeriodicPattern (respectively AperiodicPattern and SporadicPattern) enumeration literal is selected for ArrivalPattern attribute of the occKind property. The priority attribute specifies the priority

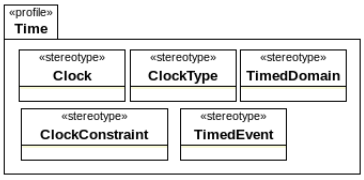


Figure 3: Time Specification in MARTE.

order of a transaction. The *realTimeOperation* stereotype factorizes also two properties: concPolicy and isAtomic. When the value of *isAtomic* is *true*, the method execution is done as one individual unit. This fact coincides with the *atomicity* property of a transaction. However, in RTDBs, this feature is relaxed in order to allow the validation of a transaction even if only a part of its actions have been executed (Agrawal et al., 1994). So, in our work, we consider that the value of the *isAtomic* property is always *false*. The *concPolicy* property specifies the concurrency policy of a transaction. The values of this property may be: *reader*, *writer* or *parallel*. A *writer* transaction implies that multiple calls from concurrent transactions may occur simultaneously and will be treated as soon as concurrency on data allows its execution. A *reader* transaction implies that multiple calls from concurrent transactions may occur simultaneously and will be executed simultaneously if there is no writer transaction using one or more data that the *reader* transaction needs. A *parallel* transaction is a transaction whose actions do not use any data of the database in reading mode nor in writing mode. In our work, we consider that for an update transaction, which can periodic or sporadic, the *concPolicy* property is *writer*.

3.2.3 Real-Time Class Stereotype

The design of a RTDB, which is by definition a database system, has to take into account the management of many components such as queries, schemas, transactions, commit protocols, concurrency control protocol, and storage management (Stankovic et al., 1999). In order to deal with time-constrained data, time-constrained operations, parallelism, and concurrency property inherent to RTDBs, we introduced the *RealTimeClass* stereotype. This stereotype specifies that instances of a class will encapsulate real-time data and real-time operations and a local concurrency mechanism. Because of the dynamic nature of the real world, more than one transaction may send requests to the same object. Concurrent execution of these transactions allows several methods to run concurrently within the same object. To handle this essential property of RTDB systems, we associate to each object a local concurrency control mechanism, named *local controller*, that manages the concurrent execu-

tion of its methods. Thus, the object receives messages in its mailbox awaking its local controller that checks the timing constraint attached to messages and selects one message following a special scheduling algorithm. The local controller verifies the concurrency constraints with the already running methods of the object. Then, it allocates a new thread to handle the message when possible. When a method terminates its execution, the corresponding thread is released and concurrency constraints are relaxed. In our work, the scheduling algorithm adopted by the mailbox is EDF (EarliestDeadlineFirst). For that purpose, we import from MARTE library the *SchedPolicyKind* enumeration which defines the most common kind of scheduling policy. Add to that, the concurrency control protocol adopted by the local controller is 2PL-HP (Two Phase Locking-High Priority). In 2PL-HP, all data conflicts are immediately resolved by aborting lower priority transactions. Thereby, we enhance our profile model library by a new enumeration type, called *ConcurrencyControlKind*, in order to define concurrency control policies (i.e. 2PL-HP, PCP (Priority Ceiling Protocol), SCC (Speculative Concurrency Control), etc.). Our *RealTimeClass* stereotype overlaps with the UML extensions presented by MARTE profile especially those relative to the mailbox and the local controller. In fact, a *RealTimeClass* may be considered as an autonomous execution resource, able to handle different messages at the same time. It can manage concurrency and real-time constraints attached to incoming messages. This has the same meaning as the *RTUnit* stereotype defined in MARTE (cf. Figure 1). Thus, we consider that the *RealTimeClass* concept as a specialization of MARTE *RtUnit* concept. Thereby, a *RealTimeClass* is a *real-time unit* but with a RTDB modeling semantics. It represents the RTDB entity which encapsulates time-constrained data and time-constrained operations. It also deals with parallelism and concurrency features.

4 MODELING OF TEMPORAL ASPECTS

4.1 Time Specification in MARTE

The sub-profile TIME of MARTE has been introduced to model timing aspects. Figure 3 depicts the MARTE extension about time modeling in UML. The *Clock* stereotype is a model element that represents an instance of *ClockType*. The *ClockType* stereotype is related indirectly to the *Clock* stereotype. Its properties specify the kind of clock (chronometric or logical), the nature (dense or discrete) of the represented time, a set of clock properties (e.g., resolution, maximal value, etc.) and a set of accepted time units. The *ClockConstraint* is a stereotype of the UML *Constraint* concept. The clock constraints are used to specify the time structure relations of a time domain. A *ClockConstraint* is a constraint that imposes dependency between clocks or between clock types.

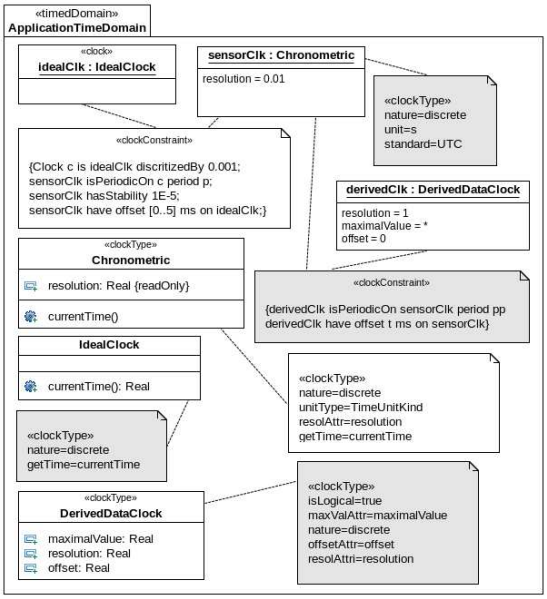


Figure 4: Clocks Specification in our RTDB Profile.

4.2 Modeling of RTDB Temporal Properties

Time is present in different RTDB features and more specially in real-time data. Real-time data are divided into two classes: sensor data and derived data. As previously mentioned in Section 3.2.1, we have defined for each sensor data a periodic method (i.e. *periodicUpdate*) which is periodically executed to update the values of the corresponding sensor data. In the same way, we have defined for each derived data a sporadic method (i.e. *sporadicUpdate*) which is sporadically executed to update the values of the corresponding derived data. In this section, we define for the sensor and derived data clocks as well as their associated properties in a high level specification using the TIME sub-profile of MARTE. The temporal unit that we are looking for can be used in two ways: (i) to reference the physical time and adopt a chronometric clock for sensor data, (ii) to reference a logical time that is incremented each time sensor data have been updated.

The MARTE *TimeLibrary* provides a model for the *ideal time* used in physical laws: *idealClk*, which

is an instance of the class *IdealClock*, stereotyped by *ClockType*. The *IdealClock* represents the time evolution. This time is expressed in seconds in the international system units. Starting with *idealClk*, we define new discrete chronometric clock (cf. Figure 4). First, we specifies *Chronometric* (a class stereotyped by *ClockType*) which is discrete, not logical (therefore chronometric), and with a read only attribute (resolution). Clocks belong to timed domains. In Figure 4, a single time domain is considered. It owns two clocks: *idealClk* and *sensorClk*, an instance of *Chronometric* that both uses the second (s) as a time unit; and whose resolution is 0.01 s. The two clocks are a priori independent. A clock constraint specifies relationships among them. The first statement of the constraint defines a clock c local to the constraint. C is a discrete time clock derived from i*dealClk* by a discretization relation. The resolution of this clock is 1 ms. The next statement specify that *sensorClk* is subclock of c with a rate *p* times slower than c. The fourth statement indicate that *sensorClk* is not a perfect clock. Flaws are characterized by non functional properties like stability and offset. Its rate may have small variations (a stability of 10^{-5} implicitly measured on *idealClk*). The last statement claims that the clocks is out of phase, with an offset value between 0 and 5 ms measured on *idealClk*.

Figure 4 depicts the definition of a logical clock dedicated for derived data. We have created a class *DerivedDataClock* with *ClockType* as stereotype.This class has three attributes: *maximalValue*, *offset* and *resolution*. *MaximalValue* specifies the maximal value of the associated clock, value at which the clock rolls over. The *offSet* property determines the initial instant of the associated clock. The *resolution* attribute defines the resolution of the associated clock. Now that we have defined the clock, we need to instantiate it. Figure 4 depicts the instantiation of the clock *DerivedDataClock* called *derivedClk*. The unit of *DerivedClk* is the sensor data update.

Sensor data are periodically updated causing the update of each derived data that uses those sensor data. This is equivalent to say that the derived data update operation is activated every tick of the clock associated with sensor data. We can deduce an affine relation between *sensorClk* and *derivedClk* as specified in the clock constraint specified in Figure 4: *derivedClk isPeriodicOn sensorClk*, period = pp, offset=t. Such a constraint states that each pp^{th} occurrence of *sensorClk* there will be an occurrence of *derivedClk* and the first occurrence of *derivedClk* appears at the t^{th} instant of *sensorClk*.

5 MODELING OF BEHAVIORAL ASPECTS

5.1 Behavior Specification in MARTE

State machines are often used in the real-time and embedded domain. They allow the description of a system behavior in terms of states and transitions between these states. MARTE introduces mainly two time-related concepts that can be used to improve the usage of the UML behaviors, such as state machines, for developing real-time applications: timed processing and timed events (cf. Figure 3). The *TimedProcessing* stereotype enables modelers to specify duration for a behavior. The *TimedProcessing* stereotype represents activities that have known start and finish times or a known duration, and whose instants and durations are explicitly bound to clocks. It can also reference events triggered when a behavior or processing starts and ends. The duration can be specified using the VSL language, which supports time expressions.

The *TimedEvent* stereotype represents events whose occurrences are explicitly bound to a Clock. It extends the *TimeEvent* concept of UML. It allows to characterize the logical or physical clock on which a time event relates.

5.2 Modeling of RTDB Behavior

A RTDB is a collection of real-time objects which are used to manage time-critical dynamic systems in the real world. RTDB behavior model building consist of describing the behavior with state-transition diagram for each real-time object of the RTDB. Stereotypes *TimedProcessing* and *TimedEvent* are used to specify behavior, duration of a behavior, and events bounded to a clock. The tagged value *queueSchedPolicy*, defined in MARTE, are associated to the state-transition diagram in order to indicate the queue scheduling policy of a behavior. We apply the *TimedProcessing* stereotype on the state machine itself and we specify the corresponding clock using the meta-attribute *on*. We employ the *TimedEvent* stereotype in order to characterize the logical and physical clocks on which a time event relates.

6 CASE STUDY

Throughout, this section, we will use a running example to illustrate the use of our profile. We illustrate our proposal on an air traffic control system. The aim of the air traffic control is to separate aircrafts,

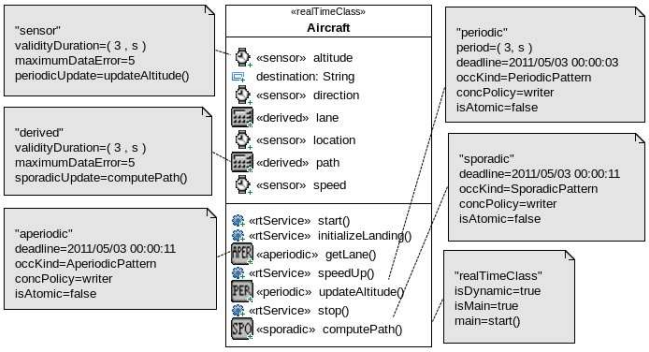


Figure 5: Aircaft real-time class.

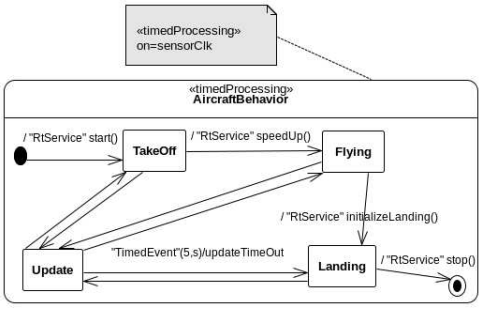


Figure 6: Aircraft behavior modeling.

to avoid collisions and to organize the flow of traffic. It consists of a large collection of data describing the aircrafts, their flight plans, and data that reflect the current state of the controlled environment (Locke, 2001). This includes flight information, such as aircraft identification, speed, altitude, origin, destination, route and clearances. In fact, each aircraft has three sensor data which are speed, altitude and location and two derived data which are path and lane. Sensor data are periodically updated to reflect the state of an aircraft. The derived data is calculated based on altitude, location, and direction values, in order to verify if the aircraft deviates from a predetermined path or lane. All these data values are published periodically by sensors supervising the aircrafts controlled elements.

For sake of clarity, only the *Aircraft* class and a subset of its methods are specified. Each *Aircraft* in the airspace is stereotyped by *realTimeClass*. We characterize the *Aircraft* class by three sensor attributes: direction, location, and altitude. Each attribute is periodically updated in order to closely reflect the real world state of the application environment. Thereby, we associate to each sensor attribute a periodic method: updateAltitude(), which periodically updates the value and the timestamp of the altitude. We characterize also the Aircraft class by two derived attributes: lane which is calculated from location and altitude values, and path, which is calculated from location and direction values. Each derived attribute has its own sporadic method: computePath(), which sporadically updates the value and the timestamp of the path. Figure 5 depicts a simplified view of the Aircraft class. We indicates that the Aircraft class is the main unit of the application (property *isMain* is set to true) and it starts a main *rtService*, called *start()*. Its *isDynamic* property is set to true. In this case, the schedulable resources are created dynamically when required. The attribute *altitude* has a validity duration equal to (*3, s*) and its maximumDataError is 5. Additionally, it is updated periodically using the *updateAltitude()* operation, which has a period equal to (*3, s*). The *concPolicy* attribute of that operation is *writer* and its *isAtomic* attribute is set to *false*. In the same way, we characterize the derived attribute path by a *validityDuration* and a *maximumDataError* and it is sporadically updated by means of the *computePath()* operation.

Figure 6 depicts a state-machine diagram that provides a simplified view of the *Aircraft* behavior. The Aircraft has four states: TakeOff, Flying, Update, and Landing. Periodically , it enters in the Update state for updating the Aircraft sensors values. The sensors values update has to be done with a period of 5 s and lasts 2 s. Then it returns to the state which activated the update transition. We apply the TimedProcessing stereotype on the state machine itself. We use the former to indicate the scheduling policy associated with the Aircraft behavior (i.e EDF). The latter is used to assign the sensorClk clock to the model. All the elements of the state machine have the same time reference. We apply the TimedEvent stereotype on the UML timed event that triggers the UpdateTimeOut transition.

7 CONCLUSIONS

The richness of MARTE profile in terms of concepts offers an interesting common modeling basis to adequately specify many design features of RTDBs. In This paper, we showed how the concepts of the MARTE standard profile can serve to model RTDB features. We used the HLAM package to describe both quantitative and qualitative properties, and the TIME sub-profile to specify temporal properties. Moreover, we proposed UML/MARTE-based extensions for a complete and powerful RTDB modeling. Additionally, the proposed extensions not only capture the structural aspects of RTDB features, but also the temporal and behavioral aspects. We have invested some effort on ensuring that all concepts have a well defined basis semantics. This has been illustrated on a case study in the context of Air Traffic Control System.

We are currently working on the the integration of the RTDB development process in the context of a Model Driven Architecture. This way, the complex task of designing the whole RTDB is tackled in a systematic, well-structured and standard manner.

REFERENCES

Agrawal, D., Bruno, J. L., Abbadi, A. E., Krishnaswamy, V., El, A., and Krishnaswamy, A. V. (1994). Relative serializability: An approach for relaxing the atomicity of transactions. In *In Proceedings of the 13th ACM Symposium on Principles of Database Systems*, pages 139–149. ACM.

Amirijoo, M., Hansson, J., and Son, S. H. (2006). Specification and management of qos in real-time databases supporting imprecise computations. *IEEE Trans. Computers*, 55(3):304–319.

Debnath, N., Riesco, D., Montejano, G., Grumelli, A., Maccio, A., and Martellotto, P. (2003). Definition of a new kind of uml stereotype based on omg metamodel. In *Computer Systems and Applications, 2003. Book of Abstracts. ACS/IEEE International Conference on*, pages 49–54.

DiPippo, L. C. and Ma, L. (2000). A uml package for specifying real-time objects. *Comput. Stand. Interfaces*, 22:307–321.

Idoudi, N., Duvallet, C., Sadeg, B., Bouaziz, R., and Gargouri, F. (2008a). Structural model of real-time databases. In *ICEIS (3-2)*, pages 319–324.

Idoudi, N., Duvallet, C., Sadeg, B., Bouaziz, R., and Gargouri, F. (2008b). Structural model of real-time databases: An illustration. In *ISORC*, pages 58–65.

Idoudi, N., Louati, N., Duvallet, C., Sadeg, B., Bouaziz, R., and Gargouri, F. (2010). A framework to model real-time databases. *International Journal of Computing and Information Sciences (IJCIS)*, 7(1):1–11.

Locke, C. D. (2001). Applications and system characteristics. In *Real-Time Database Systems*, pages 17–26.

OMG (2009). A UML Profile for MARTE: Modeling and Analysis of Real-Time Embedded systems, version 1.0, formal/2009-11-02.

Ramamritham, K. (1993). Real-time databases. *Distributed and Parallel Databases*, 1(2):199–226.

Ramamritham, K. and Pu, C. (1995). A formal characterization of epsilon serializability. *IEEE Trans. on Knowl. and Data Eng.*, 7:997–1007.

Ramamritham, K., Son, S. H., and DiPippo, L. C. (2004). Real-time databases and data services. *Real-Time Systems*, 28(2-3):179–215.

Stankovic, J. A., Son, S. H., and Hansson, J. (1999). Misconceptions about real-time databases. *IEEE Computer*, 32(6):29–36.

Database Schema Elicitation to Modernize Relational Databases

Ricardo Pérez-Castillo[1], Ignacio García Rodríguez de Guzmán[1], Danilo Caivano[2] and Mario Piattini[1]
[1]*Instituto de Tecnologías y Sistemas de Información (ITSI), University of Castilla-La Mancha, Paseo de la Universidad 4, 13071, Ciudad Real, Spain*
[2]*Department of Informatics, University of Bari, Via E. Orabona, 4, 70126 Bari, Italy*
{ricardo.pdelcastillo, ignacio.grodriguez, mario.piattini}@uclm.es, caivano@di.uniba.it

Keywords: Database Modernization, Legacy Systems, ADM, KDM, SQL, Metamodel, Model Transformations, QVT.

Abstract: Legacy enterprise systems mainly consist of two kinds of artefacts: source code and databases. Typically, the maintenance of those artefacts is carried out through re-engineering processes in isolated manners. However, for a more effective maintenance of the whole system both should be analysed and evolved jointly according to ADM (Architecture-Driven Modernization) approach. Thus, the ROI and the lifespan of the legacy system are expected to improve. In this sense, this paper proposes the schema elicitation technique for recovering the relational database schema that is minimally used by the source code. For this purpose, the technique analyses database queries embedded in the legacy source code in order to remove the dead parts of the database schema. Also, this proposal has been validated throughout a real-life case study.

1 INTRODUCTION

Today, many organizations have huge legacy systems supported by relational databases (Blaha, 2001), and these systems are not immune to software ageing (Visaggio, 2001). Nevertheless, the erosion not only affects the source code, but also databases age gradually. For instance, in order to adapt the system to new requirements, new tables and/or columns are added to the schema of the database; other tables are modified and even discarded without erasing them from the database; and so on. These changes over time generate problems related to inconsistency, redundancy and integrity among others.

Therefore, organizations must address maintenance processes in their legacy information systems. The entire replacement of these systems has a great impact in technological and economic terms (Sneed, 2005). So that, maintenance based on evolutionary reengineering processes has typically been carried out (Bianchi et al., 2003). Moreover, the typical re-engineering process has been shifted to the so-called *Architecture-Driven Modernization* (ADM) (Ulrich and Newcomb, 2010) in the last years. ADM advocates carrying out re-engineering process following the principles of model-driven development: taking into account all involved artefacts as models and implementing transformations between them.

The increasing cost of maintaining legacy systems along with the need to preserve business knowledge has turned the modernization of legacy systems into an important research field (Lewis et al., 2010). ADM provides several benefits such as ROI improvement to existing information systems, reducing development and maintenance cost and extending the life cycle of the legacy systems.

This paper proposes the elicitation of database schemas, a reverse engineering technique that follows the model-driven principles to recover a minimal relational schema from the queries embedded in the legacy source code. The reverse engineering technique to elicit relational database schema consists of two key stages:

(i) *Static analysis of source code*, which examines the Legacy source code and looks for SQL (Structure Query Language) statements embedded in the code. In this task, a model of SQL sentences is built according to a metamodel of DML (Data Manipulation Language) of SQL.

(ii) *Model transformation*, which obtains a model of the relational schema from the previous SQL sentences model. This transformation is based on a set of rules that elicits the minimal schema of the underlying database.

The objective of this proposal is to discover the minimal subset of relational elements of the legacy

database in order to improve the database to use it in the ADM process. Thus, the proposed technique rebuilds the database schema and removes the dead parts based on the embedded SQL sentences. This mechanism eradicates duplicated or unused tables, removes unused columns, and discovers new referential constraints. In this way, the software of legacy systems does not evolve in an isolated manner, but takes into account its relational databases.

The remainder of this paper is organized as follows. Section 2 summarizes related works in the database reengineering field as well as the state-of-the-art about ADM. Section 3 explains the proposed schema induction technique in detail. Section 4 presents a case study to validate the proposal. Finally, Section 5 addresses the conclusions of this work.

2 STATE-OF-THE-ART

2.1 Related Work

Research about re-engineering on applications and databases jointly is usually addressed in literature. *Reus* (Reus et al., 2006) presents a reverse engineering process based on MDA (Model-Driven Architecture) for recovering UML (Unified Modeling Language) models from PL/SQL code. *Fong* (Fong, 1997) and *Ramanathan et al.* (Ramanathan and Hodges, 1997) transform the relational models into object-oriented (OO) models to integrated them with OO applications. Also, *Cohen et al.* (Cohen and Feldman, 2003) convert parts of the application logic from the procedural style of the legacy systems to the declarative style of SQL in order to integrate them with relational databases. *Hainaut et al.* (Hainaut et al., 1996) proposes a reverse engineering process for recovering the design of relational databases and *Wu et al.* (Wu et al., 2008) discover topical structures of relational databases. Moreover, *Pérez-Castillo et al.* (Pérez-Castillo et al., 2009) propose a wrapping technique to extract Web services from relational databases that manage the data access. Finally, *Polo et al.* (Polo et al., 2007) have studied building database-driven applications.

However, related works do not address some key challenges for modernizing relational databases: (1) those works do not follow a model-driven approach in all phases of the reengineering process and (2) recovered database knowledge is not managed in an integrated and standardized manner. ADM is a potential solution for dealing with the first challenge while KDM (Knowledge Discovery Metamodel) enables optimal knowledge management for relational databases within the ADM processes, the second challenge.

2.2 Standards Involved: ADM, KDM and QVT

The reengineering processes and MDA principles converge in ADM, an OMG standard for modernizing legacy systems (Khusidman and Ulrich, 2007). ADM is the concept of modernizing existing systems with a focus on all aspects of the current systems architecture and the ability to transform current architectures to target architectures (Pérez-Castillo et al., 2011).

The *ADM Task Force* in OMG led to several standards (OMG, 2009). The cornerstone of this set of standards is KDM (*Knowledge Discovery Meta*). KDM allows standardized representation of knowledge extracted from legacy systems by means of reverse engineering. In addition, KDM has been recently recognized as the standard ISO 19506 (ISO/IEC, 2009). The KDM metamodel provides a common repository structure that makes it possible to interchange information about software artefacts in legacy systems. KDM makes it possible to represent the PIM models in the horseshoe model. KDM can be compared with the UML standard: while UML is used to generate new code in a *top-down* manner, the ADM processes that involving KDM starts from the existing code and builds a higher level model in a *bottom-up* manner (Moyer, 2009).

The KDM metamodel (ISO/IEC, 2009) is divided into four layers (each one based on a previous layer) representing both physical and logical software assets of information systems at several abstraction levels. This work focuses on (i) *program element layer*, which provide a language-independent intermediate representation for various constructs determined by common programming languages; and (ii) runtime resource layer, which enables representation of high-value knowledge about legacy systems such as databases.

3 SCHEMA ELICITATION

The proposed ADM process is organized into three stages. The first stage of the modernization process aims to obtain, through reverse engineering, a set of PSM models representing each software artefact of

the legacy system. This task involves using a specific metamodel for each artefact. After that, in the second stage, a KDM model is built (using model transformations) from the PSM models recovered from the legacy system. In this case, the KDM model plays the role of a PIM model. Therefore, this model abstracts any technology-specific aspect of the legacy system. It should be borne in mind that obtaining KDM models does not end in the restructuring stage, because it is possible to restructure the KDM model itself. For example, transformations can be tailored between KDM layers. Finally, the forward engineering stage accomplishes building new and improved information systems. This stage involves PSM models to represent specific aspects related to the technological nature of each target system. In this way, the modernization process is completed according to the horseshoe model.

The aim of this modernization process is to modernize legacy systems focused on code and database as exclusive software artefacts. These artefacts undoubtedly determine three KDM models that must be obtained in the reverse engineering stage of this modernization process: (i) *KDM Inventory Model* is based on the *Source Package* of KDM. It enumerates physical artefacts of the legacy system and defines the mechanism of traceability links between all the KDM elements and their original representation in the legacy source code. (ii) *KDM Code Model* supports both *Code Package* and *Action Package*. This model aims to represent program elements and their associations at the implementation level. It includes elements supported by several programming languages such as sentences, operators, conditions, associations, control and data flows. (iii) *KDM Data Model* represents data manipulation in legacy systems. *Data Model* is based on *Data Package* and uses the foundations provided by *Code Model* related to the representation of simple data types. Also, this model can depict the relational databases used by the legacy system.

In addition to these models, the schema elicitation technique involves other models in the reverse engineering stage of this ADM process (see the shaded part of Figure 1). The database schema is elicited from the SQL embedded in the source code by means of the proposed technique, and thus it generates an *SQL Sentences Model* by means of the static analysis of legacy source code. The static analysis activity also produces the *Inventory Model* and *Code Model*. After that, the *Database Schema Model*, a model that represents the minimal schema of the database, is obtained through the model transformation from the *SQL Sentences Model*. Finally, the needed *KDM Data Model* is obtained from the *Database Schema Model*.

At this point, both the source code and the database are represented according to KDM. Therefore the restructuring and forward engineering stages can be carried out in order to generate the modernized version of the legacy systems (see Figure 1).

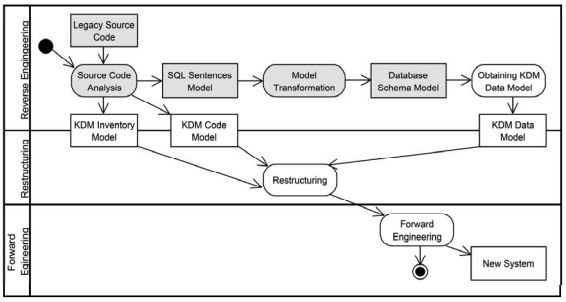


Figure 1: Schema elicitation technique based on ADM.

In order to obtain the *SQL Sentences Model* the technique analyses the legacy source code for embedded SQL sentences by means of a parser. This parser is a syntactical analyser that exhaustively scans source code. When the parser finds an SQL sentence, it translates that sentence into a model according to a metamodel of the DML (Data Manipulation Language) of SQL-92 that has been developed.

The metamodel modeling the syntax of the SQL-92 DML (ISO/IEC, 1992). It can represent the SQL operations such as *Insert*, *Select*, *Update* and *Delete* together with search conditions.

After obtaining the *SQL Sentences Model* through static analysis, the *Database Schema Model* must be obtained by mean of a model transformation. These models of relational database schemas are represented through a metamodel according to the SQL-92 standard (ISO/IEC, 1992). Deductions of the minimal database schema are based on a set of rules developed specifically for this purpose. These rules recover only a subset of the database schema elements that are handled by the SQL sentences embedded in the source code.

Rule 1. The tables that appear in any SQL sentence (Insert, Select, Update or Delete) as either source or target clauses (From, Set, Into, and so on) are created as tables in an induced database scheme.
Rule 2. The columns that are selected, added, deleted or updated in the SQL sentences are created in the corresponding tables. These tables have

previously been created through the application of Rule 1. **Rule 3.** The columns depicted through the table alias in sentences (As clause) are created in the table related to this alias. This table was created previously by means of Rule 1. **Rule 4.** The data type associated with each column can be deduced through the kind of expressions where these columns appear. For example, Like expressions →'string' data type; arithmetic expressions →'integer, decimal or numeric' data type, and so on. **Rule 5.** The Select sentences structured in Join mode suggest potential primary keys and foreign keys according to the pattern expressed in Figure 2. While the source column(s) of join select is/are related to the column(s) within the foreign key, the target column(s) of join select is/are related to the column(s) within the primary key. **Rule 6.** After applying the previous rules, it is possible that some tables are created without a primary key. In this case, a new column is attached to these tables as its primary key. This column is a sequential number that is generated automatically.

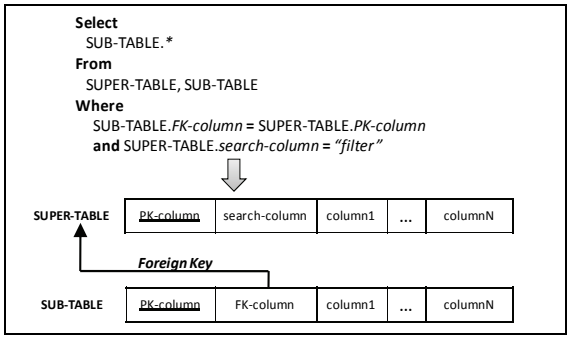


Figure 2: Pattern "join select to foreign key".

4 CASE STUDY

The case study addresses a modernization project that is currently being carried out. The subject legacy system of this project is the intranet of the *Computer Science Faculty* at the *University of Castilla-La Mancha*. This intranet was developed five years ago by several people.

The structure of the intranet consists of five modules: (i) the *Main* module is the major module of the intranet and has the standard functionalities; (ii) the *Administration* module is in charge of administrative and office tasks; (iii) the *Old Students* module manages information related to students who were members of the faculty; (iv) the *Management* module is a module for the setup of the intranet; and finally, (v) the *Quality* module measures and reports on the quality of the faculty.

The intranet has a typical Web architecture separated into three layers: presentation, business and persistence. The technology used to develop the presentation layer was *JSP* (*Java Server Pages*), *JAVA* for business layer and *ORACLE* together with *JDBD-ODBC* for the persistence layer. The total size of this legacy system is 72.68 KLOC.

The case study establishes two research questions to analyse the *Database Schema Models* obtained through the proposal:

Q1. Are the output models complete?

Q2. Are the output models minimal with regards to the original database schema?

The question *Q1* aims to assess the completeness of the obtained database schema. A specific schema is complete when: (i) all tables in this schema model have a primary key; (ii) there are no tables without columns; and (iii) the schema model does not have any duplicated elements. Moreover, the question *Q2* takes into account the minimization of the database schema. The minimization is measured by means of the size of the obtained schema regarding the size of the source database schema. In order to measure the gain between the previous and current sizes, we use two variables: the gain related to the number of tables (1) and the gain related to the number of columns in each table (2). In these formulas, T_0 is the number of tables in the legacy database schema and $C_{0\{Ti\}}$ represents the number of columns of Table i in the legacy database. Moreover, T represents the number of tables in the improved database schema and $C_{\{Ti\}}$ is the number of columns in Table i in the obtained database schema.

$$G_T = \frac{T_0 - T}{T_0} \quad (1)$$

$$G_{C\{T_i\}} = \frac{C_{0\{T_i\}} - C_{\{T_i\}}}{C_{0\{T_i\}}} \quad (2)$$

The execution of the case study was carried out by means of a tool based on the *Eclipse* platform that was developed to support the elicitation schema technique. A QVT (Queries / Views / Transformations) (OMG, 2008) model transformation is tailored in the tool from the proposed rules. The tool accomplishes several *SQL Statements Models*, a model for each source code file. Table 1 summarizes the models obtained from the legacy source code.

After that, the QVT transformations are executed through the tool using these models as input models

to obtain the output model that represents the new and minimal database schema.

Table 1: Input SQL statement models.

Module	Source Files	SQL Statements	SQL Statement Models
Main	18	23	18
Administration	8	14	8
Old Students	4	4	4
Management	1	1	1
Quality	44	60	44
Total	75	102	84

After the execution of the QVT transformations, a set of output models that depicts database segments (used for each module of the intranet) was obtained. Table 2 summarizes the results obtained for each output model. The legacy database had 140 tables and 25 tables were recovered. Table 2 shows the tables recovered for each intranet module. In addition, it presents the gain related to the tables (G_T) as well as the gain regarding the columns (G_C).

The analysis of the results obtained for these models presents several conclusions that should be considered as a response to the question *Q1*: (i) tables are usually obtained without primary keys unless Rule 6 is launch after other QVT relations; (ii) obtaining tables without columns is not usual, because any column that appears in a SQL statement is normally associated with its table.

In this case study, the QVT relations do not infer enough foreign keys, because the only QVT-implemented mechanism for inferring foreign keys is Rule 5. Indeed, the intranet source code has only two *join select* sentences due to the bad design of the legacy database.

In order to respond to the question *Q2*, the gain of the obtained database schema was also assessed. In total, 18% of the tables were recovered (25 tables) and the G_T value was 82%. With respect to the columns, the box diagram in Figure 3 shows the distributions of G_C for each intranet module. The mean per table of the G_C values was 27%, although in some modules this mean was higher. In this study, the G_C mean is lower than the G_T. However, the total gain related to the size minimization of the new database schema is significant.

5 CONCLUSIONS

This paper proposes a modernization process based on KDM. The objective of this process is the modernization of legacy source code together with legacy relational databases. For this reason, this proposal considers two complementary sources of knowledge: (i) the schema of the legacy database, and (ii) the SQL sentences embedded in the legacy source code.

The main contribution of this paper is a mechanism named 'schema elicitation' to rebuild the database schema from the SQL sentences embedded in the source code.

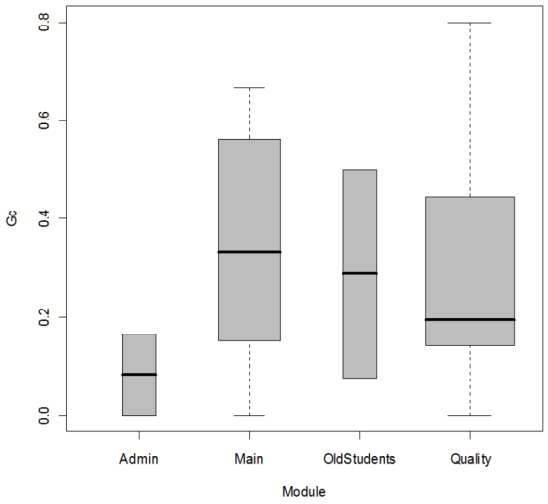


Figure 3: Box diagram of GC per module.

This mechanism removes the dead parts of the database schema such as duplicated or unused tables and unused columns. Also, this mechanism can discover new referential constraints implicit in the source code. Therefore, this proposal obtains improved and minimal database schemas used in the latter stages of the modernization process. In order to support the schema elicitation mechanism, two SQL metamodels were developed: a metamodel for representing SQL sentences embedded in source code as well as a metamodel for modelling schemas for relational databases. In addition, a set of QVT relations was tailored to transform a SQL sentences model into another database schema model.

Finally, a case study with a legacy intranet reports the main advantages of this proposal. Firstly, the obtained database schema had the adequate completeness level, thus that schema can be used as new database schema in the modernized system. Secondly, the dead parts were removed. In our case study the minimization of the size of the obtained schema was around 80% with regard to the original size.

The future extension of this research focuses on the improvement of the completeness level of the database schema models and on the detection more dead parts by means of more refined patterns. For

Table 2: The schema elements recovered through the case study and the gain of the minimal schema.

Total Tables = 140	ALUMNOSCEP	ANTIGUOSALUMNOS	ASIGNATURA	CEP	CLAVEALUMNOSOLD	CONFERENCIA	CONFIGURACION	CURSOS_CERO	CURSOS_ESP_POSTGR	CURSOS_TECNICAS_ET	DOCUMENTOS	EMPRESA	FECHAS_VISITAS	FORMACION_REGLADA_ASIG	GRAIITS	INFORMES_CALIDAD	MATRICULASCEP	NUMERO_ALUMNOS	PRESUPUESTO_INVESTIGACION	PROYECTO_INVESTIGACION	SEMINARIOS_DI	SUBCURSO_EP	TAREAS	USERS	VISITAS	Primary Keys	Foreign Keys	Tables	G_T
Main	x			x			x				x		x				x							x	x	x	1	1	9	94%
Administration															x										x		0	0	2	99%
Old Students		x			x																						0	0	2	99%
Management			x																								0	0	1	99%
Quality				x		x		x	x	x		x		x		x		x	x	x	x						1	1	13	91%
TOTAL	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**2**	**2**	**25**	**82%**
Recovered Cols	4	12	3	9	1	6	1	4	5	6	7	5	4	14	5	6	3	6	4	4	5	6	11	3	1					135
All Columns	4	13	15	28	2	19	2	7	6	7	9	9	6	18	6	6	8	6	4	7	6	7	12	3	3					213
G_C	0	0,1	0,8	0,7	0,5	0,7	0,5	0,4	0,2	0,1	0,2	0,4	0,3	0,2	0,2	0	0,6	0	0	0,4	0,2	0,1	0,1	0	0,7					27% (Mean)

this purpose, more case studies will be carried out in order to detect more information needs in the target database schema that must be obtained. In addition, the future work will address the next stages of the proposed modernization process such as the transformation from database models to KDM models, and then, the restructuring and forward engineering stages, which will use the previous knowledge.

ACKNOWLEDGEMENTS

This work has been supported by the *FPU Spanish Program*; by the R&D projects funded by *JCCM*: ALTAMIRA (PII2I09-0106-2463), INGENIO (PAC08-0154-9262) and PRALIN (PAC08-0121-1374).

REFERENCES

Bianchi, A., Caivano, D., Marengo, V. and Visaggio, G. 2003. Iterative Reengineering of Legacy Systems. IEEE Trans. Softw. Eng., 29, 225-241.

Blaha, M. Year. A Retrospective On Industrial Database Reverse Engineering Projects-Part 1. In: *Proceedings of The 8th Working Conference on Reverse Engineering (WCRE'01)*, 2001 Suttgart, Germany. Ieee Computer Society, 136-147.

Cohen, Y. and Feldman, Y. A. 2003. Automatic High-Quality Reengineering of Database Programs By Abstraction, Transformation and Reimplementation. *ACM Trans. Softw. Eng. Methodol.*, 12, 285-316.

Fong, J., 1997. Converting Relational to Object-oriented Databases. *ACM SIGMOD Record*, 26, 53–58.

Hainaut, J.-L., Henrard, J., Hick, J.-M., Roland, D. and Englebert, V. 1996. Database Design Recovery. *Proceedings of The 8th International Conference on Advances Information System Engineering*. Springer-Verlag.

ISO/IEC. 1992. ISO/IEC 9075:1992, Database Language Sql.

ISO/IEC. 2009. ISO/IEC Dis 19506. *Knowledge Discovery Meta-Model (KDM)*, V1.1 (Architecture-Driven Modernization).

Khusidman, V. And Ulrich, W. 2007. Architecture-Driven Modernization: Transforming the Enterprise. Draft V.5. andHttp://Www.Omg.Org/Docs/Admtf/07-12-01. Pdf. Omg.

Lewis, G. A., Smith, D. B. and Kontogiannis, K., 2010. A Research Agenda for Service-Oriented Architecture (Soa): Maintenance and Evolution of Service-oriented Systems. *Software Engineering Institute*.

Moyer, B., 2009. Software Archeology. Modernizing Old Systems. *Embedded Technology Journal* [Online], 1. Available: Http://Adm.Omg.Org/Docs/Software_ ArcheoLogy_4-Mar-2009.Pdf [Accessed Junio 2009].

Omg 2008. Qvt. Meta Object Facility (Mof) 2.0 Query/View/Transformation Specification. Http:// Www.Omg.Org/Spec/Qvt/1.0/Pdf. Omg.

Omg. 2009. Architecture-Driven Modernization Standards Roadmap. [Online]. Omg. Available: Http://Adm. Omg.Org/Admtf%20roadmap.Pdf [Accessed 29/10/2009].

Pérez-Castillo, R., García-Rodríguez De Guzmán, I., Caballero, I., Polo, M. And Piattini, M. 2009. Preciso: A Reengineering Process and a Tool for Database Modernisation Through Web Services *24th ACM Symposium on Applied Computing*. P. 2126-2133.

Pérez-Castillo, R., García Rodríguez De Guzmán, I. and Piattini, M. 2011. Architecture-Driven Modernization. In: Dogru, A. H. And Bier, V. (Eds.) Modern Software Engineering Concepts And Practices: Advanced Approaches. Igi Global.

Polo, M., García-Rodríguez De Guzmán, I. And Piattini, M. 2007. An Mda-based Approach For Database Re-Engineering. *J. Softw. Maint. Evol.*, 19, 383-417.

Ramanathan, S. And Hodges, J. 1997. Extraction of

Object-oriented Structures From Existing Relational Databases. *ACM SIGMOD Record*, 26, 59–64.

Reus, T., Geers, H. And Deursen, A. V. Year. Harvesting Software for MDA-based Recovering. In: European Conference on Model Driven Architecture - Foundations and Applications, 2006 Bilbao (Spain). Springer-Verlag Berlin Heidelberg.

Sneed, H. M., 2005. Estimating The Costs of a Reengineering Project, *IEEE Computer Society*.

Ulrich, W. M. and Newcomb, P. H., 2010. Information Systems Transformation. Architecture Driven Modernization Case Studies, Burlington, MA, Morgan Kauffman.

Visaggio, G., 2001. Ageing of a Data-intensive Legacy System: Symptoms and Remedies. *Journal Of Software Maintenance*, 13, 281-308.

Wu, W., Reinwald, B., Sismanis, Y. and Manjrekar, R. Year. Discovering Topical Structures of Databases. In: *ACM SIGMOD International Conference on Management of Data*, 2008 Vancouver, Canada. Acm, 1019-1030.

Data Processing Modeling in Decision Support Systems

Concepción M. Gascueña[1] and Rafael Guadalupe[2]
[1]*Department of Computing, Polytechnic of Madrid University, Carretera de Valencia Km7, 28031 Madrid, Spain*
[2]*Department of Topographic, Polytechnic of Madrid University, Carretera de Valencia Km7, 28031 Madrid, Spain*
cmgascuena@eui.upm.es, r.guadalupe@topografia.upm.es

Keywords: Multidimensional Models, Data Processing in Multidimensional Databases, Data Processing in Data Warehouses, Data Processing in Decision Support Systems, Virtual factEntity.

Abstract: Due to the advancement of both, information technology in general, and databases in particular; data storage devices are becoming cheaper and data processing speed is increasing. As result of this, organizations tend to store large volumes of data holding great potential information. Decision Support Systems, DSS try to use the stored data to obtain valuable information for organizations. In this paper, we use both data models and use cases to represent the functionality of data processing in DSS following Software Engineering processes. We propose a methodology to develop DSS in the Analysis phase, respective of data processing modeling. We have used, as a starting point, a data model adapted to the semantics involved in multidimensional databases or data warehouses, DW. Also, we have taken an algorithm that provides us with all the possible ways to automatically cross check multidimensional model data. Using the aforementioned, we propose diagrams and descriptions of use cases, which can be considered as *patterns* representing the DSS functionality, in regard to DW data processing, DW on which DSS are based. We highlight the reusability and automation benefits that this can be achieved, and we think this study can serve as *a guide* in the development of DSS.

1 INTRODUCTION

One of the challenges of Software Engineering (SE), is to propose: rules, process, guidelines and models that address Software development: quickly, efficiently, in a specific and unambiguous manner and resulting in a quality product. Methodologies are proposed continually, with varying degrees of complexity and agility; leading teams in a certain direction during the software development process, also referred to as software life cycle. In recent years, SE has acquired great importance and, increasingly, less software developments that being undertaken without prior planning. In SE the Cases of use (CU), are considered by most members of the scientific community as a technique, not necessarily object-oriented, which allows us to model the functionality of a software system at a high level of abstraction, and with no regard to the programming paradigm in which the system will be implemented.

Decision Support Systems DSS, are based upon historical databases containing large amounts of data. They try to extract the information processing the data in a certain way; allowing managers to make decisions and predict future trends. "*Predicting the future by studying the past*." However, DSS are not always based on databases built for this purpose, sometimes using transactional databases, something we don't consider efficient. We believe the DSS must be based on data warehouses (DW), or multidimensional databases (MMDB); and following specific, multidimensional (MM), data models; which reflect the multidimensional semantics and lead to analysis from the earliest stages of system development. In this work we are using MM and CU for modeling processing data in DSS.

This paper is structured as follows: Section 2 includes a study on related works in MMDB and on the representation of functionality in the development of Software Systems. In Section 3, we present our proposal. Section 4 includes an example using our proposal. In section 5, some conclusions and future work are offered.

2 RELATED WORKS

Most DSS development proposals are mainly concerned with the database on which they are built

upon, (Kimball, 1996), (Imon, 2002), (Mazón, 2006). To develop this DB, data models have been shown, as in (Tryfona, 2003), (Torlone, 2003), (Malinowski, 2004), (Luján-Mora, 2006), (Gascueña, 2006). There are authors that propose using transactional database models, as (Malinowski, 2004), (Tryfona, 2003), however other authors propose using specific models that treat the semantic MM in a specific manner, as (Kimball, 1996), (Torlone, 2003), (Gascueña, 2008c). In recent years, the importance given to MM models has increased, and there are even some proposals that try to represent spatial-temporal data behavior within them, as in (Malinowski, 2005), (Parent, 2006), (Gascueña, 2008a), (Bimonte, 2008). This leads us to stress the value that the scientific community is giving to MM models used in the development of the DW or MMDB. Regarding the processing of data, there are some works as in (Gascueña, 2008b), where an analysis is performed, while separating the concepts of basic data and derived data. They use models to represent both data types, and they propose an algorithm responsible for the automatic gathering of the data derived from the DW. However there are few proposals regarding the data processing functionalities of DSS.

The CU is the most widely employed technique to model Software systems functionalities. However, these are almost always used in a particular way for each system; they are "tailored" by the applications that they model. We think it would be desirable to propose CU "*patterns*" that could be reused by most systems that need the same functionalities. There are some initiatives that tackle generalized problems, such as in (Guttorm, 2005) who proposes using CU to represent the supposed potential threats that a system could face, modeling both the functionality and threats of systems, They name these, cases of bad use, *misuse cases*. In (Kantorowitz, 2003) a framework is proposed, oriented on CU, to build, automatically, graphical user interfaces (GUI). They also attempt to reuse these CU in different applications. In (Luján-Mora, 2006) the MM semantics are specified using class diagrams and they propose new artifacts aimed at collecting such semantics. They include an example of how to specify two data requirements by two CU. But the proposed CU, are entirely dependent upon the discussed requirements. In this paper we propose a general reusable CU, a "*pattern*", which may be used as a guide in the development of DSS to the end of modeling the data processing functionality.

3 PROPOSAL

We are framing this paper within the Software Engineering into the Analysis Phase of software life cycle. We will use data models and CU to propose a guide for development of DSS; proposing, *on one hand,* appropriate conceptual MM data models that reflect the basic starting data required to develop a DW. And *on the other hand*, we will use CU to represent the functionality of any DSS, regarding data processing, and that will allows us to obtain, dynamically and automatically derived data. The MM data models used in this study were shown in (Gascueña, 2006) and completed in (Gascueña, 2008a). To obtain dynamically derived data, we have used the algorithm presented in (Gascueña, 2008b).

3.1 Data Models

In this section we offer a brief introduction of conceptual MM model named FactEntity (FE), to better understand our proposal.

The MM models should represent the data focused to analysis at the earliest stages of the DSS development. They try to represent a *fact* object of study, from different perspectives or *dimensions* and with different *levels* of detail or granularities. *Levels* are obtained by grouping *basic data* from different criteria. With different criterion are formed different *hierarchies*. A *hierarchy* contains a set of levels grouped according to a criterion. A *dimension* can have multiple hierarchies. A *fact* consists of a set of *fact measurements*.

The *FE* model distinguishes between *basic data* (existing data) and data obtained by processing the basic data according to the analysis criteria, also called *derived data*. *Facts* and *dimensions* are combined to obtain the named *factEntities*. The *factEntities* can be *basic* and *virtual*. The *Basic factEntities BfE*, are obtained through the dimensional levels of minimum granularity (leaf levels) and *basic fact measures*. The named *Virtual factEntities VfE*, are obtained through the processing of basic data. The rules by which each factEntity contains a single level of each dimension and a set of fact measures are complied with. Though sometimes this set could be empty. In figure 1, we see the constructors, elements, relationships and functions used by the FE model, representing the MM semantics.

Hierarchies are classified according to the involvement their "*path Rollup"* (moving from a lower to a higher level) has over fact measures. Next

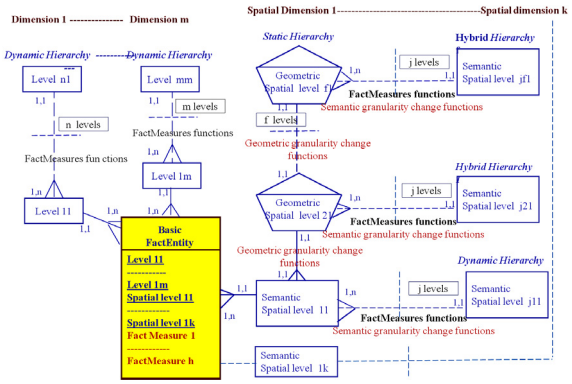


Figure 1: Basic FE model completed with the functions that will apply on fact measures when the Rollup is run.

we see these:

- *Dynamic hierarchy* (its route involves changes in fact measures).
- *Static hierarchy* (its route does not involve changes in the fact measures).
- *Hybrid hierarchy* (is a mixture of the two previous types).

As we show in Figure 1, the Static and Hybrid hierarchies represent spatial characteristics. We see that the BfE counts with representatives of the dimensional leaf levels and fact measures. Also, the diagram represents both, the functions to be applied to achieve higher levels in the hierarchy (this is of specially interest in changing spatial granularities), and the analysis functions to be applied on fact measures, once the rollup between the dimensional levels has been performed (this is necessary as to perform basic data processing and obtaining derived data).

3.2 Cases of Use

In this proposal we present a generic CU model aimed at picking up DSS functionalities in regard to the processing of basic data. This intends to be a *guide* for developers and analysts of these systems.

3.2.1 CU Diagram

In Figure 2 we can see the *To Generate Virtual factEntities* diagram, which represents a main CU named *Generate Virtual factEntity* VfE_CU, and four associated CU: Create Table, Create Materialized View, Create View, Other. All of them count with the <<extend>> label. This provides the functionality the ability to store the VfE both, inside and outside the DW, and also in various, different, ways, leaving the final choice up to the user (analyst manager).

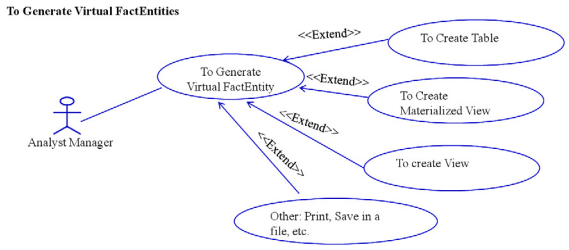


Figure 2: CU diagram: *To Generate Virtual FactEntities*. This shows how the VfE storage could be chosen in the analysis.

3.2.2 CU Description

In Tables 1, 2 and 3 we can observe the VfE_CU description. This is a generic CU that defines the minimum functionality required in any DSS, needed for the processing and gathering of derived data, from a DW. To develop this CU we have used the data model of Figures 1. Table 1 contains the principal scenario or *typical course of events,* functionalities. The head of this CU has been omitted since it is not relevant for this work.
The VfE_CU performs the following tasks: *First*, the user requests to generate VfE, the system asks for the order in which dimensions will be crossed to get all possible VfE. *Second*, the system calls the Gascueña algorithm, which in turn obtains all the possible forms of data crossing between dimensions and fact measures. *Third*, the system presents the user with a listing of the obtained VfE. *Fourth*, the user chooses to generate a certain VfE (this action

Table 1: Events typical course of Generate VfE_CU.

Events Typical Course	
USER	**SYSTEM**
1.- Requests to Generate Virtual FEs	
	2. Asks dimensions in the desired order
3. Chooses ordered dimensions by analysis to perform	
	4. Generates VfE listing applying the algorithm Gascueña
	5. Shows list of the VfE generated
6. Chooses the fact measures and the VfE by to run	
	7. Shows a listing with different options to store data: table, view, materialized view, others.
8. Chooses option to store in a table of DW.	
	9. Calls To Create Table CU
	10. Asks conformity to save data in the DW
11. Agrees	
	12.Ask if he wish to run another VfE from those obtained in step 5.
13. Chooses option: No	
	14.Ask if he wish to generate other VfE listings changing the dimensions order.
15. Chooses option: Not continue	
	16. Close option Generate VfE.

obtains and process data of basic DW, respective of the VfE structure chosen). *Fifth*, the system obtains and presents the data and requests an option towards the data treatment, as it is shown in the diagram in figure 2. *Sixth*, the system allows the execution of as many VfE as needed by the user. The system will also allow obtaining other VfE listings, taking dimensions in different order, and as many times as the user wants. All this is explained in detail in Table 1, which has 16 steps. In Tables 2 and 3 we observe some alternatives, which we have considered more important, to VfE_CU's typical course.

Table 2: Events alternative courses contemplate various options for storing structures and data of VfE.

Alternative Curse 2)	**Alternative Curse of step 8 of Events typical Curse**
USER	**SYSTEM**
	7. Shows a listing with different options to store data: table, view, materialized view, others.
8. Chooses option to store data by materialized view.	
	9. Calls To_Create Materialized View CU
	10. To return to step 10 in Events typical course
Alternative Curse 3)	**Alternative Curse of step 8 of Alternative Curse 2**
USER	**SYSTEM**
	7. Shows a listing with different options to store data: table, view, materialized view, others.
8. Chooses option to store data by view.	
	9. Calls To_Create View CU
	10. To return to step 10 in Events typical course
Alternative Curse 4)	**Alternative Curse of step 8 of Alternative Curse 3**
USER	**SYSTEM**
	7. Shows a listing with different options to store data: table, view, materialized view, others.
8. Chooses option other (not store data in DW): to print, to store data in a file, etc.	
	9. Calls Other CU
	10. To return to step 10 in Events typical course

Table 2 describes alternatives to the so called "Create table CU", (step 8 of events typical course). There are various options: Create materialized views CU, Create views CU and Others CU. Table 3 describes alternatives to run additional VfE (option: Yes, step 13 of the typical course of events); and alternatives to obtain new lists of VfE, choosing dimensions in different orders (option: Yes, step 15 of events typical course). Both, the typical course as alternative courses may contain more options, but here, they have not been considered since they do not bring greater value into our discussion.

Table 3: Events alternative courses that show the ability to implement different VfE; and the ability to obtain new lists of VfE choosing dimensions in different orders.

Alternative Curse 5)	**Alternative Curse of step 13 of Events typical Curse**
USER	**SYSTEM**
	12.Ask if he wish to run another VfE from those obtained in step 5.
13. Chooses option: Yes	
	14. To return to step 5 in Events typical course
Alternative Curse 6)	**Alternative Curse of step 15 of Events typical Curse**
USER	**SYSTEM**
	14.Ask if he wish to generate other VfE listings changing the dimensions order.
15. Chooses option: Yes	
	16. To return to step 3 in Events typical course

3.2.3 Gascueña Algorithm

Let's briefly define the Gascueña algorithm, for further details please refer to (Gascueña, 2008c). We describe it in three stages.

First: Given a set of *n* dimensions, we obtain all possible combinations, in groups of 1, 2,...,*n-1* and *n* dimensions. We apply the follow formula (1):

$$[D_i,...,D_p]/\forall i \in [1,..,n] \Lambda \forall p \in [i+1,...,n] \Lambda (p > i \text{ OR } p = \emptyset). \quad (1)$$

Second: The *Cartesian product* is applied on each of the previous subgroups, taking into account that in some application domains, the order in which we choose the elements to make up the subgroup will be significant.

Third: The Virtual factEntities are obtained by adding to the Cartesian subgroups obtained in the previous step the respective fact measures. We then apply the following formula (2):

$$VfE=([D_iX\ldots XD_p],\{G_j(me_j)\})\text{-}(BfE). \quad (2)$$

Where: $(D_iX\ldots XD_p)$ represent the *Cartesian Product.* And $(G_j(me_j)$ is the set of compatible functions G_j with the basic fact measure (me_j). It excludes the Basic fE).

4 APLICATIONS

Next we will develop a practical example in which we will apply our proposal.

We consider it desirable to study the damage caused by insect plagues in agriculture of certain Earth zones over time. The spatial area is divided into plots, and these are grouped into cities. It is necessary to store the % of extension of each plague on each plot in a given and determined moment of time. The plagues are exterminated, or attempted to, through the use of different technologies. The study requires storing existing technologies and effectiveness of such in the treatment of infected plots. The effectiveness is measured by the % of deaths caused by the treatment. The evolution of plagues on each plot is checked weekly. The spatial areas will be represented by spatial data with

geometric shapes, such as: surfaces, lines and points that can be indistinctly used. The % extension of plague and % deaths will be studied from different perspectives and details: *Time:* week, year; *Zones:* plot, city; *Technical:* technical type; *Plague:* plague type, family and order.

To offer a solution to this study we propose building a DSS, which allows us to analyze the effectiveness of anti plague treatments, and aid us in choosing the best decisions regarding the treatment of new emerging plagues. The DSS will consist of a MMDB or DW complete with spatial treatment. Furthermore, the system allows the data processing of DW on demand, in an easy and quick manner. Figure 3 shows the proposed *FE Basic model* as a solution for the storing of the input data.

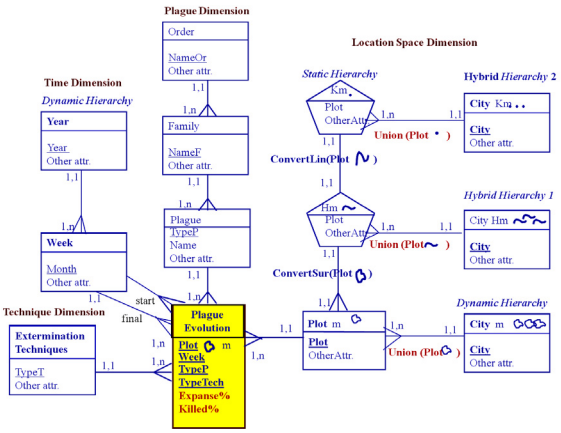


Figure 3: Basic FE model for Plagues Study.

We have identified the following dimensions: Time, Plague, Technique and Location Space. The *Time* dimension has two granularities: week, year. The *Plague* dimension has three granularities: type plague, family and order. The *Location Space* dimension has two semantic granularities: plot and city; and three geometric granularities (spatial representation): surface, line and point. Also this dimension form a *dynamic hierarchy*, a *static hierarchy* and three *hybrid hierarchies*. The "Plague Evolution" *basic factEntity* contains the primary keys inherited from the leaf level of the dimensions (underlined in the diagram). The week level has two relationships (start, final) with BfE. The *fact* under consideration contains two *fact measures*: Expanse% and Killed%. In the diagram, we can also observe the functions used to create higher levels, of both the *geometric* and *semantic* granularities, within the spatial dimension. In figure 4, we observe how the Basic FE model is completed with information regarding the functions to be used for the analysis, once the Rollup is made.

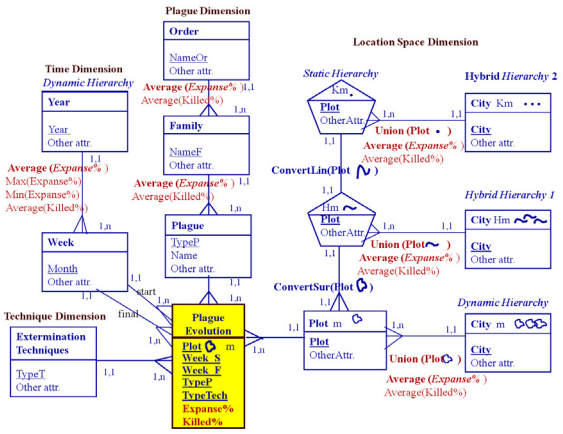


Figure 4: FE conceptual multidimensional model, prepared for processing data by "Plagues Study DSS".

Now and here we could have included the CU models presented in Figure 2 and tables 1, 2 and 3, adapted to our example. But, if we study these models in detail, we note that it is necessary to include anything new in the descriptions and diagram of the VfE_CU. We observe that the CU model proposed is valid to represent the required minimum functionality required to process the derived data in this example.

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper we have proposed a methodology, which attempts to serve as a *generalized guide* for the development of DSS following the Software Engineering guidelines. Our proposal is framed within the Analysis phase of the software development process life cycle. We have used MM data models and CU to lead the development. *On the one hand*, we offer the foundations to build a DB that collects MM semantics (to create the DW, main part of DSS). *On the other hand*, we model the data processing, defining the desired functionality through a CU model. We explain our proposal in three steps. *First*, we propose carrying out a conceptual multidimensional data model with the adequate structure required to store the basic or starting data in a DW. The model takes into account the analysis requirements. *Second*, the basic data model obtained in the previous step is completed with the operations and functions that we would want to use in the data analysis. This new model presents all the necessary elements needed for the

processing of the data, allowing us to obtain new data structures for the derived data. *Third,* data functionality processing is modeled by a CU. In particular, it is defined and developed the Virtual factEntity CU. The VfE_CU details the minimum and necessary events sequence required for the basic data processing. These VfE_CU use an algorithm that interacts with data models, collecting the information represented in them, to generate, automatically and on-demand, all the possible VfE. The steps above outlined, can be considered to have a high level of abstraction and are independent of its implementation. We believe that the proposed CU can serve as a basic *pattern* in the development of DSS; which later may be completed and adapted to each particular situation, if necessary. Finally, we have presented an example in which we develop a case study using our own proposal.

Our future research is aimed at discovering other general behavioral patterns, which could guide the development of the DSS. In addition, we are interested in developing a tool that would allow us to describe and transform, automatically, the FE data models and the VfE_CU, into real systems. The FE model transformation will be made to implement the models in commercial DB manager Systems, under different paradigms: Relational, Object Relational or Object Oriented. The VfE_CU transformation will allow us to implement a basic interface, with the features described in this proposal, while also allowing for the possibility to choose programming languages among the most popular ones.

REFERENCES

Bimonte S., Tchounikine A., Berloto M., 2008. Integration of Geographic Information into Multidimensional Models. *ICCSA 2008*: International, 2008.

Gascueña C. M., Cuadra D., Martínez P., 2006. A Multidimensional Approach to the Representation of the Spatiotemporal Multigranularity. *ICEIS 2006*.

Gascueña C. M., Guadalupe R., 2008a. Some Types of Spatio-Temporal Granularities in a Conceptual Multidimensional Model. *7th International Conference, APLIMAT* Bratislava, Slovak.

Gascueña C. 2008b. Propousal of a Conceptual Model for the Representation of Spatio Temporal Multigranularity in Multidimensional Databases. *PhD Thesis. Polytechic University of Madrid*, Spain.

Gascueña C. M., Guadalupe R., 2008c. A Study of the Spatial Representation in Multidimensional Models, *ICEIS 2008*.

Guttorm Sindre, E Andreas L. Opdahl, 2005. Eliciting security requirements with misuse cases, in *the Journal of Requirements Eng*, Issue 10, pp 34–44.

Inmon, W. 2002. Building The Data Warehouse. *Jhon Wiley & Sons*.

Kantorowitz E., Lyakas A., Myasqobsky A.. 2003. A Use Case-Oriented User Interface Framework. *Software. SwSTE '03. IEEE International Conference on.*

Kimball R. 1996. The Data Warehouse Toolkit. *John Wiley&Sons Ed*.

Luján-Mora S., Trujillo J., Song Il- Yeol. 2006. A UML profile for multidimensional modeling in data warehouses. *DKE, 59(3), p. 725–769*.

Malinowski, E. and Zimanyi, E., 2004. Representing Spatiality in a Conceptual Multidimensional Model. *Proc. of the 12th annual ACM international workshop on GIS. Washington, DC, USA*.

Malinowski E., Zimanyi E., 2005. Spatial Hierarchies and Topological Relationships in the Spatial MultiDimER model. *Lecture Notes in Computer Science, page 17, Volume 356*.

Mazón J. N., Pardillo J., Meliá S. y Trujillo J., 2006. Modelado Multidimensional de almacenes de datos con MDA. XI *JISBD 2006*.

Parent C., Spaccapietra S., Zimanyi E., 2006. The MurMur project: Modeling and querying multi-representation spatio-temporal databases. *Information Systems, Volume 31, Issue 8*, Pages 733-769.

Torlone R., 2003. Conceptual Multidimensional Models. In Multidimensional databases: problems and solutions, pages 69-90, *Idea Group Publishing, Hershey, PA, USA*.

Tryfona, N., Price, R., Jensen, C. S., 2003. Conceptual Models for Spatio-temporal Applications. In *M. Koubarakis et al. (Eds.), Spatio-Temporal DB: The CHOROCHRONOS pg. 79-11. Berlin, Heidelberg*.

Changing Concepts in Human-Computer-Interaction in Real-time Enterprise Systems

Introducing a Concept for Intuitive Decision Support in SCM Scenarios

Christian Lambeck[1], Dirk Schmalzried[2], Rainer Alt[3] and Rainer Groh[4]
[1]*Technical University Dresden, 01062, Dresden, Germany*
[2]*OR Soft Jänicke GmbH, Geusaer Str., FH, 104, 06217, Merseburg, Germany*
[3]*University Leipzig, 04109, Leipzig, Germany*
[2]*Technical University Dresden, 01062, Dresden, Germany*
dirk.schmalzried@orsoft.de, rainer.alt@uni-leipzig.de, {christian.lambeck, rainer.groh}@tu-dresden.de

Keywords: Visual Business Intelligence, Business Analytics, Real-time Supply Chain Management, Decision Support, Radial Basis Functions, Scheduling.

Abstract: In current research, Enterprise Information Systems (EIS) are increasingly based on In-Memory-Technologies, resulting in extremely fast response times for a multitude of typical system requests. In addition, up-to-date hardware configurations apply multi-core processing units, which lead to an availability of immense computing power. Instead of a single result value, a whole result set is calculated within the same period of time. Because of these dramatic changes in technology, many business processes, currently still characterized by mask and dialog oriented user interfaces, will change to interactive and simulation based approaches. This allows for the introduction of innovative, interactive and simulation based business processes instead of conventional batch oriented ones. In the combination of the described interaction concept in this contribution and the handing of result sets as described above, the authors expect a fusion of operational (e.g. supply chain management) and analytical (e.g. business intelligence) application systems. To achieve this goal, the usage of assessment functions for weighting results, multi-dimensional result space folding based on similarity measures and visualizations using 3D-landscapes based on radial basis functions is suggested.

1 INTRODUCTION

Latest research on Enterprise Information Systems (EIS) and their underlying production methods has been manifold and primarily focused on performance and real time issues (Plattner and Zeier, 2011), Service-Oriented Architectures (SOA) (Ollinger et al., 2011) as well as sensor technologies. Especially the consistent vertical interoperability of these services and standards across the levels of automation (ISA, 2012) and the application of the *Internet of Things* to the production domain are current challenges (Kortuem et al., 2010). New production methods like modular 3F factories (Buchholz, 2010) as well as increasing complexity and dynamic of supply chain processes themselves reinforce the desire for extensive simulation enabled supply chain planning with a focus on varying input parameters and resulting outcome.

As a consequence of these changed conditions in production logistics and new objectives in the field of SCM a fundamental redesign of upcoming SCM systems is required. Especially rapidly alternating influential factors such as volatile raw material and transportation costs, volatile exchange rates and other volatile cost-influencing parameters have to be taken into account. In order to derive a reliable and suitable business conclusion, simulative "What-if"-scenarios are more important than ever and have to comprise these volatile parameters comprehensively.

By these risen claims, users demand for extensive simulation based planning tools. Their ability to vary input parameters and examine their effects on the resulting outcome reveals a powerful potential. Although simulative approaches in EIS exist, current state of the art systems fail to fulfill those requirements sufficiently. Since they were designed in the middle of the 90's, stringent hardware limitations had to be considered. In contrast, future RAM-based

computers with Multi-core support enable the user to generate whole result sets instead of a single value in a fractional amount of contemporary time consumption. This trend allows the combination of operational and analytical processes. As a result, sophisticated answers for a variety of complex SCM problems can be given in almost real time.

Attendant to the increased possibilities in handling complex information sets, related user interface principles have to change accordingly. Nevertheless, user interface design principles in the field of EIS have been rarely subjected to research within the last years. While multi-touch devices and corresponding interface concepts are widespread in other domains as illustrated in (Lima, 2012), enterprise applications – especially in the upper levels of automation – are still dealing with transactional interfaces that consist of forms, tables and dashboards and are meant to be controlled by mouse and keyboard (e.g. SAP R/3 UI- History in (SAP AG, 2012)).

Due to the novelty of visual and explorative simulation and interaction techniques in EIS, related research on human-computer-interaction can be rarely found. This contribution proposes a user interface concept for the exploration of a three dimensional landscape consisting of sampling points. These "Data Landscapes" indicate a production plan's objective fulfillment through *Key Performance Indicators* (KPI). Relevant challenges such as aggregated information presentation, real time interaction and their preliminary considerations on performance and algorithms are also addressed.

2 RELATED WORK

Nowadays, production and simulation related Enterprise Resource Planning Systems (ERP) – particularly in Small and Medium Enterprises (SME) – are customarily supported by Excel-sheets and are limited to textual or diagram output (Elizandro, 2008; Gissrau and Rose, 2011). The majority of these tools visualize the production plan as a Gantt-Chart, but direct interaction is rarely supported at all. In addition, adequate presentations which give an insight to complex correlations - like the simultaneous planning of material flows and the related resource consumption - are often missing. In general, offered visualizations are subjected to reporting in most cases, whereas wide parts of the business process remain textual. This might be one of the reasons for current usability problems as described in (Topi et al., 2005).

The research project Mind Map APS (DLR, 2010) assumed an upcoming fundamental change in the handling of enterprise applications within the next years. Therefore, the three aspects *Search Engine based System Access*, *Interactive Business Process Modeling* and *Zoomable User Interface Design* were taken into account to investigate their potentials. As a primary goal, users should be able to interact with the system more intuitively through map-based, interactive and scalable process visualizations. Although the estimated breakthrough could not be fully achieved, several prototypes were conceived which deal with 3D visualizations in oil industry, mobile process assistance for healthcare scenarios or semantic search paradigms to ease the user's system access.

Real-time EIS based on In-Memory technologies allow response generation, which is faster by speed decades. Therefore, many business processes, currently characterized by sequential and iterative dialogs, are changing to simulated ones with parallel computations (Karnouskos et al., 2010). While ERP systems facilitate the concept of simulation insufficiently, additional Advanced Planning and Scheduling (APS) applications have been introduced (Stadtler and Kilger, 2008, p.109). The involved deficiencies that result from the split system landscape are different data models and potential import/export problems, time delays or problems while merging simulation alternatives with real plans.

3 BUSINESS PROCESS

The proposed design causes some challenges in the practical implementation. This primarily derives from the vast amount of data to be processed (storage issues), requirements on short response times (performance issues) and finally the novel interaction and its resulting user acceptance (interface issues). In the following, challenges regarding condensed data as well as real-time interaction on these consolidated information are discussed.

3.1 Benefits of Planning Processes based on Simulative Result Sets

To bridge the before mentioned gap in current systems, standard and sequential ERP processes could be redefined in a real-time EIS as follows:
After the adjustment of initial parameters for an overall optimization objective in a first step, the system generates a whole set of results at once. For the step of computation, optimization methods as

well as heuristics are applicable. The emerging planning alternatives are presented in a summarized visualization instead of a series of individual results in a sequential user dialog. The major benefit is an explicit and direct comparability of the suggested planning solutions.

The parameter variations in a production scheduling task might reach from different objective functions (e.g. maximized profit margin; minimal profit margin with restocking, meeting delivery dates) to additional restrictions (stock clearance, enforcing batch clearance). Thereby a combination of these restrictions is also possible, so that a composite and complex schedule optimization task is formed. Finally, specific production schedules arise which would be typically presented as Gantt-Charts. However, comparing those Gantt-Charts – or even a subset – is a challenging task for users and constitutes the sequential and iterative dialog structure mentioned before.

3.2 From Gantt-Charts to Key Performance Indicators

A more convenient way than traditional Gantt-Charts is the comparison of summarizing *Key Performance Indicators* (KPI) for each generated planning alternative, which again can be used to evaluate SCM objective satisfaction (e.g. quality, due dates, margins, flexibility and demand fulfillment). In most cases it is sufficient to choose the best fitting schedule out of the sample space and proceed with the business process. In other cases of composite result evaluation functions it might be helpful to investigate each component's impact on the composite KPI separately.

If none of the resulting production schedules satisfies the business needs or if all resulting objective functions are not satisfying, it could help to start a new simulation run with better parameterization. Therefore users have to slightly vary the parameters specifically in those regions where already promising schedules have been found. In consequence, the level of detail for this region would be increased by any of these iterations until the identified result is satisfying. To receive further reference points for regions of favorable schedules and to fully use the interactive and simulative potential of real-time EIS, a suitable visualization technique is required. In the following section, the established concept of Data Landscapes is presented and gets adapted to the field of EIS and SCM in particular.

3.3 From Key Performance Indicators to Data Landscapes

The authors suggest a projection of the generated schedules into a plane using folding algorithms based on similarity criteria. This plane uses T*ime* as one dimension and R*esource Utilization* as the other. Contrary to Gantt-Charts that use time and resource allocation as well, this plane cannot provide a specific time or resource predication. Instead it is able to illustrate the neighborhood and therefore the similarity of production schedules. Hence, similar schedules are projected closely to each other onto that plane which is caused by the multidimensional folding algorithm.

One appropriate method for neighborhood preserving multidimensional folding of production schedules are Self-Organizing Maps - so called Kohonen Maps (Kohonen, 2001) - known from neural networks. Each production schedule is uniquely defined by the set of contained production orders, which again define an unambiguous temporal allocation of resources and material flows. Thus, the proposed folding delivers reproducible nodes in the map.

This plane layer can be extended into a third dimension by applying an evaluation function on top of these nodes. The evaluation function typically results in KPIs to be used for measuring the fulfillment of the SCM targets. The resulting sampling points can be joined using radial basis functions, for example, to form three dimensional Data Landscapes (see (Carr et al., 2001)). Despite the suggested radial basis functions, equivalent construction techniques for Data Landscapes are also applicable, of course.

Besides a uniform evaluation function, the use of "*mountain stacks*" might be suitable, in which different parts of the evaluation function (e.g. separated by margin, demand fulfillment, deadlines) are added consecutively. This allows for weighting certain input parameters and also considering particular thresholds (e.g. all schedules reaching a certain margin).

3.4 Exploring Regions of Interest

In regions around a local maximum, probably more interesting production schedules can be found. By recalculating with slightly modified parameterization, the resolution of this designated area can be increased and the user might detect more interesting production schedules that are even closer to the current objective. Due to the suggested method, additional nodes will be located closely to the exist-

ing one with a high probability. However, the folding of those multidimensional schedules into a two dimensional plane cannot avoid the partial placement of sampling points outside the current region of interest. The following example illustrates the effect that might occur:

The proposed method as described above would create a landscape with 16 different sampling points (16 CPUs could deliver those simultaneously), which are distributed non-equidistantly across this map. We assume that there are two sampling points in region A and that their evaluation function (KPI) has a significant maximum here. Hence they represent promising schedules and deserve closer attention. A next recalculation run on the same region with slightly changed parameterization generates 16 additional sampling points. Due to the marginal modification of the input parameters, the majority of them would reside in this region, but some of them, as a result of the folding, might reside in a totally different region of the map. The resulting resolution has increased again and would allow for a third iteration. After three runs, 48 sampling points are distributed across the Data Landscape where most of them reside in the region of interest.

4 USER INTERFACE CONCEPT

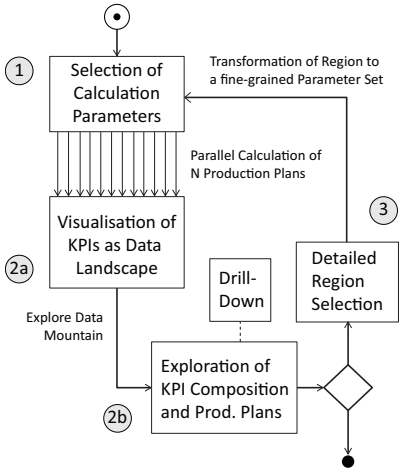


Figure 1: Business Process Dialog Model.

The preceding sections focused on the suggested business process with its benefits compared to the conventional approach. In this section, a concrete user interface concept is described, which is meant to be used on a touch-sensitive tabletop system. The described process is split into four steps as illustrated in Figure 1.

In contrast to most existing applications, the whole business process is controlled by a single view to avoid usability problems as described in (Topi et al., 2005) (identification of and access to the correct functionality, transaction execution support, overall system complexity etc.).

4.1 Selection of Calculation Parameters

As a first step, the user has to set the initial parameters (see section 3.1) which affect the selection of the simulation algorithm and adjust it according to the optimization objective. Therefore, parameters are selected in the lower left area of the screen (see Figure 2). On the right of the selection buttons, users are able to adjust the influence of a selected item by sliding the value between a minimum and maximum. Because the parameters partially affect each other, their final composition is depicted below the current slider. This way, users are always aware of the consequences during their direct manipulation. Once the parameters are selected and set as desired, the system generates the result set as described in section 3.1. Finally, a Data Landscape consisting of several sampling points gives a first overall impression of the result set's potential to satisfy the objective.

4.2 Result Presentation

Whereas conventional systems usually illustrate the simulation results in a textual manner, the Data Landscape approach has the ability to give an impression of the result set's quality at once. Each peak represents a concrete production plan which is positioned according to the axis *Resource Utilization* and *Time*. Therefore, plans with similar properties in utilization and time can be found within the same region. The height of the peak as an indicator for the achievement of objectives is build upon the sum of its *Key Performance Indicators* (KPI, see section 3.2). This means, that each KPI corresponds to a particular SCM objective and represents its partial fulfilment. Hence, the parts for quality, due dates, flexibility et cetera add up to final height and form the overall KPI for that designated production plan.

4.3 Region Selection and Drill-Down

In a next step, users might want to explore a promising area in more detail – a so called *Drill-Down*. Therefore, a top view of the Data Landscape is illustrated in the middle part of the lower control view. To select a region, users simply create a rectangle

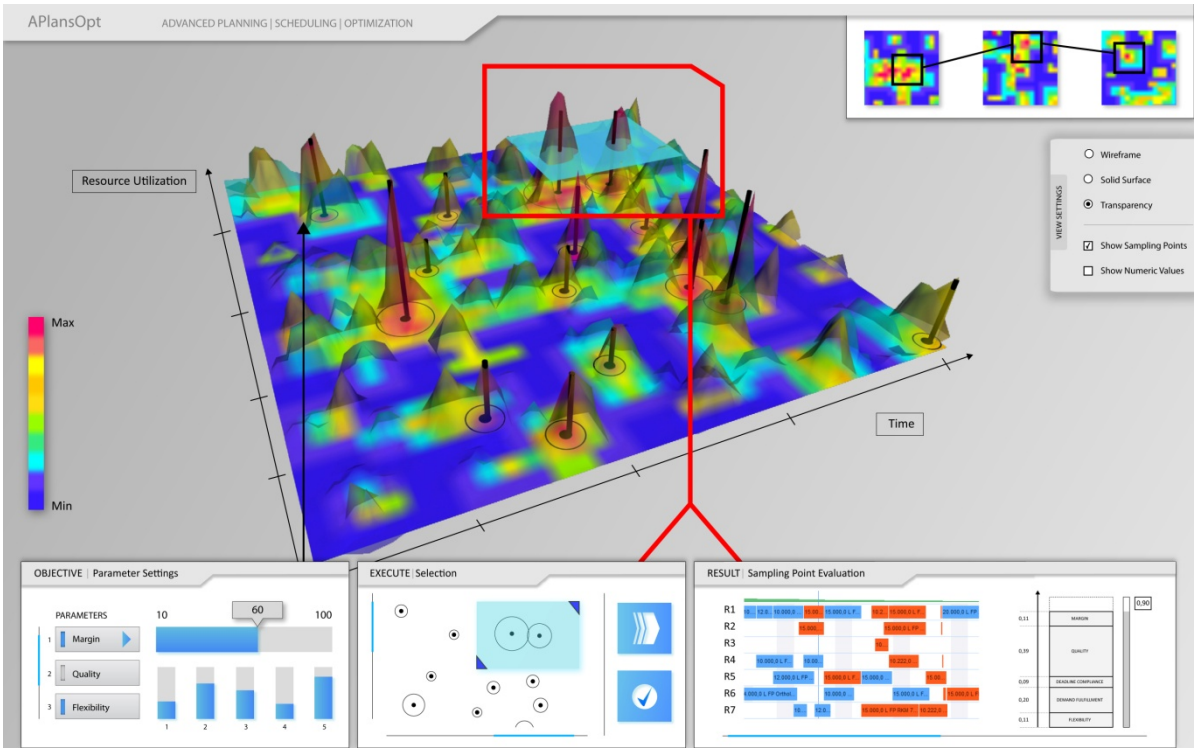


Figure 2: Suggested user interface concept with control views for parameter settings (bottom left: Objective), region selection with next iteration calculation (bottom centre: Execute) and Drill-Down with Gantt-Chart and KPI composition (bottom right: Result). The overlay indicates the Drill-Down from a selected peak to its KPI composition and the related production plan. The upper right series of snapshots illustrates the iteratively zoomed in regions in the manner of a detail-and-context. View settings (e.g. wire-frame, solid, transparent) can be adjusted in the options menu at right of the screen.

and size it to the desired dimensions. Simultaneously, a plane with same dimensions is placed in the 3D model to highlight the current area and its included peaks. However, the selection of several independent regions is not possible at present. The details of the current selection are illustrated in the lower right part of the screen, where the KPI composition and the corresponding plan's Gantt-Chart are visualized. After having examined the area peak by peak, the former selection plane might be adjusted again to restrict or enlarge the amount of included sampling points accordingly. Once the identification of valuable production plans is accomplished, a further iteration can be started which is primarily focused on the selected area. As described in section 3.2, the initial parameters are getting slightly adjusted for the next run and influence the upcoming iteration. Although not all of the computed results might be located in the area due to the parameter adjustment, its resolution is permanently increased by each iteration. In the end, the selected region gets more and more fine-grained in detail whereas the surrounding region remains widely coarse-grained. If a satisfying production plan is found, the recursive workflow ends up by applying the final production plan.

5 CONCLUSIONS

The suggested user interface concept with its related adapted business process allows for the intuitive presentation of different production schedules and their corresponding KPIs. In addition, the comparison of these schedules as well as the iterative approximation to more promising production plans is supported in a visual way.

Changing the conventional usage concept of Enterprise Applications as described in this contribution could exploit the potential of novel real-time EIS. Business analytics, business intelligence and operational design would fusion and could form a comprehensive insight into simulative information spaces. The concept of planning is transferable to other domains of operational systems, such as blend optimization, make-or-buy decisions, variations on raw material costs as well as the strategic simulation of material portfolios, geographical locations or capacity extensions. For those domains, different

simulative derived variations can be compared very rapidly and with ease. The approach of increasing a result area in resolution and its further exploration is therefore widely applicable.

6 FUTURE WORK

Although the described concept is still in a prototypical status, its potential benefits are already obvious. In further research and development, considerations on appropriate touch-sensitive hardware as well as user studies are planned. Especially the paradigm of Drill-Down with the help of multi-touch gestures on a tabletop system will be subjected to research in the future. Concerning the projection type for the 3D Data Landscape, a comparison of the current perspective projection and an isometric perspective seems to be reasonable. To support the comparability of peaks even more, the isometric projection might be more suitable. The upcoming user studies will evaluate the introduced concept by a survey with experienced users to state the major deficiencies. Due to the great demand for mobile solutions in EIS in general, further research will also have an eye on possible scenarios on mobile devices. With their numerous built-in sensors, new interaction metaphors are imaginable. One example might be the use of G-sensor abilities for suitable Drill-Down or refinement interactions.

ACKNOWLEDGEMENTS

Europa fördert Sachsen. ESF Europäischer Sozialfonds

Gefördert aus Mitteln der Europäischen Union

Christian Lambeck would like to thank the European Union and the Free State of Saxony, Germany for supporting this work. Special thanks are due to Thomas Lambeck and Frank Förster for their enthusiastic participation.

REFERENCES

Buchholz, S., 2010. Future manufacturing approaches in the chemical and pharmaceutical industry. *Chemical Engineering and Processing: Process Intensification*, 49(10), pp.993–995.

Carr, J. C. et al., 2001. Reconstruction and representation of 3D objects with radial basis functions. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. SIGGRAPH '01. New York, NY, USA: ACM, pp. 67–76. Available at: http://doi.acm.org/10.1145/383259.383266.

DLR, 2010. Mind Map APS - Project Fact Sheet. Available at: http://www.pt-it.pt-dlr.de/_media/ Infoblatt_MindMap_APS.pdf [Accessed December 12, 2011].

Elizandro, D., 2008. *Simulation of industrial systems : discrete event simulation using Excel VBA*, New York: Auerbach Publications.

Gissrau, M. & Rose, O., 2011. A DETAILED MODEL FOR A HIGH-MIX LOW-VOLUME ASIC FAB. In *Proceedings of the 2011 Winter Simulation Conference*. Winter Simulation Conference 2011. Arizona, pp. 1953–1963.

ISA, 2012. ISA | The International Society of Automation. Available at: http://www.isa.org/ [Accessed February 14, 2012].

Karnouskos, S. et al., 2010. Real-world Service Interaction with Enterprise Systems in Dynamic Manufacturing Environments. In *Artificial intelligence techniques for networked manufacturing enterprises management*. London; Heidelberg: Springer.

Kohonen, T., 2001. *Self-organizing maps*, Berlin; New York: Springer.

Kortuem, G. et al., 2010. Smart objects as building blocks for the Internet of things. *IEEE Internet Computing*, 14, pp.44–51.

Lima, M., 2012. visualcomplexity.com | A visual exploration on mapping complex networks. Available at: http://www.visualcomplexity.com/vc/ [Accessed February 10, 2012].

Ollinger, L., Schlick, J. & Hodek, S., 2011. Leveraging the Agility of Manufacturing Chains by Combining Process-Oriented Production Planning and Service-Oriented Manufacturing. In *Proceedings of the 18th IFAC World Congress. World Congress of the International Federation of Automatic Control (IFAC-2011), August 28 - September 2, Milan, Italy*. Elsevier Science Ltd.

Plattner, H. & Zeier, A., 2011. *Desirability , Feasibility, Viability – The Impact of In-Memory*, Berlin ; Heidelberg ; New York : Springer.

SAP AG, 2012. SAP Design Guild - R/3 History in Screen Shots. Available at: http://www.sapdesignguild.org/resources/r3_history.asp [Accessed December 2, 2011].

Stadtler, H. & Kilger, C., 2008. *Supply chain management and advanced planning concepts, models, software, and case studies*, Berlin : Springer,.

Topi, H., Lucas, W. T. & Babaian, T., 2005. Identifying Usability Issues with an ERP Implementation. In *Proceedings of the International Conference on Enterprise Information Systems (ICEIS)*. ICEIS'05. pp. 128–133.

MQL: A Mapping Management Language for Model-based Databases

Valéry Téguiak[1], Yamine Ait-Ameur[2], Stéphane Jean[1] and Éric Sardet[3]
[1]*LIAS, ISAE-ENSMA and Poitiers University, Futuroscope, Poitiers, France*
[2]*IRIT-ENSEEIHT, INPT-ENSEEIHT, Toulouse, France*
[3]*CRITT Informatique, Futuroscope, Poitiers, France*
teguiakh@ensma.fr, yamine@enseeiht.fr, jean@ensma.fr, sardet@ensma.fr

Keywords: Mapping, Meta-modeling, Model Transformation, Ontology Engineering, Query Languages.

Abstract: Nowadays model mapping plays a crucial role in applications manipulating various heterogeneous sources (data integration and exchange, datawarehouse, etc.). Users need to query a given data source and still obtain results from other mapped sources. If many model management systems have been proposed that support high-level operators on model mappings, a more flexible approach is needed supporting the querying of mapping models and the propagation of queries through mappings. As a solution, we present, in this paper, a mapping-based query language called MQL (Mapping Query Language). MQL extends the SQL language with new operators to exploit mappings. We show the interest of this language for the multi-model ontology design methodology proposed in the DaFOE4App (Differential and Formal Ontology Editor for Application) project.

1 INTRODUCTION

In order to deal with various heterogeneous models used to represent the same real word domain, several mapping languages (Bouquet et al., 2003; Horrocks et al., 2004) or frameworks (Jouault et al., 2008; Melnik et al., 2003; Moha et al., 2010) have been proposed. These frameworks support either model mappings or model transformations. (Bouquet et al., 2003; Horrocks et al., 2004) allow users to express correspondences between models and (Jouault et al., 2008; Melnik et al., 2003; Moha et al., 2010) describe model transformations. Both approaches aim at performing instance migration. Most of these languages run in central memory and do not address scalability when dealing with huge amount of data.

Moreover, with the emergence of the Web, the amount of models and instances is growing drastically. Managing mappings in such a context often requires writing more and more undesirable complex queries. Therefore, offering solutions for managing such mappings and instances in a convenient way becomes a necessity if one wants to address real sized problems.

Before year 2000, mappings were implemented by programs, then (Bernstein, 2003) introduced the notion of *Model Management* that aimed at reducing the amount of programming needed for the development of metadata-intensive applications. More precisely, (Bernstein, 2003) has provided model management operators (e.g, *compose*, *diff*, *merge*, *match*, *etc*) allowing to manipulate and to manage models and mappings as objects. However, to understand and to use mappings established between source models, designers need to query and to exploit them in order to express a query on a data source and to obtain data results from other sources. Thus, a more flexible approach is needed for supporting the querying of mapping model and the propagation of queries through mappings. As a solution, we propose in this paper a mapping-based query language named MQL (Mapping Query Language). This language is an extension of traditional SQL query language with new operators to exploit mappings such as crossing or filtering mappings. The interest of this language is shown on a real use case extracted from the DaFOE4App project.

This paper is organized as follows. Section 2 describes the use case set up to show the interest of our proposition. This use case is an ontology design methodology based on a multi-models approach. Section 3 discusses related work. After presenting our requirements for a new query language in Section 4, we present, in Sections 5 and 5.1, our MQL language proposal. Finally, Section 6 concludes this paper and gives some perspectives of this work.

2 CASE STUDY

In this section, we describe the ontologies design process led by the DaFOE platform (a demonstration of this platform is available at http://testcritt.ensma.fr/dafoe/demo/dafoeV1.zip), where our MQL language proposal has been applied. This platform proposes a stepwise approach for building an ontology starting from text.

2.1 Ontology Design in the DaFOE

The DaFOE platform provides a stepwise methodology for building ontologies from text analysis. The first step is dedicated to linguistic analysis (Terminology step) in which users manage linguistic information (terms and relations between terms) extracted with natural language processing tools. Then, a step for terms disambiguation (TerminoOntology step) is performed. Finally, a formalization step (Ontology step) allows users to create *classes* and *properties* of the ontologies. Each step, that is autonomous, has its own model respectively presented in Figure 1, 2 and 3 and mappings are used to establish correspondences between these models.

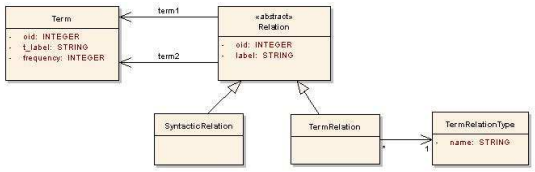


Figure 1: A subset of the Terminology model.

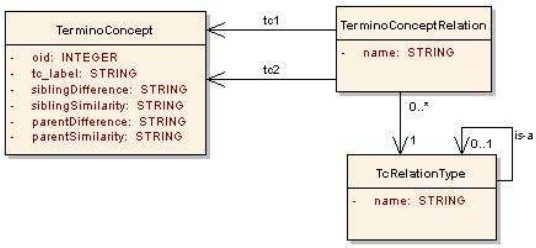


Figure 2: A subset of the TerminoOntology model.

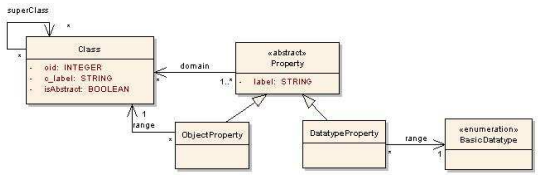


Figure 3: A subset of the Ontology model.

2.2 Persistence of Mappings

In (Téguiak et al., 2012), we argued that model-based databases (MBDB) are well adapted for handling mappings in a database context. In that proposal, we have extended MBDB with a repository for mapping representations as illustrated in Figure 4. In the resulting meta-model (named core meta-model) where models are defined by their entities and their attributes, three main constructors for creating correspondences are available. The first one, called *mLink*, is used to establish correspondences between models. The second one, called *eLink*, allows the user to establish correspondences between entities of models and finally, the *aLink* uses an *expression* to write the target attribute in term of the sources attributes.

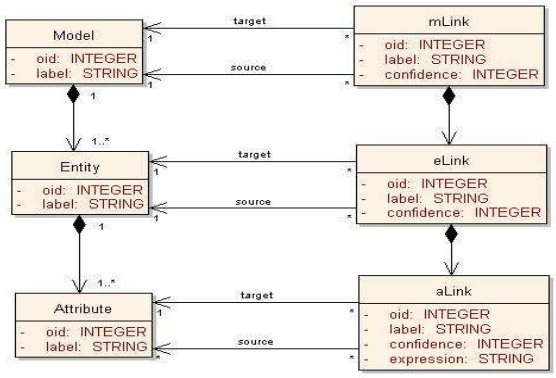


Figure 4: Core metamodel.

2.2.1 Terminology to TerminoOntology Step

Considering both *Terminology* and *TerminoOntology* models, a simplified mapping between these models consists in:

- Creating a *mLink* between the *Terminology* model and the *TerminoOntology* model;

- Creating a *eLink* from the *Term* entity and the *TerminoConcept* entity to express that instances of the *Term* entity will be transformed into instances of the *TerminoConcept* entity;

- Creating a *aLink* expressing that an instance of the *TerminoConcept* entity has the same *label* as the one of its corresponding instances of the *Term* entity, prefixed by *'tc_'*. Another *aLink* expresses that the *rate* of an instance of *TerminoConcept* entity, equals to the *frequency* of the corresponding instance in the *Term* entity divided by 100.

2.2.2 TerminoOntology to Ontology Step

For *TerminoOntology* and *Ontology* models, a simplified mapping consists in:

- Creation a *mLink* between the *TerminoOntology* model and the *Ontology* model;

- In the context of the previous created *mLink* between models, a *eLink* is created between the *TerminoConcept* entity and the *Class* entity to express that instances of the *TerminoConcept* entity will be transformed into instances of the *Class* entity;

- Creating of a *aLink* expressing that an instance of the *Class* entity has the same *label* as the one of its corresponding instance in the *TerminoConcept* entity. Another *aLink* expresses that the *relevance factor* of an instance of *Class* entity, equals to the *rate* of the corresponding instance in the *TerminoConcept* divided by 10.

As an illustration, assume that instances of the Ontology, TerminoOntology and Terminology models are represented by Tables 1, 2 and 3 respectively. Thanks to mappings, a user who queries the *Class* entity of the Ontology model could want to query both TerminoConcept of the TerminoOntology model and Term of Terminology model.

Table 1: Ontology model.

Class			
oid	c_label	relevance	isAbtract
1000	tc_car	0.01	true
1001	tc_wheel	0.002	false
1003	electric_motor	0.04	false

Table 2: TerminoOntology model.

TerminoConcept		
oid	tc_label	rate
600	tc_car	0.1
602	motor	0.08
603	motorcycle	0.8

Table 3: Terminology model.

Term		
oid	t_label	frequency
300	car	1
301	wheel	2
302	bicycle	30

Putting these mappings all together results in the MOF-like database repository (Cf. Figure 5) where M_{i+1}/M_i means that the M_i level is represented as instance of the M_{i+1} level. The meta-schema part is dedicated for managing the core metamodel while schema and instance part a dedicated for managing business models and data respectively.

3 RELATED WORK

Metadata repository systems manage metadata commonly represented as models or meta-models. Such a repository is often equipped with a MOF-based query language ((Lakshmanan et al., 2001), MSQL (Grant et al., 1993), SQL/M (Kelley et al., 1995), OntoQL (Jean et al., 2006), mSQL (Petrov and Nemes, 2008), SparQL (Konstantinos et al., 2010)) that provides capabilities to manipulate both data and meta-data.

Another query language, called mapping oriented query language ((Melnik et al., 2003), (Konstantinos et al., 2010)) provide and explicit representation of mappings between models and offer capabilities to exploit these mappings when querying data. As limitation, these languages do not allow a user to customize the mapping exploitation process. In many cases, the exploitation process is hidden to the user and all the graph of interconnected database is used even if the user wants to use only a sub-part of this graph. Furthermore, in these languages or frameworks, the representation of mappings is static and can not be extended dynamically.

As illustrated in our case study, our proposed database structure (Cf. Figure 5) is a MOF-like database that also handles mappings between models. However, as we will see in the next section, this database is a bit more complex to manage using classical SQL queries. This drawback brings us to design a query language bypassing limitations of languages presented above and that makes easier mappings exploitation using high level operators. So, requirements for such a language are needed.

4 REQUIREMENTS

(Wakeman and Jowett, 1993; Petrov and Nemes, 2008) have investigated requirements for higher-level query languages managing both data and metadata (e.g, models). In this section we introduce new requirements specific to mappings exploitation.

4.1 Handling Complex Queries

Considering the example of Section 2 and assume that a user wants to retrieve, for the ontology model, all classes of the ontology model whose relevance factor is high than 0.01. To achieve this goal, the user can write the following query:

$\mathbf{R}_1$) SELECT c_label, relevance FROM Class WHERE relevance $\geq$ 0.01.

However, if the user also want to retrieve, for each class, the corresponding object in other models

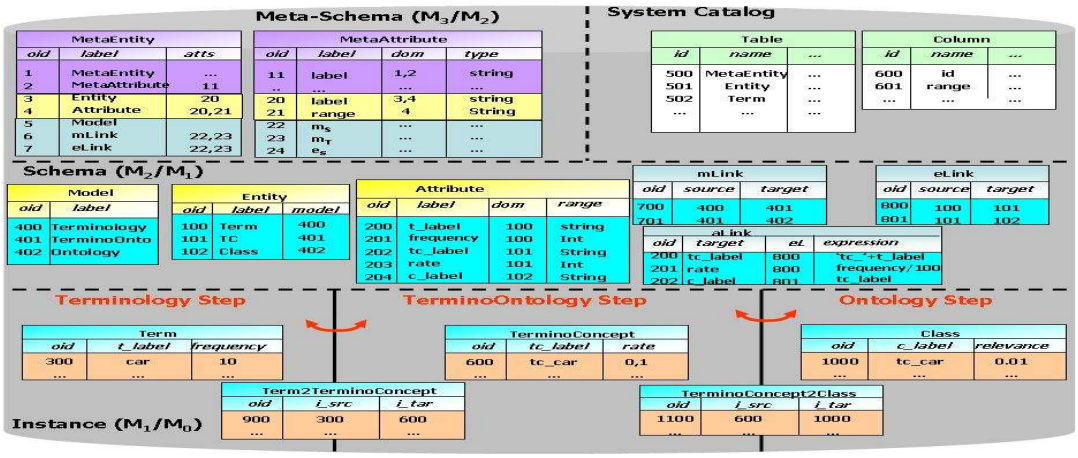


Figure 5: Mapping management in the DaFOEApp project.

mapped to the ontology model, two situations may occur.

On the one hand, if the user knows the mappings characteristics, so he/she can manually write the appropriate following SQL queries:

R$_2$) SELECT tc_label, rate FROM TerminoConcept WHERE rate/10 $\geq$ 0.01

R$_3$) SELECT t_label, frequency FROM Term WHERE (frequency/100)/10 $\geq$ 0.01

R_2 and R_3 queries are translation of the R_1 query on the TerminoOntology and Terminology models respectively according to the mappings between these models.

On the other hand, because mappings characteristics may be evolved dynamically (new mappings may be created while existing one may be deleted or updated, just as in a peer to peer system (Iraklis and Joemon, 2003)), one needs firstly to query mappings repository for characteristics retrieval, and then write the appropriate queries based on these characteristics. Such a query requires to access a repository which represented instances are model and mappings between models. According to the transitivity capability of mappings, this access may raise a syntactic complex query (Cf. Table 4). To simplify, we assume that additional data should be retrieved from the TerminoOntology model.

Table 5 represents the results of the Q_5 query. These results are exploited to generate, for the TerminoOntology model, the query for retrieving data.

As we can observe, the process of unfolding queries on target models is not easy and may become complex if one needs to integrate the complete network of mappings. In this case, the user handles by himself the transitivity capabilities of mappings. A classical approach to deal with this situation consists in writing a query translator. So, the user writes a query (R_1 query for example) and the translator generates queries for target models. This queries generation process is hidden to the user and made implicit. In other words, this approach assumes that the user does not know any mappings characteristics usable to customize the queries generation process.

Table 4: Mapping level queries.

Goals	**Queries**
Q_1) *Retrieve the Ontology model.*	SELECT M.oid FROM Entity E, Model M WHERE E.label= “Class” AND E.model= M.oid
Q_2) *Retrieve mLink where the Ontology model is involved as target*	SELECT mLink.oid FROM mLink WHERE mLink.target in Q_1
Q_3) *Retrieve the entities mapped to Class entity.*	SELECT eLink.source FROM eLink, Entity E WHERE eLink.mL in (Q_2) AND eLink.oid= E.oid AND E.label= “Class”
Q_4) *Retrieve correspondences between entities where the Class entity is involved as target*	SELECT eLink.oid FROM eLink WHERE eLink.mL in (Q_2)
Q_5) *Retrieve mapped entities and mapped attributes (through their expression).*	SELECT E.label,aLink.exp FROM Entity E, Attribute A, aLink WHERE E.oid in (Q_3) AND aLink.eL in (Q_4) AND aLink.target= A.oid AND A.dom= E.oid

Table 5: Mapping level results.

E.label	**aLink.expression**
TerminoConcept	tc_label
TerminoConcept	rate/10

4.2 Handling Mapping Navigation

Considering more closely the second situation of the requirement presented in Section 4.1 where users need to query mappings repository in order to retrieve mappings characteristics. One can ask itself how to handle transitivity with query languages such as SQL for example. This issue refers to the needs to dynamically navigate through the mappings hiding the exploitation of these mappings. So, a policy for a transitive subqueries propagation through chains of arbitrarily huge mapped models is required because these models may contain huge amount of data.

4.3 Providing Persistent Mappings

This requirement refers to the problem of *memory saturation*, that means avoiding loading into central memory big amount of data whose models are mapped together. Indeed, the mappings repository may become very huge and therefore expensive (in response time and memory consumption) for navigation purposes because, new models (says news modeling steps) could be created dynamically according to the needs of a particular user. Thus, a persistent-based approach is required.

5 OUR APPROACH

In this section, we present an overview of the MQL (Mapping Query Language), our mapping-based query language proposal for handling mappings according to previous quoted requirements. This language is highly coupled to the Model Based Database (MBDB) persistence approach presented in (Téguiak et al., 2012). For each part (meta-schema, schema, instance) of the MBDB, the MQL language provides operators to define, manipulate and query its content. Due to space limitation, we only present capibilities for MQL to query instances and mappings together. More details are available in a complete version of our unpublished internal report (Téguiak et al., 2011).

5.1 Instances and Mappings Together

To address the requirements mentioned in Section 4, we propose to extend the classical "SELECT ... FROM ... WHERE ... " query. In other words, our approach is and hybrid one that can be used even if a user knows mappings characteristics or not. As the main purpose of MQL is to facilitate navigation through mappings, we introduced optional statements useful for query propagation in order to get compact syntactic queries. The following statements are exploited in the queries translation process to customize this process.

MATCH. Specify the target models in which the MQL query is propagated at runtime.

FILTER. When propagating a MQL query from a model m_1 to another model m_2, an entity of m_1 may correspond to several entities of m_2. In this case, one may want to restrict the translation so that it applies only to part of these entities. Such a restriction is described using the FILTER clause.

CONFIDENCE. Confidence degrees are often assigned to mappings in order to handle fuzzy mappings. This clause restricts the propagation of the MQL query for the models that satisfy the specified confidence degree. When specified, this clause is used as a threshold to be respected.

With closure. If specified, the propagation of the query is achieved through the mappings repository using the *transitive closure* in the way that, instances are retrieved according to the transitivity of available mappings.

DEPTH. When a MQL query uses the *With closure* clause, it may result in a memory saturation or a bad response time according to the size of the graph of mappings. The DEPTH clause specifies the depth exploration of the graph of mappings. For example, *"DEPTH 4"* means that the MQL query will be propagated transitively on four consecutive mappings at most .

mWHERE. Unlike the classical WHERE clause of a SQL query, the mWHERE clause allows users to specify predicates to filter correspondences. In other words, the mWHERE clause is comparable to a SQL WHERE clause, but it is dedicated to mapping level.

5.2 MQL in Action

Applied to the Ontology model, the mQ_1 query returns data (Cf. Table 6) of the Ontology model (no mapping statement is used). In other words, this query is a classical SQL query. For readability purpose, all the result records are prefixed by the name of its entity.

Table 6: Results of the mQ_1 MQL query.

mQ_1	**Results**
SELECT c_label, relevance FROM Class WHERE relevance $\geq$ 0.01	Class(tc_car, 0.01) Class(electric_motor, 0.04) ...

Applied to the Ontology model, the mQ_2 query returns data (Cf. Table 7) extracted from both the Ontology and the TerminoOntology models (the

MATCH statement has been set to TerminoOntology).

Table 7: Results of the mQ_2 MQL query.

mQ_2	**Results**
SELECT c_label, relevance FROM Class WHERE relevance $\geq$ 0.01 **MATCH** TerminoOntology	Class(tc_car, 0.01) Class(electric_motor, 0.04) TC(motorcycle, 0.8) ...

Applied to the Ontology model, the mQ_3 query returns data (Cf. Table 8) extracted from both Ontology and TerminoOntology models (the MATCH statement for this query has been set to all models using the * symbol). However due to the DEPTH statement, the results are limited to 1 transitive propagation. Only the TerminoOntology model is reachable from the Ontology model with 1 propagation.

Table 8: Results of the mQ_3 MQL query.

mQ_3	**Results**
SELECT c_label, relevance FROM Class WHERE relevance $\geq$ 0.01 **MATCH** * **FILTER** * **DEPTH** 1 **With closure**	Class(tc_car, 0.01) Class(electric_motor, 0.04) TC(motorcycle, 0.8) ...

Applied to the Ontology model, the mQ_4 query returns data (Cf. Table 9) extracted from both the Ontology, TerminoOntology and Terminology models. Indeed, thanks to the * symbol of the MATCH statement and with no DEPTH limitation, mQ_4 is propagated to any model transitively reachable from the Ontology model.

Table 9: Results of the mQ_4 MQL query.

mQ_4	**Results**
SELECT c_label, relevance FROM Class WHERE relevance $\geq$ 0.01 **MATCH** * **FILTER** * **With closure**	Class(tc_car, 0.01) Class(electric_motor, 0.04) TC(motorcycle, 0.8) Term(bicycle, 30) ...

6 CONCLUSIONS

In this paper, we have presented a mapping-based query language called MQL that makes easier querying data thanks to available mappings between models. This language has a knowledge part based on a core metamodel dedicated models and mappings representation. One of the main features of our approach is that this knowledge part can be extended by evolving the core metamodel. MQL has been implemented for model-based databases, where both instance, metamodel and metametamodel level are persisted in a single database. As perspective of this work, we are working on the definition of a benchmarking scenario for improving performance of our approach.

REFERENCES

Bernstein, P. A. (2003). Applying model management to classical meta data problems. In *CIDR*.

Bouquet, P., Giunchiglia, F., Harmelen, F. V., Serafini, L., and Stuckenschmidt, H. (2003). C-owl: Contextualizing ontologies. In *ACM SIGIR'03*, pages 164–179.

Grant, J., Litwin, W., Roussopoulos, N., and Sellis, T. (1993). Query languages for relational multidatabases. In *VLDB*.

Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, S., Grosof, B., and Dean., M. (2004). Swrl: a semantic web rule language combining owl and ruleml.

Iraklis, K. and Joemon, J. (2003). An architecture for peer-to-peer information retrieval. In *ACM SIGIR'03*, pages 401–402.

Jean, S., Ait-Ameur, Y., and Pierra, G. (2006). Querying ontology based database. the ontoql proposal. In *SEKE*, pages 166–171.

Jouault, F., Allilaire, F., Bézivin, J., and Kurtev, I. (2008). Atl: a model transformation tool. In *Science of Computer Programming*, pages 31–39.

Kelley, W., Gala, S., Kim, W., Reyes, T., and Graham, B. (1995). Schema architecture of the unisql/m multidatabase system. In *Modern Database Systems*.

Konstantinos, M., Nektarios, G., Nikos, B., and Stavros, C. (2010). Ontology mapping and sparql rewriting for querying federated rdf data sources. In *OnTheMove*, pages 1108–1117.

Lakshmanan, L., Sadri, F., and Subramanian, S. N. (2001). Schemasql: An extension to sql for multidatabase interoperability. In *JTDS*.

Melnik, S., Rahm, E., and Bernstein, P. A. (2003). Developing metadata-intensive applications with rondo. In *Journal of Semantic Web*, pages 47–74.

Moha, N., Sen, S., Faucher, C., Barais, O., and Jézéquel, J.-M. (2010). Evaluation of kermeta for solving graph-based problems. In *JSTT*.

Petrov, I. and Nemes, G. (2008). A query language for mof repository systems. In *OnTheMove*, pages 354–373.

Téguiak, V., Ait-Ameur, Y., and Sardet, E. (2012). Use of persistent meta-modeling systems to handle mappings for ontology design. In *MOPAS*, page To appear.

Téguiak, V., Ait-Ameur, Y., Sardet, E., and Bellatreche, L. (2011). MQL: an extension of SQL for mappings manipulation. Technical report, LIAS.

Wakeman, L. and Jowett, J. (1993). *PCTE: the standard for open repositories*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

DISEArch
A Strategy for Searching Electronic Medical Health Records

David Elias Peña Clavijo, Alexandra Pomares Quimbaya and Rafael A. Gonzalez
Departamento de Ingeniería de Sistemas, Pontificia Universidad Javeriana, Bogotá, Colombia
{*pena-david, pomares, ragonzalez*}*@javeriana.edu.co*

Keywords: Medical Health Records, Health Data Mining, Text Mining.

Abstract: This paper proposes DISEArch, a novel strategy for searching electronic health records (EHR) of patients that have a specific disease. The objective of DISEArch is to enhance research activities on disease analysis allowing researchers to describe the disease they are interested on, and providing them the EHRs that best match their description. Its principle is to improve the precision of searching EHRs combining the analysis of structured attributes with the analysis of narrative text attributes producing a semantic ranking of EHRs with respect to a given disease. DISEArch is useful in medical systems where the information about the primary diagnosis of patients may be hidden in narrative text hindering the automatic detection of relevant records for clinical studies.

1 INTRODUCTION

Electronic health records (EHR) are a rich source of knowledge for medical research. However, their use has been limited due to the fact that important information is stored in narrative texts, intended for humans, difficult to search and analyse automatically. One of the requirements of medical research is to find the EHRs of patients that have been diagnosed with a specific disease. This task that should be easily done using classical queries (e.g. SQL) is very time-consuming because diagnosis is frequently hidden in the text (e.g. medical notes), hindering the possibility of automatically detecting relevant records and requiring the participation of an expert. Previous work on EHR systems propose strategies to improve automatic processing of narrative text in EHRs using information retrieval and data mining techniques (Han et al., 2006)(Zhou et al., 2005).

This work proposes DISEArch, a strategy for searching in EHRs those records that match a specific diagnosis, regardless of the kind of attribute (structured or non structured) that contains the information. DISEArch is composed of three phases. The first extracts the set of patient records from the medical health system. The second phase applies classical queries on structured attributes and text mining techniques over narrative text. Finally, it ranks the records by applying a semantic distance function with respect to the given disease description. DISEArch has been useful in reducing the time required for searching medical records. The structure of the paper is as follows. Section 2 presents the analysis of related works on narrative text and medical record analysis. Section 3 presents DISEArch, including its main components. Section 4 presents the main aspects of the prototype of DISEArch and the evaluation of its behaviour. Finally, Section 5 concludes this paper.

2 RELATED WORKS

Figure 1 presents a taxonomy of existing works related to text mining from EHRs. The initial categories offered are general approaches, algorithms, tools and scope. **General approaches** refer to three main bodies of work: information retrieval, natural language processing (NLP) and text/data mining. **Algorithms** are further divided into those aimed at data preparation and those aimed at data detection or classification. **Tools** offers a list of some available software tools which may support the process of text mining from EHRs. Finally, **scope** centers on work aimed at analyzing negated sentences, as opposed to work which is more generic. In our taxonomy (Figure 1) general approaches start with **text mining**, which consists of analyzing (portions of) documents typically made up of natural language. Its purpose is to uncover patterns, trends and relationships between words, meanings, terms or concepts (Spasic et al.,

Table 1: Comparative literature review.

Papers	Algorithms								Tools				Approach		Field		
	Preparation			Classification/Detection													
	Data transform.	Expert Tagging	UMLS Tagging	Regexp	Proposed	Bayesian networks	Decision trees	Hidden Markov M.	GATE	Link G. Parser	EMERSE	Own	Negated sentences	Generic	NLP	Data/text mining	IR
NegEx (Chapman et al., 2001)	√		√	√								√	√		√		
Context-Sensitive (Averbuch et al., 2004)	√		√		√	√	√					√	√				√
Negation-Recognition (Rokach et al., 2008)	√		√	√	√		√					√	√	√		√	
Genereric Extraction (Han et al., 2006)			√						√	√				√	√	√	
DM & CBR (Huang et al., 2007)	√	√			√		√					√		√		√	
Text mining in biomedicine (Spasic et al., 2005)	√		√											√		√	√
Diabetic DW (Breault et al., 2002)	√	√					√					√		√		√	
EMERSE (Seyfried et al., 2009)		√									√			√			√
ABN (Antal et al., 2001)						√						√		√			√
Decision-Making (Claster et al., 2008)	√	√			√		√					√		√		√	
Semi-structured data to knowledge (Zhou et al., 2005)	√		√				√		√	√				√	√	√	√
HMM & LSA (Ginter et al., 2009)	√				√			√				√		√		√	√

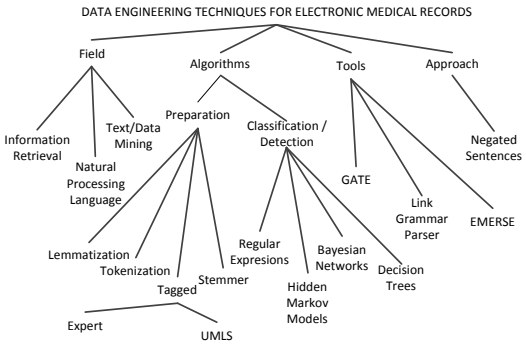


Figure 1: Taxonomy of EHRs data techniques.

2005). The second general approach is related to the field of **information retrieval (IR)**(Manning et al., 2008). The last general approach deemed useful for our purposes is **natural language processing (NLP)**, which refers to the recognition and use of information expressed in human language through computer-based systems (Hotho et al., 2005).

Our taxonomy continues by classifying specific types of algorithms that can be used as part of the three general approaches, depending on the stage of the process. With regards to **data preparation** we include four kinds of algorithms that prove useful in preparing unstructured health records prior to analysis.**Tokenization** is the process through which a flow of text is divided into segments. **Lemmatization** refers to a method in which verbs are transformed into their base form or nouns into their singular form. **Stemming** is used for removing irrelevant terms from the text. The last type of algorithm is **tagging**, which involves the interaction with a user that labels the text. In the case of EHRs, these tags are typically part of a controlled vocabulary, such as UMLS (USNLM, 2011). The second general type of algorithm in our taxonomy is grouped under **classification and detection**. **Decision trees** are a common part of the tool-belt for data mining and are useful in classifying conditions hierarchically such that a final decision is reached when a path can be followed from the root to one of its leaves. **Bayesian networks** are a powerful tool which is implemented through acyclic directed graphs that contain a set of nodes, each representing a random variable. (Antal et al., 2001).

A related kind of algorithm is called **hidden Markov model**, which represents a statistical model for linear problems and is widely used for speech recognition (Ginter et al., 2009). The third branch of related works is focused on the **tools** that support data and text mining in EHRs. Among these we find **GATE** (*General Architecture for text engineering*) which offers a general open source framework for developing or deploying software components for text engineering (Cunningham et al., 2011). Another tool is the **Link Grammar Parser**, which syntactically analyses text based on link grammar. **EMERSE** (*The Electronic Medical Record Search Engine*) (Hanauer, 2006) is specifically aimed at EHRs, acting as a search engine for free text inside such records. Using the taxonomy proposed above, Table 1 presents a comparative review of relevant literature around data / text mining in EHRs.

3 DISEArch STRATEGY

The strategies to examine narrative texts described in Section 2, provide a broad knowledge base to address the analysis of unstructured text inside EHRs. How-

ever, they are focused exclusively on the analysis of narrative text without taking into account the dependencies on other (structured) fields within the record. This work explores the combination of structured and narrative analysis to enhance the precision on the selection of relevant records for medical research. The strategy proposed in this section, called DISEArch, allows researchers to describe the disease they are interested in and provides them the set of health records that better match their description.

3.1 Phases

The process of analyzing health records in DISEArch is divided into the phases illustrated in Figure 2. In

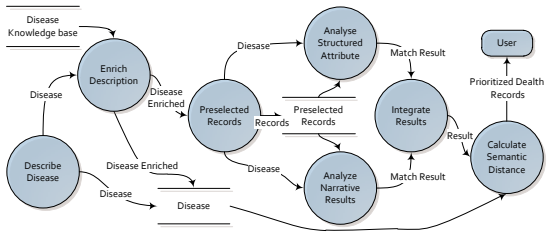


Figure 2: DISEArch Strategy.

the first phase DISEArch allows medical researchers to describe the disease they are interested in using a template. This template includes formal and informal aspects of the disease, including the scientific name, the informal name, the tests that are typically used to diagnose the disease, and the symptoms of the disease. Once the disease is described, each of the fields are enriched using knowledge about the disease. The enrichment is made using a knowledge base created in OWL (Bechhofer et al., 2009) using a MeSH based thesaurus[1].

The set of preliminary records are stored in a local database where DISEArch executes the analysis. The goal is to find the elements described in the disease within each one of the selected records, regardless of whether they are contained in an structured or a narrative text field. After this analysis, a score is given to each one and then prioritized.

3.2 Search Process

The disease description is divided into n subgroups S that are composed by m literals L (Definition 1). An example of subgroup is *Disease Name* and its literals are *Scientific name*, *Formal name*, *Informal name*, *Synonyms* and *Acronyms*. Similarly, health records M are divided into a set of p structured attributes S and q narrative text attributes T (Definition 2).

Definition 1 . ***Disease Description.*** *A Disease definition D is composed of a set of subgroups $S = \{s_1, s_2, ..., s_n\}$ that describe the main characteristics of the disease. Each subgroup s_i is specialized in a view of the disease and is composed of a set of literals $L(s_i) = \{l_{i1}, l_{i2}, ..., l_{im}\}$ where l_{ij} represents a fixed value for an atomic characteristic of the disease.*

Definition 2 . ***Health Record.*** *A health record M is composed of a set of structured attributes $C = \{c_1, c_2, ..., c_p\}$ whose domain of values is discrete and a set of narrative text attributes $T = \{t_1, t_2, ..., t_q\}$ whose domain is a natural language text.*

The goal of the search process is to detect within C_k and T_k of a record M_k, the value of each one of the literals l_{ij}. If the value of the literal l_{ij} is found in at least one attribute of the record M_k the value of the search process is changed to one (1), otherwise it is left at zero (0).

DISEArch contains two search functions in charge of detecting the occurrence of literal values into health records; the first one detects the value of a literal in structured attributes C and the second one searches within narrative text attributes T. Searching structured attributes is straightforward using classic sql queries. On the contrary, searching within narrative texts includes a previous preparation of texts and analysis that is detailed in Algorithm 1.

Algorithm 1: Narrative text search function.

Require: Record narrative text attributes
Ensure: Record search result
i,j,result $\Leftarrow 0$
for all textAttribute in record **do**
  $p \Leftarrow$ prepareText(textAttribute)
  for all subgroup in diseaseTemplate **do**
    for all literal in subgroup **do**
      ortResult $\Leftarrow$ searchValue(p).
      if $ortResult = 1$ **then**
        semResult $\Leftarrow$ searchContext(p)
        if $semResult = 1$ **then**
          result $\Leftarrow 1$
        end if
      end if
      $j \Leftarrow j + 1$
    end for
    $i \Leftarrow i + 1$
  end for
  $recordResult[i, j] \Leftarrow$ result
end for
return recordResult

[1] http://www.nlm.nih.gov/

At the end of the search process the output is the score for each literal as the matrix *Res* illustrates and the number of hits for each one of the literals. The columns of *Res* represent the literals of each subgroup and the rows the health records.

$$Res = \begin{array}{c} \\ M_1 \\ M_2 \\ \vdots \\ M_k \\ M_r \end{array} \begin{array}{c} \begin{array}{cccccc} l_{11} & l_{12}| & \cdots & |l_{n1} & l_{n2} & l_{n3} \end{array} \\ \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \\ 0 & 0 & \cdots & 1 & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 1 & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix} \end{array}$$

3.3 Integration Process

The integration process is in charge of representing the results of the search process taking into account the semantics of subgroups. This integration includes the results provided by structured and non-structured search functions. For doing this new representation the process takes into account the Definitions 3, 4, 5.

Definition 3. ***Subgroup Intensity.*** *The intensity I of a subgroup of literals s_i is the normalised percentage of matched literals within the health record M_j. If a literal has multiple possible values (e.g. multiple acronyms) each value is considered a literal (e.g. Acronym 1, Acronym 2, etc.).*

Definition 4. ***Subgroup Utility.*** *The utility U of a subgroup of literals s_i is a percentage value of the importance it has in identifying the disease diagnosed in a health record assuming that all the values of the literals are positive.*

Definition 5. ***Subgroup Level of Hits.*** *The number of hits H of a subgroup s_i is the normalised number of times that literal values were matched within M_j.*

In order to calculate the utility of each subgroup DISEArch uses a classical method of multi-criteria decision analysis where each subgroup is evaluated on multiple criteria by experts and the utility is "the average specified in terms of normalised weightings for each criterion, as well as normalised scores for all options relative to each of the criteria" (Keeney and Raiffa, 1976). The number of hits is used as an optional calibration value that takes into account the number of times that literal values are found in a health record. The intention is to assign a higher weight to records that have the same literal multiple times. This value is optional because for some subgroups it is important, but for others it is not. At the end of the integration process an integration matrix is generated (see Matrix *I*). The values of the literal in each subgroup are described in the following columns:

1. s_i is 1 if at least one of the literals of the subgroup was found in the health record M_k, otherwise its value is 0.
2. s_i^I is the intensity of the subgroup.
3. s_i^U is the utility of the subgroup to detect the disease.
4. s_i^H is the number of hits of the subgroup. This columns is optional.

$$I = \begin{array}{c} \\ M_1 \\ M_2 \\ \vdots \\ M_k \\ M_r \end{array} \begin{array}{c} \begin{array}{cccccccc} s_1 & s_1^I & s_1^U & s_1^H| & \cdots & |s_n & s_n^I & s_n^U \end{array} \\ \begin{pmatrix} 1 & 1 & 0.6 & 1 & \cdots & 1 & 1 & 0.4 \\ 0 & 0 & 0.6 & 0 & \cdots & 1 & 0.66 & 0.4 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0.5 & 0.6 & 0.5 & \cdots & 1 & 0.33 & 0.4 \\ 1 & 0.5 & 0.6 & 0.7 & \cdots & 0 & 0 & 0.4 \end{pmatrix} \end{array}$$

The distance function between the disease description *D* and each one of the analyzed health records *M* is calculated using a distance function (e.g. Euclidean, Manhattan). The disease description as well as each record are represented in a n-space (see Function 1 and 2, respectively), where n is the number of subgroups.

$$D = (p_{s_1}, p_{s_2}, ..., p_{s_n}). \quad (1)$$

$$M = (q_{s_1}, q_{s_2}, ..., q_{s_n}). \quad (2)$$

The value of each point p is equivalent to s_i^U and the value of each point q is calculated using the product of $s_i^U \times s_i^I \times s_i^H$. The record with the shortest distance is the first one in the prioritized list and so on.

4 IMPLEMENTATION AND VALIDATION

In order to evaluate DISEArch and validate its improvement on the selection of the most relevant health records given a disease, a prototype has been constructed and used to evaluate its precision and recall. This section presents the main results obtained during this evaluation.

4.1 Prototype

For evaluating the behaviour of DISEArch we developed the components presented in Figure 3. These components are written in Java. The template of the disease can be filled using the GUI or directly using an XML file. The Dictionary Manager handles the knowledge base that allows the enrichment of the description of the disease. The knowledge base is implemented in OWL (Bechhofer et al., 2009). The Extraction Manager is in charge of the extraction and initial

preprocessing of medical records from the EHR system. This component is parametrized according to the characteristics of the system and extracts the records according to the definition of initial parameters, such as date of admission, gender or age of patients. Persistent Manager and the DataStore Manager store the required tables to perform the search process inside a database. These tables are used to create a single view with all the unstructured and structured data. The component Text Mining is the core of the anal-

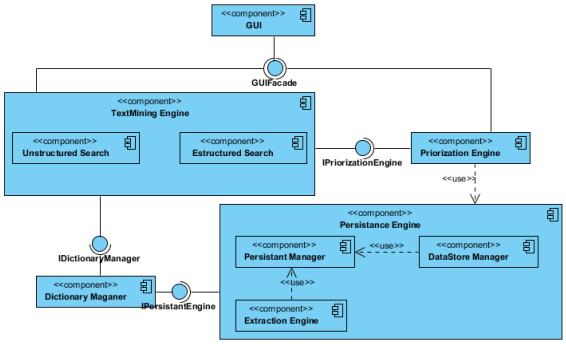


Figure 3: DISEArch component diagram.

ysis and implements Stemming using Porter Stemmer algorithm, simple string tokenisation, sentence splitting, POS tagging using Probabilistic Part-of-Speech Tagging Using Decision Trees (Schmid, 1994) for annotating text with part-of-speech and lemma information and finally gazeteer lookup using regular expressions. This component has a coordinator that calls each of the search engines. The Narrative Search Engine is in charge of the analysis of natural language and was developed using the GATE API (Cunningham et al., 2011). This API enables the inclusion of all the language processing functionality within DISEArch. In addition, we use Treetagger (Schmid, 1994), a Pearl implementation which provides tokenization and Part of the Speech tagger. The Structured Search Engine is in charge of searching the disease over the structured attributes. Finally the Integrator component integrates the results using the semantic rules and prioritizes the set of records.

4.2 Experiment Context and Results

Pulmonary Embolism (EP) was chosen to test DISEArch. A medical expert provided the subgroups and literals that describe it. Preliminary selection parameters for EHRs were defined: patients over 18 years old and records created between 2009-2011. One key item to obtain precision and recall was the prioritization process that was explained in Section 3. The results and their associated medical records were clustered according to their relevance (Lowly prioritized, Mildly prioritized and Highly prioritized medical records). The obtained results with DISEArch were 250 medical records, which correspond to records with at least one positive literal w.r.t the disease description. From these records, the prioritization process classified 30 as high, 52 as medium and 168 as low, according to the distance function.
In order to validate the precision and recall of DISEArch a medical expert analysed manually the records detecting 112 EP positive medical records. From these results DISEArch obtained 30 as high, 50 as medium and 32 as low. The precision and recall are presented in Figure 4. As expected the precision and recall is high for high and medium positive records. The low precision of the Low group is the consequence of the inclusion of records that contain few literals in common with the disease template.

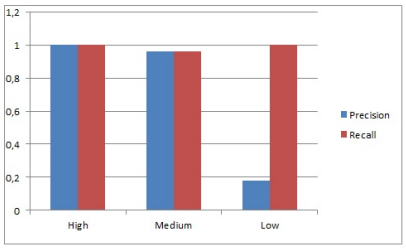


Figure 4: Precision and recall of DISEArch.

5 CONCLUSIONS

The DISEArch strategy presented in this paper enables medical researchers to identify those EHRs that include the diagnosis of a specific disease. This time-consuming and expert-dependent task can be supported by DISEArch through specific rules for identifying diseases and weights to prioritize the selected records, leaving the expert task to one of review and acceptance, rather than search and retrieval. DISEArch goes beyond classical text mining because it uses unstructured text in medical records as well as related structured fields to enrich the final results. From our first tests we found that, although the non-prioritized results are already helpful and accurate (as compared to expert selected records), prioritization still plays an important role in the classification of medical records because it adds precision and contributes to the review process by presenting the records in terms of how close they are to the disease template.

ACKNOWLEDGEMENTS

This work was supported by the project *"Identificación semiautomática de pacientes con enfermedades crónicas a partir de la exploración retrospectiva de las historias clínicas electrónicas registradas en el sistema SAHI del Hospital San Ignacio"* made by Pontificia Universidad Javeriana and Hospital Universitario San Ignacio.

REFERENCES

Antal, P., de Moor, B., and Mészáros, T. (2001). Annotated bayesian networks: A tool to integrate textual and probabilistic medical knowledge. In *Proc. of the 14th IEEE Symp. on Computer-Based Medical Systems*, CBMS '01.

Averbuch, M., Karson, T. H., Ben-Ami, O., and Rokach, L. (2004). Context-sensitive medical information retrieval. *Studies in health technology and informatics*.

Bechhofer, S., van Harmelen, F., Hendler, J., and Horrocks, I. (2009). "owl web ontology language reference". Technical report, W3C.

Breault, J. L., Goodall, C. R., and Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine*.

Chapman, W. W., Bridewell, W., Hanbury, P., and Cooper (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. of Biomedical Informatics*.

Claster, W., Shanmuganathan, S., and Ghotbi, N. (2008). Text mining of medical records for radiodiagnostic decision-making. *JCP*.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., and Aswani, N. (2011). *Text Processing with GATE (Version 6)*.

Ginter, F., Suominen, H., Pyysalo, S., and Salakoski, T. (2009). Combining hidden markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. *I. J. Medical Informatics*.

Han, H., Choi, Y., Choi, Y. M., Zhou, X., and Brooks, A. D. (2006). A generic framework: From clinical notes to electronic medical records. *Computer-Based Medical Systems, IEEE Symp*.

Hanauer, D. A. (2006). Emerse: The electronic medical record search engine. *AMIA A. Symp Proc*.

Hotho, A., Nürnberger, A., and Paass, G. (2005). A brief survey of text mining. *LDV Forum*.

Huang, M.-J., Chen, M.-Y., and Lee (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Syst. Appl.*

Keeney, R. and Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. J. Wiley, New York.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*.

Rokach, L., Romano, R., and Maimon, O. (2008). Negation recognition in medical narrative reports. *I. R.*

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Seyfried, L., Hanauer, D. A., Nease, D., and Albeiruti (2009). Enhanced identification of eligibility for depression research using an electronic medical record search engine. *Inter. J. of Medical Informatics*.

Spasic, I., Ananiadou, S., McNaught, J., and Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*.

USNLM (2011). Unified medical language system® (umls®). http://www.nlm.nih.gov/research/umls/. Noviembre 25, 2011.

Zhou, X., Han, H., Chankai, I., Prestrud, A. A., and Brooks, A. D. (2005). Converting semi-structured clinical medical records into information and knowledge. In *Proc. of the 21st Inter. C. on Data Eng. WS*.

Adaptive Data Distribution for Collaboration

Luis Guillermo Torres-Ribero and Alexandra Pomares Quimbaya
Systems Engineering Department, Pontificia Universidad Javeriana, Bogotá, Colombia
{*luis-torres, pomares*}*@javeriana.edu.co*

Keywords: Mobile Collaboration, Dissemination Adaptation, Data Distribution.

Abstract: This paper presents an adaptive data distribution model for mobile collaborative applications called *ADDOCO*. Its goal is to adapt the way data is disseminated in environments where Internet is not always available to create a collaboration network. In this kind of environments, ADDOCO allows to send and retrieve information that otherwise would be unavailable. The dynamic information dissemination model of *ADDOCO* was tested in a collaboration application using smartphones proving its utility to enhance the distribution of the required information.

1 INTRODUCTION

The increasing number of Internet-capable devices makes remote collaboration and data storage possible by enhancing the interaction with constant communication and the possibility of sharing resources and information. Storage services like iCloud, Dropbox, Box, Ubuntu One, etc. are internet based storage services that rely on the network to handle the information and the resources they manage. This kind of services allows the user the possibility of accessing his information and resources in any place at any time.

With the increase of mobile devices, these on-line services are an adequate solution for data storage and processing, overcoming some of the limitations mobile devices have. However, all of these services require an Internet connection to work properly. In a scenario where no Internet connection is available, a lot of the above mentioned services would be immediately affected by not being able to work correctly and most of them will need to wait for a new Internet connection available in order to work at all. Lack of Internet can be found in developing countries where Internet coverage is not as widespread as in other countries; even in developed countries in some conditions (e.g. high mountain roads, isolated places, etc.) might have no Internet connection whatsoever, diminishing the possibility of remote collaboration, even when the devices have technologies that make them capable of making a network of their own and collaborate. For example if there is a landslide in a mountain road where no mobile Internet is available, a network can be created with nearby smarthphone users in order to send information on the landslide, e.g. photos, location, number of injured people if any, etc. and get adequate help from nearby entities. There are tools that allow people to interact while being on a mobile environment and allow users to share information between them. However, these services rely on having an active Internet connection in order to be able to spread notifications among the users.

The requirement of having Internet independent collaborative networks motivates the creation of this work. This paper presents *ADDOCO*: an adaptive data distribution framework that supports collaborative mobile environments making a dynamic transition on the data dissemination method they use to communicate and allowing to change dynamically the entity that plays a role in a collaborative task. These dynamic properties are based on the context of the entities, taking into account their abilities, location, collaboration phases and other relevant contextual information. The structure of this paper is as follows: Section 2 describes some of the basic concepts around data dissemination; then Section 3 shows *ADDOCO*, a framework that considers user context and collaboration phase in order to dynamically disseminate data. Section 4 shows related works and their contribution to this work, afterwards Section 5 shows a prototype of *ADDOCO* and its functional evaluation. Finally, Section 6 shows the conclusions and future works.

2 PRELIMINARY CONCEPTS

In order to have an appropriate understanding of the problem context, the key elements involved in this

work are introduced in this section.

Contextual information is the information that considers environmental elements, location description and interaction of people, among other characteristics, that help defining a complete scenario (Bellavista et al., 2013). Contextual information can be very wide and may take into account a lot of elements in order to correctly describe the particular scenario in which some action is taken. In this particular case, we restrain contextual information as a series of elements that are relevant to the scenarios in which the presented framework will be used, such as medical contexts, emergency management contexts, etc. There are mainly two contexts that are considered in this work: the Collaborative Context and the Mobile context.

The **Collaborative context** determines in a logical way the elements involved in a scenario. It includes the elements that will determine whether a person is available to collaborate and the different roles, which that particular person could assume during a collaborative scenario. For example, if a person is a medical doctor and its willing to collaborate during an accident, the collaboration context for that person will include that he is willing to assume the role of a doctor if it is required.

The **Mobile context** determines the involved elements in a physical way, this is, considering elements that can be interpreted as physical signals or data; for example, if there is any Internet connection available, the speed at which the user is moving, etc. Location is an important contextual element that is present in both Collaboration Context and Mobile Context. However, location is expressed differently in each context. In the Collaboration Context location will be represented as a Hierarchical Location, which describes location as a topology or symbolic place, e.g. a room X inside building Y (Prayogi et al., 2007)(Zhang et al., 2006). In the Mobile Context location is represented as a Cartesian Location which describes locations as a set of coordinates or GPS assisted geometric calculations e.g. 04 degrees 00' N and 63 degrees 00' W. Other important elements are Participants, which collaborate sending and receiving information and executing actions according to their knowledge and role.

It is important to notice that Participants are the main elements in the model because they are mobile entities that share and store information. Additionally, a Participant is able to act as a bridge between two participants. Each participant will have a set of roles that he is willing to assume.

For example, in the landslide scenario, if there were injured people and a doctor is quickly required, *Participant A* could look for the Doctor role in participants nearby. If a nearby *Participant B* has in his set of roles the Doctor role, then he will be asked to assume that role in healing the injured.

Each participant has relevant data that must be distributed; to achieve this different variables must be taken into account, such as, the availability of Internet connection and the location, both physical and logical. For example, in our landslide scenario, information about the time when the landslide occurred, the location where it happened and the number of injured people must be distributed to Participants to whom that information is relevant (e.g Firemen, Doctors, Road Safety Department, etc.)

Since these two variables are dynamic, this is, are prone to change rapidly, the data distribution method must also by dynamic. In order to address this issue, different data dissemination methods have been selected and are the basis of the dynamic transition of data distribution according to the context of the participants. For example, if in the landslide scenario, the accident occurred in a road and the people who are driving through that road do not know each other, the first step in the dissemination of data would be to ask if a required role is nearby. If, on the contrary, people already know each other along with role information, a communication protocol can be established and a dynamic communication protocol with a basic structure will be generated according to the scenario needs. There are different data dissemination models (Bellavista et al., 2013) that are used differently according to the requirements of the environment that uses them. Since a highly-dynamic environment is being evaluated, then dissemination methods must also be highly-dynamic.

Some of the data dissemination models are: i) Sensor Direct Access dissemination: Distributing the data directly to a specific place, ii) Flooding-Based dissemination: Distributing the data to a specific group, iii) Gossip-Based dissemination: Distributing data randomly among available nodes and iv) Selection Based dissemination: Distributing data creating a backbone of nodes.

The next section will present ADDOCO, a framework that dynamically changes data dissemination methods in mobile environments according to contextual information and available participants.

3 ADDOCO

Due to the requirements of data sharing in highly- dynamic environments, this is, mobile and distributed environments; ADDOCO is created as a framework that aids in the distribution of data between partici-

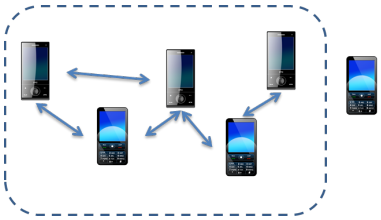


Figure 1: Smartphones creating a smartphone-network.

pants among a collaborative process. To achieve this, ADDOCO adapts, according to the context, the model used for data distribution between the participants involved in a collaboration.

3.1 Distribution Principle

When mobile devices are connected to the internet they can access on-line services that allow them to communicate, obtain and even process information. Some mobile devices nowadays are smartphones, defined as *"a mobile phone that can be used as a small computer"* by the Cambridge dictionary and are capable of doing a lot of data processing. Smartphones take more advantage of on-line services, by having synchronized contacts, configuration files, etc. Most of them have the ability to create networks between them in order to share information and communicate. It is possible to think in a smartphone network that would act like a small scale Internet, with smartphones providing services to other smartphones while communicating and sharing data. This capacity of making a network between them can be very useful when collaboration is required, but there is no Internet connection that allows smartphones to do so. In a scenario where no Internet is available the smartphones could create a network between them, with no guarantee that all of the smartphones connected to the internet will be inside the smartphone network due to technical restrictions like location, i.e. too far to be included in the network, or permissions, i.e. the device has not allowed smartphone networking, as seen in Figure 1. Within the smartphone network there may be smartphones that act as a bridge connecting two smartphones that are within its range, but not within the range of each other. For example, if three devices A,B and C are making a network and the device A needs to send some data to C, but C is out of range, then B, who is in range of both A and C will act as a bridge transmitting information between both A and C. This scenario is illustrated in Figure 2. Taking advantage of this possibility, different dissemination methods can be dynamically selected according to the context. For example, if a landslide occurs in a road where no internet access or mobile

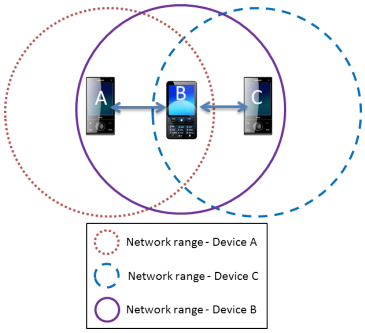


Figure 2: Smartphone Network Bridge.

network is available, someone affected by the landslide could start making a network with nearby drivers in order to send information on the landslide like pictures, location, if there were accidents, etc. In this scenario the device of the user will look for other devices and send information, either randomly to a group of those nearby users using gossip-based dissemination, or selecting a range to send to all users within that range using flooding-based dissemination. When a pattern has been established, the dissemination method changes to a selection based method, creating a backbone based on the interactions of each participant.

3.2 Dynamic Dissemination Process

In order to describe the dynamic dissemination process some basic concepts must be formally defined. A *Participant* has a physical *Location*. Note that the location here is represented as a Cartesian coordinate in order to be interpreted appropriately by the device. The *Participant* also has a set of *Permission* that allows him to request, obtain and manipulate *Data* within the device. It is important to remember that *Data* will be stored as closely as possible to the node that generated it. Therefore the *Data* modelled in this context is not a replicated copy of *Data* available someplace else, but represents the *Data* that a mobile device is responsible for creating, updating and disseminating. Note that the Mobile context also has a *Connection Status* component. This component has information regarding its environment, for example if the *Participant* is connected to a network, it must search within near devices in order to perform a collaborative task, and will aid in selecting the best dissemination method that a requested *Data* needs to follow in order to arrive at its destination.This contextual information is a key element used to dynamically adapt the dissemination method. The "dissemination service" algorithm is described in Algorithm 1. In this algorithm a directory is first used in order to obtain the *dataHolder*, this is, the entity responsible for hand-

Algorithm 1: *ADDOCO* Dissemination Process.

Require: Data requirements, Access Permissions specified, Max number of jumps.
Ensure: Data obtained or stored
1: nJump ⇐ Max number of jumps
2: **if** Directory is active **then**
3: dataHolder ⇐ **call** retrieveDataHolder **with** Task,Location,Role
4: **else**
5: **for all** Entity in nearEntities **do**
6: **if** Entity ≠ dataHolder and nJump > 0 **then**
7: dataHolder ⇐ **call** Dissemination Process **with** Entity,Task,Location,Role, (nJumps-1)
8: **end if**
9: **end for**
10: **end if**
11: **return** dataHolder

ling the data requested. A number of jumps are specified (1) in order to know how many neighbours must be queried to try to obtain the requested data. If there is an active *Directory* with the information of the data handler (2) the *dataHolder* is then acquired. If the Directory is not available then the query is made to near entities. If a near entity does not have the data, it will recursively ask its neighbours for the *dataHolder*, having one jump less than the original caller (7) until it reaches zero (6). The element *nearEntities* are dynamically located according to both contextual and collaborative information according to the messaging protocol if: a) the collaboration model relies on a centralized role and someone who plays that role is nearby, most likely a sensor direct access will be used. b) the collaboration model relies mainly on message or event interchange in a publisher-subscriber way, the selected dissemination method would be selection based. c) the collaborative model relies on roles, but there are not many entities that play the role in range a flooding based dissemination may be adequate, and d) There are some of the required roles nearby a gossip based dissemination could obtain the data.

3.3 ADDOCO Architecture

The main components of the architecture are described in Figure 3. The **BPMN parser** component is in charge of obtaining a BPMN model described in XPDL (XML Process Definition Language) and interpreting them so that both; the mobile context and the collaboration context obtain the data in their own terms in order to correctly execute the tasks described. The **Mobile Context Manager** component handles the data, location and connectivity capabilities and status of the mobile device. The Collaboration Context Manager manages all the elements related to the collaborative process such as the sequence of the tasks

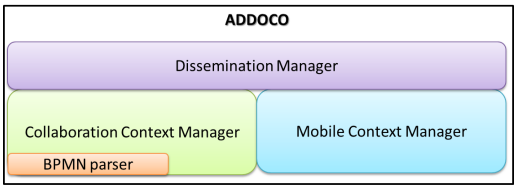


Figure 3: ADDOCO Architecture.

to be executed, the role that will perform each task, the abilities of each user, etc. This manager handles the dynamic role assignment in *ADDOCO*. The **Dissemination Manager** will take into account the information provided by both (the Mobile and Collaboration context managers) in order to determine the right dissemination model. This manager handles the dynamic dissemination in *ADDOCO*. A collaborative process using *ADDOCO* will start by defining the collaborative services involved in collaboration. This involves the definition of roles, their required abilities and the contextual elements that are going to be considered. After these basic components have been defined, a BPMN model is made in order to link those services and determine which role will execute them. The translation of a BPMN model into a collaborative service is not detailed due to the scope of this paper.

4 RELATED WORKS

This section presents an analysis of the main limitations of the architectures and frameworks related to the intention of *ADDOCOM*. MoCA (Sacramento et al., 2004) is an architecture oriented towards mobile collaboration, by communicating each node with another directly and obtaining information through a proxy. This architecture models effective collaboration architecture, but it does not define how data is managed and highly depends on the availability of the proxy for new interactions. Other solutions are distribution applications as Solar, (Chen et al., 2008) a middleware that aids in the creation of data-centred applications, by letting the program request data to the middleware and handles the data distribution for the application, storing and retrieving it using different distribution models. However, once a distribution model has been selected it distributes data in a specific way and it does not adapt according to the changing context of the application. There are also frameworks and middlewares for building collaborative applications that take into account the above mentioned concepts and act as a guideline for new or existing programs and determine how they distribute data. SALES (Corradi et al., 2010) is a middleware that

aims to distribute information in a context-oriented mobile based application. The SALES model takes into account the individual context in each node and notifies others of the contextual information it handles. SALES aids in data distribution by reducing the weight of data transferred and the amount of data transmitted while at the same time increasing the relevance of the disseminated information. However, an algorithm for intelligently distributing data is yet to be done. Another work establishes a publisher - subscriber dissemination, which is a selection based dissemination method (Wu et al., 2010), between Vehicular Ad-Hoc Networks (VANETs). This project proposes the transmission of real time information on the traffic to surrounding vehicles while taking into account their behaviour order to predict their future location and whether the information is relevant or not. This work aids in disseminating traffic data but only as information to other vehicles and it cannot be applied to a complex collaboration scenario. Zimbra (Zimbra, 2011) is another collaborative application that provides real time information and aids in the management of schedules, tasks and calendars. Mobile collaboration in these applications is well managed, however, it depends on a server-side to obtain and manage the data required in order to function properly, being highly dependent on an Internet connection. The *ADDOCOM* framework considers dynamic roles of the users, their location and context information, takes into account collaboration models and has a Dynamic Data Dissemination method that adapts to the environment, addressing the data management requirements of a distributed mobile collaboration environment.

5 PROTOTYPE

In order to make tests of the *ADDOCO* model, a mobile prototype was made. This prototype aims to validate the usefulness of the collaboration model and how dynamic data dissemination may aid in obtaining the information required to collaborate in different environments and contexts. The selected devices for the test were Android powered devices running 2.3.3 version of the OS (Gingerbread) with no cellular data internet plan, with an initially active Wi-Fi connection to connect to an initial directory, and Bluetooth in order to make a collaboration network that tests the data dissemination model. The prototype first attempts a connection with a pre-established server in order to obtain the data and begins downloading it. If that server is unreachable, whether at the beginning of the process or the connection fails in the middle of the process, an alternate dissemination method will be selected using Bluetooth networks. If the connection with the directory fails the user will be notified that the connection has been lost and that nearby users are being queried for the information. A list of nearby users who have the requested information is displayed and the user selects one of the available devices.

Once a user selects a data source who will provide the information, the provider user's device will display a message with the request of information and additional data on the user who wants to obtain it. The provider user may accept or reject the request. If the request is accepted then the information will be provided to the user that requested it, if the request is rejected a message will show up in the requester's device informing that the connection could not be established.

6 CONCLUSIONS AND FUTURE WORK

The implementation of the *ADDOCO* model aims to take full advantage of Ad-Hoc networks in collaborative environments. Functional tests proved that collaboration is enhanced with *ADDOCO* by having dynamic methods for obtaining the data while being aware of the context, thus reducing the time of a potentially non-executable task. Bluetooth technology was used in order to make closed-range networks to adequately test the dissemination model; however, we found that this technology is not the best choice for making effective data dissemination and collaborative networks due to its restrictions and characteristics. However, it shed light in how the functionality of a smartphone network could work and the effectiveness of the collaboration and data distribution model of *ADDOCO*. As a future work the prototype is going to be enhanced in order to work in more complex collaborative tasks, taking into account not only the connection status but also the specific location of the user and making possible the execution of parallel tasks. Additionally, *ADDOCO* will make use of an emerging technology called Wi-Fi Direct^{TM}(WiFi, 2011) in order to overcome technical drawbacks seen during the use of Bluetooth technology in this context. The Wi-Fi Direct technology will be supported by Android 4.0 (Ice Cream Sandwich) and due to the implementation of the prototype using the Android SDK *ADDOCO*'s prototype will most likely adopt this technology in a near future for further tests and improved functionality.

ACKNOWLEDGEMENTS

This work was supported by the project *"Identificación semiautomática de pacientes con enfermedades crónicas a partir de la exploración retrospectiva de las historias clínicas electrónicas registradas en el sistema SAHI del Hospital San Ignacio"* made by the Pontificia Universidad Javeriana in conjunction with the Hospital Universitario San Ignacio.

REFERENCES

Bellavista, P., Corradi, A., Fanelli, M., and Foschini, L. (2013). A survey of context data distribution for mobile ubiquitous systems. *Accepted in ACM Computing Surveys (CSUR), ACM Press*, expected to appear in Vol. 45:1–49.

Chen, G., Li, M., and Kotz, D. (2008). Data-centric middleware for context-aware pervasive computing. *Pervasive Mob. Comput.*, 4:216–253.

Corradi, A., Fanelli, M., and Foschini, L. (2010). Towards adaptive and scalable context aware middleware. *IJARAS*, 1(1):58–74.

Prayogi, A., Park, J., and Hwang, E. (2007). Selective role assignment on dynamic location-based access control. In *International Conference on Convergence Information Technology, 2007.*, pages 2136 –2135.

Sacramento, V., Endler, M., Rubinsztejn, H. K., Lima, L. S., Goncalves, K., Nascimento, F. N., and Bueno, G. A. (2004). Moca: A middleware for developing collaborative applications for mobile users. *IEEE Distributed Systems Online*, 5:2–.

Waze (2011). Waze,2011. waze. http://world.waze.com.

WiFi (2011). Wifi direct, 2011. wifi alliance. http://www.wi-fi.org.

Wu, L., Liu, M., Wang, X., and Gong, H. (2010). Dynamic distribution-aware data dissemination for vehicular ad hoc networks. In *2nd International Conference on Future Computer and Communication (ICFCC), 2010.*, volume 2, pages V2–353 –V2–360.

Zhang, H., He, Y., and Shi, Z. (2006). Spatial context in role-based access control. In Rhee, M. and Lee, B., editors, *Information Security and Cryptology - ICISC 2006*, volume 4296 of *Lecture Notes in Computer Science*, pages 166–178. Springer Berlin / Heidelberg. 10.1007/11927587-15.

Zimbra (2011). Zimbra, 2011. vmware. http://www.zimbra.com/products/mobility.html.

Optimizing Data Integration Queries over Web Data Sources (OPTIQ)

Muhammad Intizar Ali
Database and Information Systems Group,
Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan
intizarali@ciitlahore.edu.pk

Keywords: Data Integration, Distributed Query Processing, Query Optimization, Xsparql, Sparql, Xquery, Rdf, Xml, Web Data Sources.

Abstract: Nowadays data integration must deal with a far more in-congruent environment than in the pre-Web era. There are multiple efforts to design new query languages (which we hereafter call "data integration query languages"), that combine the features of two or more existing well established query languages to integrate data. However modern data integration applications still face many challenges while executing data integration queries over web data sources. Currently data integration queries are just focused on data integration while optimized query plan generation for such queries has not been extensively studied. Higher query processing time, data shipping and transformation on the fly, complicated syntax, efforts required to learn a new query language and limitation imposed on query languages are also a big hindrance in the emergence and wide adoption of the data integration query languages. In this paper, we have optimized data integration queries by parallel and distributed query processing over heterogeneous web data sources. We combine the data integration capabilities of the XSPARQL (a data integration query language for XML and RDF data sources) with the distributed and parallel query execution capabilities of DeXIN (a framework for highly distributed large web data sources integration). Parallel query processing suits very well for data integration of web data sources because of the highly distributed nature of the web. Experimental results are also evident of better performance of optimized query processing by parallelizing the data integration query processing. In the later part of this paper we argue that availability of an easy to use visual editor for data integration queries will greatly help in wide adoption and usage of data integration query language for web data sources integration.

1 INTRODUCTION

In modern business enterprises, it is frequent to develop an integrated application to provide uniform access to multiple existing information systems running internally or externally of the enterprise. Data integration is a pervasive challenge faced in these applications that need to query across multiple autonomous and heterogeneous data sources. Integrating such diverse information systems becomes a challenging task particularly when different applications use different data formats and query languages which are not compatible with each other. With the growing popularity of web technologies and availability of the huge amount of data on the web, the requirements for data integration has changed from the traditional database integration approaches. The large scale of Web has led to high levels of distribution, heterogeneity, different data formats and query languages.

Data transformation and query rewriting are common approaches to mitigate the heterogeneity among various data formats (Dan Connolly, 2007). Ample efforts are made to transform the data sources from one data format to another data format (Matthias Droop et al., 2008; Sven Groppe et al., 2008). However data transformation approaches are not viable for large or dynamic data sources, because the frequency of changes might make it cumbersome to perform the transformation over and over again. Another approach is to rewrite query from one query language to another query language while data sources maintain their original data format (Waseem Akhtar et al., 2008). In the past data, transformation from relational data to XML and query rewriting from SQL to XQUERY attracted significant research while nowadays data transformation from

XML to RDF and query rewriting from XQUERY to SPARQL and vice versa is likewise being explored (Nikos Bikakis et al., 2009). XSPARQL and embedding SPARQL into XSLT are data integration query languages to integrate XML and RDF data (Waseem Akhtar et al., 2008; Schmidt et al., 2002). The DeXIN framework is another query-side effort to integrate distributed heterogeneous Web Data sources (Muhammad Intizar Ali et al., 2009a). However complex syntax, inability to generate optimized solutions and efforts required to learn a new language were prominent obstacles in the adoption of these projects from academia to industry. Hence, unable to leverage the modern multi-core processing capabilities which are of utmost important for efficient query processing over highly distributed Web data sources.

In this paper we optimize data integration queries of XSPARQL by embedding XSPARQL queries into DeXIN framework. We categorize various features of data integration queries and then combine them to get the best query execution plan for distributed query processing over heterogeneous distributed web data sources. We perform experimental evaluation and compare query execution time of the optimize query plan with the existing data integration query languages processing time. It is evident from the experimental results that our approach performs very well while executing distributed query over multitude of data sources scattered over Web.

Our main focus is to devise strategies and algorithms for optimized query plan generation for data integration queries specifically targeted at Web data integration scenarios where data integration applications lack prior knowledge or control of the content of data sources. Optimized query plans for data integration queries over the Web of Data should be capable of dealing with exponential growth and continuous updates in terms of information sources and dynamic streaming data. Query optimization should also take into account scalability issues to avoid performance degradation and inefficient query execution plans.

In order to attract the wider community of users we focus on enabling optimized data integration query generation using visual editors and initiation of an open source data integration suite for distributed, parallelized and heterogeneous querying over Web data sources.

Rest of the paper is organized as: in Section 2, we briefly describe data integration queries for SPARQL and XQuery integration. We compare the features of various data integration queries to categorize and enhance existing data integration queries. In Section 3, we propose optimize query execution plan for data integration queries by embedding XSPARQL into DeXIN framework. Later in this section we discuss possible enhancement in the features of existing data integration queries. Experimental evaluation is performed in Section 4. We conclude this paper with possible future work on data integration queries in Section 5.

2 DATA INTEGRATION QUERY LANGUAGES

Data integration query languages combine the features of two or more existing query languages to integrate data on the fly. In this approach a new query language is designed or already existing query language is extended in order to query multiple heterogeneous web data sources in their respective query languages. Data integration query language is a broader term which can combine any arbitrary two or more query languages features. However in this paper we restrict our self to those data integration query languages which combine the features of XQuery and SPARQL. Data integration query language approach is a viable solution for Web data sources integration which not only integrates data on the fly but also can easily manage rapidly changing web data sources. In this section we will briefly describe three prominent approaches of data integration queries for XML and RDF data sources and then compare their features to highlight the advantages and disadvantages of the each approach.

2.1 Embedding SPARQL into XQUERY

Embedding SPARQL into XQUERY is one of the initiative efforts for designing data integration query to provide a uniform access over RDF and XML data sources (Sven Groppe et al., 2008). The approach is to select one query language for one data format as a base query language and embed the other query language for another format into the base query language. In embedding SPARQL into XQUERY approach, all SPARQL queries are embedded into XQuery/XSLT and automatically transformed into pure XQuery/XSLT queries to be posed against pure XML data after transformation of RDF data into XML. This embedding enables users to benefit from graph and tree language constructs of

both SPARQL and XQuery. The authors defined a formal SPARQL algebra to transform a SPARQL query into an operator tree of SPARQL algebra. The operator tree is later translated into XQuery/XSLT. Table 1 shows an example XQuery with SPARQL embedded inside XQuery.

The work presented in (Sven Groppe et al., 2008) is not an exclusive work on SPARQL embedded into XQuery. There exists some other similar work with slightly different approach also available in the literature. Contrary to embedding SPARQL into XQuery/XSLT, there are also some efforts to embed XPATH and XQuery into SPARQL queries (Matthias Droop et al., 2008).

Table 1: An example XQuery with SPARQL embedded inside XQuery.

```
( 1 ) d e c l a r e namespace f o a f =" h t t p : / / xmlns .
com/ f o a f / 0 . 1 / " ;
( 2 ) < r e s u l t s >{ f o r ( $n , $m) i n
( 3 ) SELECT ?name ?mbox
( 4 ) WHERE { ? x f o a f : name ?name .
( 5 ) ? x f o a f :mbox ?mbox .
( 6 ) FILTER r e g e x ( s t r ( ?mbox ) , "@work . example " ) }
( 7 ) r e t u r n <result><name>{$n}</name><mbox>{$m}
( 8 ) </mbox> </ r e s u l t >}< r e s u l t s >
```

2.2 XSPARQL

XSPARQL is a new query language or more precisely to say data integration query language designed with the idea to combine XQUERY and SPARQL query languages. Despite the availability of GRDDL transformation sets the translating between XML and RDF is a tedious and error prone task (Dan Connolly, 2007). In (Waseem Akhtar et al., 2008), a new query language based approach is used to transform XML into RDF and vice versa. XSPARQL is a query language combining XQuery and SPARQL for transformations between RDF and XML. XSPARQL subsumes XQuery and most of SPARQL (excluding ASK and DESCRIBE). Conceptually, XSPARQL is a simple merge of SPARQL components into XQuery. XQuery is a native query language in XSPARQL and all XQuery queries are also considered as XSPARQL queries. In order to execute SPARQL queries inside the body of XQuery, the XQuery FLWOR expression is slightly modified which is called FLWOR' expression. Concerning semantics, XSPARQL equally builds on top of its constituent languages. They extended the formal semantics of XQuery by additional rules which reduce each XSPARQL query to XQuery expressions; the resulting FLWORs operate on the answers of SPARQL queries in the SPARQL XML result format. All XSPARQL queries are rewritten in XQuery standard format while SPARQL queries are executed over SPARQL endpoints and results are returned in RDF/XML.

2.3 Distributed Extended XQuery for Data Integration (DeXIN)

DeXIN is an extensible framework for providing integrated access over heterogeneous, autonomous, and distributed web data sources, which can be utilized for data integration in modern web applications and service oriented architecture. DeXIN extends the XQuery language by supporting SPARQL queries inside XQuery, thus facilitating the query of data modelled in XML, RDF, and OWL (Muhammad Intizar Ali et al., 2009b). DeXIN facilitates data integration in a distributed web and service oriented environment by avoiding the transfer of large amounts of data to a central server for centralized data integration and avoids the transformation of a huge amount of data into a common format for integrated access.

At the heart of DeXIN is an XQuery extension that allows users/applications to execute a single query against distributed, heterogeneous web data sources or data services. DeXIN considers one data format as the basis (the so-called "aggregation model") and extends the corresponding query language to executing queries over heterogeneous data sources in their respective query languages.

Currently, DeXIN have implemented XML as an aggregation model and XQuery as the corresponding language, into which the full SPARQL language is integrated. However, this framework is very flexible and could be easily extended to further data formats (e.g., relational data to be queried with SQL) or changed to another aggregation model (e.g., RDF/OWL rather than XML).

The main highlights of the features of the DeXIN are as follows.

- DeXIN is an extensible framework for parallel query execution over distributed, heterogeneous and autonomous large data sources.
- DeXIN provides extension of XQuery which covers the full SPARQL language and supports the decentralized execution both XQuery and SPARQL in a single query
- DeXIN approach supports the data integration of XML, RDF and OWL data without the need of transforming large data sources into a common format.

- DeXIN is implemented as a web service to provide easy access using service oriented architecture.
- DeXIN can easily be integrated into existing web applications as a data integration tool.
- Experimental results show good performance and reduced network traffic achieved with DeXIN approach.

2.4 Comparing Features of Data Integration Queries

In Table 2, we compare different features of the various data integration queries designed to integrate XML and RDF data sources using combination of XQuery and SPARQL query languages. All the three approaches are mainly intended for data integration of XML and RDF. Embedded SPARQL approach performs poorly when the transformation of the large data sources is required while XSPARQL and DeXIN only transform results into uniform format rather than whole data source. Both Embedded SPARQL and XSPARQL require its user to learn a new query language while DeXIN does not make merely noticeable syntactic changes to the participating query languages. DeXIN also executes parallel and distributed queries for distributed web data sources. Variable induction is a strong feature of XSPARQL which enables it to share same variable for joining two or more heterogeneous Web data sources within a single query.

3 OPTIMIZING DATA INTEGRATION QUERIES

Existing data integration query languages are mainly focused on data integration while optimize query execution plan for such query language have not been studied. In this section we propose the combination of various features of the existing data integration query language which not only be utilized for optimize query execution plan but also can better cope with the modern data integration challenges faced because of the highly distributed and heterogeneous nature of the Web, availability of large amount of data and rapidly changing data sources over Web.

Table 2: Comparison of various features of data integration queries for XML and RDF data.

Features	Embedded SPARQL	XSPARQL	DeXIN
Data integration	Yes	Yes	Yes
Data transformation	Yes	No	No
Complicated syntax	Yes	Yes	No
Parallel/distributed query execution	No	No	Yes
Visual interface	No	No	Yes
Restriction imposed on query language	Yes	Yes	No
Variable induction	No	Yes	No
Query language definition	No	Yes	No
Flexibility for enhancement to further query languages	No	No	Yes

3.1 Parallel and Distributed Query Processing

As a first step towards the solution of the optimized query processing of data integration queries, we integrate the DeXIN framework with XSPARQL. Figure 1 shows the integration of XSPARQL into DeXIN framework. This will combine the data integration capabilities of XSPARQL with the distributed and parallel query execution capabilities of DeXIN. Optimized query plan generation for the Web of Data is integral to Web data sources integration. We devise strategies and algorithms for optimized query plan generation for data integration queries specifically targeted at Web data integration scenarios where data integration applications lack prior knowledge or control of the content of data sources. Optimized query plans for data integration queries over the Web of Data should be capable of dealing with exponential growth and continuous updates in terms of information sources and dynamic streaming data. Query optimization should also take into account scalability issues to avoid performance degradation and inefficient query execution plans.

3.2 SPARQL and XQuery Endpoints

XSPARQL and embedded SPARQL deal with the RDF and XML data source as a flat file. The general procedure followed in these applications is to fetch

the RDF/XML file, create an in-memory tree/graph and then execute queries. This procedure will become bottle neck once the size of the data source increases. Nowadays many XML and RDF databases are available, which are specifically designed to efficiently store and process Web data sources. On top of these databases SPARQL/ XQuery endpoints are available which provide direct access to the user for executing queries over these data sources. After the integration of XSPARQL queries inside DeXIN framework, data integration queries can benefit from endpoints of SPARQL and XQuery.

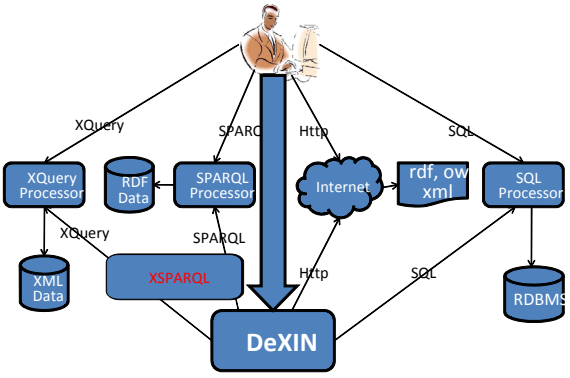


Figure 1: Parallel XSPARQL Query Processing with DeXIN.

3.3 Dynamic Data Source Selection

Data source selection is an important task while integrating distributed heterogeneous web data sources because data sources are discovered, selected and processed dynamically without any prior knowledge of the data sources. DeXIN keeps tracks of data services and their statistics for best data source selection. Moreover, DeXIN also stores user's concerns using profiling approach which helps to select the most appropriate data source for that particular user if multiple data sources are available to perform the same task (Muhammad Intizar Ali et al., 2011a; 2011b).

3.4 Visual Editor

Complicated syntax, efforts required to learn a new query language and limitation imposed on query languages are also a big hindrance in the emergence and wide adoption of the data integration query languages. We propose a query based aggregation of multiple heterogeneous data sources by combining powerful querying features of XQuery and SPARQL with an easy interface of a mashup tool for data sources in XML and RDF. Our mashup editor allows for automatic generation of mashups with an easy to use visual interface. We utilize the concept of data mashups and use it to dynamically integrate heterogeneous web data sources by using the extension of XQuery proposed in the DeXIN. All available data sources over the internet are considered as a huge database and each data source is considered as a table. Data mashups can generate queries in extended XQuery syntax and can execute the sub-queries on any available data source contributing to the mashup (Muhammad Intizar Ali et al., 2011c). We aim to design a visual editor for XSPARQL as well which can automatically generate queries from visual interface.

4 EXPERIMENTAL EVALUATION

Testbed: We have implemented parallel and distributed query processing for XSPARQL queries. For prototype development we used java, Saxon is used for XQuery processing and ARQ is used for SPARQL query processing. Our testbed includes 3 computers (Intel(R) Core(TM) 2 CPU, 2.4 GHz, 4GB RAM) running SUSE Linux with kernel version 2.6. The machines are connected over a standard 100Mbit/S network connection. An open source native XML database eXist (release 1.2.4) is installed on each system to store XML data.

Data Sets: For the evaluation of our implementation we used the XMark benchmark suite which is the most widely used benchmark suite for XQuery (Schmidt et al., 2002). XMARK benchmark suite is bundled with a data generator that produces XML data sets. The benchmark also includes a set of 20 XQuery queries over this generated data. Using the provided data generator, we created XML datasets with sizes of 1, 2, 10, 50 and 100MB. For RDF data sets we rely on XMARK and converted these XML data sets of XMARK into RDF using XSPARQL. We reformulated XMARK 20 queries into SPARQL, XSPARQL, Embedded SPARQL and DeXIN extended XQuery format.

Experimental Evaluation:

Figure 2 shows experimental result of evaluating first 10 queries of XMARK benchmark executed over data set of 2 MB. Embedded SPARQL performs better when the data size is small while XSPARQL performs well with bigger data size. However XSPARQL performs very poorly when queries contain nested join between two data sets.

DeXIN improves the performance of XSPARQL particularly when there are multiple data sets are involved in query processing.

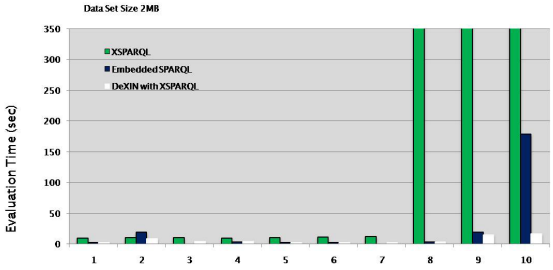


Figure 2: Query execution time comparison.

5 CONCLUSIONS AND FUTURE WORK

In this paper we have integrated XSPARQL and DeXIN capabilities to execute data integration queries in parallel. Experimental evaluation has shown promising results. By integrating XSPARQL inside DeXIN the performance of data integration queries (particularly those queries which involve highly distributed multiple data sources) has greatly increased. We believe that our proposed plan will lead to the optimized query execution plans for the data integration queries over the Web. Our framework will act as a basis for the development of data integration applications over distributed heterogeneous and continuously updating Web data sources. A visual tool to generate data integrating queries will attract a wider spectrum of the users to create data integration applications on the fly.

REFERENCES

Waseem Akhtar, Jacek Kopecký, Thomas Krennwallner, and Axel Polleres. XSPARQL: Traveling between the XML and RDF Worlds - and Avoiding the XSLT Pilgrimage. In *Proc. ESWC 2008, volume 5021 of Lecture Notes in Computer Science*, pages 432–447.Springer, 2008.

Muhammad Intizar Ali, Reinhard Pichler, Hong Linh Truong, and Schahram Dustdar, a). DeXIN: An Extensible Framework for Distributed XQuery over Heterogeneous Data Sources. In *Proc. ICEIS 2009, volume 24 of Lecture Notes in Business Information Processing*, pages 172–183. Springer, 2009.

Muhammad Intizar Ali, Reinhard Pichler, Hong Linh Truong, and Schahram Dustdar, b). On Using Distributed Extended XQuery for Web Data Sources as Services. In *Proc. ICWE 2009, volume 5648 of Lecture Notes in Computer Science*, pages 497–500. Springer, 2009.

Muhammad Intizar Ali, Reinhard Pichler, Hong Linh Truong, and Schahram Dustdar, a). Data Concern Aware Querying Using Data Services. In *Proc. ICEIS 2011*, pages 111–119. SciTePress, 2011.

Muhammad Intizar Ali, Reinhard Pichler, Hong Linh Truong, and Schahram Dustdar, b). *Incorporating Data Concerns into Query Languages for Data Services. to appear in Lecture Notes in Business Information Processing*. Springer, 2011.

Muhammad Intizar Ali, Reinhard Pichler, Hong Linh Truong, and Schahram Dustda, c). On Integrating Data Services Using Data Mashups. In *Proc. BNCOD 2011, volume 7051 of Lecture Notes in Computer Science*. Springer, 2011.

Nikos Bikakis, Nektarios Gioldasis, Chrisa Tsinaraki, and Stavros Christodoulakis. Querying XML Data with SPARQL. In *Proc. DEXA 2009, volume 5690 of Lecture Notes in Computer Science*, pages 372–381. Springer, 2009.

Dan Connolly. Gleaning Resource Descriptions from Dialects of Languages (GRDDL). *W3C Recommendation, W3C,* September 2007. http:// www.w3.org/ TR/2007/REC-grddl- 20070911/.

Matthias Droop, Markus Flarer, Jinghua Groppe, Sven Groppe, Volker Linnemann, Jakob Pinggera, Florian Santner, Michael Schier, Felix Schöpf, Hannes Staffler, and Stefan Zugal. Embedding XPath Queries into SPARQL Queries. In *Proc. ICEIS 2008*, pages 5–14, 2008.

Sven Groppe, Jinghua Groppe, Volker Linnemann, Dirk Kukulenz, Nils Hoeller, and Christoph Reinke. Embedding SPARQL into XQuery/XSLT. In *Proc. SAC 2008*, pages 2271–2278, 2008.

Schmidt, A., Waas, F., Kersten, M.L., Carey, M.J., Manolescu, I., Busse, R.: XMark: A Benchmark for XML Data Management. In: *Proceedings of the 28th international conference on Very Large Data Bases*. pp. 974–985 (2002).

An Idea for Universal Generator of Hypotheses

Grete Lind and Rein Kuusik
Informatics, Tallinn University of Technology, Raja 15, 12618, Tallinn, Estonia

Keywords: Knowledge Discovery, Data Mining, Classification, Rule, Data Description, Universal Hypotheses Generator.

Abstract: We know that the task of Machine Learning (ML) is defined as finding of rules for the class on the basis of learning examples for classification of unknown object(s). But we can use rules also for describing the class data– who/what are they? which is the task of Data Analysis and Data Mining. There are several methods for solving this task, for example, Determination Analysis (DA) and Generator of Hypotheses (GH). In the paper we describe an idea for Universal Generator of Hypotheses, the complex method which can solve the tasks of DA and GH and several new ones.

1 INTRODUCTION

In the domain of machine learning (ML) many different algorithms are in use (Mitchell, 1997), for example ID3 (Quinlan, 1986), CN2 (Clark and Niblett, 1987), CART (Breiman, Friedman, Olshen and Stone, 1984) and their derivates. There are several algorithms which try to solve the same task on a different algorithmic and pruning techniques bases. Some algorithms output rules

- as decision trees;
- some as sets of rules;
- some of them find non-intersecting rules;
- some find overlapping rules;
- some find only one system of rules;
- some algorithms find different systems of rules;
- some find a set of rules that meets certain requirements;
- etc.

This is expected, because the number of all possible rules in case of given sets of learning examples can be huge and each method for finding a set of rules tries to prune the number of rules.

We present an idea of Universal Generator of Hypotheses, which can output most of the described possibilities of output and some new possibilities for the researcher.

2 MACHINE LEARNING TASK AS A DATA MINING TASK

Machine Learning task is defined as learning from examples i.e. finding concept description (set of rules IF X THEN Y) that is both ***consistent*** and ***complete*** at the same time (Gams and Lavrac, 1987).

A description is ***complete*** if it covers all examples of all classes.

A description is ***consistent*** if it does not cover any pair of examples from different classes.

2.1 Two Directions in ML

There are two directions (subtasks) in Machine Learning:

- *Direction 1* (Main task): On the basis of learning examples to find rules for classification of unknown object(s) (*Classification task*);
- *Direction 2*: We can use the found rules for describing the data table (learning examples) under analysis: "Who/what are they?" (*Data Analysis and Data Mining task*).

The main steps of direction 1 are:

1) Finding set of rules;
2) Testing rules on test-examples;
3) Applying tested rules on new instances.

Here the main goal is to find the rules with a stably good ability of recognition. There exist several methods for solving this task.

The main steps of direction 2 are:

1) Finding set of rules;
2) Analysis of found rules;
3) Class(es) description on the basis of rules.

The main goal for direction 2 is to describe the class -"who/what they are" on the basis of found rules. The best representatives of the direction 2 are methods "Determinacy Analysis" (Chesnokov, 1980; Chesnokov, 1982) and „Generator of Hypotheses" (Kuusik and Lind, 2004). They try to answer to the questions:

- "Who are they (objects of class)?";
- "How can we describe them?";
- "What distinguishes them from the others?".

It means that on the basis of extracted rules we can describe the class. Use of rules makes possible to determine what is specific for the class and what separate different classes. Using extracted rules also the latent structure of the class can be discovered.

It is possible that the researcher is interested in dividing attributes into two parts: causes (C) and effects (E) and wants to analyze relations between them (IF C THEN E).

From the other hand it can happen that the researcher does not know what he/she seeks. It means that the use of corresponding methods provides him/her with some kind of (work) hypotheses for description and he/she must decide whether the extracted rules can help him/her to describe or understand the essence of the data. That is why we call extracted rules for data description "hypotheses". The same situation may arise also when the amount of extracted rules is very big and he/she physically cannot analyze them.

Next we present a brief description of DA and GH.

2.2 Determination Analysis

The main idea behind DA is that a rule can be found based on the frequencies of joint occurrence or non-occurrence of events. Such rule is called a determinacy or determination, and the mathematical theory of such rules is called determinacy analysis (Chesnokov, 1982).

If it is observable that an occurrence of X is always followed by an occurrence of Y, this means that there exists a rule "If X then Y", or $X \rightarrow Y$. Such correlation between X and Y is called determination (from X to Y). Here X is *determinative (determining)* and Y is *determinable*.

Each rule has two characteristics: accuracy and completeness.

Accuracy of determination $X \rightarrow Y$ shows to what extent X determines Y. It is defined as a proportion of occurrences of Y among the occurrences of X:

$A(X \rightarrow Y) = n(X\ Y) / n(X)$, where

$A(X \rightarrow Y)$ is accuracy of determination,

n(X) is a number of objects having feature X and

n(X Y) is a number of objects having both features X and Y.

Completeness of determination $X \rightarrow Y$ shows which part of cases having Y can be explained by determination $X \rightarrow Y$. It is a percentage of occurrences of X among the occurrences of Y:

$C(X \rightarrow Y) = n(X\ Y) / n(Y)$, where

$C(X \rightarrow Y)$ is completeness of determination,

n(Y) is a number of objects having feature Y and

n(X Y) is a number of objects having both features X and Y.

Both accuracy and completeness can have values from 0 to 1. Value 1 shows maximal accuracy or completeness, 0 means that rule is not accurate or complete at all. Value between 0 and 1 shows quasideterminism.

If all objects having feature X have also feature Y then the determination is (maximally) accurate. In case of accurate determination $A(X \rightarrow Y) = 1$ (100%).

Majority of rules are not accurate. In case of inaccurate rule $A(X \rightarrow Y) < 1$.

In order to make determination more (or less) accurate complementary factors are added into the first part of a rule. Adding factor Z into rule $X \rightarrow Y$ we get a rule $XZ \rightarrow Y$.

DA enables to find different sets of rules, depending on the order in which the attributes are included into the analysis. One possible set of accurate rules for well known Quinlan's data set (of eight persons characterized by height, hair colour and eye colour) (Quinlan, 1984) for example describing (persons belonging to) class "–" is following:

- Hair.red → Class. – (C=33%);
- Hair.blond & Eyes.blue → Class. – (C=67%),

The second one:

- Height.tall&Hair.red → Class. – (C = 33%)
- Height.short&Hair.blond&Eyes.blue → Class. – (C=33%)
- Height.tall&Hair.blond&Eyes.blue → Class. – (C = 33%).

2.3 Generator of Hypotheses

Generator of Hypotheses (GH) is a method for data mining which main aim is mining for patterns and association rules (Kuusik and Lind, 2004). The goal

is to describe the source data. Used evaluation criteria are deterministic (not probabilistic). The association rules it produces are represented as trees, which are easy to comprehend and interpret.

By depth-first search (from root to leaves) GH forms a hierarchical grouping tree. Such tree example is given below. Method uses effective pruning techniques.

```
(3)        0.667(2)     0.500(1)
Height.tall=>Hair  .Dark->Eyes  .Blue
                   0.500(1)
                   ->Eyes  .Brown
           0.667(2)      0.500(1)
           =>Eyes  .Brown->Hair  .Blond

 (3)       0.667(2)     0.500(1)
Hair  .Dark=>Eyes  .Blue->Height.Short
           0.333(1)
           =>Eyes  .Brown

 (3)       0.667(2)      0.500(1)
Eyes .Brown=>Hair  .Blond->Height.Short
```

The numbers above node show node's absolute frequency (in parentheses) and node's relative (to previous level) frequency (before parentheses).

Absolute frequency of node t shows how many objects have certain attribute with certain value (among objects having properties (i.e. certain attributes with certain values) of all previous levels t-1,…,1). Relative frequency is a ratio A/B, where A is the absolute frequency of node t and B is the absolute frequency of node t-1. For the first level the relative frequency is not calculated.

For example we can translate the first tree (Height.tall=>) of set of trees as "3 persons (objects/ examples) are tall, 67% of them have dark hair, and of those (with Height.tall and Hair.dark) 50% have blue eyes and 50% have brown eyes. Also, 67% of tall persons have brown eyes and 50% of those have blond hair."

GH has the following properties:

- GH guarantees immediate and simple output of rules in the form IF=>THEN;
- GH enables larger set of discrete values (not only binary);
- GH enables to use several pruning techniques;
- The result is presented in form of trees;
- GH enables to treat large datasets;
- GH enables sampling.

3 AN IDEA FOR UNIVERSAL GENERATOR OF HYPOTHESES

Here we present an idea for Universal Generator of Hypotheses (UGH), which can solve analysis task (direction 2) and which can test hypotheses (for example, whether some specific rule identifies some designated class (task of query type) i.e. can the rule open the essence of the class under description), and generate the new ones. Building of UGH is real, due to the existence of the base algorithm and special techniques on the basis of which several versions of DA (Lind and Kuusik, 2008; Kuusik and Lind, 2010; Kuusik and Lind, 2011b) and GH (Kuusik and Lind, 2004; Kuusik, Lind and Võhandu, 2004) (both direction 2) and IL task (direction 1) (Roosmann, Võhandu, Kuusik, Treier and Lind, 2008; Kuusik, Treier, Lind and Roosmann, 2009) have been realized.

The block diagram of Universal Generator of Hypotheses is shown in Figure 1.

Basically the variants divide into two:

1) The researcher (user) does not partition attributes (objects' characteristics) under consideration – presented by blocks 3..6 on the left side of the scheme;
2) The researcher divides attributes into cause and effect – blocks 7..17 on the right side of the scheme.

In the first case (blocks 3..6) simply the enumeration of analyzable attributes is given to the system, i.e. it is not required to observe all the attributes that are used for describing the objects. As a result all existing value combinations of those attributes or relations in the form of cause-and-effect where causes and effects are generated automatically can be obtained. System does not determine the causes and the effects in a relation in the same way as the user does in case of Determinacy Analysis, but offers different possibilities for that; the user has to decide what is what.

Always it is possible to define the set of observable objects (narrower than in initial data). It is shown as a logical expression (in block 2). In a sense of DA the narrowing of universal context takes place. Context is the set of qualities that describe the whole group (the ones, on the ground of which the objects are selected). The qualities common to the whole initial data set determine the universal context. In the same data set it is not possible to widen the context, it is the widest there.

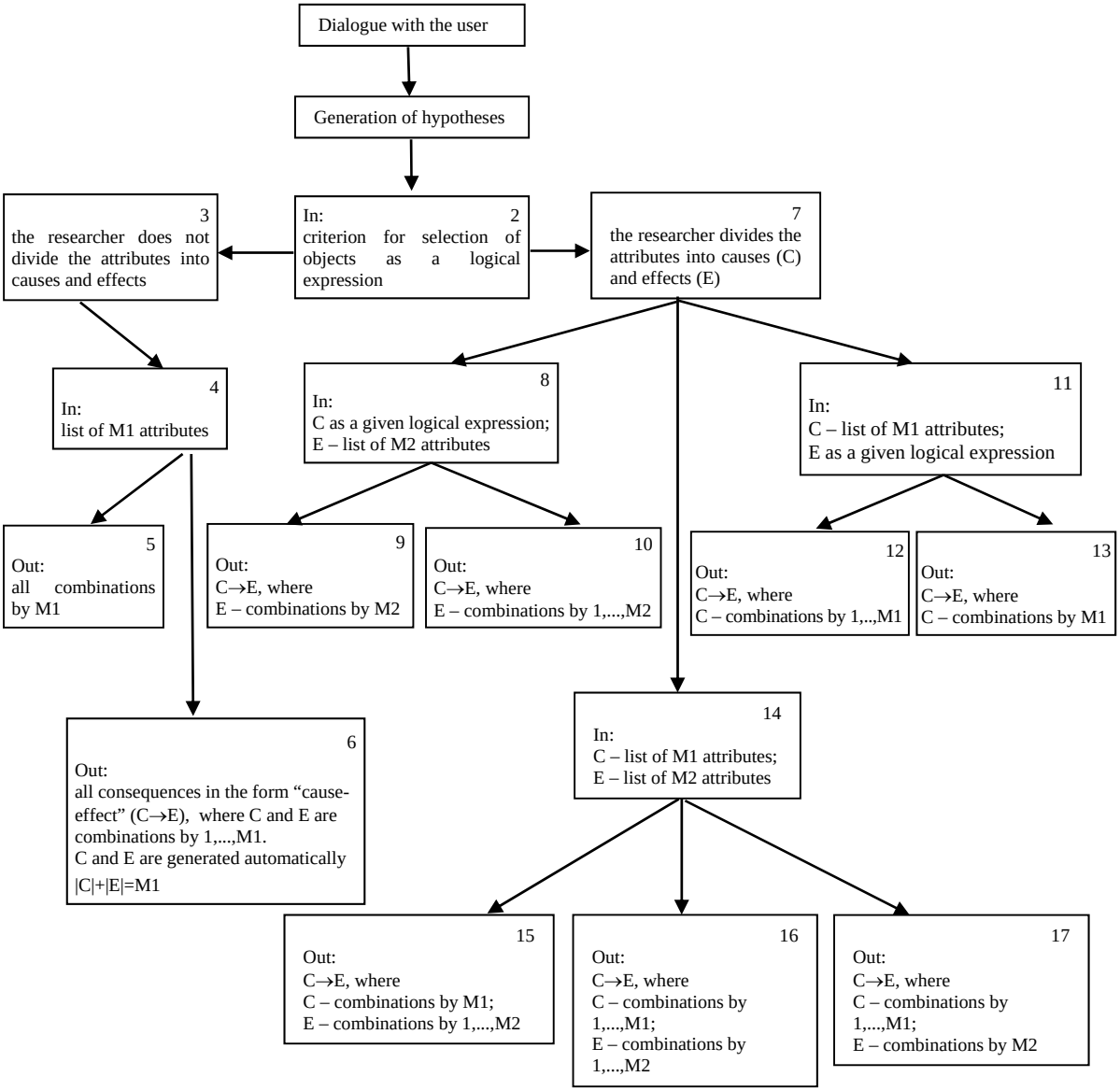


Figure 1: Block diagram of Universal Generator of Hypotheses.

Thus the context can be changed only by narrowing. For that purpose the qualities on which basis to make the restriction have to be shown. It is needless to observe the attributes that determine the context neither among causes nor among effects, since they describe the whole subset under examination.

In the second case (blocks 7..17), blocks 14..17 describe the basic cases, where the researcher distinguishes between cause-attributes and effect-attributes. Block 15 presents the case, where with each different existing combination of causes the consequences characteristic only to it are associated. In block 17 for each existing set of effects the causes inducing only it are searched for. Although these two cases are completely distinct for the user, the difference here is only in the interpretation of the data.

The case in block 16 differs from the one in block 15 so that the sets of causes for which the effects are searched for, are not restricted to the ones that contain all the cause-attributes, but also the combinations that contain only one or two etc attributes from given set of attributes are observed. In case of necessity here also the places of causes and effects can be changed.

Blocks 8..10 represent a special case of blocks 14..15, where the user investigates what are the effects resulting from specified cause(s). The set of

observable objects is determined by a logical condition over cause-attributes.

Similarly the blocks 11..13 is a special case of blocks 14&17, where the user examines what reasons lead to specified effect. The logical condition of effect-attributes determines the set of observable objects.

Again the variants in blocks 8..10 and in blocks 11..13 differ solely in the interpretation.

Basically the results findable by blocks 14..17 can be obtained by proper repeated application of simpler variants in blocks 8..13, but it is more practical to give that work to the computer. For the human user giving the different value combinations (as logical expression) one by one is arduous enough.

Usually it is reasonable to require from the user that the sets of causes and effects do not intersect. In cases (of variants) 15 and 17 the overlapping attributes are always present in the fixed-length part (C in block 15, E in block 17) and they can also appear in the other part of relations. In case of variant (in block) 16 such attributes can fall into both sides. But something that causes itself or results from itself is not very informative.

The overlapping might make sense if more than one value is allowed for the overlapping attribute(s) and objects with different values of such attribute(s) form the same cause or effect. This is possible when causes or effects are given by a logical expression (blocks 8 and 11 accordingly). Appearing in the other part of relations the overlapping attributes may provide interesting information.

The same is true for restricting the context: if more values are allowed for the attribute(s) determining a context then it makes sense to observe this(these) attribute(s) in the relations.

Generator of hypotheses does not presuppose that observable objects are classified, however it may come in handy when solving that task. (Automatic) classification occurs here as follows. The user submits a list of attributes (either causes or effects); the system finds existing value combinations of given attributes and each such combination describes a class of objects. Such classification takes place in block 15 by cause-attributes and in block 17 by effect-attributes. As mentioned, in these cases the difference (that is so important for the user) is only in the interpretation.

In blocks 8..13 the determination of interesting class by the researcher takes place on the basis of a logical condition either by causes (block 8) or by effects (block 11).

The variants on the left side of the scheme (blocks 3..6) where the attributes are not divided into causes and effects by the user is realized by Generator of Hypotheses (Kuusik and Lind, 2004). Variants on the right side are covered by machine learning methods. Generally the classes are given and rules for determining them have to be found (Roosmann et al, 2008, Kuusik et al, 2009). Usually the ML methods assume that class is shown by one certain attribute, but in essence it can be a combination of several attributes shown by a logical expression. Again, whether the given classes are cause (blocks 8..10, 14..15) or effect (blocks 11..13, 14&17), depends on the interpretation. Determinacy Analysis (DA) can be qualified as a subtask of machine learning as it finds rules for one class at a time. So it covers the variants in blocks 8..10 and 11..13. Given class can be cause (in block 8) or effect (in block 11). Output containing combinations by M attributes (as in blocks 9 and 13) can be found using DA-system (DA-system, 1998), output according to blocks 10 and 12 can be obtained using step-wise DA methods which allow rules with different length (Lind and Kuusik, 2008; Kuusik and Lind, 2010). By repeated use of DA also the variants given in blocks 14..17 can be performed.

4 CONCLUSIONS

We have presented in the paper an idea for Universal Generator of Hypotheses. We have discussed that matter with specialists of data analysis and they have mentioned that the use of DA and GH is not enough, there are several other tasks to solve and there is need for developing some additional new possibilities. All these possibilities are described in the paper. Possibilities of DA and GH are also described in the paper and they are the part of the functionality of UGH. As we have mentioned, it is possible to realize UGH, there exist the base algorithm and special pruning techniques on the basis of which the functionality of UGH is easily realizable.

REFERENCES

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. Classification and Regression Trees, Belmont, California: Wadsworth.

Clark, P., Niblett, T., 1987. Induction in Noisy Domains. In *Progress in Machine Learning: Proceedings of EWSL 87* (pp. 11-30). Bled, Yugoslavia, Wilmslow: Sigma Press.

Chesnokov, S. V., 1980. *Determination-analysis of social-economic data in dialogical regime* (Preprint). Moscow: All-Union Institute for Systems Research (in Russian).

Chesnokov, S. V., 1982. *Determinacy analysis of social-economic data.* Moscow: Nauka (in Russian).

DA-system 4.0 User's Manual Version 1.0 (1998, 1999) „Kontekst" (in Russian)

Gams, M., Lavrac, N.1987. Review of five empirical learning systems within a proposed schemata. In *Progress in Machine Learning: Proceedings of EWSL 87* (pp. 46-66). Bled, Yugoslavia, Wilmslow: Sigma Press.

Kuusik, R., Lind, G., 2004. Generator of Hypotheses – an Approach of Data Mining Based on Monotone Systems Theory. *International Journal of Computational Intelligence*, 1, 49 - 53.

Kuusik, R., Lind, G., 2010. Some Developments of Determinacy Analysis. In *Advanced Data Mining and Applications - 6th International Conference, ADMA 2010, Proceedings, Part I*. LNCS 6440 (pp. 593-602). Springer.

Kuusik, R.; Lind, G., 2011b. New Developments of Determinacy Analysis. In *Advanced Data Mining and Applications – 7th International Conference, ADMA 2011, Proceedings, Part II*. LNCS 7121 (pp. 223-236). Springer.

Kuusik, R., Lind, G., Võhandu, L., 2004. Frequent pattern mining as a clique extracting task. In *Proceedings: The 8th World Multi-Conference on Systemics, Cybernetics and Informatics* (pp. 425 - 428). Orlando, Florida, USA: International Institute of Informatics and Systemics.

Kuusik, R., Treier, T., Lind, G., Roosmann, P., 2009. Machine Learning Task as a Diclique Extracting Task. In *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 555-560). Los Alamitos, California: Conference Publishing Service.

Lind, G., Kuusik, R., 2008. New developments for Determinacy Analysis: diclique-based approach. *WSEAS Transactions on Information Science and Applications*, 5, 1458-1469.

Mitchell, T. M., 1997. *Machine Learning* McGraw-Hill.

Quinlan, J. R., 1984. Learning efficient classification procedures and their application to chess and games. In *Machine Learning. An Artificial Intelligence Approach*, Springer-Verlag, 463-482.

Quinlan, J. R., 1986. Induction of decision trees. *Machine Learning*, 1, 81-106.

Roosmann, P., Võhandu, L., Kuusik, R., Treier, T., Lind, G., 2008. Monotone Systems approach in Inductive Learning. *International Journal of Applied Mathematics and Informatics*, 2, 47-56.

Investigation of Criteria for Selection of ERP Systems

Bálint Molnár[1], Gyula Szabó[2] and András Benczúr[1]
[1]*Eötvös University of Budapest, Information Systems Department, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary*
[2]*Dénes Gábor College, Department of Computing, 1119 Bp, Mérnök u. 39, Budapest, Hungary*
{molnarba, abenczur}@inf.elte.hu, bpinformatik@gmail.com

Keywords: Information System, ERP, Enterprise Resource Planning, Enterprise Architecture, Selection Criteria.

Abstract: The application and introduction of ERP systems have become a central issue for management and operation of enterprises. The competition on market enforces the improvement and optimization of business processes at enterprises to increase their efficiency, effectiveness, and to manage better the resources outside of the company. The primary task of ERP systems is to achieve the before-mentioned objectives. For this reason the selection of a particular ERP system has a decisive effect on the future operation and profitability of the enterprise, i.e. the selection phase is highly relevant step within the introduction and implementation stage of an ERP system. The issues that are worth investigating are the criteria applied at the decision. The qualitative correlation between the size of enterprises, market position, etc. and the applied selection criteria for ERP systems could be analyzed as to whether which criteria are made use of at multinational enterprises or at SMEs. Our research is grounded in a literature review and case studies of everyday practice related to introduction, implementation and roll-out of ERP systems and it tries to provide answers for the above raised questions.

1 INTRODUCTION

Surveys and practical experiences have shown that all areas of enterprise operation have been affected by cost savings including the IT related fields; the main business objectives are modified to increase the economic efficiency in spite of previous business goals. The economic crisis has resulted generally in dramatic impact on IT budgets at enterprises (Thompson, 2010).

In spite of the enduring economic and financial crisis, the introduction and adoption of ERP systems continues. We have investigated the trends in a small EU member country (Hungary) empirically and by publications related to business management and economics. There are clear tendencies that even the small and medium enterprises (SME) that had data processing systems which had been previously developed individually or tailored to the specific requirements started projects to buy ready-made or commercially available Off-the-Shelf (*COTS*) products on the market. The main reason is that the previously developed, legacy systems cannot comply with the recent requirements related to information processing, namely cost-efficiency, staffing level and other labor conditions.

We conclude that there are **individually developed** and **standard systems** within the industrial sector specific solutions.

In this paper, an **ERP system** (Enterprise Resource Planning) is understood as an enterprise-wide, comprehensive information system involving all information processing activities that covers the human resource, production, commercial, planning, inventory, material planning, management control and monitoring business processes by placing them into a unified framework.

In next sections, we analyze the phases of implementation process and providing some answers for the raised issues.

2 RESEARCH METHOD

The research approach as a *methodology* was twofold. We have grounded our investigation in BSc. / MSc theses that were created on ERP at a Hungarian College as students' research project. There was an *empirical* research on architectural approaches of subsidiaries belonging to international companies and operating in Hungary (ELTE, 2010). The research was carried out by a consortium of

Hungarian Universities and Colleges. Beside companies situated in Hungary, the investigation covered practice of ERP introduction at several German companies either based on publications or *in-depth interviews* with managers responsible for ERP systems. There was a comprehensive literature review related to ERP introduction and implementation that we will discuss in detail.

3 CONCEPT OF ERP SELECTION

ERP selection is an important decision making problem of organizations and influences directly the performance. There are a lot of ERP alternatives in market (Wei, 2004). The failure in selection of ERP system firstly leads to the failure of ERP introduction or adaptation project or secondly to degradation of company performance (Liao, 2007).

Several research studies have been conducted to identify relevant factors having impact on success of implementation and introduction at ERP systems. The major part of studies have chosen the case study paradigm, i.e. many of them focused on single case study of ‘‘how we implemented ERP systems in our company’’ (Ang, 1995); (Bingi, 1999); (Mandal, 2002); (Wilson et al., 1994); (Yusuf, 2004). Furthermore, several studies that have measured ERP implementation success used only one or two factors of ERP implementation success (Ang, 1995, 2002); (Malbert, 2003); (Umble, 2003).

The literature research shows that problems with the implementation of ERP systems emerge for a number of reasons. We can summarize briefly the reasons as follows: (1) Generally there is a need for business process change or re-engineering for fitting together the business processes and information processes of an ERP system. Leaving out the required business process alignment could lead later operational problems. (2) Lack of commitment from top management, deficiency in data accuracy, and short of user involvement can attribute to system implementation failures appearing typically during the operation phase. (3) Education and training to make use of ERP system are frequently under estimated and are given less time due to schedule pressures. (4) The synergy demanded by cross-functional business processes are not understood properly.

There are competing measurement approaches and concepts coming from research literature and practice. Some factors that can be encountered in the literature: (1) User satisfaction (Al-Mashari, 2003); (Anget, 1995; 2002); (Mandal, 2002); (Yusuf, 2004). (2) Intended business performance improvements (Al-Mashari, 2003); (Hong, 2002); (Mandal, 2002); (Markus et al., 2000); (Yusuf, 2004). (3) On time (Al-Mashari, 2003); (Hong and Kim, 2002); (Malbert, 2003). (4) Within budget (Al-Mashari, 2003); (Hong 2002); (Malbert, 2003). (5) System acceptance and usage (Ang, 1995; 2002); (Yusuf, 2004). (6) Predetermined corporate goals (Al-Mashari, 2003); (Umble, 2003); (Yusuf, 2004).

4 PHASES OF ERP INTRODUCTION

The reason why a decision may have been made to replace an operational system by an ERP solution can be concluded from several basic causes. One of the origins for such a decision is that the enterprise would like to save or strengthen its market position through acquisitions or internal growth. The IT/IS system can be adjusted to these changes flexibly if there is a “quantum leap” in IS service quality by introducing an ERP system considered as a “best practice” in the industry sector. The other important factor is the competition on the market. Beside the market competition there are other coercive requirements to optimize and to increase performance in enterprise governance and information processing.

The samples coming out of practice demonstrate that the introduction and application of ERP is a longstanding process (Feuchtinger, 2008).There is a seven phase model for ERP introduction: proposal for changeover, analysis, conceptual plan, short listing the potential solutions, selection process, decision for the designated one, and project closure (Zimmermann, 2010).

4.1 Modernization of Operational System

The companies frequently encounter a decision situation how they can modernize existing data processing system. There are three different ways: *development*, *package procurement* and *renting or leasing* the ERP services.

At the beginning it is difficult to decide whether a program package as COTS (Commercial off-the-shelf) should be procured or a vendor should be found to develop a customized solution and who can adapt its basic system to the company’s requirement. The decision is hard as the package solution cannot cover all business processes at the enterprise. A

developed system may comply with requirements and it can be tailor made for specific business processes; however it requires more resources (Ayağ, 2007).

Before the decision between the package *solution* and *development*, an analysis should be carried out on potential solutions then after the analysis a management decision could be made.

The third opportunity is *renting* or paying a fee for all or some services of an ERP system. Most recently, the **ASP (Application Service Providing)** is an appropriate, cost-effective solution for micro and small enterprises. The software as a service can be accessed through the **Cloud Computing**.

4.2 Decision Making on the Introduction of an ERP Solution

The question emerges whether what the factors are that lead companies to consider replacing the operational legacy system fully or partially with a new information system.

A Hungarian Ltd. decided to adopt an ERP (ProFinance™) system, their justification contained three items on the grounds of the underdeveloped, legacy information processing system: (1) The rapidly developing enterprise owned old, legacy information processing system that did not cover all business processes. For this reason, the introduction and implementation of a more modern enterprise management system became the must. (2) In the region, the other, concurrent companies have adopted and will have implemented various management systems gaining competitive advantage. (3) There is intention to develop and to extend the retail branch of the enterprise. For this reason, the new information system should have a steady and reliable on-line connection between the retail shops and the wholesale units.

A company from Netherland had an AS/400 based system named TOTICS and had operated for 20 years. The question "Whether does the company need a new information system and if the answer yes then why?" has been responded as it follows: (1)The new system is pre-condition to realize the business strategy plan; (2) The new IS provides higher reliability and service level for customers; (3) Within the business group, the objective is to increase efficiency and to make more transparent the business processes; (4) The system should support the business planning and consequently the cost-efficiency and serving the consumers; (5) The new IS creates the opportunity for an integrated system. (Tóth, 2008).

The subsidiary of a multinational oil company in Hungary used to employ JDE (J.D. Edwards) ERP system. The company has roughly 100 subsidiaries world-wide and they had applied a wide variety of ERP systems. The company decided to eliminate the heterogeneity of systems. The enterprises wanted one integrated solution. Considering the opportunities, the top management of multinational company made the decision for a project called Global SAP, GSAP project (Kulcsár, 2008). The Dutch company settled to introduce SAP R/3 as well.

In one of our empirical research, we have met the following approach (ELTE, 2010); (Molnár, 2011): some business administration functions are centralized at some regional headquarters as e.g. invoice processing and payment. The customization primarily meant specific parameters that reflect the country specific legal environment. Consequently, a business function is covered totally by a single ERP module introduced during the changeover.

4.3 Objectives of ERP Selection and Practical Approaches

The difficulties in selection of ERP system did not originate from the fact that too few ERP systems is available on market, in spite of it there are multitude of ERP systems. There are hundred vendors beside the major players in Germany (Grandjean, 2010). The primary vendor selection could be based on the market position within the specific ERP sector. (Meyer, 2011).

The investigation of potential ERP solution should take into account business and financial consideration beside the information technology viewpoints (e.g. software and programming environment, information system function etc.). A Hungarian Ltd. had as selection goals for ERP the following criteria: (1) The system supplier should be a *domestic* vendor, the vendor should commit itself for satisfying the users' request for change; (2) User friendly system, easy handling of user interface and ability for customization; (3) Capability for *integration* and *interoperation* with other systems; (4) The IT *stability* of IS should be high. (Csete, 2008)

4.4 Business Case

One of the major objectives during ERP selection is to mitigate the risks inherent in the selection process. Besides the business and technical criteria and risks there are financial ones too. Evaluation methods

include Net-Present-Value, Cost–Benefit Analysis, Payback, Return on Investment, etc. To assess the financial parameters one of the analysis models is the ROI (Return on Investment) that can be applied.

There is an elaborated method that consists of several hundred questions. However, the extensive questionnaire does not solve the problem deriving from lack of information at stakeholders. There is a dearth of reliable information on the following subjects (Gronau, 2010): (1) Knowledge of the actual functions within the ERP system; (2) The applied software and –generally – information technology; (3) The market position, the economic capability, viability of the potential vendor; (4) The comprehensive view of the alternative, competing solutions existing on market; (5) The potential improvement of information processing; (6) The comparative analysis of references for alternative solutions and their implemented instances. ROI is a good compromise for assessing the financial risks of an ERP adaption process and other socio-technical viewpoints. (Lindemann, 2007).

The comparison of the potential alternatives as procuring, renting, leasing or paying per usage for services through Cloud Computing can be carried out by TCO approaches. At a Hungarian Ltd. the TCO model was employed to analyze the costs for introduction and operation (Csete, 2008).

4.5 Soft Criteria for Selection

Besides the service quality and financial criteria, there are lots of other objectives that should be taken into account during the selection process. The compliance to the requirements of the company is one of the most important criteria. To clarify and to define accurately the compliance criteria, a business process modeling exercise should be carried out to discover and to map the whole business process that will be involved in the ERP introduction. To explore the discrepancies between the existing processes and the processes of potential ERP systems, a *gap analysis* should be performed. T

The new ERP system may fulfill the recent requirements; however the ERP system should be prepared for future demands (Lotto, 2006). The *stability* of information systems means the adaptability to changes of technology, business processes and business environment.

The experiences shows that if the set of functions to be automated is minimized for several reasons – financial, compliance, project timing, resources etc. – then later on, the enhancement and evolutionary development to react to the changing environment may cause extra costs and other operational difficulties as against of maximization of set of functions for automation (Grandjean, 2010).

The flexibility of ERP systems is a success criterion within the corporate and SME world (Feuchtinger, 2008). In this context, the *flexibility* is an overarching concept that involves the simultaneous use of various languages carrying out even the same task, at the same time, furthermore adaptation to the changing business and market environment. The top management at the center of enterprises has various opportunities to find a satisfactory solution among the potential ERP systems (ELTE, 2010); (Molnár, 2011). The concrete implementation is situated in the centralization-decentralization continuum - both horizontally and vertically according to the Zachmann architecture - to provide the support that is required the top management of enterprises.

Other uncertainty factor is the structure of business processes and organization and the capability for adjustment to the processes provided by an ERP system. The ERP system adaptation and transformation of business processes has as outcome a solid market position. The ERP system adaptation may have as a side-effect stronger market position, efficient internal business processes and a profound transformation of whole activities in the enterprise.

On selecting an EPR system to support globalized business activities, so-called country specific features should be taken into account. Such features include as follows: Custom and excise handling; Tax, revenue handling; Commercial code; Financial and cost accounting; Banking, rules for bank accounts; Local legal environment, jurisdiction.

The potential ERP system may or may not contain the above listed, country specific features. The required customization needs extra implementation effort generally.

Some examples for the difficulties that occurred (Contini, 2010): Country specific, compulsory Chart of Account (Belgium); Accounting the transfer prices (Brazil); Handling and accounting the billing credit (Bulgaria); Country specific Payroll (Chile).

5 FUTURE RESEARCH DIRECTIONS

The future research should deal with the changing IT environment, especially the proliferation of *Cloud Computing*, the Software as a Service (SaaS), the application as a service, namely the ERP system

Table 1: Factors effect on ERP implementation.

Factors effect on ERP implementation	**Occurence in case studies**						
	weak positive impact	average positive impact	strong positive impact	Neutral	average negative impact	weak negative impact	strong negative impact
Top management support	8	5	6	6	2	3	3
Company-wide support	8	5	4	4	4	4	4
Business process reengineering	1	2	6	6	6	6	6
Effective project management	7	7	7	3	3	3	3
Organizational culture	1	2	4	4	9	6	7
Education and training	2	7	12	3	3	3	3
User involvement	1	2	6	6	6	6	6
User characteristics	2	3	7	3	3	7	8
ERP software suitability	8	8	12	2	1	1	1
Information quality	8	8	12	2	1	1	1
System quality	8	8	12	2	1	1	1
ERP vendor quality	2	8	12	7	1	2	1
Total :	56	65	100	48	40	43	44

services. in this situation, it will be worth investigating how the notion of asp (application service provider) changes and what the particular features may have regarding the erp services.

6 CONCLUSIONS

The multiple case studies and financial analysis models presented in this paper provide assistance for the decision making processes at enterprises where the changeover issue is reviewed.

The results of research can be summarized in a table as a conclusion (Table 1). The assessment of each single factor (Table 1.) is grounded in working up the in-depth interviews, BSc. / MSc theses and other reports, overall *40 companies* were involved in the research.

The project management during implementation and introduction typically followed the traditional pattern, the disciplined project controlling and efficient team organization was a pre-condition of success. The commitment from the top management considered generally a crucial aspect along with a comprehensive support from personnel of the organization.

The result of re-engineering is evaluated by stakeholders with mixed feelings. Nevertheless, the education and training is regarded as having positive influences on the final success of ERP implementation. The user characteristics and education are intimately related so that the technical and business skill of staff contributes to the success of projects. The high level of information quality at the customer organization makes easier the introduction of ERP system as the organization has been already accustomed to provide accurate, timely, reliable and consistent data.

The ERP vendor quality appears in the form of services that are provided together with ERP system implementation and long-term operation. These services includes response time of help desk; knowledgeable consultants with experiences in both enterprise's business processes and information technology including vendor's ERP system. The participation and support of vendor's consultant in implementation and introduction is a significant factor. The services provided by consultants can be characterized by the level of knowledge in both customer's business processes and functions of the particular ERP system.

ACKNOWLEDGEMENTS

The Project is supported by the European Union and co-financed by the European Social Fund (grant agreement no. TÁMOP-4.2.2/B-10/1-2010-0030).

REFERENCES

Al-Mashari, M., Al-Mudimigh, A., Zairi, M., 2003. Enterprise resource planning: A taxonomy of critical factors. European, *Journal of Operational Research 146*(2), 352–364.

Ang, J. S. K., Sum, C. C., Chung, W. F., 1995. Critical success factors in implementing MRP and government assistance, *Information and Management 29*(2), 63–70.

Ang, J. S. K., Sum, C. C., Yeo, L. N., 2002. A multiple-case design methodology for studying MRP success

and CSFs. *Information and Management 39*(4), 271–281.

Ayağ, Z.; ÖzdemIr, R. G., (2007): An intelligent approach to ERP software selection through fuzzy ANP, *International Journal of Production Research*, *45*(10), 2169-2194, 26p.

Bingi, P., Sharma, M. K., Godla, J. K., 1999. Critical issues affecting an ERP implementation, *Information Systems Management 16*, 7–14.

Contini, L., 2010. *Einfluss nationaler Charakteristika in internationalen Projekten zur Einführung von ERP Systemen*. (*Influences of country-specific features on projects for implementation of ERP systems*) Unpublished PhD, Universität Passau, Retrieved January 4, 2012 from http://www.opus-bayern.de/uni-passau/volltexte/2011/2249/pdf/Contini_Nemmert_Luisa.pdf .

Csete, G., 2008. Preparation for introduction of an ERP system at a Hungarian Ltd., *Unpublished BSc thesis at College of Dénes Gábor*, in Hungarian (Csete, G., 2008. Az ERP integrált vállalatirányítási rendszer bevezetésének előkészítése a TMF Magyarország Kft -nél. (GDF azon: 844/2007)).

ELTE (Research Group at Eötvös University) 2010. Research report about the effect of globalization on ERP Systems and their deployment structure at local companies of international enterprises (in Hungarian), *Eotvos University of Budapest (ELTE),* Retrieved 15 June 2010 from, http://www.mtaita.hu/hu/Publikaciok/ERP_Kutatasi_Beszamolo_2010__05__10_.pdf.

Feuchtinger, H., 2008. ERP Auswahl und Einführung. *ERP Management*, 12/2008, 17-19. Retrieved December 30, 2011 from http://www.erpmanager.de/magazin/artikel_1963_projekt_planung_phasen.html, 2008.

Grandjean, W., 2010. Die 10 Gebote der ERP-Auswahl. *ERP Management*, 6/2010, 59-60.

Gronau, N., 2010. ERP-Auswahl mittels RoI-Analyse-Risikoreduzierung und Nutzensteigerung, *ERP Management*, *6* (3), 18-20.

Hong, K. K., Kim, Y. G., 2002. The critical success factors for ERP implementation: An organizational fit perspective, *Information and Management 40*, 25–40.

Kulcsár L., 2008. GSAP project at Shell Hungary Plc. and the requirements for infrastructure, Unpublished BSc thesis at College of Dénes Gábor, in Hungarian (Kulcsár L., (Kulcsár 2008). GSAP project a Shell Hungary Rt-nél és annak infrastruktúra vonzata. (GDF azon: 623/2006)).

Liao, X., Li, Y. and Lu, B., (2007). A model for selecting an ERP system based on linguistic information processing, *Information Systems*, *32*, 1005–1017.

Lindemann, M., Schmid, S., Gronau, N., 2007. Wirtschaftlichkeitsbewertung der Einführung von Manufacturing Execution Systems. *VDMA Nachrichten*; 02/2007, 60-61. Retrieved April 12, 2010 from http://wi.uni-potsdam.de/hp.nsf/0/C2078E0F89C8DE99C12573E300721C72/$FILE/roi_analyzer_fuer_mes.pdf

Malbert, V. A., Soni, A., Venkataramanan, M.A., 2003. Enterprise resource planning: Managing the implementation process, *European Journal of Operational Research 146* (2), 302–314.

Mandal, P., Gunasekaran, A., 2002. Application of SAP R/3 in on-line inventory control, *International Journal of Production Economics*, *75*(1-2), 47–55.

Markus, M. L., Axline, S., Petrie, D., Tanis, C., 2000. Learning from adopters' experiences with ERP: Problems encountered and success achieved, *Journal of Information Technology 15*, 245–265.

Meyer, J., Gronau, N., 2011. Nutzung der Branchenstärke in der ERP-Auswahl. *ERP Management*, *7*, 7/2011, 48-50.

Molnár, B., Szabó, Gy., 2011. Information Architecture of ERP Systems at Globalised Enterprises in a Small EU Member State, *Proceedings of the ITI 2011 33rd, Int. Conf. on Information Technology Interfaces*, June 27-30, 2011, Cavtat, Croatia, ISBN 978-953-7138-20-2, ISSN 1330-1012,

Thompson, E., 2010. Customer Management Summit, *CRM Trends*, Gartner.

Tóth, P., 2008. Introduction and implementation of SAP Accounting and Finance modules, *Unpublished BSc thesis at College of Dénes Gábor*, in Hungarian (Tóth, P., 2008. Az SAP rendszer pénzügyi és értékesítési moduljának bevezetése a Fabory Közép-kelet Európai szervezetében, annak globális gazdasági és informatikai hatásai. (GDF azon: 1285/2008)

Umble, E. J., Haft, R. R., Umble, M. M., 2003. Enterprise resource planning: Implementation procedures and critical success factors, *European Journal of Operational Research 146* (2), 241–257.

Wei, C.-C., Chien, C.-F. Wang, M.-J. J., 2005, An AHP-based approach to ERP system selection, *International Journal of Production Economics*, *96* (1), 1, 47-62.

Wei, C.-C., Wang, M.-J. J., 2004.A comprehensive framework for selecting an ERP system, *International Journal of Project Management*, 22, 161–169.

Yusuf, Y., Gunasekaran, A., Abthorpe, M. K., 2004. Enterprise information systems project implementation: A case study of ERP in Rolls-Royce, *International Journal of Production Economics 87*(3), 251–266.

Zimmermann, M., Fobbe, A. H., 2010. Grundpfeiler einer ERP-Auswahl. *ERP Management*, 6/2010, 56-58.

POSTERS

Modeling Dynamic Systems for Diagnosis
PEPA/TOM4D Comparison

I. Fakhfakh[1], M. Le Goc[2], L. Torres[2] and C. Curt[1]
[1]*IRSTEA, 3275 route de Cézanne - CS 40061, Aix-en-Provence, France*
[2]*Aix-Marseille Univ, LSIS, 13397 Marseille, France*
{*ismail.fakhfakh, corinne.curt*}*@irstea.fr,* {*marc.legoc, lucile.torres*}*@lsis.org*

Keywords: Multi Modeling, Model Based-reasoning, Dynamic System, Process Algebras, Timed Observation Theory.

Abstract: Researchers have long been seeking the most suitable formalism and method to build models of dynamic systems for diagnostic tasks. In this paper, we claim that the main difficulty stems from the lack of global formalism capable of taking into account structural, functional and behavioral knowledge. To illustrate this point, we propose a comparison between two modeling approaches.

1 INTRODUCTION

In the last two decades model-based diagnosis has been an important area of research in which numerous new methodologies and formalisms have been proposed, studied and subjected to experiments (Console et al., 2000) and (Le Goc et al., 2008). This is motivated by the practical need for ensuring the correct and safe operation of large complex systems. Since (Reiter, 1987), most of frameworks have been based on logic formalism. Despite major contributions in the domain of temporal logic, a difficulty remains in taking observation time into account in diagnosis reasoning. Therefore many works have been proposed to define more or less specific formalisms to overcome this limit to the logical representation of timed knowledge, such as the discrete event system (D.E.S) formalism and the multi-modeling approach of (Chittaro et al., 1993). Moreover, these approaches have seldom been used in the context of diagnosis. More recently, PEPA formalism (Performance Evaluation Process Algebra) (Console et al., 2000) and the TOM4D methodology (Timed Observation Modeling for Diagnosis) (Le Goc et al., 2008) have been proposed to provide expressive languages to enable efficient modeling of dynamic systems for diagnosis, comprising a component centered modeling paradigm.

The goal of this paper is to bridge research into process algebras and timed observation modeling (Le Goc et al., 2008) by providing a comparison between PEPA and TOM4D. This comparison is performed with a concrete example(Section 2).

2 A HYDRAULIC SYSTEM

The dynamic system studied in (Console et al., 2000) is described in Figure 1. We use this example to compare PEPA and TOM4D.

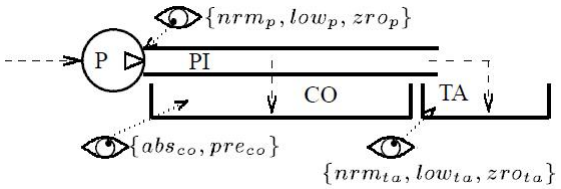


Figure 1: Hydraulic system of (Console et al., 2000).

" *The system is formed by a pump P which delivers water to a tank TA via a pipe PI; another tank CO is used as a collector for water that may leak from the pipe. For the sake of simplicity, we assume that the pump is always on and supplied with water. Pump P has three modes of behavior: OK (the pump produces a normal output flow), leaking (it produces a low output flow), and blocked (no output flow). Pipe PI can be OK (delivering the water it receives from the pump to the tank) or leaking (in this case we assume that it delivers a low output to the tank a when receiving a normal or low input, and no output when receiving no input). Tanks TA and CO are always in OK mode, i.e., they simply receive water. We assume that three sensors are available (see the eyes in Figure 1): flowp measures the flow from the pump, which can be normal (nrm_p), low (low_p), or zero (zro_p); $level_{TA}$ measures the level of the water in TA, which can be normal (m_{ta}), low (low_{ta}), or zero (zro_{ta}); $level_{co}$ records the*

presence of water in CO, which can be either present (pre_{co}) or absent (abs_{co})".

3 PEPA MODEL

The PEPA model is based on classical process algebras enhanced with timed information. Process algebras are abstract languages based on a component oriented approach (Console et al., 2000) where each component is modeled in isolation and then each of the models of the components is composed using the operators provided by the calculation in order to obtain the entire model. In PEPA, the model of a physical system is usually divided into two parts: A behavioral model(BM) and a structural model(SM).

3.1 Structural Model

SM describes the structure of the system in terms of its components. Each component is represented as an instantiation of generic model. In the example studied (cf. Figure 1), four generic behaviors are defined: the "P" behavior (Pump), the "PI" behavior (Pipe), the "TA" behavior (TA tank) and the "CO" behavior (CO tank); also four component instances can be declared: $P^{(1)} : P$; $PI^{(1)} : PI$; $TA^{(1)} : TA$; $CO^{(1)} : CO$. $P^{(1)}$ **:** P means that the component $P^{(1)}$ is an instance of a component whose behavior is P. The connection between them is ensured by the cooperation operator $\underset{L_i}{\bowtie}$ where the sets L_i define the activities on which the components must cooperate. Equation SD_1 describes the SM of the hydraulic system. The SM of the example is:

$$SD_1 \stackrel{def}{=} (P^{(1)} \underset{L_1 \cup \{end\}}{\bowtie} (PI^{(1)} \underset{L_2 \cup \{end\}}{\bowtie} (TA^{(1)} \underset{\{end\}}{\bowtie} CO^{(1)}))$$

where $L_1 = \{nrm_p, low_p, zro_p\}$, $L_2 = \{nrm_1, low_1, zro_1, abs_2, pre_2\}$, $H = \{nrm_0, nrm_1, low_1, zro_1, abs_2, pre_2\}$. The $TA^{(1)}$ tank, for example, cooperates with the $CO^{(1)}$ tank with the "*end*"

3.2 Behaviour Model

The behavior of each component type is described as a nondeterministic choice between the various modes. For example, the BM of the pipe is the following: $PI = PIok_1 + PIlk_1 + End$;
$PIok_1 = nrm_p.PIok_2 + low_p.PIok_3 + zro_p.PIok_4$;
$PIok_2 = nrm_1.abs_2.PI$;
$PIok_3 = low_1.abs_2.PI$;
$PIok_4 = zro_1.abs_2.PI$;
$PIlk_1 = nrm_p.PIlk_2 + low_p.PIlk_2 + zro_p.PIlk_3$;
$PIlk_2 = low_1.pre_2.PI$; $PIlk_3 = zro_1.abs_2.PI$;
End = end.End

For each behavior, a set of equations is defined to specify the relations between the component variables. In particular, the actions of PEPA are used to express conditions on input, output and state variables. $PI = PIok_1 + PIlk_1 + End$ means that the component PI may either be in OK behavior ($PIok_1$) or in leaking behavior ($PIlk_1$). The additional identifier End allows the component to evolve into a final state.

4 TOM4D MODEL

TOM4D is a multi-model approach that combines CommonKads templates with the conceptual framework proposed in (Zanni et al., 2006) and the tetrahedron of states (T.O.S), (Chittaro et al., 1993). These elements are merged according to the Timed Observations Theory (cf Figure 2 more details in (Le Goc, 2006)). In this theory, it is usual to define an observation class C^i=$\{(x_i, \delta^i_j)\}$ as a singleton to associate one variable x_i with a constant δ^i_j. The concept of observation class is close to the notion of discrete event in the D.E.S domain. Figure 3 describes the three main steps of the TOM4D modeling process: The Knowledge Interpretation step uses a CommonKADS template to interpret and organize available knowledge (an expert, a set of documents, etc.) of a dynamic system.

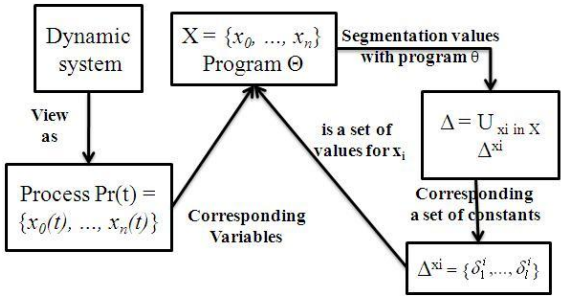


Figure 2: Timed Observation Theory: abstract.

The scenario model $M(\omega) = < SM(\omega), FM(\omega), BM(\omega) >$ of the system is consistent with knowledge available on its evolution over time. This model is necessary to provide, by using the tetrahedron of states, a physical and a logical interpretation of the terms used (variables, constants, etc.). In the example studied two physical dimensions are given for the variables: volume (m^3) and flows of water ($m^3.s^{-1}$) (leaking and normal output). This leads to using the Hydraulic T.O.S. where no pressure (Pr), no resistivity (R) or pressure moment (Pp) are evoked in the

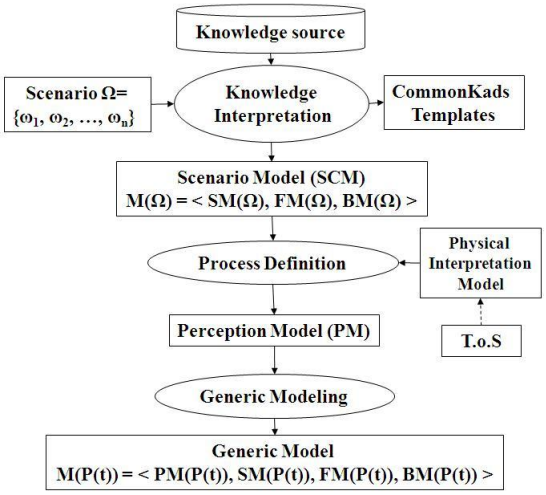


Figure 3: TOM4D Modeling Process.

available knowledge. Thus it is easy to design an abstract generic hydraulic component forming a relation between an input flow $Qi(t)$, an internal volume $V(t)$ and two output flows, a normal output flow $Qs(t)$ and an uncontrolled output flow $Qf(t)$ (Figure 4a). Such a component is generic because it can be used to model all the components of the system.

4.1 Perception Model: PM

The abstract generic hydraulic component is sufficient to define the role of each variable of the system and the associated concrete components.

Table 1 shows the component-variable-value association that can be made according to the abstract generic hydraulic component.

Table 1: component-variable-value association.

COMPS	X	dimension	Action (PEPA)	Δ
c_1	x_7	V	nrm_0, low_0,zro_0	2,1,0
	x_1	Qs	nrm_p,low_p,zro_p	2,1,0
c_2	x_6	V	nrm_{pi},low_{pi},zro_{pi}	2,1,0
	x_4	Qs	nrm_1,low_1,zro_1	2,1,0
	x_5	Qf	$pres_2$,abs_2	1,2
c_3	x_2	V	$nrm_{TA}low_{TA}$,zro_{TA}	2,1,0
c_4	x_3	V	$pres_{CO}$, abs_{CO}	1,2

4.2 Structural Model

A TOM4D structural model $SM(P(t))$ is a 3-tuple $< COMPS, R^p, R^x >$ (cf. Figure 4) where:

- $COMPS$=$\{c_1, c_2, c_3, c_4\}$ is the finite set of constants denoting the system components,
- R^p is a set of equality predicates defining the interconnections between the components. R^p=$\{out(c_1)$=$in(c_2)$, $out_1(c_2)$=$in(c_3)$, $out_2(c_2)$=$in(c_4)\}$
- R^x is a set of equality predicates linking each variable. R^x=$\{$ $out(c_1)$=x_1, $out(c_3)$=x_2, $out(c_4)$=x_3, $out_1(c_2)$=x_4, $out_2(c_2)$=$x_5\}$.

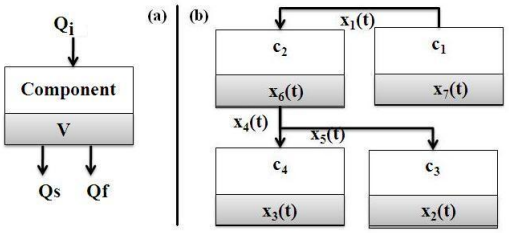


Figure 4: Structural Model $SM(P(t))$.

4.3 Behavioral Model

The behavior model $BM(P(t))$ is a 3-tuple $< S, C, \gamma >$ where $S = \{s_i\}_{i=1...l}$ is a set of states (s_0 for example corresponds on x_6=0 $\wedge$ x_4=0 $\wedge$ x_5=1), C is a set of timed observation classes C^i=$\{(x_i, \delta^i_j\}$ (C^6_1 = $\{(x_6,0)\}$ for example) and γ: $S \times C \rightarrow S$ is the state transition function that implements the state evolution in the system modeled (i.e. $\gamma(s_1, C^6_3) = s_2$).

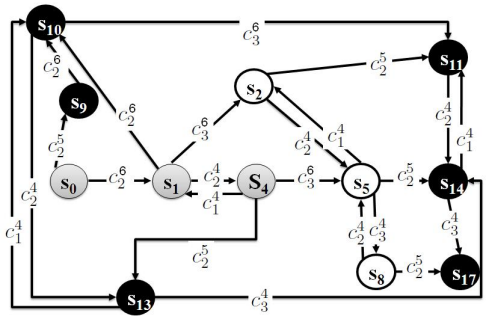


Figure 5: Behavioral Model of the Pipe.

The *ok* and *leaking* PEPA modes of the pipe correspond to the grey and black states in Figure 5, respectively.

4.4 Functional Model: FM

A functional model FM is a 3-tuple $< \Delta, F, R^f >$ where Δ is the set of values assumable by the different variables (Δ_{x_1} = $\{2, 1, 0\}$ for example), F is a set of functions (The result of the T.o.S and structural model denotes 7 functions) and R^f is a set of equality predicates defining a variable as a function of the others. The graph of $FM(P(t))$ is shown in figure 6).

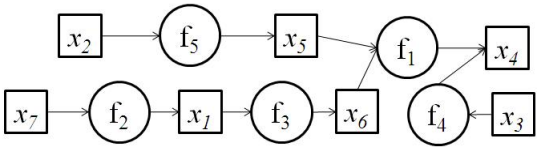


Figure 6: Functional Model of the hydraulic system.

5 DISCUSSION AND CONCLUSIONS

The example studied shows that the TOM4D structural model plays the same role as the declaration of generic component instances, the connection equations and the activities declaration in PEPA formalism.

The functional TOM4D models play the same role as the so called "behavioral" model of components in Reiter's theory. There is no equivalent in PEPA because the process algebras are centered with the description of the behavioral properties of the connected components. In this perspective, the value of a variable at a particular time depends on the different activities at work in the process. Consequently, the FM cannot be modelled in the modeling process.

Process algebras define the set of states through a set of symbols corresponding to an expert's language items, contrary to TOM4D where the states are anonymous: their meanings are provided with the value of the whole set of variables used when the system enters a state. The set of PEPA actions plays the same role as the set of timed observation classes and the behavior definition is similar to the set of transition relations of the TOM4D behavioral models. Such a behavioral model is not covered by Reiter's theory. In other words, a diagnosis model built according to Reiter's theory is formulated with a structural model and a functional model in the TOM4D meaning. A diagnosis model built according to PEPA is formulated with a structural model and a behavioral model.

On the other hand, the TOM4D methodology obliges the experts to define the way they "see" the system in order to model in terms of perception. There is no equivalent in PEPA because it considers the diagnosis model as a consequence of both the system structure and the behavior of its components. This was one of the reason for proposing TOM4D.

An important property of the TOM4D methodology is the use of T.O.S. T.O.S. facilitates the introduction of a physical interpretation to model behaviors having a physical meaning.

From the technical viewpoint, the PEPA model is more compact than TOM4D models. A compact representation is an advantage for the modeler since the lower the number of symbols there are to be defined, the better the model will be.

One the advantages of TOM4D is precisely that its makes explicit the different relations between the terms used by an expert to formulate their knowledge (variable, value, state transition condition, etc). In other words, TOM4D obliges experts to clarify their knowledge when analyzing the system to be modeled according to four points of view: perception, structure, function and behavior. From this standpoint, the graphical representations of TOM4D models are clearly an advantage for interpreting and validating them.

Finally, TOM4D methodology provides concepts and tools to help the modeler to define the correct level of abstraction for efficient diagnosis. The experiments we performed with TOM4D methodology show that this level of abstraction corresponds to that used by an expert to formulate their knowledge of diagnoses applied to dynamic systems.

We are now investigating these approaches to characterize the properties of their diagnosis algorithms (computational and pertinence properties).

ACKNOWLEDGEMENTS

The authors would like to thank the PACA region and FEDER for their funding.

REFERENCES

Chittaro, L., Guida, G., Tasso, C., and Toppano, E. (1993). Functional and teological knowledge in the multimodeling approach for reasoning about physical systems: A case study in diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics 23(6), 1718-1751.*

Console, L., Picardi, C., and Ribaudo, M. (2000). Diagnosing and diagnosability analysis using pepa. *Paper presented at the 14th European Conference on AI.*

Le Goc, M. (2006). *Notion d'observation pour le diagnostic des processus dynamiques : application Sachem et la dcouverte de connaissances temporelles*. Université Aix-Marseille III - FST.

Le Goc, M., Masse, E., and Curt, C. (2008). Modeling processes from timed observations. *(ICSoft 2008)*.

Zanni, C., Le Goc, M., and Frydmann, C. (2006). A conceptual framework for the analysis, classification and choice of knowledge-based system. In *International Journal of Knowledge-based and Intelligent Engineering Systems, 10*, pages 113–138. Kluwer Academic Publishers.

On the Efficient Construction of Query Optimizers for Distributed Heterogeneous Information Systems
A Generic Framework

Tianxiao Liu, Dominique Laurent and Tuyêt Trâm Dang Ngoc
ETIS, CNRS, ENSEA - Cergy-Pontoise University, 95000 Cergy-Pontoise, France
{*Tianxiao.Liu, Dominique.Laurent, Tuyet-Tram.Dang-Ngoc*}*@u-cergy.fr*

Keywords: Distributed Heterogeneous Information Systems, Query Optimization, Search Strategy, Cost Model.

Abstract: It is now common practice to address queries to Distributed Heterogeneous Information Systems (DHIS). In such a setting, the issue of query optimization becomes crucial, and more complex than in centralized homogeneous approaches. Indeed, the optimization processing must be as flexible as possible so as to apply to different database models, and integrate different cost models. In this paper, we present a generic framework for query optimization in the context of DHISs, with the goal of *facilitating* the implementation of *efficient* query optimizers. To this end, we identify all necessary components for building such a query optimizer and we define the basic functions that have to be implemented. Moreover, we report on experiments showing that our approach allows for an efficient query optimization in the context of DHISs.

1 INTRODUCTION

It is well known that one of the main features of relational database systems is to allow for *query optimization*. When optimizing a query Q, Q is first transformed into an initial execution plan, which is then transformed into other equivalent plans using transformation rules. These candidate plans form a search space that is explored by the query optimizer module in order to find an optimized execution plan (*i.e.*, an execution plan having a lower execution cost). As the size of the search space is generally huge, a search strategy is used to efficiently find such an optimized execution plan. Query optimization processing in a given database is based on the following information:

- **Meta-data Model:** database schema, data location, data accessibility, etc.
- **Data and Query Model:** relational, object oriented, semi-structured, services, etc.
- **Optimization Goals:** minimize runtime, minimize money cost, minimize the access to networks, etc.
- **A Search Strategy:** exhaustive, incremental, genetic, dynamic, etc.

When queries are addressed to a single database for which the information above is known in advance, query optimization allows for efficiently minimizing the computation cost. However, it is well known that changing a piece of the information mentioned above requires significant efforts in source code writing.

On the other hand it is now common practice to address queries to Distributed Heterogeneous Information Systems (DHIS). In this setting, the evaluation of a given query requires accessing different heterogeneous data sources, and so, query optimization becomes more complex. Indeed, in a DHIS, the optimization processing must be as flexible as possible so as to (1) consider databases located in different sites, (2) apply to different data models, and (3) integrate different cost models. The contribution of this paper is to propose a *generic framework* for integrating various optimization techniques in order to build efficient optimizers in the context of DHISs. In this framework, we consider the following components:

- Plug-in modules dealing with meta-data, data models, queries, search strategies and transformation rules, respectively.
- Basic functions for an easy and efficient implementation of search strategies.

We have implemented our generic framework, and the experiments reported in this paper show that our approach offers the necessary *flexibility* when designing *efficient* query optimizers for DHISs. More precisely, we show that, by using our approach:

1. the number of code lines for implementing a new

strategy or designing a new optimizer is drastically reduced, as compared with an implementation from scratch;

2. the generated optimized execution plan reduces the processing time by 28 times.

The paper is organized as follows: In Section 2 we describe our generic framework, in Section 3, we report on experiments, in Section 4, we overview related work, and in Section 5, we conclude the paper and offer research directions for future research.

2 THE GENERIC FRAMEWORK

Figure 1 shows the main components of our generic framework, namely:

1. *Source Description*, such as data schema, source location, cost model, etc.
2. *Transformation Rules*, used to transform an execution plan into another equivalent execution plan.
3. *A Collection of Search Strategies*, which are metaheuristic algorithms used for optimization.
4. *A Toolbox of Five Basic Functions*, implemented *only once* and reused to combine optimization processes for different search strategies.

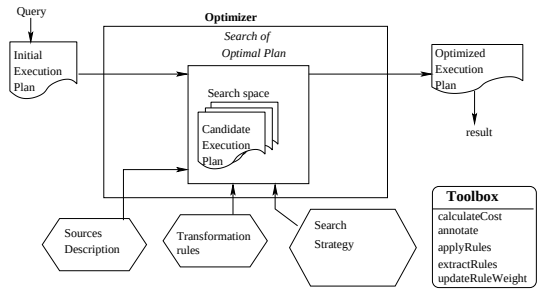


Figure 1: The main components of our framework.

2.1 Generating Execution Plans

We implemented a description module called GSD (standing for *Generic Source Description* and formerly called TGV in (Dang-Ngoc and Travers, 2007)) for annotating any kind of information about data sources. However, any module collecting data source information can be used, as long as it implements the following function, taking an execution plan as input and returning the annotated execution plan:

annotate(exec_plan) → annotated_exec_plan

Given an annotated execution plan, the optimizer should be able to calculate the cost of this execution plan. In our generic framework, this step is achieved through a call to the following function:

calculateCost(exec_plan) → cost_value

Clearly, the implementation of this function depends on the characteristics of the data source where the execution plan is to be run. This information is provided by the annotation returned by the annotate function.

On the other hand, in order to generate a new execution plan from a given one, *transformation rules* are applied. Usually, the following kinds of transformation rules are considered: (1) Logical rules, that reflect basic properties of the underlying algebra; (2) Physical rules that specify the best way a given operation can be computed; (3) User defined rules, such as specific commutation rules.

In our generic framework, such a rule manager is implemented through the following two functions:

extractRules(exec_plan) → set_of_rules
applyRule(rule, exec_plan) → exec_plan

A call to the function extractRules for a given execution plan generates all rules that can be applied for transforming this execution plan into another one. On the other hand, a call to the function applyRule, for a given rule and a given execution plan, applies the given rule to generate a new execution plan.

We emphasize that these four functions do *not* depend on the optimization strategy. Therefore, they are implemented only once for a fixed DHIS, and changing the optimization strategy does *not* require any additional implementation work in this respect.

2.2 Search Strategy

We recall that, in order to avoid searching the *whole* set of execution plans of a given query, search strategies are used. These strategies are based on well known algorithms such as Dynamic Programming (Selinger et al., 1979), Simulated Annealing (Kirkpatrick et al., 1987), or Genetic Algorithm (Goldberg, 1989). Of course each strategy has its own characteristics, and thus, changing from one search strategy to another requires some implementation work.

However, in our generic framework, only the following three functions have to be reconsidered when changing the strategy:

chooseRules(exec_plan, integer) → set_of_rules
updateRuleWeight(rule) → value
getOptimizedPlan(exec_plan) → exec_plan

The function chooseRules returns a set of rules (whose cardinality is the value of the second parameter) chosen from the set of rules returned by a call to extractRules. These rules are applied to the current execution plan, to generate new execution plans

whose costs are computed using the function calculateCost. Finally, the function getOptimizedPlan allows to compute the execution plan with lowest cost, using a given search strategy.

3 EXPERIMENTS

We now present experimental results on the development of our generic framework. The considered DHIS contains 12 distributed heterogeneous data sources (relational, xml, object-relational and Web service), and we have chosen typical queries with inter-site binary operators to be optimized.

Search strategies are compared in the following two respects: We first compare their quality with respect to the *absolute optimal* plan obtained by searching the whole search space, and second, we compare the runtimes needed by each search strategy for the computation of its output optimized execution plan. Figure 2 shows the comparison of various search strategies: Simulating Annealing, Genetic Algorithm, Incremental (Nahar et al., 1986), Dynamic Programming (Selinger et al., 1979), Random (Swami, 1989) and Ant Colony (A. Colorni, 1991).

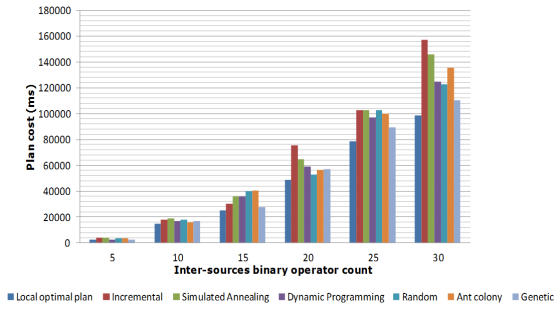


Figure 2: Quality of optimization strategies.

For queries with 10, 15 and 20 inter-site operators, we compare the quality of execution plans found by each strategy with the absolute optimal. We can see that for most strategies, the optimized execution plan is very close to the absolute optimal execution plan.

Figure 3 shows that the time spent for finding the optimized plan increases almost linearly with respect to the complexity of the query being optimized. It can also be seen that the average time for computing the optimized execution plan is in the order of 2 seconds, whereas we found that computing the absolute optimal execution plan takes about 2 minutes in average. This clearly shows that any search strategy is much more efficient than the exhaustive strategy.

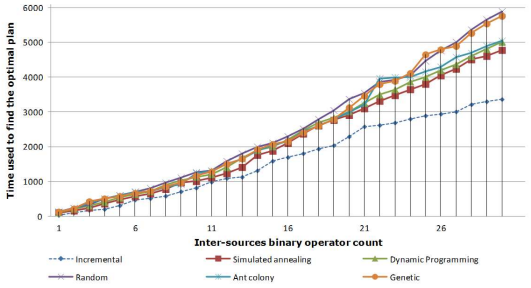


Figure 3: Runtime for the computation of the optimal plan.

We also note from Figure 3 that the Incremental strategy is the most efficient in terms of runtime, whereas the other strategies show similar and acceptable performance. However, Figure 2 shows that, among all other strategies, the efficiency of the Incremental strategy is obtained at the cost of computing the worst execution plan, in terms of quality. More generally, our experiments have shown that the runtime of the optimized execution plan is reduced by up to 28 times, as compared with the runtime of the non optimized initial execution plan. We refer to (Liu, 2011) for more details in this respect.

Regarding now the efforts in coding search strategies, we recall that applying a new search strategy only requires to modify the implementation of two the functions getOptimizedPlan and chooseRules. Our experiments show that the Exhaustive search strategy can be implemented with about 260 lines of Java code. Knowing that the total number of code lines of the whole implementation of the query optimizer is 5000, implementing the Exhaustive search strategy represents only 5.2% of these 5000 lines.

Assuming that the query optimizer has already been implemented using the Exhaustive strategy, changing from the Exhaustive search strategy to the Incremental search strategy requires to replace the 260 code lines of the Exhaustive strategy with the 280 code lines for implementing the Incremental strategy. This represents only 5.2% of the 5000 code lines of the whole implementation. This example clearly shows that our generic framework is flexible enough to allow query optimization in various environments.

4 RELATED WORK

Although search strategies are the core component of query optimization, most mediation systems use exhaustive search strategies on a portion of the search space (System R* (Daniels et al., 1982), DIOM (Liu and Pu, 1997), DISCO (Naacke et al., 1998), Garlic (Roth et al., 1999)) or dynamic programming (Star-

Burst (Haas et al., 1989), Garlic (Roth et al., 1999)). In (Josifovski and Risch, 2002), the authors propose an approach using the AMOS*II* mediator database system, in which a given query is transformed into an executable object algebraic execution plan. The optimization process is based on built-in algebraic operators and a built-in cost model for local data.

Moreover, in (Josifovski and Risch, 2002), Dynamic Programming, Simulated Annealing, or Random can be used as search strategies and so, this approach is the only system that offers the choice between different search strategies, as we do. However, the strategies proposed in (Josifovski and Risch, 2002) are hard-coded, which offers less flexibility than in our approach.

We finally mention that, in (Stonebraker 2008), the author stresses that database systems are now more and more specialized (e.g. OLPT vs. OLAP systems), and thus, that there is no hope in designing efficient common optimization techniques for all these systems. We therefore conclude that our approach of providing a generic framework for the integration of different optimization techniques can contribute in the design of *efficient* and *flexible* query optimizers in DHISs.

5 CONCLUSIONS

In this paper, we have proposed a generic framework for query optimization in the context of DHISs. Our framework is composed of a set of basic functions, whose implementation takes into account all aspects of query optimization such as transformation rule, cost estimation, construction and annotation execution plans, and search strategy. The experimental results reported in this paper show the high flexibility of our framework used to create or upgrade easily optimizers with high performance. Moreover, our experiments also show that using our generic framework allows for significant gains of runtime.

Regarding future work, we plan to investigate the following issues: (i) implementing a cache system in order to optimize the cost computation of a given execution plan, (ii) consider cost models in the context of cloud computing, and (iii) incorporate multi-criteria optimization techniques in our framework.

REFERENCES

Colorni, A. et al (1991). Distributed optimization by ant colonies. In *Conférence européenne sur la vie artificielle*, pages 134–142.

Dang-Ngoc, T.-T. and Travers, N. (2007). Tree graph views for a distributed pervasive environment. In *International Conference on Network-Based Information Systems (NBIS)*.

Daniels, D., Selinger, P. G., Haas, L. M., Lindsay, B. G., Mohan, C., Walker, A., and Wilms, P. F. (1982). An introduction to distributed query compilation in r*. In *DDB*, pages 291–309.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

Haas, L. M., Freytag, J. C., Lohman, G. M., and Pirahesh, H. (1989). Extensible query processing in starburst. In *ACM SIGMOD*, pages 377–388.

Josifovski, V. and Risch, T. (2002). Query decomposition for a distributed object-oriented mediator system. *Distributed and Parallel Databases*, 11:307–336.

Kirkpatrick, S., Gelatt, Jr., C. D., and Vecchi, M. P. (1987). Readings in computer vision: issues, problems, principles, and paradigms. chapter Optimization by simulated annealing, pages 606–615. Morgan Kaufmann.

Liu, L. and Pu, C. (1997). An adaptive object-oriented approach to integration and access of heterogeneous information sources. *Distributed and Parallel Databases*, 5:167–205.

Liu, T. (2011). *Proposition d'un cadre générique d'optimisation de requêtes dans les environnements hétérogènes répartis*. PhD thesis, Université de Cergy-Pontoise, France. (French).

Naacke, H., Gardarin, G., and Tomasic, A. (1998). Leveraging mediator cost models with heterogeneous data sources. In *ICDE*, pages 351–360.

Nahar, S., Sahni, S., and Shragowitz, E. (1986). Simulated annealing and combinatorial optimization. In *23rd Design Automation Conference*.

Roth, M. T., Ozcan, F., and Haas, L. M. (1999). Cost models do matter: Providing cost information for diverse data sources in a federated system. In *VLDB*, pages 599–610.

Selinger, P. G., Astrahan, M. M., Chamberlin, D. D., Lorie, R. A., and Price., T. G. (1979). Access path selection in a relational database management system. In *ACM SIGMOD*, pages 23–34.

Stonebraker, M. (2008). Technical perspective - One size fits all: an idea whose time has come and gone. *CACM*, 51(12):76.

Swami, A. (1989). Optimization of large join queries: Combining heuristics and combinatorial techniques. In *ACM SIGMOD*, pages 367–376.

Proactive Monitoring of Moving Objects

Fábio da Costa Albuquerque[1,3], Ivanildo Barbosa[1,2], Marco Antonio Casanova[1,3], Marcelo Tílio Monteiro de Carvalho[3] and Jose Antonio Macedo[4]
[1]*Departament of Informatics, PUC-Rio, Rio de Janeiro, Brazil*
[2]*Department of Surveying Engineering, Military Institute of Engineering, Rio de Janeiro, Brazil*
[3]*TecGraf, PUC-Rio, Rio de Janeiro, Brazil*
[4]*Department of Computing, University of Ceará, Fortaleza, Brazil*
fcosta@tecgraf.puc-rio.br, {ibarbosa, casanova}@inf.puc-rio.br, tilio@tegraf.puc-rio.br, jose.macedo@lia.ufc.br

Keywords: Moving Objects, Trajectory Analysis, Real-time Monitoring Systems, Web-based Applications.

Abstract: Positioning systems, combined with inexpensive communication technologies, open interesting possibilities to implement real-time applications that monitor moving objects and that support decision making. This paper first discusses basic requirements for proactive real-time monitoring applications. Then, it proposes an architecture to deploy applications that monitor moving objects, are pro-active, explore trajectory semantics and are sensitive to environment dynamics. The central argument is that proactive monitoring based on process models, such as workflows, is a promising strategy to enhance applications that control moving objects. Finally, to validate the proposed architecture, the paper presents a prototype application to monitor a fleet of trucks. The application uses workflows to model truck trips and features a module to extract data from the Web which helps detect changes on road conditions.

1 INTRODUCTION

Positioning systems, combined with inexpensive communication technologies, open interesting possibilities to implement real-time applications that monitor moving objects and that support decision making. An example would be an application to monitor a fleet of tank trucks that distribute fuel to gas stations in an urban environment. Every trip is carefully planned to follow pre-defined routes, avoiding sensitive areas (such as school areas) and periods of the day or routes where the transportation of dangerous cargo is banned and to pro-actively re-route the truck in case of traffic accidents and other events that might cause delays.

We may classify such applications according to different perspectives. The application may use *trajectory semantics*, such as stopping at a point of interest, or the application may use just *raw trajectory data*, such as speed and direction. We cite Alvares (2011), Siqueira and Bogorny (2011) and Moreno, Times, Renso and Bogorny (2010) and as related works in trajectory semantics.

A *reactive* application uses just the past behavior of the objects, as opposed to a *proactive* application that features models of the predicted (future) behavior of the objects and perhaps suggests alternative actions. Proactive computing is investigated in Tennenhouse (2000), which advocates a paradigm shift from human-centered to human-supervised computation. In his perspective, a system to be proactive must: (1) have a direct connection with the real world; (2) be able to execute actions in response to external stimuli; (3) execute actions faster than the human response. In other words, a system with proactive behavior must detect interesting situations before they happen and must be able to handle such situations without human supervision.

Finally, the application may be *sensitive to environment dynamics*, meaning that it monitors the current state of the environment (or even estimates future states of the environment) where the object is moving to base its decisions. Environmental facts are considered when they directly affect the moving object behavior. By contrast, the application may be *insensitive to environment dynamics*, in the sense that it has just a static model of the environment (such as a road map) where the object is moving.

In this paper, we first discuss basic requirements for proactive monitoring applications. Then, we propose an architecture for applications that monitor moving objects, are pro-active, explore trajectory

semantics and are sensitive to environment dynamics.

To achieve proactive behavior, the proposed architecture includes models of the processes behind the moving objects. The prototype application uses workflows to model truck trips. To monitor moving objects, the architecture includes support for real-time trajectory data stream processing. Finally, to account for trajectory semantics and support sensitivity to environment dynamics, the architecture features additional data sources, classified as *(geospatial) static structured data sources (SSD sources)* and *dynamic structured data sources (DSD sources)*. The prototype application uses geospatial databases and georeferenced facts posted in feeds and tweets about the road conditions that may affect the predicted behavior of the trucks.

The contributions of the paper are therefore threefold: a discussion of the basic requirements for proactive monitoring applications; a proposal for an architecture for such applications; and a prototype application to assess the proposed architecture. The central argument is that proactive monitoring based on process models, such as workflows, is a promising strategy to enhance applications that control moving objects.

The rest of the paper is organized as follows. Section 2 describes a motivating scenario. Section 3 discusses basic requirements for proactive monitoring. Section 4 introduces an architecture for proactive monitoring applications. Section 5 presents a prototype application to validate the ideas. Section 6 discusses related work. Finally, Section 7 contains the conclusions.

2 A MOTIVATING APPLICATION

Consider an application to monitor a fleet of delivery trucks, abstractly defined as follows.

Each truck is modeled as a *moving object M* and each trip is described as a *workflow* W_M that defines the customers to be serviced in the trip and the routes to be followed. Each *step p* of W_M either represents *delivering* merchandize at a customer C_p located at place L_p, or *moving* from a place O_p, called the *origin* of p, to a place D_p, called the *destination* of p, through a *route* R_p.

For each moving object M, the system receives a data stream containing the date, time, geographic position and speed. The system transforms this raw data into meaningful events with the help of a geospatial database storing the location of points-of-interest.

The application monitors several trucks, sharing the same underlying road network and the same emergency workflows. A centralized application is desired to integrate the monitoring of the individual trucks, as well as of the events that affect the road network where the trucks move. The application also reduces human interference on the monitoring process to minimize failures due to fatigue.

Consider now the problem of improving the truck monitoring application to become proactive and sensitive to the environment.

Briefly, the first change in the application design is to use the truck delivery workflows to infer their future behavior. The second change is to detect anomalies in the conditions of the roads where the trucks are expected to drive in the next steps of their trips (defined by their workflows). As an example, the system may issue an alert to the driver to proceed more carefully (or even to take an alternate route) when detected that a vehicle, carrying a flammable load, is driving along a road with wet floor ahead.

Finally, we note that we may describe similar scenarios related to other classes of moving vehicles, such as planes and ships. Workflows in this case will be abstractions for flight or sailing plans.

3 PROPOSED ARCHITECTURE

Figure 1 illustrates the proposed architecture. The *Proative Central Monitor (PCM)* is the core component that, as the name implies, coordinates the other components to pro-actively monitor moving objects. The *Planning Manager (PM)* stores and controls the workflows that model the behavior of the moving objects. The *Application Databases* contain auxiliary data such as names and addresses of customers, the road network, etc. The *Moving Objects Monitor* (*MOM*) sends to the *PCM* the structured data stream containing information relative to the real-time monitoring of moving objects: position, trajectory semantic data (i.e., interpreted trajectory data) and other signals from moving objects. The *Mediators* facilitate access to either dynamic or static external data sources.

4 A PROTOTYPE APPLICATION

This section outlines some of the features of a prototype application to monitor a fleet of delivery trucks, along the lines of the application presented in

Section 2. The prototype follows the architecture proposed in Section 4 and the discussion focuses on some aspects of the *Dynamic Structured Data Mediator* and the *Proactive Central Monitor*.

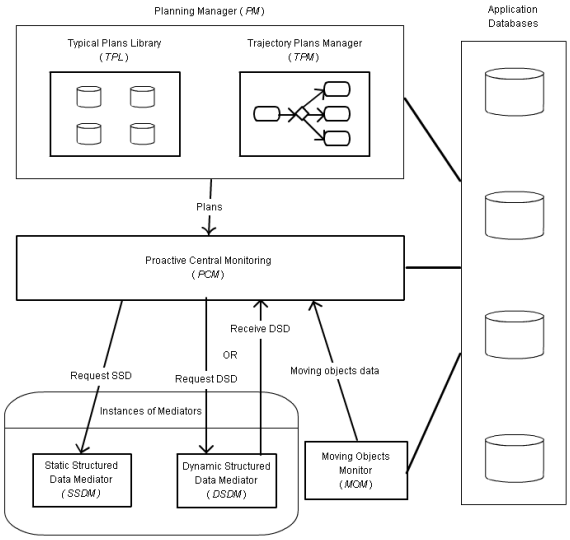


Figure 1: General view of architecture proposal.

4.1 Dynamic Structured Data Mediator

Proactivity is two-fold: situations may be detected from past behavior of the object or from external agents that affect the application.

Santos and Moreira (2010) propose an input for proactive computing by predicting the next step of moving objects based in its current location and road data. Previous moving object data is not used. The success of prediction may vary according to the scenario and variables.

The second approach to proactivity is based on the extraction of relevant facts that potentially affect the future behavior of moving objects.

The prototype implementation of the *Dynamic Structured Data Mediator* (*DSDM*) uses Twitter as the main dynamic structured data source. Similar applications were deployed by Carvalho, Sarmento and Rossetti (2010) and MacEachren et al. (2011). The prototype considers tweets from a predefined list of institutions, assessed as trustworthy sources, as well as from users related to the primary sources (e.g. followers).

The implementation follows the second strategy listed in Section 4.2, that is, the *DSDM* is responsible for post-processing the results returned by the wrappers. As illustrated in Figure 2, the *DSDM* receives raw data containing text body, source, user, location (when available), number of re-tweets, hashtags and time stamp. It then filters tweets according to their creation date and keeps only the most recent ones. At the classification step, the *DSDM* selects only the text body and the source. It classifies tweets according to the occurrence of relevant facts in the text body (e.g. car crashes, floods and road blocks). After filtering the relevant tweets, the *DSDM* extracts the spatial reference for the reported fact, with the help of a street gazetteer stored in the *SSDM*. Finally, the *DSDM* transforms the extracted data into a predefined structure before sending the data to the *PCM*.

4.2 Proactive Central Monitor

The prototype implementation of the *Proactive Central Monitor (PCM)* processes facts and events it receives from the *DSDM* and the *MOM* as follows.

For each moving object M, with workflow W_M, the *PCM* uses the events the *MOM* sends to monitor the step c that W_M is executing. It then simulates the steps of W_M that may follow c, up to a certain depth, and collects the routes that M may traverse.

Next, the *PCM* verifies if such routes are affected by a fact that the *DSDM* has already sent. If this is the case, the *PCM* warns the (human) controller or the driver, or both, that future steps planned for M may have to be changed or aborted.

For simple facts, the *PCM* just generates warnings both to the controller and the driver, but it does not recommend that W_M be necessarily changed. For example, a fact reporting heavy traffic in a route generates just a delay warning to the driver or even suggests an alternative route.

However, some facts may imply restrictions to traffic, even if temporarily. In this case, the *PCM* recommends to the controller that W_M be changed or aborted. The controller then invokes the route planning component (outside the scope of this paper) to create a new version of W_M.
The route planning component is prepared to create routes that consider a list of traffic restrictions (usually maximum load and maximum height permitted, forbidden cargo traffic hours, etc…).

Finally, the *PCM* may also receive events from the *MOM* that represent incidents involving M (e.g. a mechanical problem with M). It then invokes workflows, stored in the *TPL*, to mitigate the incident and eventual damages to the environment (e.g. to clean up an oil spill).

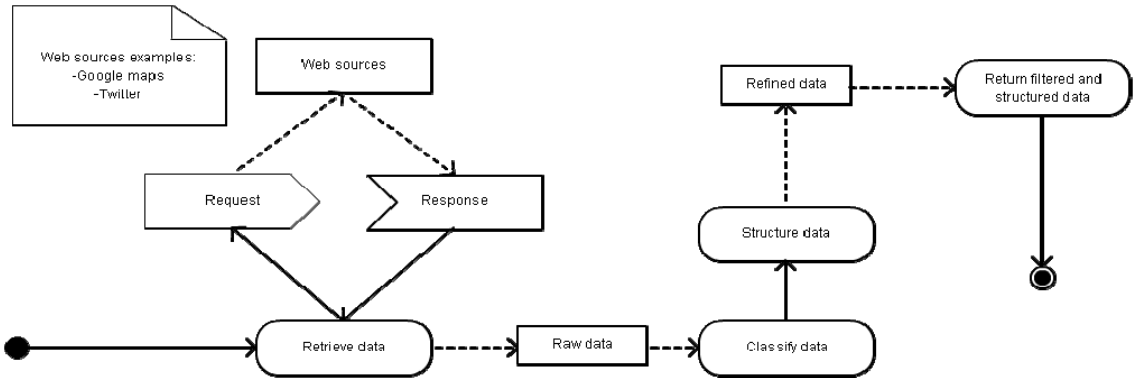


Figure 2: Data flow of the *DSDM*.

5 CONCLUSIONS

In this paper, we first discussed basic requirements to achieve proactive monitoring of moving objects. Then, we proposed an architecture that meets the requirements. The first key point of the discussion is to model the process behind a moving object as a workflow to be able to infer future actions. The second key point is to monitor or even to predict changes in the environment by exploring dynamic data sources.

Finally, we outlined some of the features of a prototype application to monitor a fleet of delivery trucks. In particular, the prototype uses Twitter as a viable dynamic data source to detect changes in the current road conditions, as well as to register future, planned changes that may affect the traffic in certain roads.

We plan to improve the prototype application in several directions. In particular, we intend to explore a supervised strategy to address the problem of classifying facts extracted from tweets. We also plan to explore RSS feeds as a dynamic data source (Chen et al, 2007) and to automatically analyze Web site containing news and weather reports as a viable source of dynamic information.

REFERENCES

Alvares, L. O., Loy, A. M., Renso, C., and Bogorny, V., 2011. An algorithm to identify avoidance behavior in moving object trajectories. In: *Journal of the Brazilian Computer Society 11*.

Carvalho, S., Sarmento, L. and Rossetti, R., (2010). Real-Time Sensing of Traffic Information in Twitter Messages. In: *ATSS @ IEEE ITSC 2010 Proceedings of 4th Workshop on Artificial Transportation Systems and Simulation*, Madeira, Portugal

Chen, Y. F., Di Fabbrizio, G., Gibbon, D., Jora, S., Renger, B., Wei, B., 2007. Geotracker: geospatial and temporal RSS navigation. In *16th International Conference on World Wide Web*. pp. 41-50. Alberta.

MacEachren, A. M., Robinson, A. C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J., Mitra, P., 2011. Geo-Twitter analytics: applications in crisis management, In 25th *ICC, International Cartographic Conference* [available online]

Moreno, B., Times, V. C., Renso, C., and Bogorny, V., 2010. Looking inside the stops of trajectories of moving objects. In *XI Brazilian Symposium on Geoinformatics*, pp. 9–20, Campos do Jordão.

Santos, M. Y., and Moreira, A., (2010). GUESS: on the prediction of mobile users' movement in space, In: Wachowicz, M. (Ed.) *Movement-Aware Applications for Sustainable Mobility: Technologies and Approaches*, IGI Global Publishing.

Siqueira, F. L. and Bogorny, V., 2011. Discovering chasing behavior in moving object trajectories. In *Transactions in GIS, 15(5)*.

Tennenhouse, D., 2000. Proactive computing. In: *Comm. ACM*, vol. 43, May. 2000, pp. 43–50.

Static Parameter Binding Approach for Web Service Mashup Modeling

Eunjung Lee and Hyung-Joo Joo
Department of Computer Science, Kyonggi University, Suwom, South Korea
{ejlee, joo8214}@kyonggi.ac.kr

Keywords: Service Composition, Multiple Parameters, Parameter Binding, Mashup.

Abstract: The most essential aspect of integrating web services to create mashups is determining parameter bindings for the connected requests. However, binding multiple parameters from a large and complicated xml tree is something that has not been discussed in the literature. In this paper, we presented a multi-parameter binding algorithm for repeated and nested xml trees. Moreover, we are interested in context-based parameter bindings, for scenarios where the user selects a certain context node. The proposed binding approach allows for the automatic integration of methods, even when the binding data is inside a repeated group or deep in the nested level. As a result, we can generate navigation menus depending on the contexts for the bound methods. In addition, we present a method for generating navigation codes (context menus) for the mashup views, using the parameter bindings. To demonstrate the usability of the proposed approach, we present an example of a course registration system.

1 INTRODUCTION

Service compositions and mashups have become one of the most important technologies in the development of new web applications and services. With the increasing availability of web services and the dynamic nature of these services, user-centric client-side mashups have attracted considerable attention (Pietschmann, 2010). On the other hand, a difficulty of client-side mashup pages is that they often have to interact with many services and resources.

To support a dynamic service environment, it is necessary to support the automatic generation of codes from a given set of service methods. In addition, the design of the client mashup page navigation may be complicated when it comes to handling several service requests and responses. To support the generation of navigational code for a mashup page, this paper aims to detect possible service compositions for a method's output data, as well as data bindings for the corresponding parameter passing.

In a previous paper (Lee, 2010), we introduced the concept of parameter binding the process of deciding data elements for parameters of the next request. We also introduced the concept of repeat binding, i.e., deterministic binding for the current context of the repeated part of the output tree. However, evaluating bindings for a context node is challenging if the tree has a complicated and nested, repeated structure.

This paper focuses on an algorithm for evaluating the parameter bindings of a nested, repeated structure xml tree. We introduce a top-down binding approach, using xml schema definitions, for the static evaluation of all possible bindings.

As an extension of the previous paper's code generation system, we implemented context menu generation for the multiple parameter bindings of each output view. Our approach can identify a useful set of mashup menus for a given client page context, minimizing user interactions. To the best of our knowledge, previous studies have not considered user interface issues that arise from such compositions.

This paper is organized as follows. Section 2 discusses related studies and provides background. Section 3 describes our models and introduces the concept of repeated bindings. Section 4 presents proposed method for context dependent XML parameter bindings. Section 5 briefs the

This work was supported by the GRRC program (GRRC Kyonggi 2012B03) of Gyeonggi province.

implementation of the parameter bindings as an extension of the code generation system. Section 6 concludes the paper.

2 RELATED WORKS

Service composition involves integrating services by connecting and relaying data. Mashups and data integration have recently been studied in depth (Pietschmann, 2010). In this paper, we are interested in service composition methods and user interface development.

There have been many pieces of research on user-interface development approaches for service compositions. Recently, several client-side service composition and execution frameworks have been published. MashArt is a framework that is intended to act as a component-based development tool for mashups, integrating all three layers of application, data, and presentation. Nestler et al. proposed a model-driven approach to develop a user interface for service compositions. Several main research frameworks have been compared and discussed, in terms of their application composition at the presentation layer (Pietschmann and Waltsgott, 2010). Lastly, the authors' of this work have published a previous paper that presented a code generation approach for client-side service navigation (Lee, 2010).

The most essential aspect of integrating web services to create mashups is determining parameter bindings for the connected requests. However, binding multiple parameters from a large and complicated xml tree is something that has not been discussed in the literature. There have been several studies on XML schema inclusion tests (Hosoya, 2003). In this paper, we presented a multi-parameter binding algorithm for repeated and nested xml trees. Moreover, we are interested in context-based parameter bindings, for the case when the user selects a certain context node.

3 MODELS AND SCENARIOS

In this paper, we are interested in finding parameter values for issuing requests to other service methods, from a given output xml tree. For an xml tree *T* and a method *m*, identifying parameter bindings in *T* means mapping values from *T* to the method parameters. Therefore, by binding parameters, a method becomes callable, since parameters are then ready to be provided. Computing a parameter binding is not straightforward if the number of parameters is more than one and if the xml tree has many repeated nodes and a complicated nested structure. In addition, our main concern is to consider the context of a selected node.

The running example in this paper is a course schedule xml tree, as shown in Figure 1, and a set of search methods for the courses and schedules and create/delete registration methods.

For example, in Figure 1, there are several repeated nodes in the tree, including grade, course, class, and slot. When a classcode node is selected, then department, year, course, and class are bound for the given context. Therefore, a user can request SearchClasses(dept, lecturer) and AddRegister(st_id, classcode) in that context. On the other hand, when the slot node is selected, we can find parameter bindings for SearchClasses(dept, day) or SearchClasses(dept, room), as shown in Figure 1.

Figure 2 shows the xml schema of the course schedule xml tree, which includes five repeated nodes, nested in depth. In this example, we assume some global values, such as student id and today's date.

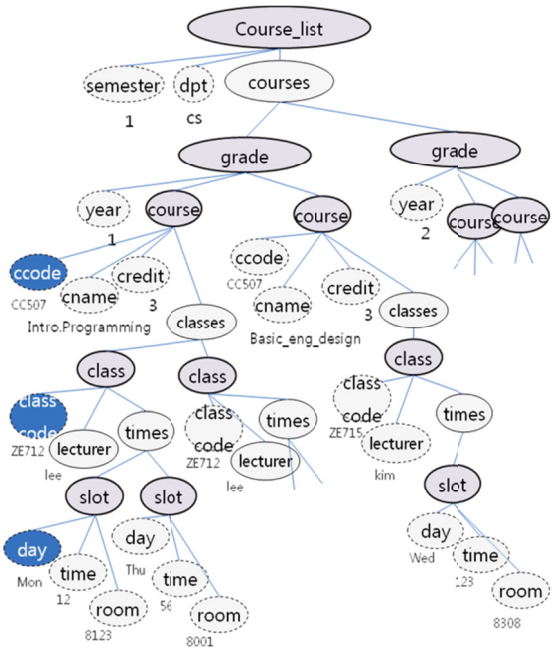


Figure 1: Tree instance and the available requests.

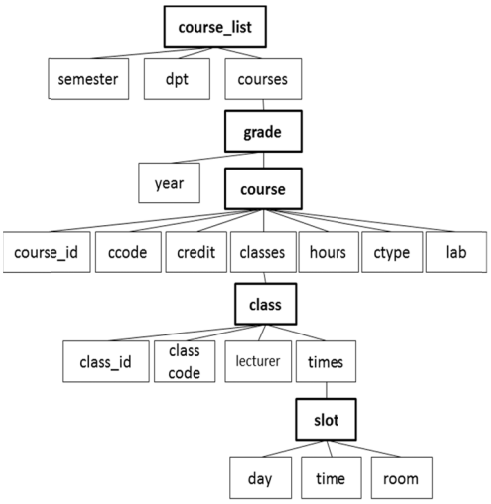


Figure 2: Schema of the xml tree in Figure 1.

4 CONTEXT-DEPENDENT XML PARAMETER BINDINGS

Composing methods by relaying output data to input parameters is an important technique in the implementation of web service mashups. For a complex, structured xml tree, it is difficult to find the corresponding parameter bindings for request methods. Moreover, dynamic parameter bindings for a selected node require inspections of every terminal element under the current context. Therefore, we object to present methods that can identify the context-dependent parameter bindings, statically.

4.1 Repeated Trees

When a schema definition is given, we can extract repeated tree structures (Lee, 2006), as shown in Figure 3. A repeated tree consists of repeated nodes; a repeated node is a node that is defined more than once in the xml schema. For computational convenience, the root node is considered a repeated node. Also, terminal nodes are grouped into repeated nodes, where each repeated node includes direct terminal descendants. Therefore, for a given repeated node r, the set of direct terminal descendants and the path from the parent repeated node are denoted by *direct_terminals(r)* and *rel_path(r),* respectively. In the example in Figure 4, repeated nodes are course_list, grade, course, class, and slot. For the class repeated node, the parent repeated node is course and the child is slot. Moreover, direct_terminals(class) = {classcode, lecturer} and rel_path(class) = course_classes/. Note that @st_id and @today are global static values.

4.2 Parameter Bindings

For the example in Figure 4, we present our approach of parameter bindings with following service methods:

- m1 = SearchCourses(dept, syear)
- m2 = SearchCourses(dept, ctype)
- m3 = SearchClasses(dept,lecturer)
- m4 = AddRegister(user_id, classcode)
- m5 = SearchClasses(dept, lecturer, day).

At each repeat node, we can find parameter mappings using direct terminals. For a given method *m* and its parameters, a binding table for m, denoted as btable(m), is defined as a tuple of elements as many as m's parameters. For the example of Figure 4, btables are shown in Table 1.

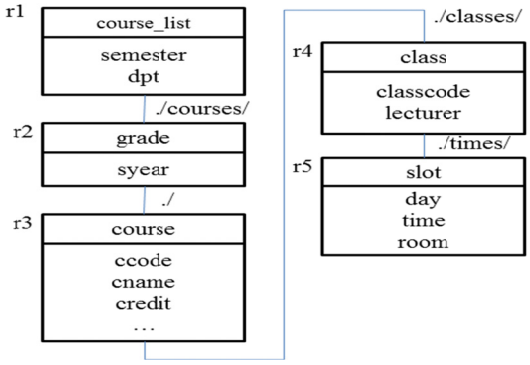


Figure 3: The repeat tree of the course schedule schema.

For each method, we can find the repeat node where the parameter binding is completed during a top down traverse. This repeated node is called "a base context node." For the above example, parameter bindings of m1 and m2 are finished at r2 and r3, respectively. Therefore, m1 and m2 have r2 and r3 as the base nodes, respectively.

Then, the binding path for each parameter is computed from the base context node. For the example m3, the binding paths for the parameters are `[//course-list/dept, ./lecturer]`. Now, we define *parameter_bindings* for a method *m* as follows.

Definition. Let $m \in Methods$ and r be a repeated node in *output(m)*. Moreover, let m have input parameter types $[p_1, p_2, ..., p_k]$. Then, the parameter bindings for m are defined as follows: *parameter_bindings(m)* = $\{(r, [\pi_1, \pi_2, ..., \pi_k] \mid r$ is a base context node of m in *output(m),* and π_i is a relative path of the node which is mapped to $p_i\}$.

Algorithm 1: Top-down computing of XML parameter bindings.

```
        Input: Methods = [m1, m2, ..., mk].
               params(mi) = [pi1, pi2, ..., pni], ni is the number of parameters of i-th method mi.
               RT: the repeat tree of the output type schema tree, root: the root node of RT.
        Output: parameter_bindings
1       Procedure findAllParamsBindings(root):
2           ∀ 1 ≤ i ≤ k,
3               Let btable[i] = [b1, b2, ..., bni], bj = null, 1 ≤ j ≤ ni.
4               parameter_bindings(mi) = null.
5           Call bind_repeat(root).
6       Procedure bind_repeat(r):
7           ∀ mi ∈ Methods, let btable[i] = [b1, b2, .., bni] and binding of mi is not finished,
8             ∀ px s.t. x-th parameter of mi where bx = null, 1 ≤ x ≤ ni,
9           ∃ tj ∈ direct_terminals(r), s.t. tj ~ px ,
10                bx = (r, tj).        // r is the base current node of bindings(mi)
11            If btable[i] is filled at this level,    // binding is now finished,
12          for 1 ≤ x ≤ ni , let bx = (r', tj ), r≠r'
13                πi = a relative path of (r', tj ) from the current repeated node r,
14              Let parameter_bindings(mi) ← add (r, [π1, π2, ..., πni]).
15          If there is any method where binding is not finished,
16              ∀ r' ∈ repeat_child(r),
17                call bind_repeat(r').
```

Table 1: Binding tables on each repeated node of Figure 3.

		r1	r2	r3	r4	r5
m1	dpt	O				
	syear		O			
m2	dpt	O				
	ctype			O		
m3	dpt	O				
	lecture				O	
m4	user_id	@st_id				
	classcode				O	
m5	dpt	O				
	lecturer				O	
	day					O

For the example of Figure 3, parameter bindings for methods m1 to m5 are as follows:

- parameter_bindings(m1) = {(r2, [../../dpt, ./year])}
- parameter_binding(m2)={(r3, [../../../dpt, ./ctype])}
- parameter_bindings(m3) = {(r3, [./cname])}
- parameter_bindings(m4)= {(r4, [../../../../dpt, ./lecturer])}
- parameter_bindings(m5)={(r4, [@st_id, ./classcode])}

Now, we are ready to present the parameter binding algorithms using the notations introduced thus far.

Algorithm 1 traverses the xml schema tree from top to bottom, to identify the mapping of data elements for the parameters of methods in *Methods*. The algorithm visits repeated nodes through their parent-child relations (line 17), gathering the node paths that are matched to method parameters (line 9). If the binding is completed at the repeated node *r* (line 11), then, r is the binding context base. Once we find the base context node of *m*, we evaluate the paths of the matched nodes (line 13) and parameter binding is completed for the method. The top down recursive call is continued while methods remain to be bound and while there are more descendants to visit.

4.3 Generating Context Menus

The *parameter_bindings(m)* refer to the parameters that are ready when the repeated node is bound. Therefore, the method *m* is callable when a user selects one of the repeated nodes. Context popup menus include all callable method requests for the selected context. Therefore, we need to compute all methods bound to a given repeated node from *Methods,* a set of available methods. Thus, we can define a set of callable methods for a given repeated node r as follows:

$$callable(r) = \{m \mid m \in Methods, (r, pb) \in parameter\_bindings(m) \text{ for some } pb\}.$$

For example, the schema and the parameter bindings in Figure 3 show that *callable(r4) = {m3, m4}*. On the other hand, the repeated node represents a repeated level, so if any of the terminal values at this level are selected, then the corresponding methods can be called. For example, selecting a node `ccode` determines the repeated node `course` and its direct terminal descendants.

For a given xml tree, we have an output view

rendered with terminal nodes, as shown in Figure 5. The context menu is provided for a request call for the bound method.

5 IMPLEMENTATION

In this section, we introduce the implementation result of the top down parameter binding methods by Algorithms 1 and 2, introduced in Section 4. In a previous study, the authors introduced the MashupBench system (Lee, 2010), which is a platform providing service selection, data mapping, and mashup code generation. Figure 4 shows the overall architecture of the system. In this paper, we extended the analyzer and the code generator allows multiple parameter bindings for complicated and repeated xml tree structures. We also enhanced the popup menu-generating algorithm to efficiently generate javascript code.

The system takes WADL (Web Application Description Language) files, which is the standard for describing REST style services (WADL, 2006), as service description inputs to specify the available services. Schema files are read by the analyzer to identify parameter bindings for the service methods.

To enable efficient code generation, we construct all popup menus when the view is created at the time of output data response. Since we statically computed parameter bindings beforehand, only menu visibility and event handling run dynamically.

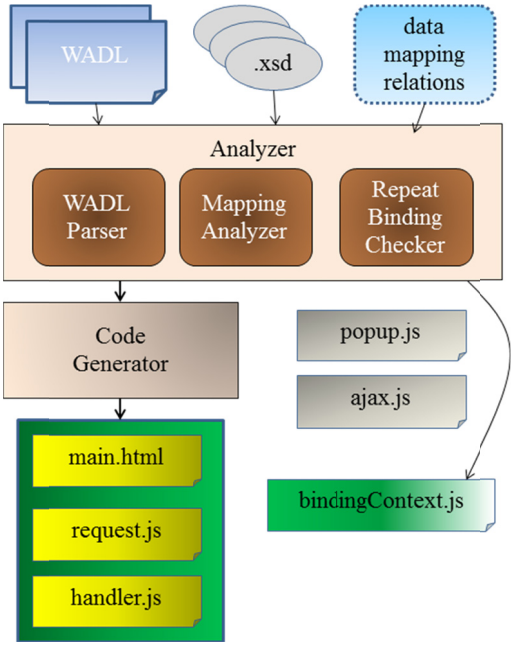


Figure 4: Architecture of code generation system using the proposed approach.

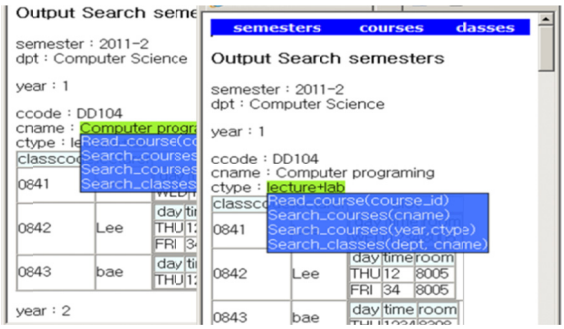


Figure 5: Static-time generated popup menus for the same repeated node.

6 CONCLUSIONS

In this paper, we presented a multi-parameter binding algorithm for repeated and nested xml trees. Moreover, we are interested in context-based parameter bindings, for scenarios where the user selects a certain context node.

The proposed binding approach allows for the automatic integration of methods, even when the binding data is inside a repeated group or deep in the nested level. As a result, we can generate navigation menus depending on the contexts for the bound methods. In addition, we presented a method for generating navigation codes (context menus) for the mashup views, using the parameter bindings.

Since current mapping approaches do not consider nested repeat structure, our methods could be applied to service mashup frameworks to enhance the mashup connections. We are working on implementing the incremental computation of the parameter bindings.

REFERENCES

Stefan Pietschmann, et al., 2010. A Thin-Server Runtime Platform for Composite Web Applications. *ICIW '10*.

Florian Daniel, et al., 2009. Hosted universal composition: models, languages and infrastructure in MashArt. *ER'2009*.

Eunjung Lee and Kyong-Jin Seo, 2010. Designing Client View Navigations Using Rest Style Service Patterns. *WEBIST'2010*.

Web application description language (WADL), http://www.w3.org/Submission/wadl.

Eunjung Lee, 2006. Inline binding of XML data. Proc. Of ICMOCCA'06.

Haruo Hosoya, 2003. Boolean operations and inclusion test for attribute-element constraint, *ICIAA'03*.

Generalized Independent Subqueries Method

Tomasz Marek Kowalski[1], Radosław Adamus[1], Jacek Wiślicki[1] and Michał Bleja[2]
[1]*Computer Engineering Department, Technical University of Lodz, Lodz, Poland*
[2]*Faculty of Mathematics and Computer Science, University of Lodz, Lodz, Poland*
{tkowals, radamus, jacenty}@kis.p.lodz.pl, blejam@math.uni.lodz.pl

Keywords: Query Optimization, Independent Subqueries, Object-Oriented Database, Stack-Based Approach, SBQL.

Abstract: The following paper presents generalisation of the independent subquery method for object-oriented query languages. A subquery is considered independent if none of involved names is bound in a stack section opened by a currently evaluated non-algebraic operator. Optimisation of such a subquery is accomplished by factoring it out from a loop implied by its query operator. We generalise the method to factor out also subqueries that are evaluated only in a context of independent subqueries of a given query. The query is rewritten to an equivalent form ensuring much better performance. Our research bases on the Stack-Based Architecture of query languages having roots in semantics of programming languages. The paper illustrates the method on an comprehensive example and finally presents the general rewriting rule.

1 INTRODUCTION

The ODRA system (Lentner and Subieta, 2007) is an environment facilitating development of object-oriented data-intensive and distributed applications. The main component of ODRA is SBQL (Stack-Based Query Language) (Lentner and Subieta, 2007; Subieta, 2008 and 2009). SBQL evolved from a pure database query language to a fully-fledged object-oriented programming language with a lot of features such as an UML-like object model, collections constrained by cardinalities, processing semi-structured data, static type-checking, closures, etc. As a query language, SBQL is supported by a query optimiser, which contains a set of optimisation methods, including query rewriting (Płodzień, 2000), indices (Kowalski et al., 2008). We have adapted and generalised some of them from relational database systems, but in majority they are totally new. In this paper we propose one of such new powerful optimisation methods that has not been presented yet in any source.

Analysing query evaluation in the Stack-Based Approach (SBA) (Subieta, 2008) one can notice that some subqueries are processed multiple times in loops implied by non-algebraic operators, despite the fact that in subsequent loop cycles their results are the same. Such subqueries should be evaluated only once and their result reused in next loop cycles. This observation is a basis for an important rewriting optimisation technique called the method of independent subqueries (Płodzień and Kraken, 2000, Płodzień, 2000). This method is more general than classical pushing of a selection/projection known from relational system and SQL (Ioannidis, 1996). In SBA it works for any kind of a non-algebraic query operator and for any object-oriented database model.

The generalised independent subqueries method belongs to the group of optimisation methods based on query rewriting. Rewriting means transforming a query Q_1 into a semantically equivalent query Q_2 providing much better performance. It is accomplished according to rewriting rules based on locating parts of a query matching some pattern. These parts are to be replaced by other parts according to these rules. The main benefit from rewriting is that algorithms are fast, optimisation is performed before a query is executed and resulting performance improvement can be very significant, sometimes several orders of magnitude (concerning queries' response times).

Presented method includes cases where an independent query is divided into two or more parts (within a larger query), which makes more difficult to detect and factor out. We show that there is an efficient rewriting rule to factor an independent subquery out of a non-algebraic operator together with its dependent subqueries that are also independent of this operator. For example, consider a query – *for pairs being a Cartesian product of all*

company employees and departments, taking into consideration only departments whose bosses earn more then 2000, return a reference to an Emp object together with a communicate indicating whether a salary of the employee is greater than an average salary calculated for employees working in a given department:

```
(Emp as e) join (((Dept where
boss.Emp.sal > 2000) as d).
(e.fullName() + (if (e.sal >
avg(d.employs.Emp.sal)) then "earns"
else " doesn't earn ") + " more than
an average salary of " + d.name + "
department."))                    (1)
```

In this case the subquery (Dept where boss.Emp.sal > 2000) as d) is independent from the join operator hence it will be factored out of this operator by the method of independent subqueries. In the result of transformation performed by this method we obtain the following query:

```
(((Dept where boss.Emp.sal > 2000)
as d) groupas aux1).(Emp as e) join
(aux1. (e.fullName() + (if (e.sal >
avg(d.employs.Emp.sal)) then "earns"
else " doesn't earn") + " more than
an average salary of " + d.name + "
department."))                    (2)
```

Unfortunately, form (2) terminates the optimisation action – no further optimisation by means of the independent subqueries method is possible any more. This method cannot factor the subquery ***avg***(*d.employs.Emp.sal*) out of the join operator, despite none of its names (d, employs, Emp, sal) being bound in the stack section opened by this operator. The reason is that this subquery is not independent of its parent non-algebraic operator (the dot operator after second *aux1*). However, it is possible to factor out also the subquery ***avg***(*d.employs.Emp.sal*) in (1), because it depends only on the independent subquery ((*Dept* **where** *boss.Emp.sal* > 2000) **as** *d*). Such a transformation will result in limiting the number of its evaluations. This paper explains how such cases can be generally formalised and what a corresponding rewriting algorithm should be.

The rest of the paper is organised as follows. Section 2 describes the overall idea of the generalised independent subqueries method. Section 3 presents the results of simple experiments with the method. Section 4 presents conclusions.

2 THE GENERALIZED METHOD

To present SBA and SBQL in the following examples, we use an object store realising a class diagram (schema) presented in Fig.1. It defines three collections of objects: *Person*, *Emp*, and *Dept*. *Person* is the superclass of the classes *Emp*. Names of classes (attributes, links, etc.) are followed by cardinality numbers (cardinality [1..1] is dropped).

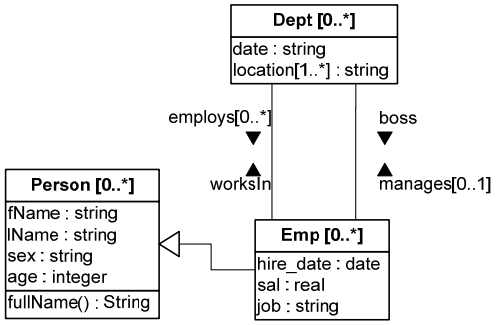


Figure 1: A schema of an example database.

2.1 The General Idea of the Optimisation Method

The starting point for the optimisation process is the method of independent subqueries. If this method detects an independent subquery like the following subquery of (1):

```
((Dept where boss.Emp.sal > 2000) as
d)                                (3)
```

that is a left-hand subquery of some non-algebraic operator then we must analyse the right-hand subquery of this operator. The subquery *((Dept* ***where*** *boss.Emp.sal > 2000)* ***as*** *d)* in (1) is connected by the first *dot* operator with the subquery (4):

```
(e.fullName() + (if (e.sal >
avg(d.employs.Emp.sal)) then "earns"
else " doesn't earn ") + " more than
an average salary of " + d.name +
"department.")                    (4)
```

Because all the names occurring in *avg*(*d.employs.Emp.sal*) are bound in the stack section opened either by first *dot* (name *d*) or in sections opened by non-algebraic operators of this query (the other names), this query is dependent only of the subquery (3). The subquery (3) and *avg*(*d.employs.Emp.sal*) are parts of the right-hand subquery of the *join* operator. Since the subquery *avg*(*d.employs.Emp.sal*) is also independent of this operator so it can be factored out of this operator together with the independent subquery (3). To achieve it we construct a query involving the subqueries like (3) and *avg*(*d.employs.Emp.sal*) connected by the *join* operator and factor them out

of the *join* operator. Moreover, it concerns also the subquery: " more than an average salary of " + *d.name* + " department.".

Our algorithm operates on the query (2) transformed by the factoring independent subqueries method and rewrites it to the following optimised form:

```
(((((Dept where manager.Emp.salary >
2000) as d) as aux1_c) join
(aux1_c.(avg(d.employs.Emp.sal)
groupas aux1_1, (" more than an
average salary of " + d.name + "
department.")  groupas aux1_2)))
groupas aux1).((Emp as e) join
(aux1.(aux1_c.((e.fullName() + (if
(e.sal > aux1_1) then " earns" else
" doesn't earn") + aux1_2)))))      (5)
```

Unique names *aux1_c, aux_1_1, aux1_2* are automatically assigned by the optimiser. In the first three lines of (5), before the last *dot*, the query returns a bag named *aux1* consisting of structures. Each structure has three fields:

- *aux1_c* – with a binder *d* holding a reference to a *Dept* object,
- *aux1_1* – the average salary calculated for employees of the given department,
- *aux1_2* – the string "more than …. department" with the name corresponding to the given department.

The last *dot* in the third line puts on top of ENVS a binder *aux1* containing those structures. It is then used to calculate the query in the following lines. In this way, average salaries are calculated for each department once and they are used in the final query, as required.

Detecting subqueries like *avg*(*d.employs.Emp.sal*) is accomplished by analysing in which section of the environment stack the names occurring in a subquery are to be bound. The binding levels for names are compared to the scope numbers of non-algebraic operators.

2.2 The General Rewriting Rule

Let us consider the query in the form (6) (denotes string concatenation and α_i denotes a part of an arbitrary query):

```
α0 ° Q1 groupas N ° α1 ° N ° α2      (6)
```

Such a query pattern is a result of applying the independent sub-query method. The Q1 sub-query represents the part that was factored out and grouped under the name N. Referring to name N in the further part of the query (shown in (6)) represents the use of the result. Obviously to enable binding name N, the α1 query part must assure the appropriate ENVS state, but from the perspective of our method this is irrelevant because of the compilation error that appears otherwise.

It is worth noticing that the (6) form can be also a result of some other query transformation or a direct query writing.

Let us now consider the situation where (6) has the form (7):

$$\alpha_0 \circ Q_1 \textbf{ groupas } N \circ \alpha_1 \circ (N \textbf{ } \Theta_1 \textbf{ } (\alpha z_0 \circ z_1 \circ \alpha z_1 \circ z_2 \circ \ldots \circ \alpha_{zn-1} \circ z_n \circ \alpha_{zn})) \circ \alpha_3 \quad (7)$$

The query (7) contains Θ_1 – a non-algebraic operator whose left hand operand is a single name query and the right-hand operand includes subqueries z_i ($i \epsilon 1..n$). The characteristics of z_i subqueries is that they depend only on an environment introduced by the Θ_1 operator and some global (from the point of view of the evaluation of (6) query) environment. In other words they are independent of any non-algebraic operators that appear in the α_1 query part.

The query string (7) represents the general state that is a starting point for our rewrite algorithm that can be described as follows.

The z_i subqueries are factored out from the Θ_1 operator (and any other that appear in α_1) and joined with Q_1 query (with the use of non-algebraic operator **join**).

$$((Q_1 \textbf{ as } N\_C) \textbf{ join } (N\_C.(z_1 \textbf{ groupas } N\_1, z_2 \textbf{ groupas } N\_2, \ldots, z_n \textbf{ groupas } N\_n))) \textbf{ groupas } N \quad (8)$$

The Q_1 query results are named N_C with use of as operator instead of groupas because each of the Q_1 result should be processed separately by the join operator and should become a context for the subsequent z_i queries evaluation. The use of join operator as well as naming the Q_1 result and subsequent z_i results (N_i, I ϵ 1..n) preserves all the partial results (for further use) in the form of a bag of structures named N. In the query (7) the z_i subqueries are replaced with name queries that refers to names N_i. The query is additionally modified with introducing another dot operator that creates an environment containing binders with N_C and N_i.

The modified query takes the form (9):

$$\alpha_0 \circ (((Q_1 \textbf{ as } N\_C) \textbf{ join } (N\_C.(z_1 \textbf{ groupas } N\_1, z_2 \textbf{ groupas } N\_2, \ldots, z_n \textbf{ groupas } N\_n))) \textbf{ groupas } N) \circ \alpha_1 \circ (N.(N\_C \textbf{ } \Theta_1 \textbf{ } (\alpha_{z0} \circ N\_1 \circ \alpha_{z1} \circ N\_2 \circ \ldots \circ \alpha_{zn-1} \circ N\_n \circ \alpha_{zn}))) \circ \alpha_3 \quad (9)$$

This is the final result of the main algorithm process. Notice that the result query contains similar patterns to the one that appear in the initial state (6). Each pair of the subqueries: z_i groupas N_i and α_{zi-1} ° N_i ° α_{zi} are similar to the structure of the (6). Consequently (9) can be represented as follows:

$$\alpha_0' \circ z_i \textbf{ groupas } N\_i \circ \alpha_1' \circ N\_i \circ \alpha_2' \quad (10)$$

If the (10) has the form corresponding to (7):

$$\alpha_0' \circ z_i \textbf{ groupas } N\_i \circ \alpha_1' \circ (N\_i \ \Theta_1' (\alpha_{z0}' \circ z_1' \circ \alpha_{z1}' \circ z_2' \circ \ldots \circ \alpha_{zn-1}' \circ z_n' \circ \alpha_{zn}')) \circ \alpha_3' \quad (11)$$

then the optimisation can be recursively applied to the query.

All the described transformations are, in reality, performed on an abstract syntax tree (AST) query representation. The description use string representation due to conciseness.

3 OPTIMISATION GAIN

The method has been experimentally tested within the ODRA system. Fig.2 presents the performance gain after optimisation of query (1) according to the Generalised independent subqueries method, i.e. to the form (5). For instance, on a collection of 10000 employee objects, execution of the optimised one is approximately 64 times faster. In contrast, our tests have shown that the standard factoring out method applied to the query (1) (i.e., transforming it to the form (2)) does not introduce optimisation gain greater than 2 in all tested cases. The advantage of the proposed method is being able to correctly factor out the most expensive part of the query (1), i.e. **avg***(d.employs.Emp.sal)*

Figure 2: Optimization gain between evaluations of query (1) and (10).

According to the expectations, correct factoring out of a complex subquery results in improving of a query performance by orders of magnitude.

4 CONCLUSIONS

The presented generalised version of the independent subquery method is an effective complement to the original method. Applied repeatedly (after factoring), it detects and resolves subsequent independent subqueries in a query. Our rewriting rule is general, it works for any non-algebraic operator and for any data model (assuming that its semantics would be expressed in terms of SBA). The rule makes also no assumptions concerning what type an independent subquery returns: it may return a reference to an object, a single value, a structure, a collection of references, a collection of values, a collection of structures, etc. Finally the rule enables rewriting for arbitrarily complex nested subqueries, regardless of their left and right contexts.

Nevertheless, preliminary studies show possibilities of designing methods based on factoring out in cases that are still not covered, e.g. when the left-hand operand of operator Θ_1 in the query (7) is a complex subquery containing name *N*.

ACKNOWLEDGEMENTS

This research work is funded from the Polish Ministry of Science and Higher Education finances in years 2010-2012 as a research project nr N N516 423438.

REFERENCES

Cluet, S., Delobel, C., 1992, A General Framework for the Optimization of Object-Oriented Queries. *Proc. SIGMOD Conf.*, 383-392

Ioannidis Y. E., 1996 Query Optimization. *Computing Surveys*, 28(1), 121-123

Kowalski, T., et al., 2008, Optimization by Indices in ODRA. *Proc. 1st ICOODB Conf.*, 97-117

Lentner, M., Subieta, K., 2007, ODRA: A Next Generation Object-Oriented Environment for Rapid Database Application Development. *Proc. 11th ADBIS Conf.*, Springer LNCS 4690, 130-140

Płodzień, J., Kraken, A., 2000, Object Query Optimization through Detecting Independent Subqueries. *Information Systems 25*(8), 467-490

Płodzień, J., 2000, Optimization Methods in Object Query Languages. Ph.D. Thesis. *Institute of Computer Science, Polish Academy of Sciences*, http://www.

sbql.pl/phds/PhD Jacek Plodzien.pdf
Subieta, K., 2008, Stack-Based Approach (SBA) and Stack-Based Query Language (SBQL). http://www.sbql.pl
Subieta, K., 2009, Stack-based Query Language. *Encyclopedia of Database Systems 2009. Springer US*, 2771-2772

Benchmarking with TPC-H on Off-the-Shelf Hardware
An Experiments Report

Anna Thanopoulou[1], Paulo Carreira[2,3] and Helena Galhardas[2,3]
[1]*Department of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece*
[2]*Department of Computer Science and Engineering, Technical University of Lisbon, Lisbon, Portugal*
[3]*INESC-ID, Lisbon, Portugal*
anna.thanopoulou@gmail.com, paulo.carreira@ist.utl.pt, helena.galhardas@ist.utl.pt

Keywords: Database Benchmarking, Database Performance Tuning, Decision Support.

Abstract: Most medium-sized enterprises run their databases on inexpensive off-the-shelf hardware; still, they need quick answers to complex queries, like ad-hoc Decision Support System (DSS) ones. Thus, it is important that the chosen database system and its tuning be optimal for the specific database size and design. Such choice can be made in-house, based on tests with academic database benchmarks. This paper focuses on the TPC-H database benchmark that aims at measuring the performance of ad-hoc DSS queries. Since official TPC-H results feature large databases and run on high-end hardware, we attempt to assess whether the test is meaningfully downscalable and can be performed on off-the-shelf hardware. We present the benchmark and the steps that a non-expert must take to run the tests. In addition, we report our own benchmark tests, comparing an open-source and a commercial database server running on off-the-shelf hardware when varying parameters that affect the performance of DSS queries.

1 INTRODUCTION

In the day-to-day operations of a medium-sized enterprise, two types of queries are executed: *Online Transaction Processing (OLTP)* and *Decision Support (DSS)*. The former are basic information-retrieval and -update functions. The latter are aimed at assisting management decisions based on historical data. In addition, DSS queries can be further categorized into reporting and ad-hoc queries, depending on whether they are executed routinely or in a spontaneous fashion, respectively. As one would expect, the most challenging queries are the DSS ones as they are more complex and deal with a larger volume of data; even more so, ad-hoc DSS queries are challenging as they do not allow for prior system optimization. Hence, it is highly important to facilitate their execution.

The time needed to execute ad-hoc DSS queries is above all related to the database design and size. Furthermore, for a given database, time depends on the choice of the RDBMS and its tuning. Given the wide offer of database systems as well as their great complexity, it is crucial yet not trivial for the enterprise to determine the best choice for its needs, both in terms of price and in terms of performance. Therefore, it would be very helpful to realize a quantitative comparison of database systems performance under various comparable configurations, possibly using a benchmark.

The Transaction Processing Performance Council (TPC) benchmark TPC-H sets out to model a business database along with realistic ad-hoc DSS questions. It has been extensively used by database software and hardware vendors as well as researchers (Somogyi et al., 2009; Guehis et al., 2009). However, TPC-H officially published results refer to very large databases running on high-end hardware that are difficult to compare to the reality of a small enterprise. Moreover, understanding TPC-H requires significant technical expertise and, to the best of our knowledge, no step-by-step guide exists in literature, apart from generic guidelines for benchmark execution (Oracle, 2006; Scalzo, 2007).

This paper examines whether TPC-H can be used as a tool by small enterprises as well as which would be the best way to do so. Specifically, our contributions are: *(i)* a comparison of the performance of a commercial and an open-source database system executing a small-scale TPC-H test under various comparable configurations on off-the-shelf hardware; and *(ii)* insights into the tuning parameters that influence DSS performance at this scale.

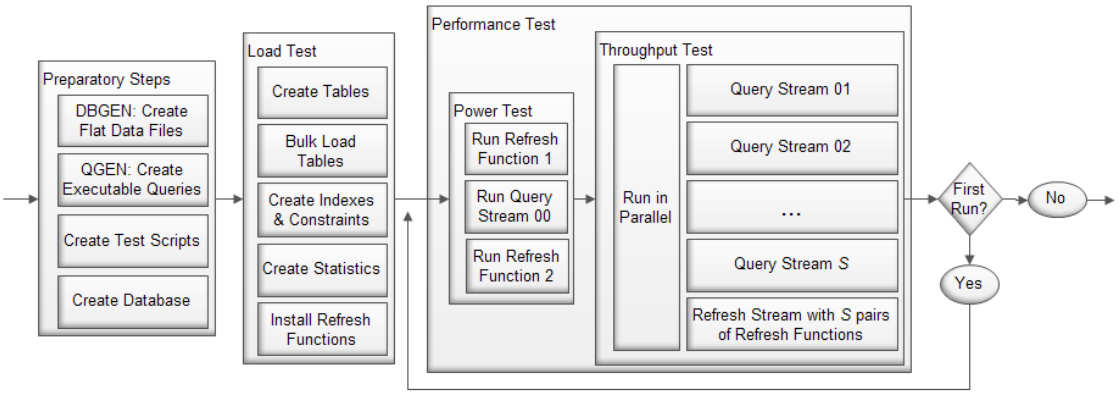


Figure 1: Complete process for running the TPC-H tests. The term *query stream* refers to a sequential execution of each of the 22 TPC-H queries, in the order specified by TPC.

2 AN OVERVIEW OF TPC-H

The TPC-H benchmark models the activity of a product supplying enterprise. For that purpose, it uses a simple database schema comprised by eight base tables. Tables have different sizes that change proportionally to a constant known as *scale factor (SF)*. The available scale factors are: 1, 10, 30, 100, 300, 1000, 3000, 10000, 30000 and 100000. The scale factor determines the size of the database in GB. Tables are populated using DBGEN, a data generator provided in the TPC-H package to populate the database tables with different amounts of synthetic data.

The benchmark workload consists of 22 queries, representing frequently-asked decision-making questions, and 2 update procedures, representing periodic data refreshments. The update procedures are called *refresh functions*. From a technical standpoint, the queries include a rich breadth of operators and selectivity constraints, access a large percentage of the populated data and tables and generate intensive disk and CPU activity. The TPC-H workload queries are defined only as query templates by TPC. The syntax is completed providing random values for a series of substitution parameters, using QGEN, an application provided in the TPC-H package.

2.1 TPC-H Tests

TPC-H comprises two tests: the load test and the performance test. The former involves loading the database with data. The latter involves measuring the system performance against a specific workload. As soon as the load test is complete, the performance test, which consists of two *runs*, can start. Each run is an execution of the *power test* followed by an execution of the *throughput test*. The *power test* aims at measuring the raw query execution power of the system with a single active session. This is achieved by sequentially running each one of the 22 queries. The *throughput test* aims at measuring the ability of the system to process the most queries in the least amount of time, possibly taking advantage of I/O and CPU parallelism. Thus, the throughput test includes at least two query sessions that run in parallel. The minimum number of query streams is specified by TPC and increases together with the scale factor. Figure 1 illustrates the steps for running a complete TPC-H test.

2.2 Performance Metrics

After running the tests, we get three types of timing measurements: the *database load time*, the *measurement interval* and the *timing intervals*. The measurement interval is the total time needed to execute the throughput test. The timing intervals are the execution times for each query or refresh function. Next, these timing measurement results must be combined to produce global, comparable metrics. To avoid confusion, TPC-H uses only one primary performance metric indexed by the database size: the *composite query-per-hour performance metric* represented as $QphH@Size$, where *Size* represents the size of data in the test database as implied by the scale factor. This metric weighs evenly the contribution of the single user power metric (*processing power metric* represented as $Power@Size$) and the multi-user throughput metric (*throughput power metric* represented as $Throughput@Size$). Finally, the *price/performance metric* represented as $Price - per - QphH@Size$ serves to make a price/performance comparison between systems.

Table 1: TPC-H full test results for increasing memory size. In SQL Server, we varied the total server memory; in MySQL, the buffer and sort memories, with a 3:1 ratio as recommended by MySQL developers. Fill factor is kept at 90% for SQL Server and 15/16 (default) for MySQL. Page size is kept at 8KB (which is the default for SQL Server) for both systems.

Memory Size Test								
total server memory		16 MB	64 MB	128 MB	256 MB	512 MB	768 MB	1024 MB
load test	**SQL Server**	46min	20min	19min	17min	16min	16min	36min
	MySQL	48min	23min	20min	16min	16min	14min	57min
performance test	**SQL Server**	4h54min	1h13min	1h	52min	41min	40min	1h9min
	MySQL	5h32min	1h28min	1h13min	1h2min	56min	54min	1h44min
QphH@1GB	**SQL Server**	19.13qph	78.55qph	90.20qph	102.30qph	130.76qph	131.80qph	86.03qph
	MySQL	17.41qph	75.70qph	79.84qph	89.77qph	103.67qph	105.10qph	70.63qph
Price-per-QphH@1GB	**SQL Server**	73.08$	17.80$	15.49$	13.67$	10.69$	10.61$	16.25$
	MySQL	28.72$	6.60$	6.26$	5.57$	4.82$	4.76$	7.80$

3 EXPERIMENTS

The goal of our experiments was to showcase a set of useful TPC-H tests that any small enterprise could perform in order to choose the database system and tuning configurations that offer optimal ad-hoc DSS performance in their system. In addition, we ran these tests ourselves on off-the-shelf hardware, aiming at some take-away rules-of-thumb for choosing between a commercial (SQL Server 2008) and an open-source (MySQL 5.1) database system and optimizing tuning for DSS queries at this scale.

We are interested in the characteristics of ad-hoc DSS workloads and the tuning parameters that affect their performance, for a given database. Since DSS queries deal with large amounts of data within scans, sorts and joins, the size of the buffer pool and the sort buffer play an important role. Following the same logic, the fill factor and the page size can also influence performance, as they can contribute to more rows per page thus keeping more sequential data in the data cache.

However, not all these parameters can be set by the user in each of the database systems at hand. In SQL Server, it is not possible to set the size of the buffer pool or the sort buffer; only the total size of memory that the system can use can be set, by determining its minimum and maximum values. MySQL, on the other hand, allows to set a specific size for the buffer pool and the sort buffer. Also, while SQL Server operates with a fixed page size of 8 KB, in MySQL the user can set the page size to 8, 16, 32 or 64 KB. Finally, in SQL Server it is possible to specify the fill factor for each page, while MySQL manages the free space automatically, with tables populated in sequential order having a fill factor of $^{15}/_{16}$.

In light of these differences, we decided to run two general types of tests: the *memory size test* and the *number of rows per page test*. Tables 1, 2 and 3 show the test results. For the number of rows per page test, note that the resulting range of number of rows per page is different for the two database systems, but that serves exactly the purpose of verifying whether allowing the user to specify much larger page sizes gives MySQL an advantage.

Table 2: TPC-H full test results for increasing fill factor in SQL Server. Page size is kept at default value of 8KB. Memory size is set at a medium value of 128KB.

MS SQL Server- Number of Rows per Page Test				
fill factor	40%	60%	80%	100%
load test	27min	22min	20min	19min
perf. test	2h2min	1h9min	1h3min	59min
QphH@1GB	34.59qph	80.10qph	89.34qph	91.58qph
PPQphH@1GB	40.42$	17.45$	15.65$	15.23$

Table 3: TPC-H full test results for increasing page size in MySQL. Fill factor is kept at default value of 15/16. Total memory size is set at a medium value of 128KB, with a buffer/sort memory ratio of 3:1 as recommended by MySQL developers.

MySQL- Number of Rows per Page Test				
page size	8 KB	16 KB	32 KB	64 KB
load test	20min	18min	17min	17min
perf. test	1h13min	59min	52min	50min
QphH@1GB	79.84qph	92.41qph	106.20qph	109.38qph
PPQphH@1GB	6.26$	5.41$	4.71$	4.57$

In the interest of simulating the environment of a smaller enterprise, we chose inexpensive off-the-shelf hardware (an AMD Athlon processor with 1GB of RAM and a SATA 80 GB hard disk) and the lowest possible scale factor (yielding a 1 GB database). We find it interesting to provide some results with a lower scale factor, as the only available ones to date are the official TPC-H results starting at 100 GB. Finally, for the price/performance metric calculations, we considered the hardware cost to be approximately 500$ and the software cost to be the current price of 898$ for SQL Server 2008 (circa 2010) and 0$ for MySQL 5.1.

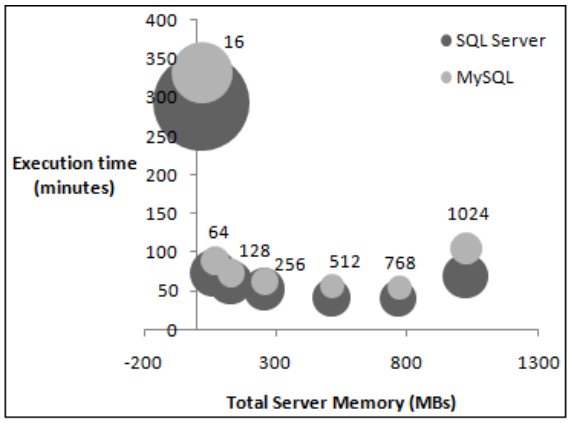


Figure 2: Influence of memory size. Larger bubbles represent greater price per query per hour.

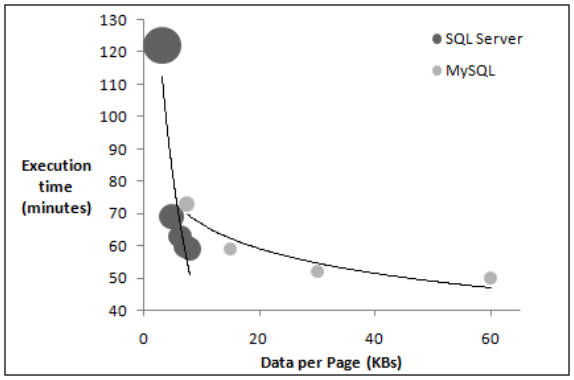


Figure 3: Influence of data per page (product of page size and fill factor). Larger bubbles represent greater price per query per hour.

3.1 Discussion

As illustrated in Figure 2, in the case of memory size test, for both database systems performance improves dramatically as we move from 16 to 768 MB of memory. The system ends up reaching its full potential around 512 MB; moving to 768 MB does not make much difference, and reaching 1024 MB actually leads to a performance decrease. In this case, the server allocates all physical memory to the cache causing part of the latter to be on virtual memory thus triggering further I/O operations.

For the same memory size, increasing either the page size or the fill factor improves performance, as illustrated in Figure 3. This makes sense because in full scans relevant data are next to each other; thus, the more data per page the less I/O operations and the better the performance. Increasing the page size is less effective than increasing the fill factor, as seen by the trendlines steepness in Figure 3 for MySQL and SQL Server respectively. In any case, increasing the memory size has an influence that exceeds both those of increasing the page size and the fill factor.

In addition, it is clear that, for approximately the same configurations, the performance of MySQL is slightly worse. This may mean that there are other tuning parameters that cause performance deterioration when left in their default values. Most likely, however, this performance difference indicates the superiority of SQL Server query optimizer when dealing with complex queries.

Finally, even though the tests run faster in SQL Server, the price/performance metric favors MySQL by far. The additional 898$ for SQL Server do not seem to be worthy for such low-scale needs.

4 CONCLUSIONS

We can conclude that the TPC-H test is meaningfully downscalable. Even with a low scale factor, we could still observe differences between different systems and configurations. However, our intuition is that its set-up time and complexity make the benchmark an unlikely choice for a medium-sized enterprise without a team of experts.

Running TPC-H motivated us to look into the factors that influence the performance of DSS queries. We concluded that the most influential tuning option is undoubtedly the memory size. Yet, other parameters (ie. page size, fill factor) also influence performance.

Since the systems do not have identical tuning options, it is hard to ascertain whether we tuned them fairly. For similar tuning, MySQL is consistently slower than SQL Server. We think this is due to different query optimizer philosophies. Yet, MySQL may take a little longer to execute the TPC-H tests but it has a higher price/performance ratio. If not chasing optimal performance, it is a viable alternative.

REFERENCES

Guehis, S., Goasdoue-Thion, V., and Rigaux, P. (2009). Speeding-up data-driven applications with program summaries. In *IDEAS'09, 2009 Int'l Database Engineering and Applications Symposium*. ACM Press.

Oracle (2006). Conducting a data warehouse benchmark.

Scalzo, B. (2007). *Top 10 Benchmarking Misconceptions*. Quest Software, 121007 edition.

Somogyi, S., Wenisch, T., Ailamaki, A., and Falsa, B. (2009). Spatio-temporal memory streaming. In *ISCA'09, 36th Annual Int'l Symposium on Computer Architecture*. ACM Press.

Business Intelligence
Definitions, Managerial Effects and Aspects: A Systematic Literature Review

Dalia Al-Eisawi and Mark Lycett
School of Information Systems, Computing and Mathematics, Brunel University, Kingston Lane, Uxbridge, U.K.
{Dalia.Al-Eisawi, Mark.Lycett}@brunel.ac.uk

Keywords: Business Intelligence, Systematic Literature Review, Decision Making, Business Performance Management, Data Management.

Abstract: This paper presents findings from the Systematic Literature Review (SLR) on Business Intelligence (BI), to clarify key definition alongside managerial effects resulting from its implementation in organizations. In doing this, the paper aims to assist organizations, decision makers, managers and information system researchers to validate the existing state of research in BI motivation. The review highlights gaps in the presented body of existing literature, contradictory answers in relation to BI definition and aspects, in addition, uncovers themes significant to BI implementation that are not well addressed in the literature. The need for empirical research is also highlighted, as the majority of the articles analyzed are at the conceptual and/or theoretical level. In addition, the research recognized a connection between a set of different managerial aspects affected by BI.

1 INTRODUCTION

Given the emerging importance of BI in organizations, this paper presents a Systematic Literature Review (SLR) on the core aspects of BI, and their effect(s) on certain managerial and organizational aspects. The SLR followed a course of action derived from Brereton (2011) based on accumulating a representative pool of articles, classifying them according to research questions, evaluating and synthesizing that literature in relation to the research questions and, finally, documenting the review and its outcomes. The review addressed two key research questions:

RQ1: from a definitional perspective, what are the core aspects of BI?

RQ2: From a managerial perspective, how are these aspects affected by BI?

The review, once rationalized, examined 65 studies spanning from January 2001 to December 2012. And also proposed a novel and comprehensive definition of BI that includes a coherent relation between BI and a set of key managerial elements that should be mentioned

When defining BI. The coherent relation indicated that "Decision making", "Business Performance Management", and "Data Management" are interrelated and cohesive managerial and key organizational aspects that can be affected positively when applying and implementing BI within organizations.

In presenting the review, the paper is organized as follows. Section 2 provides an explanation of the method used for the Systematic Literature Review (essentially following rules in a protocol that is autonomously validated). Section 3 presents results of the synthesis of the literature, consisting of chronological and sequential aspects, alongside publication details. Section 4 reports the results and background of the analysis process in relation to the research questions. Last, Section 4 presents the conclusions of the exercise.

2 METHODS

In accordance with systematic review guidelines (Brereton, 2011), the following steps were undertaken:(1) Recognizing the need for a systematic literature review; (2) formulating a set of research questions; (3) accumulating a representative pool of articles; (4) evaluating and synthesizing the gathered articles; (5) dividing the papers according to research questions; and (6), documenting the review and outcomes.

The remainder of this section will present detail in relation to these steps.

2.1 List of Searched Resources

A primary set of key words were used for the literature search, these being 'Business Intelligence', 'Decision Making' and 'Business Performance Management' ,as these specific words were drawn up for each research question. Initially, the following databases were searched: Scopus; Science Direct; ABI Inform; Academic Search Complete (ASC); and the IEEE/IET electronic library. These sources were supplemented with selected conference proceedings and specific journals including The International Journal of Business Intelligence Research (IJBIR), Institute of Electrical and Electronic Engineers proceedings (IEEE) and the European Conference on Information Systems proceedings (ECIS). Overall, we established an ultimate list of 65 papers that matches our search requirements.

2.2 The Process for Including Study Papers, Data extraction and Synthesis

The process for including and excluding gathered studies is a crucial step in the methods, as it provides and assures a strong backbone to generating a quality based literature review. All published studies that answer the author's research questions and are published within the years 2001 –2012 was integrated in the inclusion list. Moreover, the included research study must be published in conference proceedings or journal paper, In order to insure that all references included will be recorded in a fully organized structure Refworks system (www. refworks.com) was used to document reference information and details for each study. We synthesized data through classifying themes derived from the findings and results documented in each accepted paper. The categorized themes consequently revealed the creation of the categories and segments for the results section. We also conducted a type of analysis called sensitivity analysis; it is a technique for assessing the riskiness of a certain investment. For the given research purposes the sensitivity analysis was used to test how certain factors affected the field of BI Research. Key factors analyzed were based on year of publication, type of study, and finally based on which Journal or conferences preceding these papers were published. The sensitivity analyses gave us a clear idea and explicit information on where to find prejudiced and biased data. The sensitivity analysis is also reported in the results section.

3 RESULTS – BACKGROUND OF THE ANALYSIS PROCESS

3.1 Types of Study Papers

From the 65 studies, 47% were found to be theoretical or conceptual, and 37% empirical in nature. A small number of studies (16%) presented literature reviews. Empirical and literature review related to BI were less found within the pool of BI research, most of the studies were either conceptual or theoretical.

As for the data collection methods used in the case studies and empirical studies, they were primarily questionnaires/surveys, interviews by telephone or face-to-face interviews and, lastly, field studies. 46 % of the empirical research papers used questioners and surveys.

3.2 Sequential View of Publications

A statistical analysis for studies engaged in the review based on almost ten year period from 2001 until 2012 was performed, it showed that within the last 6 years there is an observed raise in published papers related to BI implementation in organizations, and its effects on decision making, We also noticed that before the year 2000 studies on BI were almost not present. The observed increase in BI research is in-line with emergent and increasing organizational awareness of the significance role of BI (and spends on technology). Alternatively, this increase might perhaps just counterpart a common rise in recent published papers in Information Systems and Decision Support Systems (Fitriana et al., 2011).

4 RESULTS – ANSWERING RESEARCH QUESTIONS

This part of the research illustrates how the literature provides answers to the research questions. The current research questions act jointly to provide an absolute explanation of the research focus. Information relating BI definitions was collected for (RQ1) to expand the understanding of traditional definitions, and extract key managerial aspects

embedded within those definitions. Papers were then analyzed to provide a more detailed understanding of BI in relation to those (addressing RQ2).

4.1 The Core Aspects of BI

Definitions within the analyzed papers were recognized as answering RQ1, these papers emphasized, or had a direct relation to a certain attributes which relates to 'Definitions' of BI such as Decision Making ,BPM, Data Management ,Knowledge Management, and finally better organizational relations. A closer assessment noticed from analyzing a set of definitions resulted in proposing that when defining BI it is always linked with any of the linked aspects as follows:

- BI Definition can contain a direct link with BI role in DM, decision making is defined as process that assist managers to make a choice about a course of action, decisions can be categorized as structured or unstructured; they also can be classified according to managerial levels such as strategic decisions, and tactical.
- BI Definition can contain a direct link with BI role in BPM. "(BPM) is a key business initiative that enables companies to align strategic and operational objectives with business activities in order to fully manage performance through better informed decision making and action"(Shi and Lu, 2010).
- BI Definition can contain a direct link with BI role in Data Management and control, data management and control refers here to how BI can assist organization in controlling the large amount of data generated daily, monthly, or annually.
- BI Definition can contain a direct link with one or more of the above given attributes.
- BI Definition did have a weaker direct relation contained by its definitions with aspects such as business knowledge, and effective organizational relationships. However, these two aspects might be required as very important facets, which are indirectly affected by BI, and the benefit of BI on them is required to be as an intangible benefit sometimes impossible to enumerate. They are however significant, and often unseen sources of business value.

The current section aimed at delivering an initial level of transparency by presenting and scrutinizing the results of analyzing a number of definitions available in the literature of the BI concept, as in the Table in the Appendix , covering the years 2001-2012 it was looked at a sample of 12 different definitions . The content of column (Direct relation of BI) denotes the significant attributes that were proposed from the authors understanding of BI definitions and that can present a direct relation as a role or effect on specific managerial and organizational aspects. Moreover, revealing these relations will have an impact in assisting the author finding answers for RQ2. A number of these definitions were obviously stated in the article, whereas others were implicit in the text. Since the current review is concept-centric explicitly, the author performed a qualitative content analysis on the collected sample that answers RQ1; the content analysis is explained as in the following definition; "A research method for the subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes or patterns" (Zhang and Wildemuth, 2009).The content analysis of the definitions revealed the following outcomes,

- It was not comprehensible withier BI is required to be a 'Process' or a "Product" ,the "process is composed of methods that organizations use to create useful information or intelligence that will support companies and organizations succeed and have a competitive advantage in the global economy" (Jordan, Rainer& Marshall, 2008).And a "product is information that will permit organizations to forecast and expect the performance and behaviours of their competitors, suppliers, customers, technologies, acquisitions, markets, products and services" (Jordan, Rainer& Marshall, 2008). As for other indicated that BI can be both a process and a product.
- Only two definitions out of the 12 given, pointed within their content to a direct relation between BI role with all the three roles defined previously by the author, within the definitions almost 90% of definitions mentioned a direct relation between BI and DM, and few had a relation with data management, BPM, both, or all three attributes.
- Since most of study papers collected for the purpose of this research found to be theoretical based, the author noticed that few of the given definitions were extracted form an empirical and observed practice. Therefore, this will lead to a delay in the understanding of what BI characterizes to business leaders and researchers.
- And finally, It was not clear weather BI is required be a technological or managerial concept, or it can be both.

As a result, applying a content analysis on the collected definitions realized a main and general concern as follows; "BI did not yet reach a standardized and unified definition".

4.2 How are these Aspects Affected by BI?

Another synthesis for the purpose of answering RQ2 was undertaken; the synthesis shows that 55% of papers illustrate a role of BI in (DM), 22 % Role on BI in Data Management and Control, and 23% are related to (BPM), and these statistics can be more explained as follows:

(BI effect on DM): most of the papers searched agreed that BI has a direct effect on DM; according to literature BI has turn out to be a vital aspect of decision-making, not only at the top, but at each particular hierarchical level. That is the reason behind the needs for it to be associated with the business and organizational strategy in order to be capable to support analytical decision-making. Nevertheless, this relation is turning to be as a known fact rather than just a proposition since it has been researched and proven in large number of study papers.

(BI effect on BPM): as defined earlier BPM" is a key business initiative that enables companies to align strategic and operational objectives with business activities in order to fully manage performance through better informed decision making and action"(Shi and Lu, 2010). According to the synthesized literature, BPM will start within the coming future to be required as being the last constituent of BI, and the following stage in the growth of BI, organization, and information systems. If BPM is a consequence of BI and better decision making, and contains many of its technologies, tools and techniques, then BI itself can play as a key role and deliver the insight needed to improve overall business performance .This was hypothesized by the authors from a theoretical viewpoint and sensible perspective.

(BI effect on Data Management): Data Management and control refers here to how BI can assist organization in controlling the large amount of data generated daily, monthly or annually. This effect was related in most papers to the use of Data Warehouse technology, that have the ability to assist the transformation of organizational operational data system into an analytical data system construction, and that can sustain business requirements and needs. Thus, this formation enables business executive to attain a chronological view of operational data, moreover, eliminating the load on organizational Information technology assets and enabling mangers to create positive decisions instead of unconsidered ones.

As a conclusion for answering RQ2: "Decision Making", "Business Performance Management", and "Data Management" are interrelated and cohesive managerial and key organizational aspects that can be affected positively when applying and implementing BI within organizations as the bellow figure.

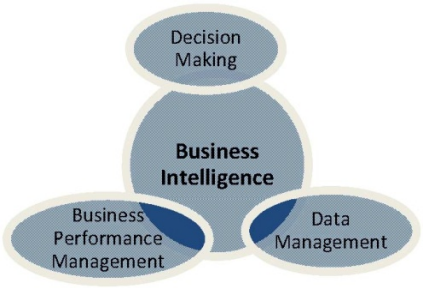


Figure 1: The cohesive effects of BI on managerial aspects.

5 CONCLUSIONS

The paper presented a global view of BI definition, and a global understanding of its effects and roles derived from a state of the art (SLR) process. Thus, the research presented clear face validity for researchers, managers, and decision makers to help them understand the managerial facets of BI. Within the (SLR) the author researched a total of 65 published papers. The results from analyzing papers showed that the predominance of the papers were published by Journals that are only dedicated for BI research, such as the International Journal of Business Intelligence (IJBIR) ,Other papers may be related to special interest group on computer personnel research and conference preceding. Also, it was observed that there is a noticed increase in BI published research, and this increase might be an indication of the emergent organizational awareness of the significance role of BI. The review process investigated that the empirical and case studies related to BI were scarcely obtainable within the pool of BI research; most of the studies were either conceptual or theoretical.

The research results from answering the first research question revealed that the existing definition of BI extracted from the literature, and extracted from applying a content analysis on the set of definitions did not yet reach a standardized and unified definition. Therefore, the author proposed a coherent relation between BI and a set of key managerial elements which are; 1. (DM) 2. (BPM) and 3. Data Management, that should be all

mentioned when defining BI as follows:

"Business intelligence is a combination of processes, products, and technologies that have the ability in supporting organization, and can have a direct key role in Data management by storing, and analyzing the data collected from internal and external sources ,and on Decision Making by creating knowledge ,and finally on Business Performance management "

The systematic review process also collected papers that investigated the effect of BI on those managerial aspects. 55% of papers illustrated a direct effect of BI on decision making, 22 % of the papers illustrated the effect of BI on Data Management and 23% showed an effect on BPM. Consequently, and as a results the author concluded a general understanding from the second research question as follows:

Decision Making, Business Performance Management, and Data Management are interrelated and cohesive managerial and key organizational aspects that can be positively affected when applying and implementing BI within organizations.

As future lines of work, we will expand the analysis of organizational features recognized by BI and its implementations this is, defining more substantial and insubstantial effects of its presence in organizations and where exactly inside an organization they are playing their significant roles, yet how can they be evaluated and quantified.

REFERENCES

Ariyachandra, T. and Frolick, M. H., R.T., (2011). "10 Principles to Ensure Your Data Warehouse Implementation is a Failure", pp. 37-47.

Böhringer, Martin; Gluchowski, Peter; Kurze, Christian; and Schieder, Christian,(2010) "A Business Intelligence Perspective on the Future Internet" *AMCIS2010 Proceedings.* Paper267. htt://aisel.aisnet. org/amcis2010/267.

Brereton, P (2011), "A Study of Computing Undergraduates Undertaking a Systematic Literature Review", *IEEE Transactions on Education*, 54, 4, pp. 558-563, Academic Search Complete, EBSCOhost, viewed 6 December 2011.

Fitriana1,R., Eriyatno, and Djatna,t(2011)," Progress in Business Intelligence System research : A literature Review", *International Journal of Basic & Applied Sciences IJBAS-IJENS* Vol: 11 No: 03.

Habul, A. & Pilav-Velic, A. (2010) "Business intelligence and customer relationship management", *Information Technology Interfaces (ITI), 2010 32nd International Conference on*, pp. 169.

Lida Xu, Li Zeng, Zhongzhi Shi, Qing He & Maoguang Wang (2007) "Research on Business Intelligence in enterprise computing environment", *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pp. 3270.

Pirnau, M. & Botezatu, C.P. (2010) "General information on business Intelligence and OLAP systems architecture", *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, pp. 294.

Yan Shi and Xiangjun Lu (2010) "The Role of Business Intelligence in Business Performance Management", Information Management, *Innovation Management and Industrial Engineering (ICIII), 2010 International Conference on*, pp. 18.

Zack Jourdan, R. Kelly Rainer Jr., Thomas E. Marshall, (2008). "Business Intelligence: An Analysis of the Literature." *IS Management* 25(2): 121-131.

Zhang, Y., and Wildemuth, B. M., (2009). "Qualitative analysis of content. In B. Wildemuth (Ed.), *Applications of Social Research Methods to Questions in Information and Library Science* "(pp.308-319). Westport, CT: Libraries Unlimited.

APPENDIX

A Review of BI Traditional Definitions							
Author	Year	BI Explanation	Is there a Direct relation with the following within the definition?				
			DM	BPM	Data control	Better relations management	Knowledge Management
Ortiz, 2003	2003	*"(BI) is a set of products, which are sets of tools and technologies designed to efficiently extract useful information from oceans of data"*	No	No	Yes	No	Yes
Dharan & Swami, 2004	2004	*"BI is a term that encompasses a broad range of analytical software and solutions for gathering, consolidating, analysing and providing access to information in a way that is supposed to let an enterprise's users make better business decisions"*	Yes	No	Yes	No	No
Xu, Zhang & Jiang, 2005	2005	*"The concept of Business Intelligence (BI) is brought up by Gartner Group since 1996. It is defined as the application of a set of methodologies and technologies, that improve enterprise operation effectiveness, support management/decision to achieve competitive advantages."*	Yes	Yes	No	No	No
Xie & Zhou, 2008	2008	*"Business intelligence systems are interactive computer-based structures and subsystems intended to help decision makers use communication technologies, data, documents, knowledge, and analytical models to identify and solve problems. The new generation of BIS offers the potential for significantly improving operational and strategic performance for organizations of various sizes and types".*	Yes	Yes	No	No	Yes
Viaene, 2008	2008	*"BI refers to a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data that helps Decision making process"*	Yes	No	Yes	No	No
Yoav & Kolodner, 2009	2009	*"BI is a system that supports activities such as data analysis, managerial decision making, and business-performance measurement".*	Yes	Yes	Yes	No	No
Wixom & Watson, 2007	2010	*"Business intelligence (BI) is an umbrella term that is commonly used to describe the technologies, applications, and processes for gathering, storing, accessing, and analysing data to help users make better decisions."*	Yes	No	Yes	No	No
Foley & Guillemette, 2010	2012	*"A combination of processes, politics, culture, and technologies for gathering, manipulating, storing, and analysing the data collected from internal and external sources in order to communicate information, create knowledge and inform decision making. BI helps report business performance, uncover new business opportunities and make better business decisions "*	Yes	Yes	Yes	No	Yes
Patrick; Christian; Christian; Peter & Martin, 2010	2010	*"Business Intelligence (BI) a concept provides a means to obtain crucial information to improve the decision making process"*	Yes	No	No	NO	NO
Hill, Ariyachandra & Frolick, 2011 10 prince	2011	*"(BI) is seen as the ultimate solution that will help organizations leverage information to make informed, intelligent business decisions"*	Yes	No	Yes	No	No
Glancy & Yadav, 2011	2011	*"Business intelligence (BI) a system that provide relevant competitive intelligence, combine it with a business 'internal information, provide expert information, incorporate advanced analytical decision techniques, and are able to inform the executive of the relevance of the knowledge created from the system."*	Yes	Yes	Yes	No	Yes
Gluchowski & Gluchowski, 2011	2011	*"BI is a data driven decision support system (DSS) that combines data gathering, data storage, and knowledge management with analysis in the interests of better managerial decision making"*	Yes	No	Yes	No	Yes

ARTIFICIAL INTELLIGENCE AND DECISION SUPPORT SYSTEMS

FULL PAPERS

Design of Human-computer Interfaces in Scheduling Applications

Anna Prenzel and Georg Ringwelski
Hochschule Zittau / Görlitz, Department of Electrical Engineering and Computer Science,
Obermarkt 17, 02826, Görlitz, Germany
{aprenzel, gringwelski}@hszg.de

Keywords: Human Factors, Planning and Scheduling, Decision Support System, Automation, Interactive Scheduling.

Abstract: There are many algorithms to solve scheduling problems, but in practice the knowledge of human experts almost always needs to be involved to get satisfiable solutions. In this paper, we describe a set of decision support features that can be used to improve human computer interfaces for scheduling. They facilitate and optimize human decisions at all stages of the scheduling procedure. Based on a study with 35 test subjects and overall 105 hours of usability testing we verify that the use of the features improves both quality and practicability of the produced schedules.

1 INTRODUCTION

Scheduling solutions to support human decisions are widely asked for in several application domains. Very often these solutions turn out in practice to work as sociotechnical or mixed initiative systems. Numerous (human) agents and stakeholders as well as software systems are involved in decision making (Burstein and McDermott, 1997), (Wezel et al., 2006).

Problem Description. In this paper we focus practical scheduling problems. A fleet scheduling system serves as an example. It is to be included in an information system for water suppliers. The final product is sold to several companies, which have similar, but never uniform problems and workflows. The customers require interactive scheduling features including

- adapting schedules during execution due to accidents that must be resolved immediately
- adapting future schedules due to expert knowledge which was not included in the model a priori
- allowing manual adaptation in order to evaluate different scenarios for parts of a future schedule.

Another problem is the acceptance of the product by end-users. In interviews with human schedulers we have observed that

- they fear that a system could replace their work and are reluctant to accept push-the-button-optimizers
- consequently they tend to find problems in the produced schedules, which can hardly be solved a priori through better modeling
- it is inevitable that expert knowledge on the scheduling process is maintained in a company.

From this point of view we must find appropriate ways to incorporate human factors in the computer-supported scheduling process.

Contribution. In order to target these requirements we define several human-computer interaction models based on an analysis of human decision-making. They can be distinguished by their level of automation that varies between manual and fully automatic.

- We deduce a set of decision support (DSS) features from this analysis that can be combined to different human-computer interaction models.
- We show that human operators should be able to choose the level of automation for each scheduling problem individually.
- We compare the models based on an empirical study we carried out in 105 hours of usability testing with 35 test subjects. Our study shows that the quality of the produced schedules correlates with use and availability of the regarded features.

2 A SHORT INTRODUCTION INTO PRACTICAL SCHEDULING

2.1 The Common Structure of Scheduling Problems

The main concern of scheduling is the assignment of *jobs* to *resources*. Jobs are services that must be carried out by the resources, for example, items for production, items for transport or shifts in a hospital. Machines, vehicles and employees can be considered as resources. Scheduling systems are expected to solve combinatorial problems such as finding sequences or start times of jobs, good resource utilization, minimal makespan and many more. Solving these problems is complex (often NP-complete) because solutions have to satisfy numerous constraints including

Start Time Constraints:
For individual jobs, such as "each job has a time window that restricts earliest and latest possible start time".
Among several jobs, such as "jobs must not overlap in time if they are assigned to the same resource".

Resource Constraints:
For individual jobs, such as "each job has a set of resources it can be assigned to".
Among several jobs, such as "a limited set of resources can be used at a time".

Our case study in fleet scheduling is based on a formal model described by Kallehauge, Larsen, Madsen and Solomon (2005). In addition to meeting the constraints the goal of scheduling is to keep costs low and to minimize the execution time. The calculation of the costs is again application-specific. The objective functions of our fleet scheduling system are:

a) The total travel time between each two jobs in the schedule (cost function)
b) The time between the beginning of the first and the end of the last job in the schedule (execution time)

The latter also addresses the common requirement of balancing the workload of the resources. Scheduling aims to find an arrangement of jobs that optimizes the current objective values and provides a good tradeoff between them.

2.2 Preferences and Modifications

We have gathered information about scheduling issues in several projects with domain experts in scheduling. Each company has its specific technical requirements on their schedules. For example, a manufacturing company will define the sequence, in which items are processed on the assembly line. The individual start time and resource constraints reflect the physical conditions of the production system and thus have to be enforced as hard constraints.

However, the dispatchers also know the criteria that make their schedules practicable or impracticable and prefer certain schedules over others. Their preferences arise from dynamic changes in the operational requirements. Consider the following types of preferences:

Start Time Preferences: "start this job not until 10 o'clock"; "start this job as early as possible"

Resource Preferences: "use resource X (not) for this job"; "use only half of the jobs for this resource"

Optimization Preferences: "reduce the travel time for this resource"; "reduce the overall execution time"; "change the weight of this objective function"

Preferences like these are based on the experience of the human operators in their field of work (Fransoo et al., 2011). They have an idea of what an "optimal" schedule looks like in a particular situation. This also means that they are able to find optimization preferences in automatically produced schedules. In the most cases it is not obvious how to set the weight of multiple optimization goals in advance of the scheduling. Therefore humans derive them from existing schedules and use them for subsequent adaptations of parts or the whole schedule.

In contrast to the hard constraints preferences include some uncertainty. It is not clear from the start whether and to what extent they can be incorporated. This depends on the impact they have on the overall schedule and particularly on how much the remaining jobs are changed. For example, if a preference is known *before* scheduling, the remaining jobs can be scheduled within the bounds of their hard constraints. However, this is more complicated, if the preference is applied to an existing schedule which only allows partial changes.

In addition to preferences subsequent modifications of schedules play a big role in practical scheduling as well. For different reasons there might be unanticipated changes to schedules being carried out. For example, a schedule has to be adapted if a resource breaks down or a new job has to be included in case of an event. Again, there might be preferences about the best way to perform modifications.

2.3 Abstraction Levels of Scheduling Actions

Human operators tend to have an intuition about how to adapt a schedule such that a preference is considered. They use mental models containing as much details of the system as needed to plan the scheduling actions that lead to the desired state of the schedule (St-Cyr and Burns, 2001), (Wezel et al., 2006), and (Turban et al., 2010). The possible levels of detail a schedule provides can be represented in an abstraction hierarchy.

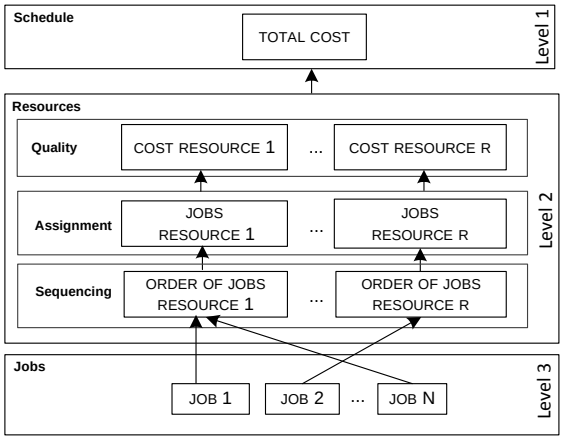


Figure 1: Abstraction levels for scheduling tasks.

The hierarchy we chose is shown in Figure 1. From top to bottom, it reveals different levels of detail of a general schedule. At level 1 the only information used is the objective value of the overall schedule. The underlying level 2 reveals details of the sub-schedules for each resource including the assignment of jobs to vehicles and, zooming in further, the order of the particular jobs. The lowest level 3 contains the individual jobs that hold their start times and resources as properties.

A scheduling action at a certain level can be defined without information of the underlying levels. Consider for instance the goal of changing the resource affiliation of a job. It is irrelevant for the human operator where the job is positioned within the sequence of jobs or at which time it starts. However, for the preference to take effect a decision about the start time has to be made in order to obtain a schedule that does not violate any hard constraints. That means, the level a preference targets and the level at which it is implemented can be different. We describe this with the term “loss of abstraction”.

3 INVESTIGATING THE HUMAN CONTRIBUTION TO SCHEDULING

Manual optimization of schedules is a monotonous job unsuitable for humans (Burstein and Holsapple, 2008). Due to the structure of the problems the number of valid positions for jobs is exponential (Burke and Kendall, 2005) which makes it difficult for the human to find the optimal costs. In contrast, it is important for the user to collect and interpret the data of schedules to find preferences and modifications. Having identified them, he participates in the adaptation of the schedule.

3.1 Making Decisions

The decisions about how identified preferences and modifications are incorporated should be left to the human in order to prevent problems of the kind we have described in section 1.

3.1.1 Decision-making in General

Scheduling can be modeled as decision process (Higgins, 1999) consisting of *intelligence*, *design* and *choice* (Turban et al., 2010). The intelligence phase involves the recognition of the problem at the start of the decision process. After that, possible solutions are evaluated in the design phase. The best alternative is finally selected in the choice step. We add a *completion* step, if the selected solution yet has to be completed. If the completion step is still complex, a new decision process is triggered. The decision processes are chained that way until the task is accomplished.

The decision process is influenced by skills and knowledge of the human. We distinguish skill-based (SBB), rule-based (RBB) and knowledge-based reasoning (KBB) (Rasmussen, 1983). As shown in Figure 2 RBB and SBB shorten the decision process.

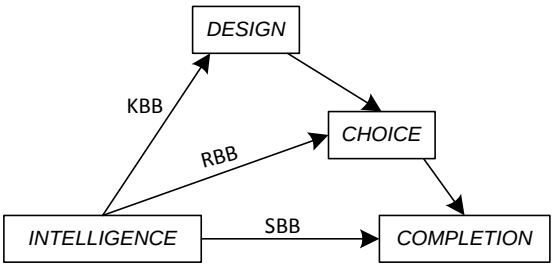


Figure 2: Stages in decision-making and shortcuts.

Table 1: Types of reasoning.

KBB	No pattern can be used. Intelligent reasoning is required. KBB coincides with the design phase.
RBB	Familiar patterns in the data map to a rule that implies the action.
SBB	Perception is mapped to action directly.

3.1.2 Decision-making in Scheduling

The decision stages can be directly applied to human scheduling activities.

Design: In the design stage the human operator compares alternative solutions for the task. Depending on the abstraction level this involves comparing

- different schedules (level 1)
- different assignments of jobs to resources(level 2)
- different orders of jobs within a resource (level 2)

Each considered alternative is evaluated with regard to optimality and practicability.

However, only valid schedules can be evaluated. Due to the earlier mentioned "loss of abstraction" the human operator has to make decisions about the details below the abstraction level of the task. This leads to a new decision process in order to find a valid implementation of the solution to be considered. The original decision process is compromised, as the human must keep track of nested design stages at different levels.

Choice and Completion: The human operator chooses the best suited schedule. If complete schedules are compared in the design stage the completion step can be omitted.

It depends both on the experience of the human operator and on the characteristics of the task whether the decision process can be shortened by SBB or RBB.

SBB: The scheduling task is a pure optimization of cost functions if no alternative solutions exist or if the preference is formulated as a hard constraint. Furthermore, typical modifications such as the addition of jobs sometimes do not require an evaluation in terms of practicability but only in terms of optimality and thus are skill-based.

RBB: Applies, if the human operator deals with the task repeatedly or if there are best practices, such that the best suited alternative is known from experience. The human operator has to implement the chosen alternative in the completion step.

4 DESIGN OF INTERACTIVE SCHEDULING INTERFACES

4.1 Hypothesis for Optimal Decision Support

It is an important issue for decision support to keep the human operator at the level of abstraction, that is related to his preference and to the current type of reasoning. For SBB and RBB the computer can undertake the whole work of optimizing at level 1. In KBB the scheduler should be able to test the outcome of decisions in the design phase while disregarding low-level constraints. To overcome the loss of abstraction the system has to provide the level of automation, that is needed for a particular action.

We define the levels of automation according to the levels of abstraction shown in Figure 2.

Level 3: This level requires the least amount of automation, as the human operator undertakes all decisions about start times, orders, resources and other properties of jobs. However, to prevent faulty decisions, the system should supervise the compliance with the underlying constraints. In doing so it is not sufficient to show an error message as soon as a constraint is violated. We rather suggest to visualize the scope of action already when the human is about to make a decision. According to the types of constraints in section 2.1 this means highlighting valid properties for the considered job that

a) meet its individual constraints
b) meet its constraints in relation to other jobs

with regard to the state of the current schedule. This way the human does not have to make the effort to withdraw a faulty decision.

Level 2: The human makes decisions on *some selected* properties of either individual jobs or the schedule only. The computer is required to solve the remaining properties such that

a) all constraints are satisfied
b) the schedule is optimal or at least good with regard to the cost function.

This is especially important for KBB, as it allows the human operator to try and evaluate several assignments that are based on his manual decision. The portion of work of the computer increases with the sublevels as shown in Table 3. At the quality sublevel the human defines the cost function for the scheduling of one or more jobs. In case all jobs are chosen the decision support is equal to level 1.

Table 2: Properties assigned by human and computer at different sublevels of level 2.

Sublevel	Human	Computer
Sequencing	resource, relative position, cost function	start time
Assignment	resource, cost function	relative position, start time
Quality	cost function	resource, relative position, start time

Level 1: Full automation is applied at this level. The human operator is only concerned about the cost function the computer should use to optimize the whole schedule.

To sum up, the human operator decides, how much details he contributes to a change of the schedule.

4.2 Interactive Decision Support Features

We have designed a set of interaction features that can be used to build a scheduling interface providing the recommended decision support. They are described in Table 4. We neglect commonly used features like Undo/Redo, as they can be found in the standard literature about successful user interface design (Shneidermann, 2010).

At level 3 we use colors to visualize the domain of the property of a job in the current schedule. For level 2 we suggest the use of controls that allow the human operator to select a group of jobs for optimization. This is a simple way to deal with optimization preferences, as different objective functions can be chosen for different groups. The FO-feature is suited for tasks at level 1.

Fixation covers all three levels. It is the prerequisite for all other features, as it deals with the way the human operator enters a condition for a certain property in the interface. Having done this the computer considers the condition in optimizing or constraint highlighting. Properties that are not fixed to a certain value can be automatically resolved with level 2 and level 1 features.

Furthermore, fixation allows keeping decisions made at lower levels when using features at higher levels. For example, if the human operator modifies some jobs with the help of ECH and FIT, he can fix their properties at level 3. If FO is applied afterwards, the modified jobs are not changed anymore. Figure 3 shows the abstraction levels the features belong to.

Table 3: Decision support features.

Full Optimization (FO)	A control to optimize the whole schedule. It allows choosing from various built-in cost functions.
Single Job Optimization (SJO)	The interface allows to select a single job in the schedule and triggers automatic optimization of its position. *Remaining jobs in the schedule are kept unchanged.*
Resource Optimization (RO)	Like SJO. All jobs belonging to the same resource can be selected at once.
Group Optimization (GO)	Like SJO. Any group of jobs from different resources can be selected.
Fit-in (FIT)	The interface allows the user to define the position of a job within the sequence and looks for a valid start time.
Constraint Highlighting (CH)	The interface recognizes the intention to change a property of a job and colors possible values *red*, if they are invalid *green*, if they are valid with regard to constraints of *the individual job*.
Enhanced Constraint Highlighting (ECH)	Additional to CH: values of properties, that violate constraints *in relation to other jobs* are colored *yellow*, if the value can be applied as soon as the properties of conflicting jobs are adapted *grey*, if the value can never be applied in conjunction with the conflicting jobs.
Fixation (FIX)	The interface allows the direct input of the desired properties of one or more jobs. They are turned into additional constraints to be considered by all features.

4.3 Example Interfaces

Our hypothesis does not include recommendations about how to support the *identification of preferences and modifications.* This is an issue for the graphical information visualization of the specific scheduling application. It should follow the principles of Ecological Interface Design (Vicente, 2002), (Vicente and Rasmussen, 1992) and display information according to the abstraction hierarchy. We show two example interfaces that include our recommended DSS features.

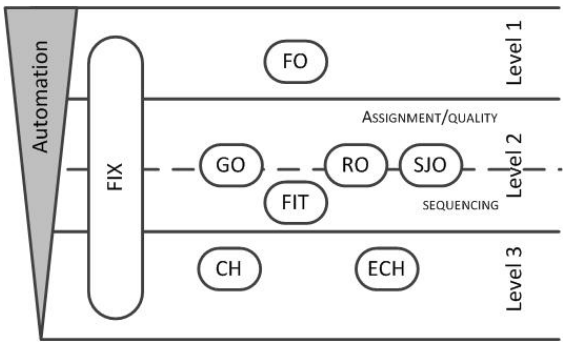


Figure 3: DSS features at different abstraction levels.

4.3.1 Fleet Scheduling

The interface used in our experiments is sketched in Figure 4. We decided to use a Gantt chart, as it clearly shows the sequence of jobs in time and the travel times between them. This makes it easy to analyze start times and resources of jobs in order to derive certain preferences. For further support we provide a map.

The human operator can move the jobs per Drag and Drop. If he starts dragging constraint highlighting is applied to the Gantt chart: the colors of the positions show whether there are time window conflicts or overlaps with other jobs in case the job is dropped there. A job can be dropped at any position colored green or yellow, in the latter case the fit-in feature can be used to put the job correctly in the sequence.

Furthermore SJO, RO and GO are available through context menus and provide the two cost functions introduced in section 2.1. Scheduling preferences can be defined in property dialogs and by using the pin (FIX) that fixes both start time and resource of a job. A button to create schedules from scratch (FO) is also provided.

4.3.2 Nurse Rostering

A possible interface for nurse rostering is shown in Figure 5. In contrast to the vehicle routing interface the jobs are not grouped by their resource (nurse), but by the shift they belong to. Each shift requires a certain number of nurses which corresponds to the number of jobs that must be included. The cost function usually deals with considering the preferences of the individual nurses.

The start time of a shift determines the start times of the associated jobs. Their resources can be chosen from a drop-down menu, whose entries are colored according to CH and ECH. For example, if a nurse had a night shift the day before it must not be assigned to the early shift due to legal requirements.

However, if the selection of this nurse is colored yellow, the human operator is able to ask the system to reschedule the day before such that the early shift becomes valid. Furthermore the interface contains features to select a group of jobs (SJO, GO) or the whole schedule (FO) for automatic optimization. In this case fixed nurses (FIX) are kept unchanged.

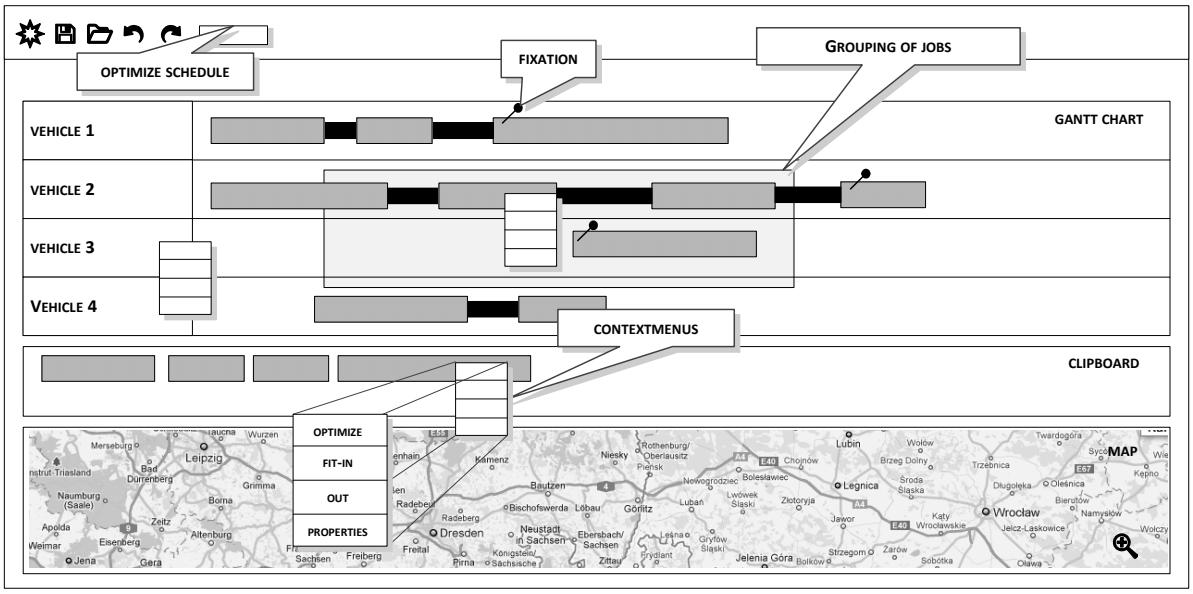


Figure 4: Interface Design for Vehicle Routing.

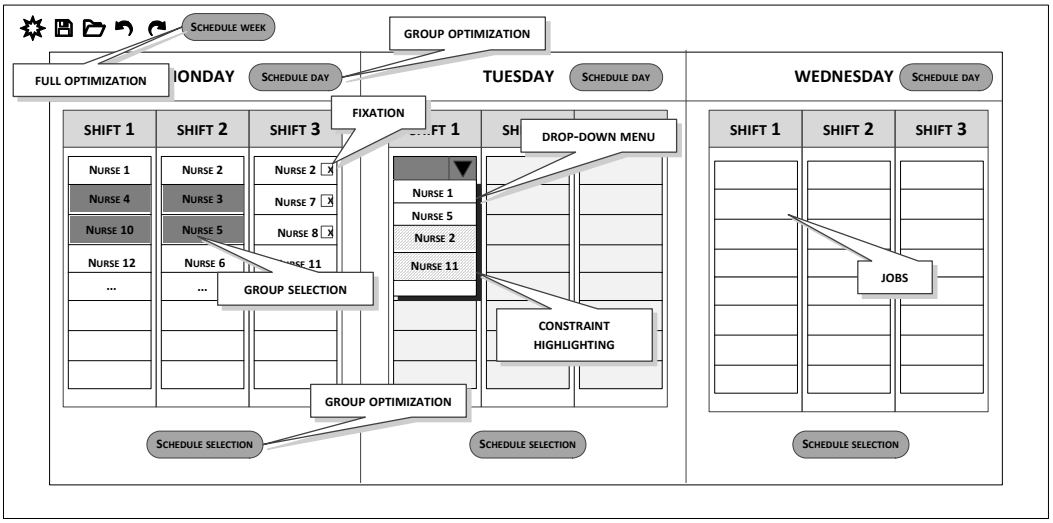


Figure 5: Interface design for Nurse Rostering.

5 EVALUATION OF THE DECISION SUPPORT

5.1 Combining DSS Features to Interaction Models

In order to prove our claims from section 1 it remains to provide an empirical evaluation of
a) the suitability of the features for performing scheduling tasks at different abstraction levels
b) the quality that can be achieved in terms of the cost function.

For this we combine DSS features to 5 interaction models located at different abstraction levels. They are shown in Table 5.

Model 1/2: manual scheduling at level 3

Model 3: FO at level 1, subsequent manual modifications at level 3 are allowed, fixation is not allowed

Model 4: like model 3, fixation is allowed

Model 5: level 2, fixation can be achieved indirectly by excluding manually positioned jobs from optimization groups.

Several test tasks with scheduling preferences at different abstraction levels are carried out by peer groups. Each model is used for each task.

5.2 Setup of the Usability Test

We have formed 5 test groups each consisting of 7 students from different faculties of our institution. The subjects were asked to perform 6 scheduling tasks. The models available for the particular tasks were dependent on the test group. We determine the best model for each task by comparing the average performance and confidence interval in the following metrics: accumulated travel time, task completion, time effort, number of undo operations and number of manual interactions. The tests took 3 hours per participant including a briefing of 30 minutes at the start. The maximum duration for each task was set to 15 minutes.

5.2.1 Design of the Test Tasks

The participants had no experiences in scheduling. Therefore the relevant scheduling preferences that would otherwise arise from the expert knowledge of the scheduler had to be predefined for each task.

1. Schedule a set of jobs such that the total travel time is minimized and the workload[1] is balanced between the resources. For some jobs there are precedence constraints **(level 2 sequencing)**.
2. Schedule a set of jobs such that the total travel time is minimized and the workload is balanced. For some jobs fixed start times and resources are given **(level 3)**.
3. An additional vehicle is to be utilized. Change the given schedule such that some suitable jobs are assigned to it **(level 2 assignment)**.
4. An event occurs and requires an additional job. The working schedule must include the job as early as possible, but it has to remain unchanged until 10 o'clock **(level 3)**.

[1] The workload corresponds to the total number of jobs that a resource has to carry out.

5. Schedule a set of jobs such that the total travel time is minimized and the workload is balanced. Jobs beyond the German-Polish border must be carried out in one piece **(level 2 sequencing).**
6. Change the current schedule such that vehicle 3 finishes work at 12 o'clock. Remaining jobs have to be assigned to other vehicles **(level 2 assignment)**.

The tasks are to be carried out with 4 vehicles and about 25 predefined jobs. All jobs have time window and resource constraints. The participants always have to strive for a compromise between low travel time and balanced workload (**level 2 quality/level 1**).

5.2.2 Assignment of Test Groups to Interaction Models

The table below shows the distribution of test persons to different models. The models are divided into two areas: manual optimization (model 1 and 2) and automated optimization (models 3, 4 and 5). The participants first carried out their tasks manually and then repeated them with the help of automatic features.

The assignment of models to groups changes from task to task. This ensures that each group deals at least one time with each interaction model. We assigned fewer participants to models that were expected to be very difficult (model 1 and the model without any features) or discouraging for the test subjects.

Table 4: Example peer groups and models for task 1.

Model	Features	Persons	Group (Task 1)
-	-	7	1
Model 1	CH	7	2
Model 2	ECH	21	3,4,5
Model 3	FO + ECH	7	1
Model 4	FO + FIX + ECH	14	2,3
Model 5	SJO + GO + RO + FIX + FIT + ECH	14	4,5

5.3 Results

5.3.1 Usability Metric 1: Travel Time

In Figure 9-13 the achieved qualities of the schedules are shown for each particular task. The average qualities are influenced by the number of successfully completed tasks. Both task 6 and task 1 turned out to be insoluble for our testers in 15 minutes if no decision support was provided. Consequently, we cannot present further results.

Level 3 Tasks: The results for task 2 and 4 are shown in Figures 6 and 7. The schedules created with level 3-features only were worse than those created with higher-level-features. This confirms the assumption that skill-based scheduling tasks should be carried out by the computer. CH and ECH help the human to find a scheduling decision for *some* jobs, but are not sufficient for creating *complete* schedules.

Comparing models 3 and 4, the quality decreases if fixation is not allowed. This suggests that preferences should be incorporated in advance (FIX) rather than after automated optimization.

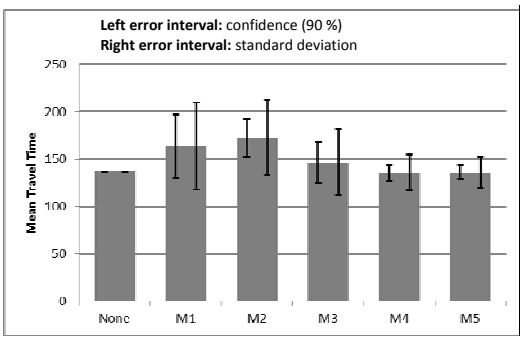


Figure 6: Mean travel time – task 2.

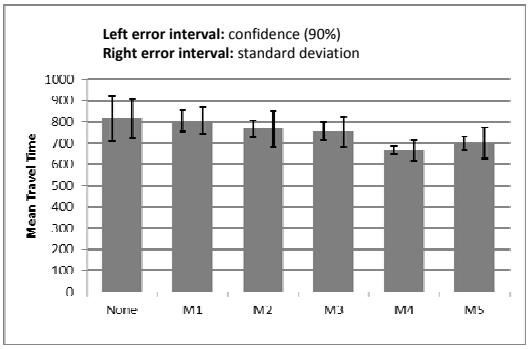


Figure 7: Mean travel time – task 4 (task 5 is very similar).

Level 2 Tasks: The results for tasks 1, 3, 5 and 6 are shown in Figures 7, 8, 9 and 10. They are similar to those for the level 3 tasks. The best schedules mostly result from models 4 and 5. There is no significant difference in the performance of the two models, which applies to *all* test tasks too.

The overall ranking of the models is shown in Figure 11 (1 is the best, 6 the worst rank). It confirms the assumption that models 4 and 5 generally provide the best decision support.

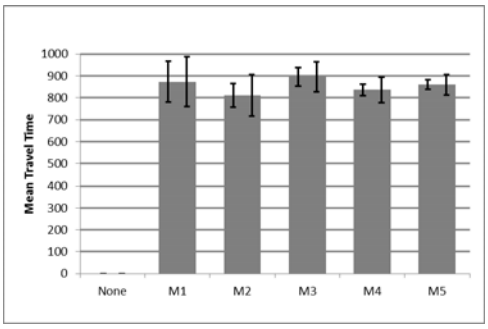


Figure 8: Mean travel time – task 1.

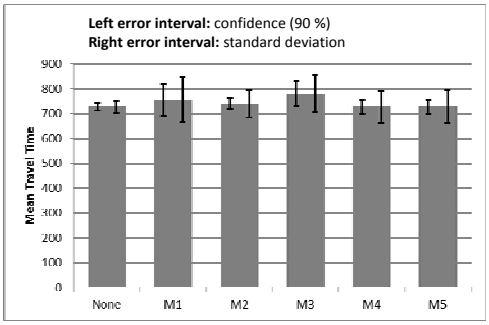


Figure 9: Mean travel time – task 3.

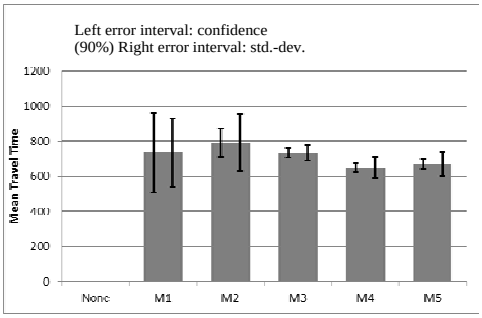


Figure 10: Mean travel time – task 6.

5.3.2 Usability Metric 2: Task Success

The number of participants that have managed to obtain a solution is shown in Figure 12. A task was considered successful, if the schedule did not violate any time window or resource constraints and the scheduling preferences were fulfilled.

With models 1, 2 and "None" many participants ran into dead-ends, where they were not able to insert further jobs in the clipboard. In this case model 2 merely depicted a grey Gantt chart background. They would have to manually backtrack former decisions. However, testers would rather give up at this point.

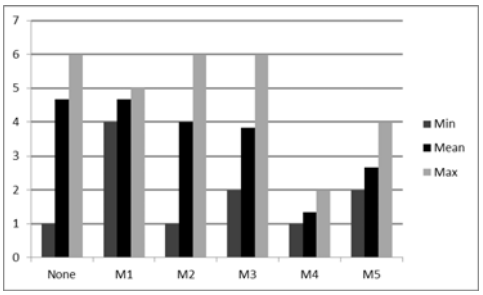


Figure 11: Ranking of the models averaged over the tasks.

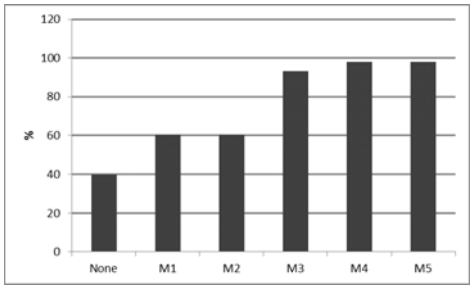


Figure 12: Rate of successful task completion.

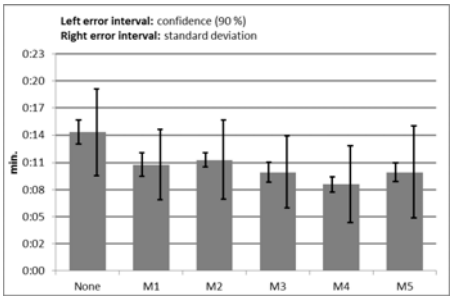


Figure 13: Average task duration.

5.3.3 Usability Metric 3: Task Duration

The average time, users required to solve the tasks (deadline was 15 minutes) is shown in Figure 13. Although the time needed with no model is particularly high, in general the models have a high variance in their execution time. How much time a test person spent to fulfill a task was strongly dependent on his motivation and ideas to improve the schedule. The runtime of the system to solve the scheduling problem was negligible.

5.3.4 Metric 4: Interaction Frequency

Figure 14 shows the number of undo operations averaged over the number of participants. Models 4 and 5 have a strikingly high occurrence of undo, which refers to the general behavior in the design phase, if there are high-level scheduling features. It

consists of alternately applying and reversing automated scheduling features until a satisficing solution is found. Model 1 has a small peak in undo-operations, as there is no aid to predict if an operation will be feasible. Model 2 compensates for this with the background-color grey.

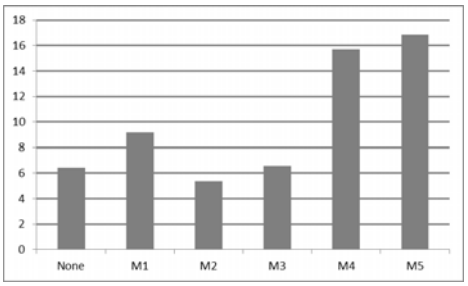


Figure 14: Average number of undo operations.

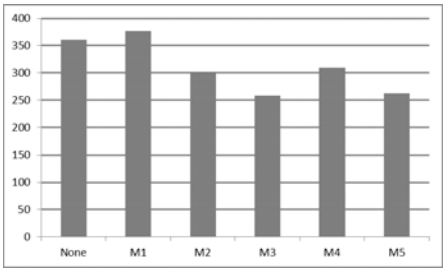


Figure 15: Average number of manual operations.

Figure 18 shows the average number of manual operations (drag and drop of jobs). As expected the manual effort is the higher, the less support is provided. However, manual scheduling is not completely replaced by automated features, as the user performs subsequent changes or sets certain jobs according to his ideas.

6 CONCLUSIONS

We proposed 8 interaction features to enhance human interaction in scheduling. These features were evaluated in a quantitative study (usability test) with regard to 4 relevant metrics. The results are:

1. The practicability of resulting schedules improves with features to manually fixate, reorder and optimize groups of jobs.
2. The success rate (solved tasks in given time) is highly influenced by the availability of automated scheduling features.
3. Automated scheduling features encourage the user to explore his scope of action on the basis of trial and error (optimize - undo).

REFERENCES

Burstein, F. and Holsapple, C. W. (Eds.). (2008). *Handbook on Decision Support Systems 1 and 2.* Springer.

Burstein, M. H. and McDermott, D. V. (1997). Issues in the development of human-computer mixed initiative systems. *Cognitive Technology*, 285-303.

Burke, E. K. and Kendall, G. (2005). *Search methodologies: introductory tutorials in optimization and decision support techniques.* Heidelberg: Springer.

Fransoo, J. C., Waefler, T. and Wilson, J. R. (2011). *Behavioral Operations in Planning and Scheduling*. Springer.

Higgins, P. G. (1999). Job Shop Scheduling: Hybrid Intelligent Human-Computer Paradigm. Ph.D. diss., Department of Mechanical and Manufacturing Engineering, The University of Melbourne, Melbourne, Australia.

Kallehauge, B., Larsen, J., Madsen, O. and Solomon, M. (2005). Vehicle Routing Problem with Time Windows. In Desaulniers, G., Desrosiers, J. and Solomon, M. M. (Eds.), *Column Generation* (pp. 67-98). Springer US.

Rasmussen, J. (1983). Skills, rules, and knowledge; Signals, Signs and Symbols, and Other Distinctions in Human Performance Models. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(3), 157-266.

St-Cyr, O. and Burns, C.M. (2001). Mental Models and the Abstraction Hierarchy: Assessing Ecological Compatibility. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting,* 297-301.

Shneidermann, B. (2010). *Designing the User Interface.* Pearson Addison-Wesley.

Turban, E., Sharda, R. and Delen, D. (2010). *Decision Support and Business Intelligence Systems.* Prentice Hall, Pearson.

Vicente, K. J. (2002). Ecological Interface Design: Progress and Challenges. *The Journal of the Human Factors and Ergonomics Society*, 44, 62-78.

Vicente, K. J. and Rasmussen, J. (1992). Ecological interface design: theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics,* 22(4), 589-606.

Wezel, W., Jorna, R. and Meystel, A. (Eds.). (2006). *Planning in Intelligent Systems.*Wiley Interscience.

Unified Algorithm to Improve Reinforcement Learning in Dynamic Environments
An Instance-based Approach

Richardson Ribeiro[1], Fábio Favarim[1], Marco A. C. Barbosa[1],
André Pinz Borges[2], Osmar Betazzi Dordal[2], Alessandro L. Koerich[2] and Fabrício Enembreck[2]
[1]*Graduate Program in Computer Engineering, Federal Technological University of Paraná, Pato Branco, Paraná, Brazil*
[2]*Post-Graduate Program in Computer Science, Pontificial Catholical University of Paraná, Curitiba, Paraná, Brazil*
{*richardsonr, marcocb, favarim*}*@utfpr.edu.br,* {*andre.borges, osmarbd, alekoe, enembreck*}*@ppgia.pucpr.br*

Keywords: Intelligent Agents, Reinforcement Learning, Dynamic Environments.

Abstract: This paper presents an approach for speeding up the convergence of adaptive intelligent agents using reinforcement learning algorithms. Speeding up the learning of an intelligent agent is a complex task since the choice of inadequate updating techniques may cause delays in the learning process or even induce an unexpected acceleration that causes the agent to converge to a non-satisfactory policy. We have developed a technique for estimating policies which combines instance-based learning and reinforcement learning algorithms in *Markovian* environments. Experimental results in dynamic environments of different dimensions have shown that the proposed technique is able to speed up the convergence of the agents while achieving optimal action policies, avoiding problems of classical reinforcement learning approaches.

1 INTRODUCTION

Markov Decision Processes (MDP) are a popular framework for sequential decision-making for single agents, when agents' actions have stochastic effect on the environment state and need to learn how to execute sequential actions. Adaptive intelligent agents emerge as an alternative to cope with several complex problems including control, optimization, planning, manufacturing and so on. A particular case is an environment where events and changes in policy may occur continuously (*i.e.*, dynamic environment). A way of addressing such a problem is to use Reinforcement Learning (RL) algorithms, which are often used to explore a very large space of policies in an unknown environment by trial and error. It has been shown that RL algorithms, such as the Q-Learning algorithm (Watkins and Dayan, 1992), converge to optimal policies when a large number of trials are carried out in stationary environments (Ribeiro, 1999).

Several works using RL algorithms and adaptive agents in different applications can be found in the literature (Tesauro, 1995; Strehl et al., 2009; Zhang et al., 2010). However, one of the main drawbacks of RL algorithms is the rate of convergence which can be too slow for many real-world problems, *e.g.* traffic environments, sensor networks, supply chain management and so forth. In such problems, there is no guarantee that RL algorithms will converge, since they were originally developed and applied to static problems, where the objective function is unchanged over time. However, there are few real-world problems that are static, *i.e.* problems in which changes in priorities for resources do not occur, goals do not change, or where there are tasks that are no longer needed. Where changes are needed through time, the environment is dynamic.

In such environments, several approaches for achieving rapid convergence to an optimal policy have been proposed in recent years (Price and Boutilier, 2003; Bianchi et al., 2004; Comanici and Precup, 2010; Banerjee and Kraemer, 2010). They are based mainly on the exploration of the state-action space, leading to a long learning process and requiring great computational effort.

To improve convergence rate, we have developed an instance-based reinforcement learning algorithm coupled with conventional exploration strategies such as the ε-greedy (Sutton and Barto, 1998). The algorithm is better able to estimate rewards, and to generate new action policies, than conventional RL algorithms. An action policy is a function mapping states

to actions by estimating a probability that a state s' can be reached after taking action a in state s.

In MDP, algorithms attempt to compute a policy such that the expected long-term reward is maximized by interacting with an environment (Ribeiro, 1999). The approach updates into state-action space the rewards of unsatisfactory policies generated by the RL algorithm. States with similar features are given similar rewards; rewards are anticipated and the number of iterations in the Q-learning algorithm is decreased.

In this paper we show that, even in partially-known and dynamic environments, it is possible to achieve a policy close to the optimal very quickly. To measure the quality of our approach we use a stationary policy computed previously, comparing the return from our algorithm with that from the stationary policy, as in (Ribeiro et al., 2006).

This article is organized as follows: Section 2 introduces the RL principles and the usage of heuristics to discover action policies. The technique proposed for dynamic environments is presented in Section 3 where we also discuss the Q-Learning algorithm and the k-Nearest Neighbor (k-NN) algorithm. Section 4 gives experimental results obtained using the proposed technique. In the final section, some conclusions are stated and some perspectives for future work are discussed.

2 BACKGROUND AND NOTATION

Many real-life problems such as games (Jordan et al., 2010; Amato and Shani, 2010), robotics (Spaan and Melo, 2008), traffic light control (Mohammadian, 2006; Le and Cai, 2010) or air traffic (Sislak et al., 2008; Dimitrakiev et al., 2010), occur in dynamic environments. Agents that interact in this kind of environment need techniques to help them, *e.g.*, to reach some goal, to solve a problem or to improve performance. However because individual circumstances are so diverse, it is difficult to propose a generic approach (heuristics) that can be used to deal with every kind of problem. Environment is the world in which an agent operates.

A dynamic environment consists of changing surroundings in which the agent navigates. It changes over time independent of agent actions. Thus, unlike the static case, the agent must adapt to new situations and overcome possibly unpredictable obstacles (Firby, 1989; Pelta et al., 2009). Traditional planning systems have presented problems when dealing with dynamic environments. In particular, issues such as truth maintenance in the agent's symbolic world model, and replanning in response to changes in the environment, must be addressed.

Predicting the behavior (*i.e.*, actions) of an adaptive agent in dynamic environments is a complex task. The actions chosen by the agent are often unexpected, which makes it difficult to choose a good technique (or heuristic) to improve agent performance. A heuristic can be defined as a method that improves the efficiency in searching a problem solution, adding knowledge about the problem to an algorithm.

Before discussing related work, we introduce the MDP which is used to describe our domain. A MDP is a tuple $(S, A, \partial^a_{s,s'}, R^a_{s,s'}, \gamma)$ where S is a discrete set of environment states that can be composed by a sequence of state variables $< x_1, x_2, ..., x_y >$. An episode is a sequence of actions $a \in A$ that leads the agent from a state s to s'. $\partial^a_{s,s'}$ is a function defining the probability that the agent arrives in state s' when an action a is applied in state s. Similarly, $R^a_{s,s'}$ is the reward received whenever the transition $\partial^a_{s,s'}$ occurs and $\gamma \in \{0...1\}$ is a discount rate parameter.

A RL agent must learn a policy $Q : S \rightarrow A$ that maximizes its expected cumulative reward (Watkins and Dayan, 1992), where $Q(s,a)$ is the probability of selecting action a from state s. Such a policy, denoted as Q^*, must satisfy Bellman's equation (Sutton and Barto, 1998) for each state $s \in S$ (Equation 1).

$$Q(s,a) \leftarrow R(s,a) + \gamma \sum_{s'} \partial(s,a,s') \times \max Q(s',a) \quad (1)$$

where γ weights the value of future rewards and $Q(s,a)$ is the expected cumulative reward given for executing an action a in state s. To reach an optimal policy (Q^*), a RL algorithm must iteratively explore the space $S \times A$ updating the cumulative rewards and storing such values in a table $\hat{Q}$.

In the Q-learning algorithm proposed by Watkins (Watkins and Dayan, 1992), the task of an agent is to learn a mapping from environment states to actions so as to maximize a numerical reward signal. The algorithm approaches convergence to Q^* by applying an update rule (Equations (2)(3)) after a time step t:

$$v \leftarrow \gamma \max Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \quad (2)$$

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha[R(s_t, a_t) + v] \quad (3)$$

where V is the utility value to perform an action a in state s and $\alpha \in \{0,1\}$ is the learning rate.

In dynamic environments such as traffic jams, it is helpful to use strategies like ε-greedy exploration (Sutton and Barto, 1998) where the agent selects an action with the greatest Q value with probability $1-\varepsilon$.

In some Q-Learning experiments, we have found that the agent does not always converge during training (see Section 4). To overcome this problem we have used a well known Q-Learning property: actions can be chosen using an exploration strategy. A very common strategy is random exploration, where an action is randomly chosen with probability ε and the state transition is given by Equation 4.

$$Q(s) = \begin{cases} \max Q(s,a) & ,if \quad q > \varepsilon \\ a_{random} & ,otherwise \end{cases} \qquad (4)$$

where q is a random value with uniform probability in $[0,1]$ and $\varepsilon \in [0,1]$ is a parameter that defines the exploration trade-off. The greater the value of ε, the smaller is the probability of a random choice, and a_{random} is a random action selected among the possible actions in state s.

Several authors have shown that matching some techniques with heuristics can improve the performance of agents, and that traditional techniques, such as ε-greedy, yield interesting results (Drummond, 2002; Price and Boutilier, 2003; Bianchi et al., 2004). Bianchi (Bianchi et al., 2004) proposed a new class of algorithms aimed at speeding up the learning of good action policies. An RL algorithm uses a heuristic function to force the agent to choose actions during the learning process. The technique is used only for choosing the action to be taken, while not affecting the operation of the algorithm or modifying its properties.

Butz (Butz, 2002) proposes the combination of an online model learner with a state value learner in a MDP. The model learner learns a predictive model that approximates the state transition function of the MDP in a compact, generalized form. State values are evaluated by means of the evolving predictive model representation. In combination, the actual choice of action depends on anticipating state values given by the predictive model. It is shown that this combination can be applied to increase further the learning of an optimal policy

Bianchi et al. (Bianchi et al., 2008) improved action selection for online policy learning in robotic scenarios combining RL algorithms with heuristic functions. The heuristics can be used to select appropriate actions, so as to guide exploration of the state-action space during the learning process, which can be directed towards useful regions of the state-action space, improving the learner behavior, even at initial stages of the learning process.

In this paper we propose going further in the use of exploration strategies to achieve a policy closer to the Q^*. To do this we have used policy estimation techniques based on an instance learning, such as the k-Nearest Neighbors (k-NN) algorithm. We have observed that is possible to reuse previous states, eliminating the need of a prior heuristic.

3 *K*-NR: INSTANCE-BASED REINFORCEMENT LEARNING APPROACH

In RL, learning takes place through a direct interaction of the algorithm with the agent and the environment. Unfortunately, the convergence of the RL algorithms can only be reached after an exhaustive exploration of the state-action space, which usually converges very slowly. However, the convergence of the RL algorithm may be accelerated through the use of strategies dedicated to guiding the search in the state-action space.

The proposed approach, named k-Nearest Reinforcement (k-NR), has been developed from the observation that algorithms based on different learning paradigms may be complementary to discover action policies (Kittler et al., 1998). The information gathered during the learning process of an agent with the Q-Learning algorithm is the input for the k-NR. The reward values are calculated with an instance-based learning algorithm. This algorithm is able to accumulate the learned values until a suitable action policy is reached.

To analyze the convergence of the agent with the k-NR algorithm, we assume a generative model governing the optimal policy. With such a model it is possible to evaluate the learning table generated by the Q-Learning algorithm. To do this, an agent is inserted into a partially known environment with the following features:

1. **Q-Learning Algorithm:** learning rate (α), discount factor (γ) and reward (r);
2. **Environment E:** the environment consists of a state space where there is an initial state ($s_{initial}$), a goal state (s_{goal}) and a set of actions A=$\{\uparrow, \downarrow, \rightarrow, \leftarrow\}$, where $\uparrow$, $\downarrow$, $\rightarrow$, $\leftarrow$ mean respectively *east*, *south*, *north* and *west* (Figure 1).

A state s is an ordered pair (x,y) with positional coordinates on the axis X and Y respectively. In other words, the set of states S represents a discrete city map. A status function $st : S \rightarrow ST$ maps states and traffic situations where ST = {-0.1, -0.2, -0.3, -0.4 , -1.0, 1.0)}, where -0.1, -0.2, -0.3, -0.4, -1.0 and 1.0 mean respectively *free*, *low jam*, *jam* or *unknown*, *high jam*, *blocked*, and *goal*. After each move (transition) from state s to s' the agent knows whether its

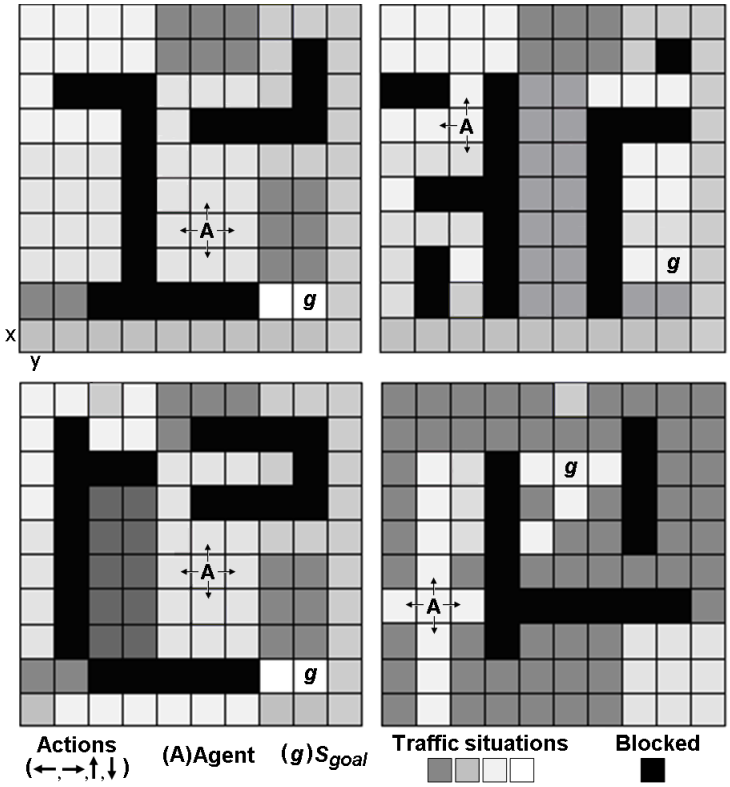


Figure 1: Environment: An agent is placed at random positions in the grid, having a visual field depth of 1.

action is positive or negative through the rewards attributed by the environment. Thus, the reward for a transition $\partial^{a}_{s,s'}$ is $st(s')$ and Equations (2) (3) is used as update function. In other words, the agent will know if its action has been positive if, having found itself in a state with traffic jam, its action has led to a state where the traffic jam is less severe. However, if the action leads the agent to a more congested status then it receives a negative reward.

The pseudocode to estimate the values for the learning parameters for the Q-Learning using the k-NR is presented in Algorithm 1. The following definitions parameters are used in such an algorithm:

- a set $S = \{s_1, \ldots, s_m\}$ of states;
- an instant discrete steps $step = 1, 2, 3, ..., n$;
- a time window T_x that represents the learning time (cycle(x) of steps);
- a set $A = \{a_1, \ldots, a_m\}$ of actions, where each action is executable in a step n;
- a status function $st : S \rightarrow ST$ where $ST = \{-1, -0.4, -0.3, -0.2, -0.1\}$;
- learning parameters: α=0.2 and γ=0.9;
- a learning table $QT : (S \times A) \rightarrow \mathbb{R}$ used to store the accumulative rewards calculated with the Q-Learning algorithm;
- a learning table $kT : (S \times A) \rightarrow \mathbb{R}$ used to store the reward values estimated with the k-NN;
- #*changes* is the number of changes in the environment.

3.1 k-NN and k-NR

The instance-based learning paradigm determines the hypothesis directly from training instances. Thus, the k-NN algorithm saves training instances in the memory as points in an n-dimensional space, defined by the n attributes which describe them (Aha et al., 1991; Galvn et al., 2011). When a new instance must be classified, the most frequent class among the k nearest neighbors is chosen. In this paper the k-NN algorithm is used to generate intermediate policies which speed up the convergence of RL algorithms. Such an algorithm receives as input a set of instances generated from an action policy during the learning stage of the Q-Learning. For each environment state, four instances are generated (one for each action) and they represent the values learned by the agent. Each training instance has the following attributes:

1. attributes for the representation of the state in the way of the expected rewards for the actions: north (N), south (S), east (E) and west (W);
2. an action and;
3. reward for this action.

Table 1 shows some examples of training instances.

Algorithm 2 shows that the instances are computed to a new table, denoted as kT, which stores the values generated by the k-NR with the k-NN. Such values represent the sum of the rewards received with the interaction with the environment. Rewards are computed using Equation 6 which calculates the similarity between two training instances $\vec{s_i}$ and $\vec{s_m}$.

Algorithm 1: Policy estimation with k-NR.

Require: Learning Table: $QT(s,a)$;
$kT(s,a)$; S=$\{s_1, ..., s_m\}$; A=$\{a_1, ..., a_m\}$
st: S $\rightarrow$ ST;
Time window T_x;
Environment E;
Ensure:
1. **for all** $s \in$ S **do**
2. **for all** $a \in$ A **do**
3. $QT(s,a) \leftarrow 0$;
4. $kT(s,a) \leftarrow 0$;
5. **end for**
6. **end for**
7. **while** not stop_condition() **do**
8. **CHOOSE** $s \in$ S, $a \in$ A
9. **Update rule:**
10. $Q_{t+1}(s_t,a_t) \leftarrow Q_t(s_t,a_t)+\alpha[R(s_t,a_t)+v]$ where,
11. $v \leftarrow \gamma \max Q_t(s_{t+1},a_{t+1}) - Q_t(s_t,a_t)$
12. step $\leftarrow$ step + 1;
13. **if** step $< T_x$ **then**
14. **GOTO**{8};
15. **end if**
16. **if** changes are supposed to occur **then**
17. **for** I $\leftarrow$ 1 to #changes **do**
18. **Choose** $s \in$ S
19. $st(s) \leftarrow$ a new status $st \in$ ST;
20. **end for**
21. **Otherwise** continue()
22. **end if**
23. k-NR(T_x,s,a); *// Algorithm 2*
24. **for** $s \in$ S **do**
25. **for** $a \in$ A **do**
26. $QT(s,a) \leftarrow kT(s,a)$
27. **end for**
28. **end for**
29. **end while**
30. **return** (.,.);

Algorithm 2: k-NR(T_x,s,a).

1: **for all** s $\in$ S and s $\neq s_{goal}$ **do**
2: costQT $\leftarrow$ cost(s, s_{goal}, QT)
3: costQ^* $\leftarrow$ cost(s, s_{goal}, Q^*)
4: **if** $costQT_s \neq costQ*_s$ **then**
5:

$$kT(s,a) \leftarrow \frac{\sum_{i=1}^{k} HQ_i(.,.)}{k} \quad (5)$$

6: **end if**
7: **end for**
8: **return** $(kT(s,a))$

$$f(\vec{s_i},\vec{s_m}) = \frac{\sum_{x=1}^{x}(s_{i_x} \times s_{m_x})}{\sum_{x=1}^{x} s_{i_x}^2 \times \sum_{x=1}^{x} s_{m_x}^2} \quad (6)$$

The cost function (Equation 7) calculates the cost for an episode (path from a current state s to the state s_{goal} based on the current policy).

$$cost(s,s_{goal}) = \sum_{s \in S}^{s_{goal}} 0.1 + \sum_{s \in S}^{s_{goal}} st(s) \quad (7)$$

Equation 5 used in Algorithm 2 shows how the k-NN algorithm can be used to generate the arrangements of training instances: here, $kT(s,a)$ is the estimated reward value for a given state s and action a, k is the number of nearest neighbors, and $HQ_i(.,.)$ is the i-th existing nearest neighbor in the set of training instances generated from $QT(s,a)$.

Using the k-NR, the values learned by the Q-Learning are stored in the kT table. This contains the best values generated by the Q-Learning and the values that have been estimated by the k-NR.

We have evaluated different ways of generating the arrangements of instances for the k-NN algorithm with the aim of finding the best training sets. First, we used the full arrangement of instances generated throughout runtime. Second, instances generated inside n time windows were selected, where $A_{[T(n)]}$ denotes an arrangement of $T(n)$ windows. In this core, each window generates a new arrangement and previously instances are discarded. We have also evaluated the efficiency rate considering only the arrangement given by the last window $A_{[T(last)]}$. Finally, we have evaluated the efficiency rate of the agent using the last arrangement calculated by the k-NN algorithm - $A_{[T(last),T(k-NN)]}$. The results on these different configurations for generating instances are shown in Section 4.

Table 1: Training instances.

State (x,y)	Reward (N)	Reward (S)	Reward (E)	Reward (W)	Action Chosen	Reward Action
(2,3)	-0.875	-0.967	0.382	-0.615	(N)	-0.875
(2,3)	-0.875	-0.968	0.382	-0.615	(S)	-0.968
(2,3)	-0.875	-0.968	0.382	-0.615	(W)	0.382
(2,3)	-0.875	-0.968	0.382	-0.615	(E)	-0.615
(1,2)	-0.144	1.655	-0.933	0.350	(N)	-0.144
...	...	...	...	...	...	...

4 EXPERIMENTAL RESULTS

In this section we present the main results obtained from using the k-NR and Q-Learning algorithms. The experiments were carried out in dynamic environments with three different sizes as shown in Figure 2: 16 (4×4), 25 (5×5) and 64 (8×8) states. Note that a number of states S can generate a long solution space, in which the number of possible policy is $|A||^s|$.

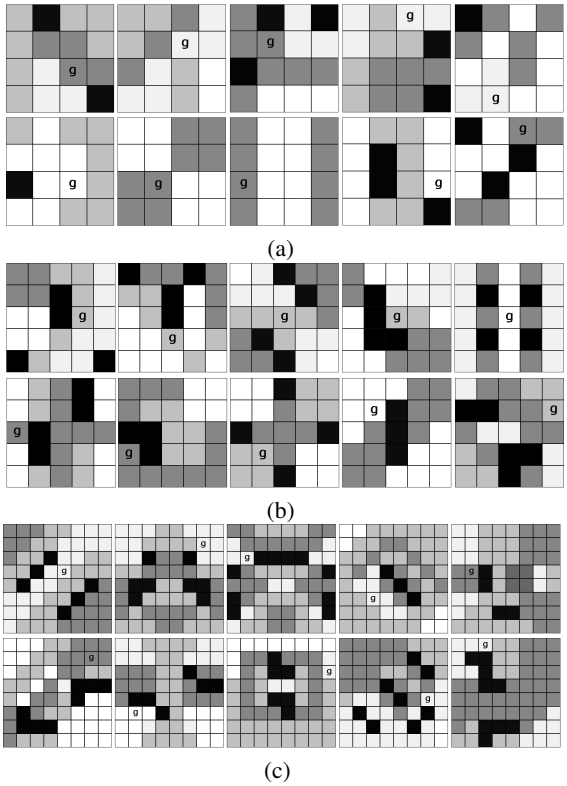


Figure 2: Simulated environments: (a) 16-state, (b) 25-state, (c) 64-state.

For each size of environment, ten different configurations were arbitrarily generated to simulate real-world scenarios. The learning process was repeated twenty times for each environment configuration to evaluate the variations that can arise from the agent's actions which are autonomous and stochastic. The results presented in this section for each environment size (16, 25, and 64) are therefore average values over twenty runs. The results do not improve significantly when more scenarios are used (≈2.15%). The efficiency of the k-NR and Q-Learning algorithms (Y axis in figures) takes into account the number of successful outcomes of a policy in a cycle of steps. We evaluated the agent's behavior in two situations:

1. ♯percent of changes (10, 20, 30) in environment for a window T_x=100. In 64-state environments the changes were inserted after each 1000 steps (T_x=1000) because in dynamic environments such large environments require many steps to reach a good intermediary policy.
2. ♯percent of changes in environment after the agent finds its best action policy. In this case, we use the full arrangement of learning instances $A_{[T_x]}$, because it gave the best results.

The changes were simulated considering real traffic conditions such as: different levels of traffic jams, partial blocking and free traffic for vehicle flowing. We also allowed for the possibility that unpredictable factors may change traffic behavior, such as accidents, route changing or roadway policy, collisions in traffic lights or intersections, and so on. Changes in the environment were made as follows: for every T_x window, the status of a number of positions is altered random. Equation 8 is used to calculate the number of altered states (♯changes) in T_x.

$$\sharp changes_{(T_x)} = \left(\frac{\sharp states}{100}\right) \times \sharp percent \qquad (8)$$

We observed in initial experiments, that even with a low change rate in the environment, the agent with Q-Learning has trouble converging without the support of exploration strategies.

To solve this problem, we used the Q-Learning together with the ε-greedy strategy, which allows the agent to explore states with low rewards. With such a strategy, the agent starts to re-explore the states that

underwent changes in their status. More detail of the ε-greedy strategy in others scenarios are given in (Ribeiro et al., 2006). It can be seen from this experiment that the presence of changed states in an action policy may decrease the agent's convergence significantly. Thus, the reward values that would lead the agent to states with positive rewards can cause the agent to search over states with negative rewards, causing errors. In the next experiments we therefore introduce the *k*-NR.

4.1 *k*-NR Evaluation

We used the *k*-NR to optimize the performance of *Q*-Learning. The technique was applied only to the environment states where changes occur. Thus, the agent modifies its learning and converges more rapidly to a good action policy. Figure 3 shows how this modification in the heuristic affects convergence of *Q*-Learning. However using *k*-NR, the agent rapidly converges to a good policy, because it uses reward values that were not altered when the environment was changed.

It is seen that the proposed approach may accelerate convergence of the RL algorithms, while decreasing the noise rate during the learning process. Moreover, in dynamic environments the aim is to find alternatives which decrease the number of steps that the agent takes until it starts to converge again. The *k*-NR algorithm causes the agent to find new action policies, for the states that have had their status altered by the reward values of unaltered neighbor states. In some situations, the agent may continue to convergeeven after a change of the environment. This happens because some states have poor reward values (values that are either too high or too low) as a consequence of too few visits, or too many. Therefore, such states must be altered by giving them more appropriate reward values.

To observe the behavior of the agent in other situations, changes were introduced into the environment only after the agent finds its near-optimal policy (a policy is optimal when the agent knows the best actions). The aim is to analyze the agent's performance when an optimal or near-optimal policy has been discovered, and to observe the agent's capacity then to adapt itself to a modified environment.

Enembreck et al. (Enembreck et al., 2007) have shown that this is a good way to observe the behavior of an adaptive agent. We have analyzed the agent's adaptation with the *k*-NR and *Q*-Learning algorithms. The *Q*-Learning presents a period of divergence (after some changes were generated), usually a decreasing performance (Figure 4). However, after a reasonable

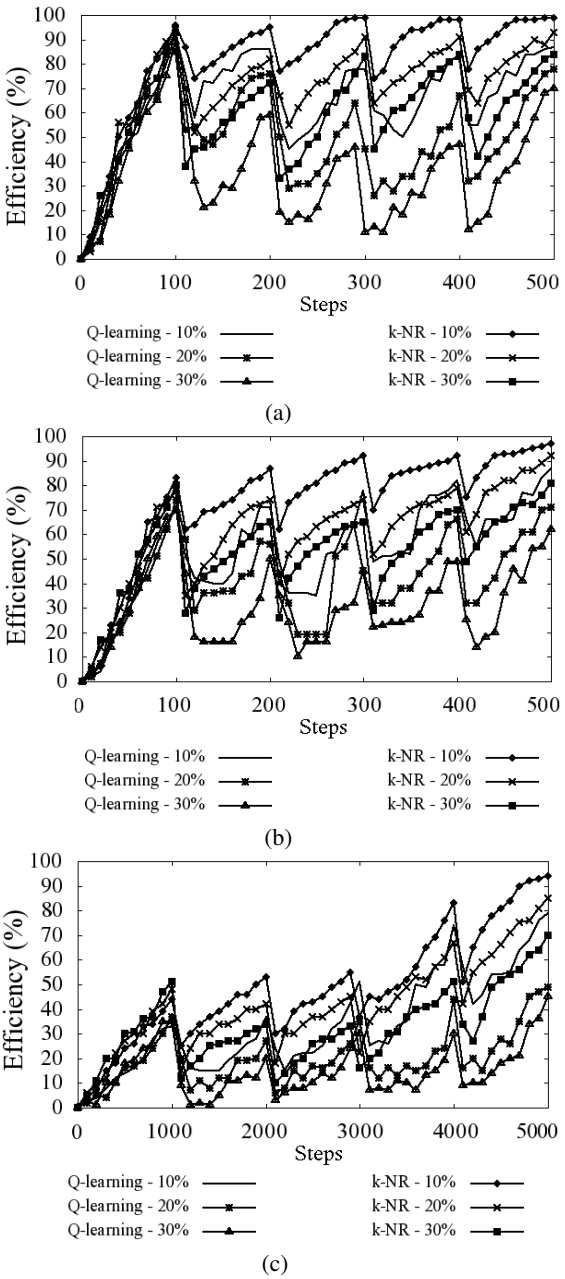


Figure 3: Performance of the *K*-NR and *Q*-Learning algorithms in a: (a) 16-state environment, (b) 25-state environment, (c) 64-state environment.

number of steps, it is seen that there is again convergence to a better policy, as happens when learning begins and performance improves. The decreasing performance occurs because *Q*-Learning needs to re-explore all the state space, re-visiting states with low rewards to accumulate better values for the future. The ε-greedy strategy helps the agent by introducing random actions so that local maxima are avoided. For

example, a *blocked* state that changed to *low jam* must have negative rewards and would no longer be visited.

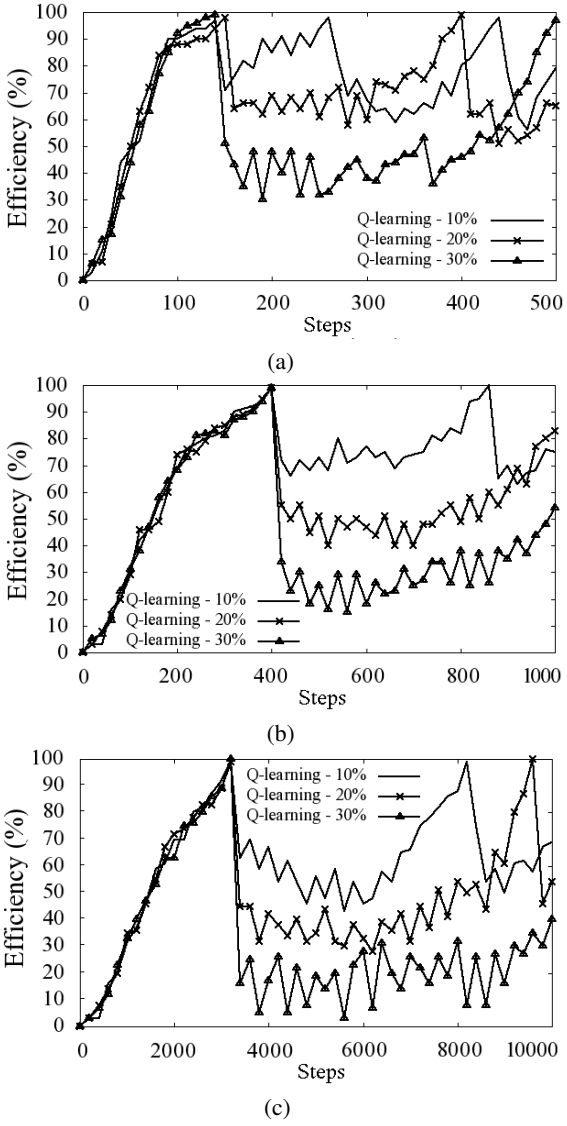


Figure 4: Agent adaptation using the *Q*-Learning in a: (a) 16-state environment, (b) 25-state environment, (c) 64-state environment.

We used the *k*-NR algorithm with heuristic to optimize agent performance with the methodology discussed in Section 2, which uses instance-based learning in an attempt to solve the problem described in the previous subsection. The heuristic has been applied only to the environment states where changes occur. Thus, the heuristic usually caused the agent to modify its learning and converge more rapidly to a good action policy.

Figure 4 also shows that the *Q*-Learning does not show uniform convergence when compared with *k*-NR (Figure 5). This occurs because the *k*-NR algorithm uses instance-based learning, giving superior performance and speeding up its convergence. The *k*-NR is able to accelerate the convergence because the states that have had their status altered were estimated from similar situations observed in the past, so that states with similar features have similar rewards.

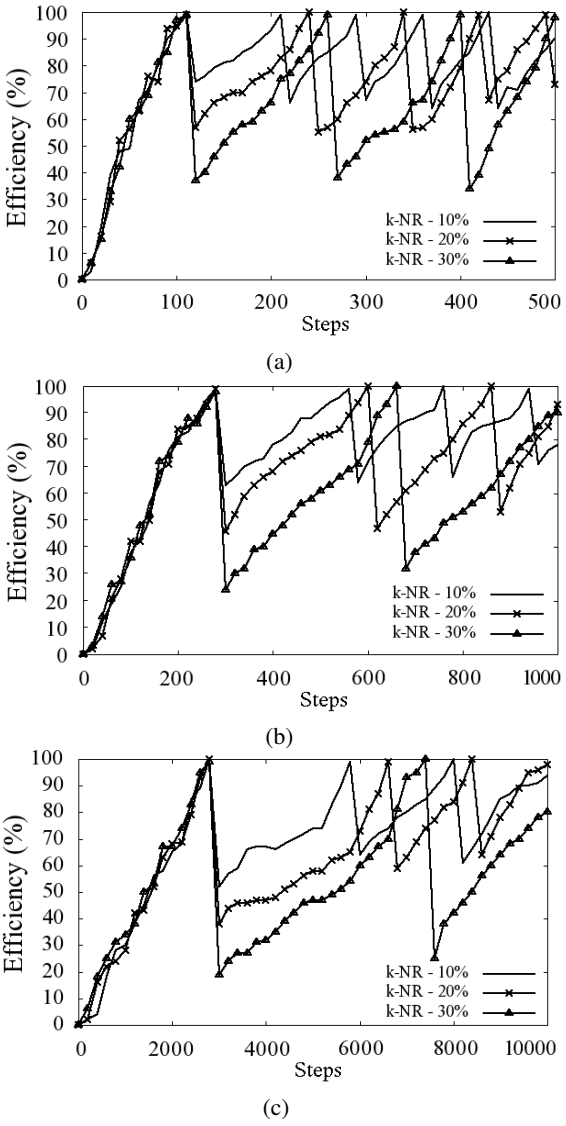


Figure 5: Agent adaptation using the *k*-NR in a: (a) 16-state environment, (b) 25-state environment, (c) 64-state environment.

Table 2 shows the number of steps needed for the agent to re-find its best action policy. It is seen that *k*-NR performs better than standard *Q*-Learning. In 16-state environments, the agent finds its best policy of actions with 150 steps using the *Q*-Learning algorithm and 110 steps with *k*-NR. After changing the en-

Table 2: Number of steps needed for the agent to find its best policy after changes.

	Before changes		After changes					
			10%		20%		30%	
♯ states	Q	k	Q	k	Q	k	Q	k
16	150	110	130	90	240	140	350	150
25	400	280	440	290	1130	320	2140	380
64	3,430	2,830	5,450	2,950	6,100	3,850	13,900	4,640

vironment with 10%, 20% and 30% the k-NR needed 47% fewer steps on average before it once again finds a policy leading to convergence. For 25-state environments the agent finds its best action policy in approximately 400 steps using Q-Learning, and in 280 steps with k-NR. In this environment, k-NR uses an average of 30% fewer steps than Q-Learning, after alteration of the environment. For 64-state environments, the agent needed an average of 3,430 steps to find its best action policy with Q-Learning and 2,830 with k-NR. The k-NR used in average 18% fewer steps than Q-Learning after environmental change. It is seen that k-NR is more robust in situations where the reward values vary unpredictably. This happens because the k-NN algorithm is less sensitive to noisy data.

5 DISCUSSION AND CONCLUSIONS

This paper has introduced a technique for speeding up convergence of a policy defined in dynamic environments. This is possible through the use of instance-based learning algorithms. Results obtained when the approach is used show that RL algorithms using instance-based learning can improve their performance in environments with configurations that change. From the experiments, it was concluded that the algorithm is robust in partially-known and complex dynamic environments, and can help to determine optimum actions. Combining algorithms from different paradigms is an interesting approach for the generation of good action policies. Experiments made with the k-NR algorithm show that although computational costs are higher, the results are encouraging because it is able to estimate values and find solutions that support the standard Q-Learning algorithm.

We also observed benefits related to other works using heuristic approaches. For instance, Bianchi et al. (Bianchi et al., 2004) proposes a heuristic for RL algorithms that show a significantly better performance (40%) than the original algorithms. Pegoraro et al. (Pegoraro et al., 2001) use a strategy that speeds up the convergence of the RL algorithms by 36%, thus reducing the number of iterations compared with traditional RL algorithms. Although the results obtained with the new technique are satisfactory, additional experiments are needed to answer some questions raised. For example, a multi-agent architecture could be used to explore states placed further from the goal-state and in which the state rewards are smaller. Some of these strategies are found in Ribeiro et al. (Ribeiro et al., 2008; Ribeiro et al., 2011). We also intend to use more than one agent to analyze situations as: i) sharing with other agents the learning of the best-performing one; ii) sharing learning values among all the agents simultaneously; iii) sharing learning values among the best agents only; iv) sharing learning values only when the agent reaches the goal-state, in which its learning table would be unified with the tables of the others. Another possibility is to evaluate the algorithm in higher-dimension environments, that are also subject to greater variations. These possibilities will be explored in future research.

REFERENCES

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Amato, C. and Shani, G. (2010). High-level reinforcement learning in strategy games. In *Proc. 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, pages 75–82.

Banerjee, B. and Kraemer, L. (2010). Action discovery for reinforcement learning. In *Proc. 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, pages 1585–1586.

Bianchi, R. A. C., Ribeiro, C. H. C., and Costa, A. H. R. (2004). Heuristically accelerated q-learning: A new approach to speed up reinforcement learning. In *Proc. XVII Brazilian Symposium on Artificial Intelligence*, pages 245–254, So Luis, Brazil.

Bianchi, R. A. C., Ribeiro, C. H. C., and Costa, A. H. R. (2008). Accelerating autonomous learning by using heuristic selection of actions. *Journal of Heuristics*, 14:135–168.

Butz, M. (2002). State value learning with an anticipatory learning classifier system in a markov decision process. Technical report, Illinois Genetic Algorithms Laboratory.

Comanici, G. and Precup, D. (2010). Optimal policy switching algorithms for reinforcement learning. In *Proc. 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, pages 709–714.

Dimitrakiev, D., Nikolova, N., and Tenekedjiev, K. (2010). Simulation and discrete event optimization for automated decisions for in-queue flights. *Int. Journal of Intelligent Systems*, 25(28):460–487.

Drummond, C. (2002). Accelerating reinforcement learning by composing solutions of automatically identified subtask. *Journal of Artificial Intelligence Research*, 16:59–104.

Enembreck, F., Avila, B. C., Scalabrini, E. E., and Barthes, J. P. A. (2007). Learning drifting negotiations. *Applied Artificial Intelligence*, 21:861–881.

Firby, R. J. (1989). *Adaptive Execution in Complex Dynamic Worlds*. PhD thesis, Yale University.

Galvn, I., Valls, J., Garca, M., and Isasi, P. (2011). A lazy learning approach for building classification models. *Int. Journal of Intelligent Systems*, 26(8):773–786.

Jordan, P. R., Schvartzman, L. J., and Wellman, M. P. (2010). Strategy exploration in empirical games. In *Proc. 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10), v. 1*, pages 1131–1138, Toronto, Canada.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Le, T. and Cai, C. (2010). A new feature for approximate dynamic programming traffic light controller. In *Proc. 2th International Workshop on Computational Transportation Science (IWCTS'10)*, pages 29–34, San Jose, CA, U.S.A.

Mohammadian, M. (2006). Multi-agents systems for intelligent control of traffic signals. In *Proc. International Conference on Computational Inteligence for Modelling Control and Automation and Int. Conf. on Intelligent Agents Web Technologies and Int. Commerce*, page 270, Sydney, Australia.

Pegoraro, R., Costa, A. H. R., and Ribeiro, C. H. C. (2001). Experience generalization for multi-agent reinforcement learning. In *Proc. XXI International Conference of the Chilean Computer Science Society*, pages 233–239, Punta Arenas, Chile.

Pelta, D., Cruz, C., and Gonzlez, J. (2009). A study on diversity and cooperation in a multiagent strategy for dynamic optimization problems. *Int. Journal of Intelligent Systems*, 24(18):844–861.

Price, B. and Boutilier, C. (2003). Accelerating reinforcement learning through implicit imitation. *Journal of Artificial Intelligence Research*, 19:569–629.

Ribeiro, C. H. C. (1999). A tutorial on reinforcement learning techniques. In *Proc. Int. Joint Conference on Neural Networks*, pages 59–61, Washington, USA.

Ribeiro, R., Borges, A. P., and Enembreck, F. (2008). Interaction models for multiagent reinforcement learning. In *Proc. 2008 International Conferences on Computational Intelligence for Modelling, Control and Automation; Intelligent Agents, Web Technologies and Internet Commerce; and Innovation in Software Engineering*, pages 464–469, Vienna, Austria.

Ribeiro, R., Borges, A. P., Ronszcka, A. F., Scalabrin, E., Avila, B. C., and Enembreck, F. (2011). Combinando modelos de interao para melhorar a coordenao em sistemas multiagente. *Revista de Informtica Terica e Aplicada*, 18:133–157.

Ribeiro, R., Enembreck, F., and Koerich, A. L. (2006). A hybrid learning strategy for discovery of policies of action. In *Proc. International Joint Conference X Ibero-American Artificial Intelligence Conference and XVIII Brazilian Artificial Intelligence Symposium*, pages 268–277, Ribeiro Preto, Brazil.

Sislak, D., Samek, J., and Pechoucek, M. (2008). Decentralized algorithms for collision avoidance in airspace. In *Proc. 7th International Conference on AAMAS*, pages 543–550, Estoril, Portugal.

Spaan, M. T. J. and Melo, F. S. (2008). Interaction-driven markov games for decentralized multiagent planning under uncertainty. In *Proc. 7th International Conference on AAMAS*, pages 525–532, Estoril, Portugal.

Strehl, A. L., Li, L., and Littman, M. L. (2009). Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research (JMLR)*, 10:2413–2444.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Tesauro, G. (1995). Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3/4):279–292.

Zhang, C., Lesser, V., and Abdallah, S. (2010). Self-organization for cordinating decentralized reinforcement learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS'10, pages 739–746. International Foundation for Autonomous Agents and Multiagent Systems.

SHORT PAPERS

Semantic Similarity between Queries in QA System using a Domain-specific Taxonomy

Hilda Kosorus[1], Andreas Bögl[2] and Josef Küng[1]
[1]*Institute of Applied Knowledge Processing, Johannes Kepler University, Altenbergerstraße 69, Linz, Austria*
[2]*MEOVI, Hagenberg, Austria*
hkosorus@faw.jku.at, andreas.boegl@meovi.com, jkueng@faw.jku.at

Keywords: Query recommendation, Semantic similarity, Short text similarity, Taxonomy.

Abstract: Semantic similarity has been extensively studied in the past decades and has become a rapidly growing field of research. Sentence or short text similarity measures play an important role in text-based applications, such as text mining, information retrieval and question answering systems. In this paper we consider the problem of semantic similarity between queries in a question answering system with the purpose of query recommendation. Our approach is based on an existing domain-specific taxonomy. We define innovative three-layered semantic similarity measures between queries using existing similarity measures between ontology concepts combined with various set-based distance measures. We then analyse and evaluate our approach against human intuition using a data set of 90 questions. Further on, we argue that these measures are taxonomy-dependent and are influenced by various factors: taxonomy structure, keyword mappings, keyword weights, query-keyword mappings and the chosen concept similarity measure.

1 INTRODUCTION

Current implementations of QA systems that incorporate a recommendation mechanism are based on (i) methods using external sources, like user profiles, (ii) methods based on expectations (e.g. query patterns, models) or (iii) methods using query logs (Marcel and Negre, 2011). These methods do not take into account the semantic meaning of queries. In the past two decades researchers have been studying semantic similarity in order to improve information retrieval and develop intelligent semantic systems.

A semantic sentence similarity measure can have an important role in the development of a query recommender system. Nevertheless, such measures can be successfully used in other directions, like query clustering for discovering "hot topics" or to find the query that best represents a cluster, pattern recognition for identifying user groups or in web page retrieval to calculate page title similarities.

Studies of semantic similarity in the past decades has been focusing on two extremes: either measuring the similarity between single words or concepts or between documents. However, there is a growing need for an effective method to compute short text similarity. Web search technologies incorporate tasks, such as query reformulation, query recommendation, sponsored search and image retrieval, that rely on accurately computing similarity between two very short segments of text. Unfortunately, traditional techniques for detecting similarity between documents and queries fail when directly applied to these tasks. Such methods rely on analysing shared words or the co-occurence of terms in both the query and the document.

In this paper we define innovative three-layered semantic similarity measures between queries using existing similarity measures between ontology concepts combined with various set-based distance measures. We then analyze and evaluate our approach against human intuition using a dataset of 90 questions. The goal of this paper is to present semantic query similarity measures that can be successfully integrated into query recommender systems and to evaluate and compare them in terms of human judgement.

The rest of the paper is structured as follows. In section 2 we review related work in the area of semantic similarity measures between concepts, between sets of concepts and the area of short text similarity. In section 3 we present and define the domain-specific taxonomy on which our semantic similarity measures are based. In section 4 we introduce similarity measures between queries as a combination of topic similarity and keyword similarity using the defined ta-

xonomy. In section 5 we analyze and evaluate these similarity measures. Finally, in section 6 we summarize the contents of this paper, drawing some important conclusions and present our future work.

2 RELATED WORK

The problem of similarity is a heavily researched subject in particular in information retrieval, but also in general in computer science, artificial intelligence, philosophy and natural language processing. Measuring similarity between documents has a long tradition in information retrieval, but these approaches compare only vectors of document features (Burgess et al., 1998; Landauer et al., 1998a; Landauer et al., 1998b), usually single words or word stems, by counting their occurrence in the document.

There is extensive literature on measuring similarity between concepts within a taxonomy (Rada et al., 1989; Lee et al., 1993; Wu and Palmer, 1994; Resnik, 1995; Jiang and Conrath, 1997; Leacock and Chodorow, 1998; Lin, 1998; Resnik, 1999; Li et al., 2003; Bouquet et al., 2004; Haase et al., 2004; Cordì et al., 2005; Al-Mubaid and Nguyen, 2006; Wang et al., 2006; Lee et al., 2008; Dong et al., 2009; Bin et al., 2009), while there are few publications that cover the area of short text semantic similarity (Li et al., 2006; O'Shea et al., 2010; Oliva et al., 2011) and some related to semantic similarity between sets of concepts (Bouquet et al., 2004; Haase et al., 2004; Cordì et al., 2005). In (Li et al., 2006) it is argued that existing long text similarity measures have some limitations and drawbacks and their performance is unsatisfactory when applied to short sentences.

In the following we will briefly present the related research in the domain of semantic similarity between concepts and between sets of concepts.

2.1 Semantic Similarity between Concepts using Taxonomies

There are basically two ways of using an ontology or taxonomy to determine the semantic similarity between concepts: the *edge-based approach* and the *information content-based approach* (Resnik, 1995; Resnik, 1999; Lin, 1998). In the following we will make a short overview of the edge-based approaches.

Intuitively, the similarity of different concepts in an ontology is measured by computing the distance within the ontology. Namely, if two concepts reside closer in the ontology, then we can conclude that they are more similar. When computing the ontology distance we actually use the specialization graph of objects and we define it as being the shortest path between the two concepts (Rada et al., 1989).

Rada, Mili, Bicknell and Blettner (1989) defined the conceptual distance as

$$sim(c_1,c_2) = minimum \quad number \quad of \quad edges \quad separating \quad c_1 \quad and \quad c_2,$$

where c_1 and c_2 are the node representation of the two concepts in the ontology. Wu and Palmer (2004) redefined the edge-based similarity measure taking into account the depth of the nodes in the hierarchical graph:

$$sim(c_1,c_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3}, \qquad (1)$$

where N_1 and N_2 are the number of nodes from c_1 and c_2, respectively, to c_3, the *least common superconcept* (LCS) of c_1 and c_2, and N_3 is the number of nodes on the path from c_3 to the root node.

Li et al. (2003) defined the similarity between two concepts as:

$$sim(c_1,c_2) = \begin{cases} e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} & \text{, if } c_1 \neq c_2 \\ 1 & \text{, otherwise} \end{cases} \qquad (2)$$

where, similarly, the parameters α and β scale the contribution of the two values $l = N_1 + N_2$ and $h = N_3$. Based on the benchmark data set, they obtained the optimal parameters $\alpha = 0.2$ and $\beta = 0.6$.

2.2 Semantic Similarity between Sets of Concepts

Defining a semantic similarity measure between sets of concepts was the next step in computing semantic similarity mainly for information retrieval purposes.

In (Bouquet et al., 2004) the ontological distance between sets of concepts is computed by summing up the distances between every pair (c_1,c_2), where $c_1 \in C_1$ and $c_2 \in C_2$. Haase et al. (2004) used the edge-based similarity measure between concepts defined by Li et al. (2006) (see 2) to introduce the similarity between sets of concepts as:

$$Sim(C_1,C_2) = \frac{1}{|C_1|} \cdot \sum_{c_1 \in C_1} \max_{c_2 \in C_2} sim(c_1,c_2), \qquad (3)$$

which computes an average of distances between $c_1 \in C_1$ and the most similar concept in C_2.

In (Cordì et al., 2005) a new similarity measure between sets of concepts was introduced, which gives more weight to keyword pairs with a higher similarity, but still allowing lower values to contribute to the final outcome.

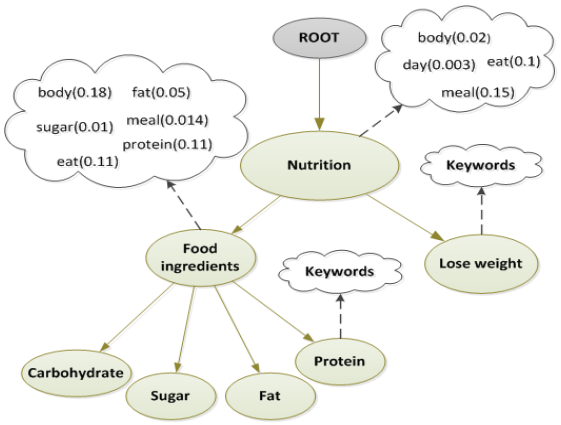


Figure 1: Snapshot of the topic-tree with keywords and their weights.

3 THE DOMAIN-SPECIFIC TAXONOMY

Before introducing our proposed semantic query similarities, it is important to understand the structure of the underlying domain-specific taxonomy. While most of the previously described similarity measures make use of the english lexical taxonomy WordNet[1], our similarity measures are based on a new domain-specific (nutrition) taxonomy with a tree-like structure, where the links between nodes represent IS-A relationships. In the following we will refer to this structure as "topic-tree".

Our topic-tree is composed of a set of *topics*:

$$\mathcal{T} = \{t_1, t_2, ..., t_n\},$$

an IS-A relationship between topics:

$$\mathcal{L} \subset \mathcal{T} \times \mathcal{T}, (t_p, t_q) \in \mathcal{L} \iff t_p \quad parent \quad of \quad t_q,$$

a set of *keywords*:

$$\mathcal{K} = \{k_1, k_2, ..., k_m\},$$

a *mapping* relationship between topics and keywords:

$$\mathcal{M} \subset \mathcal{T} \times \mathcal{K}, (t_p, k_q) \in \mathcal{M} \iff k_q \quad mapped \quad to \quad t_p,$$

and the corresponding mapping *weights*:

$$w : M \to (0, 1],$$

where the value $w(t_p, k_q)$ represents how relevant is keyword k_q for topic t_p.

Figure 1 shows a partial snapshot of the above defined taxonomy. The *topics* represent selected categories and sub-categories in the specified domain (i.e. nutrition), the mapped keywords are frequent relevant words occuring within these topics which were obtained by crawling related websites and/or documents. The corresponding weights were calculated using the TF-IDF method (Salton and Buckley, 1988).

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

4 PROPOSED SEMANTIC SIMILARITY MEASURES

Let $Q = \{q_1, q_2, ..., q_N\}$ be a set of queries in the nutrition domain. We want to define a semantic similarity measure $sim_q : Q \times Q \to [0, 1]$ between these queries using the topic-tree defined in section 3. We assume that to each query $q \in Q$ we can assign a set of keywords $\mathcal{S}_q \subset \mathcal{K}$, where $\mathcal{S}_q$ was extracted from q using some natural language processing methods (HaCohen-Kerner et al., 2005; Turney, 2000; Hulth, 2003). For example, for

$q =$ *"What type of food can I eat and at what time in order to lose weight?"*

$$\mathcal{S}_q = \{food, eat, time, lose\ weight\}.$$

In the following we will define the semantic query similarity sim_q using three other similarity measures: between topics, between keywords and between sets of keywords, each incorporating the one before.

4.1 Semantic Similarity between Topics

Let $sim_t : \mathcal{T} \times \mathcal{T} \to [0, 1]$ be the *topic similarity* function where $sim_t(t_p, t_q)$ represents the semantic similarity between two topics $t_p, t_q \in \mathcal{T}$ using the structure of the topic-tree. For our experiments, we defined sim_t using the similarity measures (1) and (2).

4.2 Semantic Similarity between Keywords

Let $sim_k : \mathcal{K} \times \mathcal{K} \to [0, 1]$ be the *keyword similarity* function where $sim_k(k_p, k_q)$ represents the semantic similarity between two keywords $k_p, k_q \in \mathcal{K}$. We define sim_k in the following way:

$$sim_k(k_p, k_q) = \frac{w_p + w_q}{2} sim_t(t_p, t_q) \qquad (4)$$

where

$$w_i = \max_{(t, k_i) \in \mathcal{M}} w(t, k_i), \quad i \in \{p, q\}$$

and

$$t_i = arg \max_{(t, k_i) \in \mathcal{M}} w(t, k_i), \quad i \in \{p, q\}.$$

4.3 Semantic Similarity between Sets of Keywords

Let $sim_{ks} : \mathcal{P}(\mathcal{K}) \times \mathcal{P}(\mathcal{K}) \to [0, 1]$ be the *keyword-set similarity* function where $sim_{ks}(\mathcal{S}_p, \mathcal{S}_q)$ represents the semantic similarity between two sets of keywords $\mathcal{S}_p, \mathcal{S}_q \subset \mathcal{K}$ and $\mathcal{P}(\mathcal{K})$ contains all subsets of $\mathcal{K}$. In the following we will introduce several possible definitions of sim_{ks} using well-known set distance measures from the literature.

4.3.1 The Sum of Maximum Similarities

The sum of minimum distances measure was originally defined by Niiniluoto (1987) to measure truthlikeness in belief revision theory. We apply the same concept to define the similarity measure sim_{ks} between sets of keywords (the *sum of maximum similarities*):

$$sim_{ks}(\mathcal{S}_p, \mathcal{S}_q) = \frac{1}{2}\left(\frac{1}{|\mathcal{S}_p|}\sum_{k_p \in \mathcal{S}_p} Sim(k_p, \mathcal{S}_q) + \frac{1}{|\mathcal{S}_q|}\sum_{k_q \in \mathcal{S}_q} Sim(k_q, \mathcal{S}_p)\right) \quad (5)$$

where

$$Sim : \mathcal{K} \times \mathcal{P}(\mathcal{K}) \to [0,1], \;\; Sim(k, \mathcal{S}) = \max_{k_s \in \mathcal{S}} sim_k(k, k_s).$$

is the semantic similarity between a keyword $k \in \mathcal{K}$ and a set of keywords $\mathcal{S} \subset \mathcal{K}$.

4.3.2 The Surjection Measure

The surjection measure was introduced by Oddie (1979), who suggested defining the distance between two sets by considering surjections that map the larger set to the smaller one. We applied this concept to measure similarity between sets of keywords, and defined *surjection similarity measure*, sim_{ks}, as

$$sim_{ks}(\mathcal{S}_p, \mathcal{S}_q) = \max_{\eta} \frac{1}{|\eta|} \sum_{(k_p, k_q) \in \eta} sim_k(k_p, k_q). \quad (6)$$

where the maximum is taken over all surjections η that maps the larger set to the smaller one.

4.3.3 The Maximum Link Similarity Measure

The minimum link distance measure was proposed in (Eiter and Mannila, 1997) as an alternative to the previously mentioned distance measures between point sets. First, let us define the *linking* between $\mathcal{S}_p$ and $\mathcal{S}_q$ as a relation $\mathcal{R} \subseteq \mathcal{S}_p \times \mathcal{S}_q$ satisfying

(a) for all $k_p \in \mathcal{S}_p$ there exists $k_q \in \mathcal{S}_q$ such that $(k_p, k_q) \in \mathcal{R}$

and

(b) for all $k_q \in \mathcal{S}_q$ there exists $k_p \in \mathcal{S}_p$ such that $(k_p, k_q) \in \mathcal{R}$.

We now apply this concept to define the *maximum link similarity* between sets of keywords as

$$sim_{ks}(\mathcal{S}_p, \mathcal{S}_q) = \max_{\mathcal{R}} \frac{1}{|\mathcal{R}|} \sum_{(k_p, k_q) \in \mathcal{R}} sim_k(k_p, k_q), \quad (7)$$

taking the maximum over all relations $\mathcal{R}$.

4.4 Semantic Similarity between Queries

Finally, we define the *query similarity* measure $sim_q : \mathcal{Q} \times \mathcal{Q} \to [0,1]$ as

$$sim_q(q_a, q_b) = sim_{ks}(\mathcal{S}_{q_a}, \mathcal{S}_{q_b}) \quad (8)$$

where $\mathcal{S}_{q_a}, \mathcal{S}_{q_b} \subset \mathcal{K}$ are the corresponding set of keywords extracted from q_a and q_b, respectively.

5 COMPARISON AND EVALUATION

In order to evaluate these similarity measures we conducted a survey with 15 persons, men and women, age between 25 and 60. We randomly sampled 50 pairs from a dataset of 90 different questions in the nutrition domain and asked the survey participants to compare and measure the relatedness of each pair by ranking them with a value between 0 and 4 (0=not related at all, 1=somehow related, 2=related, 3=very related, 4=similar).

Finally, we compared the participants' ranking against six different semantic similarity measures: the one defined by Haase et al. (3), the sum of all similarities (Bouquet et al., 2004), the one introduced by Cordì (2005), the cosine similarity (Li et al., 2003), the sum of maximum similarities (5), the surjection similarity (6) and the maximum link similarity (7).

While some question pairs were ranked almost the same by all participants (low variance), there were some cases where participants answered very differently (high variance). This reflects how *diversely* is the "relatedness" of two questions perceived by humans. Table 1 contains the mean, maximum and minimum variances calculated by question pairs rankings.

Table 2 contains the correlation values of each semantic similarity method with the average participant ranking values.

Table 1: Survey results - Variances calculated by question pair rankings.

Mean variance	0.93
Maximum variance	2.14
Minimum variance	0

Based on our experiments and the above results we make the following observations:

- The semantic similarity measures depend on the structure of the taxonomy (Bernstein et al., 2005). In our case, the topic hierarchy, the keyword-topic mappings and the assigned keyword weights affect the computed similarity.

Table 2: Correlation between survey results and the semantic similarity measures.

Method	Correlation
Haase	0.605
Sum of All	0.597
Cordì	0.563
Cosine	0.563
Sum of Maximum	0.617
Surjection	**0.634**
Maximum Link	0.626

- The similarity measure between sets of keywords, and therefore between queries, depends on the chosen topic similarity (edge-based or information content-based) and on the keyword similarity. In our experiments we used the edge-based similarity measures defined by Wu and Palmer (1994) and Li et al. (2003).

Table 3: Types of question pairs based on ranking variance and difference between average survey ranking and semantic similarity values.

Type	Var.	Diff.	Percentage
A	low	low	48%
B	high	low	20%
C	low	high	12%
D	high	high	20%

- Although the correlation between the participants' ranking and the evaluated measures are rather low (see table 2), this can be explained by the following factors:
  - the queries are selected from a specific and narrow domain (nutrition),
  - the concepts that appear in the queries are rather complex,
  - the participants' ranking for some question pairs was very diverse,
  - the participants tend to understand the ranking values or the question pair "relatedness" differently.
- The correlation results (between 0.563 and 0.634) do not contradict the fact that the semantic similarity measures reflect on some level the human perception. Most of the question pairs were evaluated by the participants and the semantic similarity measures almost the same. In our evaluation, compared to the surjection measure, 48% of the question pairs were of type A and 20% of type B (see table 3).

6 CONCLUSIONS AND FUTURE WORK

In this paper we introduced innovative three-layered semantic similarity measures between queries using a domain-specific taxonomy. We evaluated our measures by conducting an on-line survey and comparing them and other four existing semantic similarity measures against the participants' intuition. The results show that our similarity measures have a higher correlation with the average survey ranking than the other four measures. We believe that measuring semantic similarity between concepts using taxonomies can improve significantly the results retrieved by recommender systems. We also argue that these measures depend on the structure of the underlying taxonomy (hierarchy, keyword-topic mappings, keyword weights, etc.) and on the chosen concept-to-concept similarity measure. In the future, we plan to analyze the aspects that alter the behavior of the semantic similarity measures.

In this context, we distinguish two types of recommendations. The first type can be directly obtained by using the semantic similarity measure and retrieving the queries with the highest similarity to the user's last query. These recommendations will be rather "general" and maybe "too similar" to the last query (i.e. predictions with low probability). The second type of recommendations requires a much elaborate analysis (extracting patterns, clustering) of all users' history and then comparing the learned query patterns to the current user's history. With this type of recommendations we can predict the user's next set of questions (with a high probability) and, on the long run, his interests and goals. In the future we intend to focus on the second type of recommendations. We also plan to test the goodness of the semantic recommendations by analyzing users' feedback.

ACKNOWLEDGEMENTS

The authors would like to thank MEOVI[2] for the financial support during their research that lead to the findings presented in this paper.

REFERENCES

Al-Mubaid, H. and Nguyen, H. A. (2006). A cluster-based approach for semantic similarity in the biomedical domain. In *Proceedings of the 28th IEEE EMBS*

[2]www.meovi.com

Annual International Conference, pages 2713–2717, New York, USA.

Bernstein, A., Kaufmann, E., Bürki, C., and Klein, M. (2005). How similar is it? Towards personalized similarity measures in ontologies. In *7. Internationale Tagung Wirtschaftsinformatik*, pages 1347–1366.

Bin, S., Liying, F., Jianzhuo, Y., Pu, W., and Zhongcheng, Z. (2009). Ontology-based measure of semantic similarity between concepts. In *World Congress on Software Engineering*, volume 2, pages 109–112.

Bouquet, P., Kuper, G., Scoz, M., and Zanobini, S. (2004). Asking and answering semantic queries. In *Proceedings of Meaning Coordination and Negotiation Workshop (MCNW-04) in conjunction with International Semantic Web Conference*.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25(2-3):211–257.

Cordì, V., Lombardi, P., Martelli, M., and Mascardi, V. (2005). An ontology-based similarity between sets of concepts. In *6th Joint Workshop "From Objects to Agents": Simulation and Formal Analysis of Complex Systems*, pages 16–21, Camerino,Italy.

Dong, H., Hussain, F. H., and Chang, E. (2009). A hybrid concept similarity measure model for ontology environment. In *Proceedings of the Confederated International Workshops and Posters on the Move to Meaningful Internet Systems*, pages 848–857.

Eiter, T. and Mannila, H. (1997). Distance measures for point sets and their computation. *Journal Acta Informatica*, 34:103–133.

Haase, P., Siebes, R., and Harmelen, F. V. (2004). Peer selection in peer-to-peer networks with semantic topologies. In *International Conference on Semantics of a Networked World: Semantics for Grid Databases*.

HaCohen-Kerner, Y., , Gross, Z., and Masa, A. (2005). Automatic extraction and learning of keyphrases from scientific articles. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 3406 of *Lecture Notes in Computer Science*, pages 657–669. Springer Berlin / Heidelberg.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiang, J. and Conrath, W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics*, pages 19–33, Taiwan.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998a). Introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Landauer, T. K., Laham, D., and Foltz, P. (1998b). Learning human-like knowledge by singular value decomposition: A progress report. In *Advances in Neural Information Processing Systems 10*, pages 45–51. MIT Press.

Leacock, C. and Chodorow, M. (1998). *Combining Local Context and WordNet Similarity for Word Sense Identification*, pages 305–332. In C. Fellbaum (Ed.), MIT Press.

Lee, J. H., Kim, M. H., and Lee", Y. J. (1993). Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, 49(2):188–207.

Lee, W. N., Shah, N., Sundlass, K., and Musen, M. (2008). Comparison of ontology-based semantic-similarity measures. In *AMIA Annual Symposium Proceedings*, pages 384–388.

Li, Y., Bandar, Z. A., and McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4).

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

Marcel, P. and Negre, E. (2011). A survey of query recommendation techniques for data warehouse exploration. *7èmes Journées Francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA)*, B-7.

Oliva, J., Serrano, J. I., del Castillo, M. D., and Iglesias, A. (2011). Sysmss: A syntax-based measure for short-text semantic similarity. *Data and Knowledge Engineering*, 70:390–405.

O'Shea, J., Bandar, Z., Crockett, K., and McLean, D. (2010). Benchmarking short text semantic similarity. *International Journal of Intelligent Information and Database Systems*, 4(2):103 – 120.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI-95*, pages 448–453, Montreal, Canada.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal or Artificial Intelligence Research*, 11:95–130.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336.

Wang, G. H., Wang, Y. D., and Guo, M. Z. (2006). An ontology-based method for similarity calculation of concepts in the semantic web. In *Proceedings of the 5th International Conference on Machine Learning and Cybernetics*, pages 1538–1542, Dalian, China.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138.

Efficient Multi-alternative Protocol for Multi-attribute Agent Negotiation

Jakub Brzostowski[1] and Ryszard Kowalczyk[2]
[1]*Institute of Mathematics, Silesian University of Technology, ul. Kaszubska 23, Gliwice, Poland*
[2]*Faculty of Information and Communication Technologies, Swinburne University of Technology, John St, Hawthorn, Australia*
jakub.brzostowski@polsl.pl, rkowalczyk@swin.edu.au

Keywords: Negotiation, Negotiation Protocol, Negotiation Offer.

Abstract: In this paper we present a novel multi-alternative negotiation protocol for multi-attribute agent negotiations. It allows for improvement of negotiation outcomes in terms of time needed to reach an agreement and the Pareto optimality of the outcome. By allowing the agent to offer a proposal comprising a set of alternatives we eliminate the problem of making trade-offs in the negotiation. We experimentally evaluate the proposed approach to show how it performs in comparison to a typical negotiation protocol.

1 INTRODUCTION

In this work we propose a novel negotiation protocol for multi-attribute agent negotiations allowing agents to improve the negotiation outcome both in terms of time needed to perform a successful negotiation and Pareto efficiency of agreements. Typical negotiation protocols used for solving the multi-attribute agent negotiations are based on exchanging single offers. This means that in the consecutive rounds of negotiation an agent can only propose a single agreement alternative. Such a negotiation protocol requires an agent to trade-off between multiple attributes of an object under negotiation in order to improve the negotiation outcome in terms of Pareto efficiency. However, making trade-offs in multi-attribute negotiations is a difficult problem since it is hard to determine the direction of trade-offs that guarantees the optimal outcome.

Works of John Nash (Nash, 1950) formulate the negotiations as cooperative games and propose a solution in the form of an arbitration scheme, which underpins mediation in negotiations. Based on the Nash bargaining solution a negotiation protocol needs to allow the agents to truthfully reveal their preferences to a trusted third party, i.e. a mediator. The preferences are aggregated by the mediator to determine a solution satisfying a series of axioms. The problem in such an approach is the assumption of truthful revelation of preferences. Therefore, instead of revealing the full structure of preferences, a number of negotiation protocols assume the agents can exchange single offers repeatedly until they reach an agreement. The method based on multiple exchange of offers is more practical and realistic since the structure of preferences is not revealed.

In general, most existing approaches are based either on the assumption of knowledge about the opponents preferences or the use of a trusted third party, i.e. mediator that can guide the negotiation agents in making efficient trade-offs and reaching an agreement. In the mediation approach the parties submit some knowledge about their preferences to the mediator that fuses the knowledge of both parties and proposes solutions. Ethamo et. al (Ethamo et al., 1999) present a constraint proposal method to generate a Pareto frontier of a multi-attribute negotiation. The mediator generates a constraint in each consecutive step and asks the parties to find an optimal solution satisfying this constraint. If the feedback from the agents coincide then a solution is found, otherwise the mediator updates the constraint based on the received feedback and the procedure continues. The approach proposed by Klein et al. (Klein et al., 2003) addresses mediation in the case of complex contracts where the values of issues are binary (either 0 or 1). In each stage of mediation the unbiased mediator generates an offer and proposes it to the parties. In the next stage the agents vote whether to accept the offer or not according to their private strategies. If both agents vote to accept the proposed offer it is mutated in the next stage (values of some issues are switched) and the procedure is repeated. In the case one of the agents votes to reject an offer, the last acceptable of-

fer is mutated and proposed again to the parties. In the work of Li et al. (Li et al., 2011) the authors present an approach for supporting mediation with the use of the Conditional Preference (CP) Networks. Similarly to the approach of Klein et al. (Klein et al., 2003) the approach is applicable for issues with low level of options. The agents build their CP networks that encode their preferences and then submit them to the third party which fuses the preferences by the use of majority rule-based aggregation.

The approaches to trade-off performed by individual agents, rather than the mediator, include a mechanism proposed by Faratin et al. (Faratin et al., 2002) that uses similarity criteria. The trade-off is performed according to a similarity measure between the last offer proposed by the counterpart and the current proposal of the negotiation agent. In making a trade-off the indifference curve is considered. An alternative located on the indifference curve that maximizes the similarity to the last offer proposed by the counterpart is selected for a proposal. Other approaches modify protocols of negotiation allowing the agents to include in the proposal different type of knowledge apart form the negotiation alternative, i.e. the agent can also send to the counterpart arguments aiming at convincing the partner to change his beliefs. A belief that can be influenced by such a kind of persuasion is typically the utility function of the counterpart. Sycara (Sycara, 1991) proposes an approach incorporating argumentation into negotiation and illustrates the merit of argumentation-based reasoning in negotiation dialogues.

Some works also consider protocols dependent on the shape of preferences. Ito et all. (Ito et al., 2007) consider non-linear utility functions, and propose a protocol where the agents employ adjusted sampling to generate proposals and use a bidding-based mechanism to find social welfare maximizing deals. However, in their work they also assume that the bids are submitted to the mediator, which again is an issue since such a protocol assumes revelation of private information. Similarly as in other works (Hattori et al., 2007), (Fujita et al., 2010a) (Fujita et al., 2010b) there is an issue with assuming a central authority to which the information about utilities is revealed. Such solutions require the presence of a third trusted party that is unbiased, independent and capable of carrying out intensive computation.

In the work (Bichler and Segev, 2001) authors present an approach towards establishing a toolset for the design of negotiation protocols on electronic markets focusing rather on mechanism of auction and assuming single-offer bids.

The work of Lai et al. (Lai et al., 2008) presents a decentralized model for self-interested agents aiming at reaching win-win solutions in the multi-attribute negotiation. At each negotiation round an agent proposes a multi-alternative offer, namely it offers several alternatives in one round. When making a counter-offer the partner uses heuristic search in order to propose an offer located on the indifference curve that is closest to the best alternative contained in the set proposed by its counterpart in the previous offer. This alternative is used as a seed and the remaining alternatives that will be send together with the seed are chosen from the neighbourhood of this alternative. However in that approach there is no guarantee of reaching a Pareto optimal solution.

In this paper we propose a protocol that uses multi-alternative offers and allows the agents to reach a Pareto optimal solution. In our approach the agents exchange offers consisting of sets of alternatives determined by α-cuts of the search space in each round of negotiation. Therefore, we do not need to apply any heuristic to search the space of alternatives that can be proposed. The proposed protocol is presented in Section 2. Its experimental evaluation and discussion of the results is presented in Section 3. The concluding remarks are presented in Section 4.

2 THE NOVEL NEGOTIATION PROTOCOL

Typical protocols used in agent negotiation are based on exchanging single alternative proposals. Namely in each consecutive round of negotiation an agent sends to its counterpart an offer consisting of a single alternative. In such an approach the agents are forced to perform trade-offs while looking for agreement that can satisfy the preferences of both negotiation parties. In this paper we consider an approach in which instead of single alternative offers the agents can use multiple alternatives enclosed in one negotiation proposal. In such a situation the sending agent assumes that all alternatives enclosed in the offer are acceptable with the same value of utility. This means that all alternatives proposed in one round of negotiation are indifferent to the proposing agent. The counterpart receiving the offer can check each of the alternatives contained in the offer to what extend its preferences are satisfied. In such a situation the receiver can select the alternative maximizing its utility and decide if such an alternative is suitable to form an agreement. It is intuitive that in the case of multiple alternatives forming one proposal the chance of finding an agreement is higher than in the case of a protocol where a single alternative is proposed. Indeed, as we will

show later in this paper an agreement is reached faster and its value is more efficient than in the case of a typical protocol. More specifically the proposed protocol is realized as follows. The preferences of a negotiator are encoded by an utility functions assigning to each feasible alternative a score. The agent concedes during the negotiation process in the space of utility according to its negotiation strategy. At each negotiation round the agent proposes a full set of alternatives (in a discrete space of alternatives) exceeding the current value of utility. The offer comprising all alternatives exceeding particular value of utility that eliminates the need of using trade-offs since the offer contains the whole indifference curve.

2.1 Negotiation Thread

The negotiation thread is a sequence of proposals and counter-proposals of two negotiation parties. As said above the elements of the sequence are subsets of the acceptance sets of two negotiation parties. Let us assume that the agents defined its utility functions u^a and u^b over the sets of feasible two-attribute alternatives D^a and D^b (acceptance sets) of agent a and agent b, respectively.

Definition 1. *A Negotiation thread between agents $a,b \in Agents$ at time $t_n \in Time$ is any finite sequence of length n of the form $(C^{t_1}_{a\to b}, C^{t_2}_{a\to b}, \ldots, C^{t_n}_{a\to b})$ with $t_1, t_2, \ldots, t_{n-1} \leq t_n$, where:*

1. $t_{i+1} > t_i$
2. *Each offer $C^{t_i}_{a\to b}$ proposed by agent a is determined in the following way: $C^{t_i}_{a\to b} = \{(x,y) \in D^a \,|\, u^a(x,y) \geq f^a(t_i)\}$ where $f^a(t_i)$ is the concession in utility space in time point t_i for agent a*
3. *The analogous offer $C^{t_i}_{b\to a}$ proposed by agent b is determined in the following way: $C^{t_i}_{b\to a} = \{(x,y) \in D^b \,|\, u^b(x,y) \geq f^b(t_i)\}$ where $f^b(t_i)$ is the concession in utility space in time point t_i for agent b*

The negotiation thread is active if none of the agents accepted the offer or withdrew from the negotiation.

2.2 Evaluation Decisions

The evaluation decision says when the negotiation agent can propose its next offer, accept the counterpart's offer or withdraw from the negotiation. When the offer that an agent a is going to propose in the next round overlaps with the last offer of counterpart b the agent a is ready to accept the partners last proposal. The existence of non-empty overlap is equivalent to the condition that the utility function u^a of the agent a exceeds the current level of its concession over the last proposal of the counterpart b. When the overlap is empty the agent a proposes the next offer. In the case of exceeding the time given for negotiation the agent a withdraws.

Definition 2. *For the agent a and its associated utility function u^a, a's interpretation (I) at time t' of the counterpart offer $C^t_{b\to a}$ proposed at time $t < t'$, is defined as:*

$$I_a(t', C^t_{b\to a}) = \begin{cases} withdraw(a,b) & if \quad t' > t_{max} \\ accept(a,b,p(C^t_{b\to a} \cap C^{t'}_{a\to b})) & if \quad f(t',\beta^a) \in u^a(C^t_{b\to a}) \\ offer(a,b,C^{t'}_{a\to b}) & otherwise \end{cases} \quad (1)$$

where f is a decision function and β^a is the parameter determining the shape of concession curve generated with function f and p is a function choosing any point from the set. The equivalent definition of interpretation is of the following form:

$$I_a(t', C^t_{b\to a}) = \begin{cases} withdraw(a,b) & if \quad t' > t_{max} \\ accept(a,b,p(C^t_{b\to a} \cap C^{t'}_{a\to b})) & if \quad C^t_{b\to a} \cap C^{t'}_{a\to b} \neq \emptyset \\ offer(a,b,C^{t'}_{a\to b}) & otherwise \end{cases} \quad (2)$$

According to the above interpretation the negotiation outcome is one point taken from the set $C^t_{b\to a}$. The agent a will accept such a point if its current acceptance threshold $f(t',\beta^a)$ lies in the image of last opponents offer $C^t_{b\to a}$ under the utility function u^a of agent a. Equivalently, the agent a will accept the point $p(C^t_{b\to a})$ if the intersection of sets $C^t_{b\to a}$ and $C^{t'}_{a\to b}$ is not empty.

2.3 Concession Generation Decisions - Tactics

In order to compute the counter-offer $C^{t'}_{a\to b}$ in the form of a set an agent uses functions called tactics. The tactics allow for computing concessions in the utility space $[0,1]$ that then are used in computation of the proposal.

2.3.1 Time-dependent Tactics

When an agent uses the time-dependent tactic it generates its offers according to time that elapses from the beginning of negotiation. In other words the predominant factor influencing the value of concession is the current point in time. The decision function generating offers in the case of time-dependent tactic is dependent on deadline. The agent is conceding in the utility space down to the lowest value 0 when it is approaching the deadline.

The set proposed at time t, with $0 < t < t^a_{max}$, is determined by a function $\alpha^a(t)$ specifying the current level of utility concession.

$$C^t_{a\to b} = \{(x,y) \in D^a \,|\, u^a(x,y) \geq (1 - \alpha^a(t))\}$$

Table 1: The results of experiment - comparison of the classical approach and the efficient approach for different negotiation strategies. The Table contains utility values obtained by the first agent for two approaches.

β^a	0.1		0.2		0.5		1		2		5		10	
β^b	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a
0.1	0.23	0.28	0.41	0.49	0.74	0.74	0.8	0.8	0.84	0.88	0.92	0.92	0.92	0.95
0.2	0.23	0.28	0.41	0.49	0.63	0.63	0.73	0.73	0.81	0.81	0.86	0.89	0.89	0.92
0.5	0.20	0.23	0.33	0.33	0.48	0.48	0.6	0.6	0.71	0.71	0.73	0.78	0.83	0.83
1	0.14	0.14	0.23	0.23	0.36	0.36	0.46	0.46	0.59	0.59	0.67	0.73	0.76	0.76
2	0.087	0.10	0.15	0.15	0.26	0.26	0.33	0.36	0.46	0.46	0.59	0.59	0.65	0.65
5	0.05	0.05	0.08	0.08	0.26	0.14	0.2	0.23	0.30	0.30	0.40	0.40	0.49	0.49
10	0.02	0.03	0.04	0.05	0.08	0.08	0.13	0.13	0.19	0.19	0.29	0.29	0.49	0.49

Table 2: The results of experiment - comparison of the classical approach and the efficient approach for different negotiation strategies. The Table contains utility values obtained by the second agent for two approaches.

β^a	0.1		0.2		0.5		1		2		5		10	
β^b	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a
0.1	0.23	0.28	0.23	0.28	0.23	0.23	0.14	0.14	0.08	0.10	0.05	0.05	0.02	0.03
0.2	0.41	0.49	0.41	0.49	0.33	0.33	0.23	0.23	0.15	0.15	0.07	0.08	0.04	0.05
0.5	0.68	0.74	0.63	0.63	0.48	0.48	0.36	0.36	0.26	0.26	0.12	0.14	0.08	0.08
1	0.8	0.8	0.73	0.73	0.6	0.6	0.46	0.46	0.36	0.36	0.2	0.23	0.13	0.13
2	0.84	0.88	0.81	0.81	0.71	0.71	0.55	0.59	0.46	0.46	0.3	0.3	0.19	0.19
5	0.92	0.92	0.89	0.89	0.78	0.78	0.67	0.7	0.59	0.59	0.40	0.40	0.29	0.29
10	0.89	0.95	0.89	0.92	0.83	0.83	0.76	0.76	0.65	0.65	0.49	0.49	0.49	0.49

Table 3: The results of experiment - comparison of the classical approach and the efficient approach for different negotiation strategies. The Table contains numbers of rounds used to reach agreement in case of two approaches.

β^a	0.1		0.2		0.5		1		2		5		10	
β^b	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a	u_c	u_a
0.1	2	1	2	1	5	2	12	6	20	10	36	18	46	22
0.2	2	1	2	1	8	4	16	8	28	13	40	19	46	23
0.5	5	2	8	4	14	8	22	12	30	16	42	22	50	25
1	10	6	16	8	22	12	30	16	38	19	46	23	50	26
2	18	10	26	13	32	16	38	19	42	22	50	25	54	27
5	34	18	37	19	44	22	46	23	50	25	54	27	54	28
10	44	22	45	23	49	25	50	26	54	27	54	28	54	28

The offer defined above includes all alternatives from the acceptance set D^a of the agent a that exceed in terms of utility the current level of concession $1-\alpha^a(t)$. The function $\alpha^a(t)$ can be defined in variety of ways under the condition that $0 \leq \alpha^a(t) \leq 1$. This range is universal since it can be rescaled to fit the space in which the agent is conceding. Faratin (Faratin et al., 2002) proposed two families of functions, namely the polynomial decision functions and exponential decision functions. Both families are parametrized by a value of $\beta \in R^+$ specifying the shape of the concession curve.

- **polynomial** $\alpha^a(t) = k^a + (1-k^a_j)(\frac{min(t,t^a_{max})}{t^a_{max}})^{\frac{1}{\beta}}$
- **exponential** $\alpha^a(t) = e^{(1-\frac{min(t,t^a_{max})}{t^a_{max}})^{\beta} lnk^a}$

where the parameter k^a specifies the first concession, β^a is responsible for the shape of a curve, t^a_{max} is the deadline of the agent a and t is the current point in time. In the next sections we extend the negotiation tactics proposed by Faratin (Faratin et al., 2002) to fit the proposed protocol.

2.3.2 Behaviour-dependent Tactics

The behaviour-dependent tactic computes the next offer imitating the behaviour of the negotiation partner. The concession in the utility space may be determined based on the previous concessions of the negotiation partner. The agent may imitate the concession in different ways. It may imitate the behaviour proportionally, in absolute terms or it may compute the concession as an average of proportions in a number of previous offers. Hence, given the negotiation thread:

$$\ldots, C^{t_n-2\delta}_{b\to a}, C^{t_n-2\delta+1}_{a\to b}, C^{t_n-2\delta+2}_{b\to a}, \ldots, C^{t_n-2}_{b\to a}, C^{t_n-1}_{a\to b}, C^{t_n}_{b\to a},$$

1. **Relative Tit-for-Tat.** The agent imitates the opponent relative value of concession proposed $\delta > 1$ steps ago. The imitative offer is determined by multiplying previous offer of the decision-maker by the relative concession of the counterpart. The relative concession is the quotient of the two consecutive offers of the opponent proposed δ steps ago. The condition of applicability is $n > 2\delta$.

$$C^{t_{n+1}}_{a\to b} = \{(x,y) \in D^a \mid u^a(x,y) \leq min(max(\frac{max\,u^a(C^{t_n-2\delta}_{b\to a})}{max\,u^a(C^{t_n-2\delta+2}_{b\to a})} (1-\alpha^a(t_{n-1})), 0), 1)\}.$$

The value $max\,u^a(C^{t_n-2\delta}_{b\to a})$ is the utility of the best alternative from the set $C^{t_n-2\delta}_{b\to a}$ from the viewpoint of agent a. Therefore, the coefficient $\frac{max\,u^a(C^{t_n-2\delta}_{b\to a})}{max\,u^a(C^{t_n-2\delta+2}_{b\to a})}$ is the proportion of utility by which

the negotiation partner conceded between the round $n-2\delta$ and the round $n-2\delta+2$ from the viewpoint of agent a. The proportion is multiplied by the last level of utility concession $1-\alpha^a(t_{n-1})$ what results in the utility level to which the agent a is conceding in the next round of negotiation. The next offer is computed as all alternative exceeding this level of utility in terms of utility function of the agent a.

2. **Random Absolute Tit-for-Tat.** The agent imitates the concession of the opponent in absolute terms. This means that for example if the concession of the opponent was 0.2 of utility then the agent also concedes by 0.2. Additionally, the concession is modified by a random value in order to enable an agent to avoid a loop of non-improving contract offers or a local minima in the social welfare function (Faratin et al., 1998). The condition of applicability is again $n > 2\delta$.

$$\begin{aligned} C^{t_{n+1}}_{a\to b} &= \{(x,y)\in D^a \mid u^a(x,y) \leq \\ &\leq min(max(max\,u^a(C^{t_{n-1}}_{a\to b}) - max\,u^a(C^{t_{n-2\delta}}_{b\to a}) + (1-\alpha^a(t_{n-1})) + \\ &+ (-1)^s R(M),0),1)\} \end{aligned}$$

where

$$s = \begin{cases} 0 & \text{If } u^a \text{ is decreasing} \\ 1 & \text{If } u^a \text{ is increasing} \end{cases}$$

and $R(M)$ is a random value from the interval $[0,M]$. M is the maximal value by which an agent can change its imitative behaviour.

As in the case of previous tactic the value $max\,u^a(C^{t_{n-1}}_{a\to b})$ is the utility of the best alternative from the set $C^{t_{n-1}}_{a\to b}$ from the viewpoint of agent a. The difference $max\,u^a(C^{t_{n-1}}_{a\to b}) - max\,u^a(C^{t_{n-2\delta}}_{b\to a})$ is the absolute value of concession of the negotiation partner in utility space from the viewpoint of agent a. This difference is summed with the last value of utility concession $1-\alpha^a(t_{n-1})$ of agent a what results in the current utility level to which the agent a is going to concede. All alternatives exceeding this value are included in the next negotiation offer.

3. **Average Tit-for-Tat.** The agent imitates the overall concession of the opponent proposed in $\gamma > 1$. steps. When $\gamma = 1$ then the offer is the same as in the case of Relative Tit-for-Tat with $\delta = 1$. The condition of applicability is $n > 2\gamma$.

$$C^{t_{n+1}}_{a\to b} = \{(x,y)\in D^a \mid u^a(x,y) \leq min(max(\frac{max\,u^a(C^{t_{n-2\gamma}}_{b\to a})}{max\,u^a(C^{t_n}_{b\to a})}(1-\alpha^a(t_{n-1})),0),1)\}.$$

The above tactics can be combined together to form negotiation strategies ((Faratin et al., 2002)).

3 EXPERIMENTAL EVALUATION AND DISCUSSION OF RESULTS

In this section we present results of an experiment illustrating the efficiency of the proposed multi-alternative protocol of reaching negotiation agreement in comparison with a typical single-alternative negotiation approach with similarity-based trade-off (Faratin et al., 2002). We simulate a number of negotiations in a two-attribute scenario.

We consider the following negotiation setup involving two agents, a client agent and a provider agent. For the client agent the acceptance range is a Cartesian product of the ranges corresponding to two attributes:

$$D^a = [0,1]\times[0,1]$$

Therefore, the range for the first and second attribute is $[0,1]$. The acceptance range for the second agent in the role of provider is defined in the same way. Over the sets D^a and D^b the utility functions for both the agents are defined in the additive form. The weights corresponding to the importance levels of the attributes are set to 0.5. Therefore the function for the client is defined as follows:

$$u^a(x_1,x_2) = 0.5u^a_1(x_1) + 0.5u^a_2(x_2)$$

where the functions u^a_1 and u^a_2 are defined as follows:

$$u_k(x_k) = \begin{cases} 1 & \text{if } x_k < 0.25 \\ \frac{0.75-x_k}{0.75-0.25} & \text{if } 0.25 \leq x_k \leq 0.75 \\ 0 & \text{if } x_k > 0.75 \end{cases} \quad (3)$$

For the provider agent the additive utility function is defined in similar way as for the client agent:

$$u^b(x_1,x_2) = 0.5u^b_1(x_1) + 0.5u^b_2(x_2)$$

However, the single-attribute utility functions are defined with reversed monotonicity compared to the functions of the client agent.

$$u_k(x_k) = \begin{cases} 1 & \text{if } x_k > 0.75 \\ \frac{x_k-0.25}{0.75-0.25} & \text{if } 0.25 \leq x_k \leq 0.75 \\ 0 & \text{if } x_k < 0.25 \end{cases} \quad (4)$$

As described above, the preferences of both agents do not change during the negotiation experiment. What varies in the experiment are the negotiation strategies. We use the time-dependent tactics encoded by the parameter beta indicating how sharp the concession curve is. We apply a wide range of time-dependent tactics varying from the value of 0.1 to 10. We consider seven types of tactics with following values of β parameter:

$$\beta \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$$

For the values of β lower then 1 the strategy belongs to Conceder strategy type. For the value of β equal to 1 the shape of concession curve is linear. For the values of β higher then 1 the negotiation strategy resulting from the usage of such β values belongs to Boulware strategy type. The variety of negotiation strategies used in our experiment aims at investigating how the two approaches for negotiation perform. In the Tables 1,2 we present the results of the experiments. For various negotiation strategies we simulate 49 negotiation settings. In the Table 1 we present the utility values (pay-offs) obtained by the first agent using the traditional (column u_c) and proposed (column u_a) approaches. As we can see, the utilities obtained in the scenario where the second approach was used are not worse or better than utilities obtained in the scenario where the first approach was used. The situation is similar for the second agent - the utilities obtained in the scenario where the second approach was used are at least as good as the utilities obtained in scenario where the first approach was used. In the case of scenario where the second approach was used the obtained results are best, and can not be further improved (in terms of Pareto efficiency) under the assumption of particular preferences and negotiation strategies. The reason for this observation is the application of a specific negotiation protocol which allows the agents to propose the full α-cuts. Such a protocol leads to Pareto efficient outcomes since in a particular round of negotiation the agents propose all feasible alternatives exceeding the particular level of utility allowed at this stage of negotiation. Therefore, the second approach results in Pareto efficient outcomes and therefore outperforms slightly the first approach which does not guarantee the Pareto efficiency. In the third Table 3 we present the comparison of numbers of rounds used to reach agreement in scenarios where the first and second approach was used (columns u_c and u_a, respectively). As we can see the number of rounds resulting in agreement in the case of classical approach is approximately twice larger as the number of rounds used to reach agreement in the case of proposed approach and therefore it outperforms the typical, single-alternative approach.

4 CONCLUDING REMARKS AND FURTHER WORK

The paper presents a novel negotiation protocol for multi-attribute agent negotiations based on using α-cuts to determine multi-alternative offers. As shown in the experiments it allows for improvement of negotiation outcomes in the terms of time needed to reach an agreement and the Pareto optimality of the outcome. In addition by allowing the agent to offer a proposal comprising a set of alternatives we eliminate the problem of making trade-off in the negotiation.

In the future work the proposed approach will be tested in scenarios involving different overlaps of acceptance ranges and different deadlines of the negotiating parties. We will also consider a number of issues, higher than two in further experiments.

REFERENCES

Bichler, M. and Segev, A. (2001). Methodologies for the design of negotiation protocols on e-markets. *Computer Networks*, 37:137–152.

Ethamo, H., Hamalainen, R. P., Heiskanen, P., Teich, J., Verkama, M., and Zionts, S. (1999). Generating pareto solutions in a two-party setting: Constraint proposal methods. *Management Science*, 45:1697–1709.

Faratin, P., Sierra, C., and Jennings, N. R. (1998). *Negotiation among groups of autonomous computational agents*. University of Londond.

Faratin, P., Sierra, C., and Jennings, N. R. (2002). Using similarity criteria to make issue trade-offs in automated negotiations. *Artificial Intelligence*, 142:205–237.

Fujita, K., Ito, T., and Klein, M. (2010a). Representative-based protocol for multiple interdependent issue negotiation problems. *Web Intelligence and Intelligent Agents*.

Fujita, K., Ito, T., and Klein, M. (2010b). A secure and fair protocol that addresses weaknesses of the nash bargaining solution in nonlinear negotiation. *Group Decision and Negotiation*, pages 1–19.

Hattori, H., Klein, M., and Ito, T. (2007). A multi-phase protocol for negotiation with interdependent issues. In *Proc. of IAT*.

Ito, T., Hattori, H., and Klein, M. (2007). Multi-issue negotiation protocol for agents: exploring nonlinear utility spaces. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1347–1352.

Klein, M., Faratin, P., Sayama, H., and Bar-Yam, Y. (2003). Protocols for negotiating complex contracts. *IEEE Intelligent Systems*, 18:32–38.

Lai, G., Sycara, K., and Li, C. (2008). A decentralized model for automated multi-attribute negotiations with incomplete information and general utility functions. *Multiagent and Grid Systems*, 4:45–65.

Li, M., Vo, Q. B., and Kowalczyk, R. (2011). Majority-rule-based preference aggregation on multi-attribute domains with cp-nets. In *Proceedings of Autonomous Agents and Multi-Agent Systems*.

Nash, J. (1950). The bragaining problem. *Econometrica*, 18:155–162.

Sycara, K. (1991). Problem restructuring in negotiation. *Management Science*, 37:1248–1268.

Construction of Fuzzy Sets and Applying Aggregation Operators for Fuzzy Queries

Miroslav Hudec[1] and František Sudzina[2]
[1]*Institute of Informatics and Statistics, Dubravská cesta 3, Bratislava, Slovakia*
[2]*School of Business and Social Sciences, Aarhus University, Falstersgade 50, Aarhus, Denmark*
hudec@infostat.sk, fransu@asb.dk

Keywords: Fuzzy Queries, Construction of Fuzzy Sets, Aggregation Operators, Database.

Abstract: Flexible query conditions could use linguistic terms described by fuzzy sets. The question is how to properly construct fuzzy sets for each linguistic term and apply an adequate aggregation function. For construction of fuzzy sets, the lowest value, the highest value of attribute and the distribution of data inside its domain are used. The logarithmic transformation of domains appears to be suitable. This way leads to a balanced distribution of tuples over fuzzy sets. In addition, users' opinions about linguistic terms as well as current content in database are merged. The second investigated issue is selection of an adequate aggregation operator. Usual t-norm functions as well as compensatory γ – operator have been examined. Finally, the interface for managing these issues has been proposed. A user can obtain an overview about stored data before running a query; that may reduce empty or overabundant answers.

1 INTRODUCTION

Users query databases in order to obtain data needed for analysis or decision making. The common way how to realise such a query is to formulate a logical condition. In general, a logical condition consists of several atomic (elementary) conditions connected with logical *and* or *or* operators. Querying with imprecision allows users to implement linguistic terms to better qualify data they wish to obtain. An example of such a query is *select small departments with high turnover*. The linguistic terms clearly suggest that there is a smooth transition between acceptable and unacceptable records.

The fuzzy set theory (Zadeh, 1965) is a rational option which offers the solution. It brings a paradigm in dealing with the graduation, uncertainty and ambiguity described by linguistic terms. Main reasons to use fuzzy logic in queries are discussed in (Dubois and Prade, 1997) and advocated in (Kacprzyk and Zadrożny, 2001).

The matching degree critically depends on constructed membership functions of all linguistic terms (Klir and Yuan, 1995); (Meier et al., 2005) and chosen logical aggregation function. The former issue has been examined in (Kacprzyk and Zadrożny, 2001); (Tudorie, 2008); (Tudorie, 2009). There exist many different operators which calculate conjunctions and disjunctions of membership values (Zimmermann, 2001). Usually, in practical realisations, the minimum t-norm is used as an aggregation function for *and* operator.

Our paper is focused on these two issues of fuzzy queries. Section 2 shortly presents basic concepts of fuzzy queries. Section 3 is devoted to construction of membership functions of linguistic terms used in queries. Section 4 is focused on calculation of query matching degree by aggregation functions. Section 5 presents suggested user interface for managing examined issues of fuzzy queries. Finally, some conclusions are drawn in section 6.

2 PRELIMINARIES OF FUZZY QUERYING

Let R be a table or relation of a relational database. A set of tuples t is then defined as relation on Cartesian product in the following way:

$$R \subseteq \{t \mid t \in Dom(A_1)\, x \cdots x\, Dom(A_n)\} \quad (1)$$

where A_i is the database attribute (table column) and $Dom(A_i)$ is its associated domain. In our case, domains are set of real numbers or its subsets.

In queries based on fuzzy logic, the database

record (tuple) can fully or partially satisfy the intent of a query *Q*. Let *A(Q)* be the set of answers to query *Q* defined in the following way:

$$A(Q) = \{(t, \mu(t)) \mid t \in R \wedge \mu(t) > 0\} \quad (2)$$

where $\mu(t)$ indicates how well the selected tuple *t* satisfies a query criterion. It is expressed as a number from the [0, 1] interval.

Several fuzzy query implementations have been proposed such as FQUERY (Kacprzyk and Zadrożny, 1995), SQLf (Bosc and Pivert, 2000), FQL (Wang et al, 2007), FuzzyKAA (Tudorie, 2009) and fuzzy generalized logical condition (Hudec, 2009). "Although there are variations according to the particularities of different implementations, the answer to a fuzzy query sentence is generally a list of records, ranked by the degree of matching" (Branco et al, 2005, p. 21). The value of matching degree depends on membership functions constructed for each elementary query condition and on chosen aggregation function.

3 CONSTRUCTION OF MEMBERSHIP FUNCTIONS

If the system uses badly defined membership functions, it will not work properly. These functions have to be carefully defined (Galindo, 2008). This issue has two main aspects. In the first aspect, users define parameters of membership functions according to their reasoning and preferences. The second aspect is devoted to calculation of these parameters from data stored in a database.

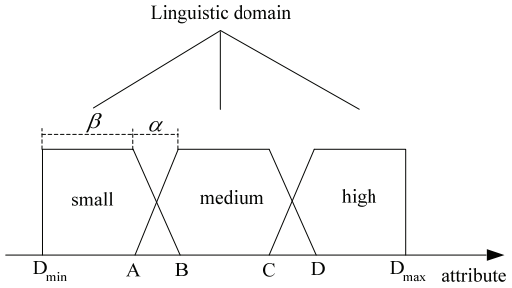


Figure 1: Linguistic and crisp domain.

Let a linguistic domain consists of linguistic terms {*small, medium, high*}. Linguistic domain covers crisp domain of attribute in a way shown in Figure 1. Let D_{mim} and D_{max} be the lowest and the highest domain values of attribute A i.e. Dom(A) = [D_{min}, D_{max}]. Let *L* be the lowest boundary value and *H* be the upper boundary value of attribute in current content of a database; that is, [L, H] $\subseteq$ [D_{min}, D_{max}]. In case of attribute number of days with empty supply shelves, the domain is the [0, 365] interval of integers. For example, empty shelves for all spare parts are noticed between 7 and 75 days i.e. L=5 and H=75.

3.1 Users Create Fuzzy Sets Parameters

In this approach, users are required to choose parameters *A*, *B*, *C* and *D* (Figure 1) according to their reasoning and preferences. Therefore, these parameters are applied in a query realization phase. Detailed discussion on how to cope with this issue can be found in (Klir and Yuan, 1995). Users usually consider their preferences on the whole domain of attributes. Let's have the attribute A defined on domain Dom(A) = [D_{Amin}, D_{Amax}]. Let values for all records be non-uniformly distributed inside domain in such a way that majority of records are concentrated near value *L* whereas few records have value of the attribute A near the value *H* (Figure 2). If a user decide to set parameters *C* and *D* for the condition *attribute A is high* as is depicted in the Figure 2 only few records meet the condition.

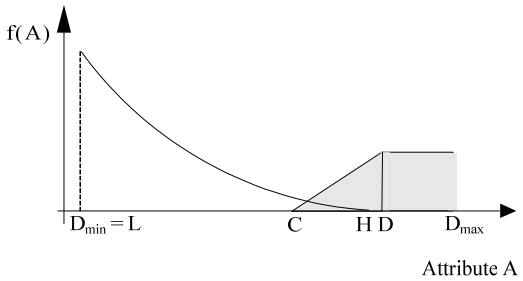


Figure 2: Fuzzy set high for the attribute A.

If the query is more restrictive (conjunction of several atomic conditions) and distribution of values is highly unbalanced, it may easily end up with an empty answer.

3.2 Fuzzy Sets Construction from the Current Content of a Database

This problem was initially examined for fuzzy queries where the second elementary condition depends on the result of the first elementary condition. It means that the second elementary criterion requires taking into account sub domains of the attributes domains limited by tuples already selected by the first elementary condition (Tudorie,

2008). Two ways of fuzzy sets construction are offered: the uniform domain covering method and the statistical mean based algorithm. In our research, we have examined these methods for fuzzy queries where overall query condition consists of atomic conditions connected by *and* operator.

3.2.1 Uniform Domain Covering Method

At the beginning, this method requires the values of L and H. These values are obtained form a database content. Length of fuzzy set core β and length of fuzzy set slope α (Figure 1) are created in the following way (Tudorie, 2008):

$$\alpha = \frac{1}{8}(H - L) \tag{3}$$

$$\beta = \frac{1}{4}(H - L) \tag{4}$$

Consequently, it is easy to calculate required parameters A, B C and D from values L, H, α and β.

The uniform domain covering method reduces the issue depicted in Figure 2, if the distribution of attribute values in the domain is more or less uniform.

If it is not the case, the uniform domain coverage could lead to a highly unbalanced distribution of tuples over fuzzy sets. It implies that only few tuples are in one fuzzy set, while majority of tuples is in another one. It might lead to a conclusion that the meaning of the linguistic term is far from real data.

Let's have a query, which looks for sellers with a high amount of sold items. The query condition has to consider parameters of each region where sellers operate. The meaning of the term high differs among regions.

3.2.2 Statistical Mean based Algorithm

A possible solution is adding the statistical mean into construction of fuzzy sets. The middle of the medium fuzzy set core is the statistical mean of attribute. In this approach, cores of all three fuzzy sets (β) have equal size; lengths of fuzzy sets slopes are different. Experiments on altitude above sea level for 2877 municipalities in Slovakia reveal a limitation of this approach. Many municipalities are close to the value L, whereas only few municipalities are close to H. It is similar to the distribution depicted in Figure 2. Moreover, the value of β is smaller in comparison with the uniform domain covering method. This causes that only two municipalities fully belong to the fuzzy set high. In order to solve this limitation, we have realised experiments with a logarithmic transformation.

3.2.3 Logarithmic Transformation

In many cases, values of attributes are close to e.g. the value L, whereas only few are close to H and therefore belong to the fuzzy set high or contrary. An illustrative example is population density of municipalities where only few big cities have high population density. This kind of data distribution where only few tuples highly determine fuzzy set parameters cannot be properly evaluated by uniform domain covering method or by the linear transformation used in (Kacprzyk and Zadrożny, 2001). The logarithmic transformation is a rational option which might provide a solution. After a logarithmic transformation, the values of α and β are not equally long for all fuzzy sets. The interval [L, H] is transformed into the interval [log(L), log(H)]. Consequently, in this interval, logarithms of α, β and A, B, C and D are calculated using equations (3) and (4). Finally, obtained values are delogarithmised into real values.

4 CALCULATION OF MATCHING DEGREE

The most used operators are t-norm and t-conorm functions; they are specialized for the aggregation under uncertainty (Detyniecki, 2001). In this paper, other aggregation operators are mentioned.

4.1 T-norm Functions

They are generalizations of the two-valued logical aggregation operators. The associative axiom (Klir and Yuan, 1995) ensures that all t-norm and t-conorm functions can be used for *and* and *or* operators respectively. Actually, it is not easy to aggregate all these functions to arbitrary number of elementary conditions. The following t-norm functions can be easily aggregated for cases when more than two attributes are used (Siler and Buckley, 2005):

- minimum

$$\mu(\mathrm{t}) = \min(\mu_i(\mathrm{a_i})) \qquad \mathrm{i} = 1,...,\mathrm{n} \tag{5}$$

- product

$$\mu(\mathrm{t}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} (\mu_i(\mathrm{a_i})) \tag{6}$$

- Lukasiewicz

$$\mu(\mathrm{t}) = \max(0, \sum_{i=1}^{n} \mu_i(a_i) - n + 1) \quad (7)$$

where $\mu_i(a_i)$ denotes the membership degree of the attribute a_i to the i-th fuzzy set.

It is obvious that different t-norm functions calculate different matching degrees. In addition, they do not meet all axioms of Boolean logic. It is consequence of generalization of {0, 1} logic into many-valued logics (including fuzzy logic) based on truth functionality. The two-valued logic meets all axioms of Boolean algebra, namely excluded middle, contradiction and idempotency whereas in fuzzy logic it is not the case (Radojević, 2008).

From the above mentioned t-norms, only minimum (5) is an idempotent t-norm what makes it the most acceptable for users accustomed to the crisp logic. On the other hand, this t-norm does not meet the contradiction axiom. The product t-norm (6) takes into account all membership degrees and balances the query truth membership value across each of elementary conditions. But the query matching degree could be significantly lower than the matching degree of the lowest value of elementary conditions. In addition, this t-norm does not meet the contradiction and the idempotency axioms. The Lukasiewicz t-norm (7) is a nilpotent t-norm. This t-norm satisfies the contradiction axiom but does not satisfy the idempotency axiom.

Let's have two records which satisfy the first elementary condition (A) and the second elementary condition (B) as is shown in Table 1.

Table 1: Example of matching degrees using t-norms.

tuple	A	B	Min (5)	Prod (6)	Luk (7)
1	0.11	0.2	0.11	0.02	0
2	0.1	0.9	0.1	0.09	0

It is obvious that the min-t-norm prefers the tuple 1. This contradicts the human decision-making process. Although the tuple 1 is only slightly better according to the first elementary condition and significantly worse according to the second elementary condition, it is preferred. The product t-norm prefers the second record with the membership degree lower than 0.1. Lukasiewicz t-norm calculates membership degrees of 0 for both records because they do not significantly satisfy both atomic conditions.

4.2 Other Aggregation Functions

"Several authors noticed that t-norms and t-conorms lack compensational behaviour" (Detyniecki, 2001, p.28). This issue can be solved using compensatory operators to model the fuzzy or linguistic *and* operator. The compensation of a bad value of one attribute by a good value of another attribute can be achieved e.g. by the γ - operator (Zimmermann and Zynso, 1980) adapted to the fuzzy queries in the following way:

$$\mu(t) = (\prod_{i=1}^{n} \mu_i(a_i))^{1-\gamma} (1 - \prod_{i=1}^{n} (1 - \mu_i(a_i)))^{\gamma} \quad (8)$$

where $\gamma \in [0,1]$, other elements have the same meaning as in (5) – (7). Applying the γ - operator with the value of 0.5 implies that all attributes are equally relevant in the calculation of the matching degree. A short discussion of applicability of γ - operator can be found in (Werro et al, 2005).

Let's look at the query containing two elementary conditions. The matching degrees of all above mentioned t-norm functions and γ - operator are presented in Table 2.

Table 2: Matching degrees using t-norms and γ = 0.5.

tuple	A	B	min (5)	prod (6)	L (7)	γ (8)
1	0.1	0.1	0.1	0.01	0	0.04
5	0.11	0.2	0.11	0.02	0	0.08
3	0.1	0.9	0.1	0.09	0	0.29
8	0.33	0.42	0.33	0.14	0	0.29
4	0.1	1	0.1	0.1	0.1	0.32
7	0.2	0.9	0.2	0.18	0.1	0.41
9	0.55	0.45	0.45	0.25	0	0.43
11	0.5	0.5	0.5	0.25	0	0.43
12	0.51	0.55	0.51	0.28	0.06	0.47
10	0.9	0.5	0.5	0.45	0.4	0.65
13	0.85	0.77	0.77	0.65	0.62	0.79
14	0.9	0.9	0.9	0.81	0.8	0.9
15	1	1	1	1	1	1

The product t-norm and the γ – operator give us the same ranking of records except the records 8 and 4. The γ – operator requires double time in comparison with the product t-norm. On the other hand, the product t-norm often gives values which are significantly lower than ones obtained from the minimum t-norm. For users, it seems that the compensation of bad and good values is worse than the bad value. If a user cares about "which objects does the system get me first" the product t-norm is a better solution. In other cases, like data examination in official statistics "how the system does internally rate its answers" the γ – operator is more informative. According to results in Table 2, the γ - operator is the most appropriate one.

Other aggregation operators could be applied, such as Choquet integral or Ordered Weighted Averaging (OWA) operators in order to create more

sophisticated queries. The later one is examined in (Zadrożny and Kacprzyk, 2009). Prioritized fuzzy constraint satisfaction problem can be applied in queries which handle fuzzy conditions. The value with the biggest priority has the largest impact on the result given by the priority t-norm (Takači and Škrbić, 2008).

5 QUERY REALIZATION

In (Bordogna and Psaila, 2008), the following drawback of fuzzy query languages is recognized: The proposals defined so far usually assume that fuzzy predicates are defined "a priori" and included in a query at need. Even when user-defined fuzzy predicates can be specified, there are not specific commands in the query language itself to customize the meaning of terms. One solution to this issue is examined in (Tudorie, 2009). The FuzzyKAA is able to assist a user in defining linguistic terms according the content in database.

A direct user input is an ideal case (Gurský et al, 2008). It assumes that a user has a clear idea what data he wants to select. Moreover, it reduces computational burden. This is often not the case and a user needs some information about stored values before he creates query conditions.

5.1 Proposed Interface

In order to manage querying across the approach examined in sections 3 and 4, the interface for desktop application depicted in Figure 3 is proposed. The interface is decomposed into three main parts. The first part deals with the navigation through a list of query-able attributes (in this case, adapted to attributes from the municipal database).

The second part is focused on creation of flexible query conditions. All chosen attributes for the fuzzy part of a query are situated inside the tab control. Each tab page contains one indicator. The user can directly input parameters of linguistic terms (A, B, C and D) or ask for the suggestion by one of methods recommended above (the uniform domain covering method or the logarithmic transformation).

Third part is devoted to selection of aggregation function (γ – operator, minimum and Lukasiewicz) and presenting results in a tabular form.

Finally, the user request is translated into the SQL query and processed by the database management system. At the end of this process, the answer is presented through the interface.

In the suggested approach, users obtain overview of stored data before a query realization, so they have a possibility to adjust parameters of fuzzy sets inside each elementary condition. The suggested approach could reduce empty answer and overabundant answer problems. The empty answer problem simply means that there is no data matching the overall query condition. The query Q results in an empty answer if $Q(t) = Ø$. (Bosc et al, 2008). The overabundant answer problem is defined as an answer where the cardinality of $Q(t)$ is too large (Bosc et al, 2008).

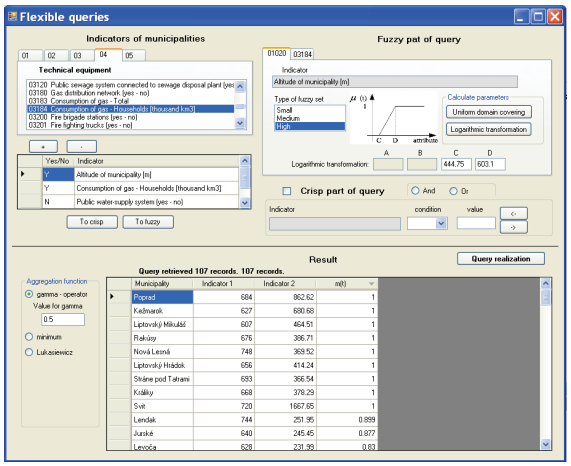


Figure 3: Proposed interface.

6 CONCLUSIONS

Although fuzzy set theory has been already established as an adequate framework to deal with flexible queries, there are still many ways how to improve fuzzy queries. In our paper, we focused on the issue of fuzzy sets construction and examination of adequate aggregation functions.

The first issue can be satisfactorily solved if we merge a user's opinion about linguistic terms with the current content in database. A user can directly input fuzzy sets parameters or ask for suggestions. The uniform domain coverage method is appropriate when attribute values are more or less uniformly distributed inside its domain. In the other case, a logarithmic transformation is more suitable. This information helps to reduce empty or overabundant answer problem.

For the second issue, t-norm functions used in fuzzy queries are discussed. As a result, the γ – operator is suggested. This operator takes into account all membership degrees and compensates a bad value of one attribute with a good value of another attribute.

Finally, both above examined issues have been incorporated into the proposed querying interface.

Integration of approaches of membership functions construction from current content in database and selection of appropriate aggregation operators could bring more sophisticated querying tool for end users.

The topic for further research is how to recognize directly from data whether the uniform domain method is more suitable than the logarithmic transformation and how to offer most suitable aggregation operator to meet users' needs.

REFERENCES

Bordogna, G., Psaila, G., 2008. Customizable Flexible Querying for Classical Relational Databases. In: Galindo J. (Ed.), *Handbook of Research on Fuzzy Information Processing in Databases* (pp. 191-217). IGI Global, London.

Bosc, P., HadjAli, A., Pivert, O., 2008. Empty versus overabundant answers to flexible relational queries. *Fuzzy Sets and Systems,* 159, 1450-1467.

Bosc, P., Pivert, O., 2000. SQLf query functionality on top of a regular relational database management system. In: Pons, M., Vila, M. A., Kacprzyk, J. (Eds.), *Knowledge Management in Fuzzy Databases* (pp. 171-190). Physica-Verlag, Heidelberg.

Branco, A., Evsukoff, A., Ebecken, N., 2005. Generating fuzzy queries from weighted fuzzy classifier rules, In *ICDM workshop on Computational Intelligence in Data Mining*. IOS Press.

Detyniecki, M., 2001. Fundamentals on Aggregation Operators, In *AGOP International Summer School on Aggregation Operators*. Asturias.

Dubois, D., Prade, H., 1997. Using fuzzy sets in flexible querying: Why and how? In: Andreasen, T., Christiansen, H., Larsen H.L. (Eds.), *Flexible Query Answering Systems* (pp. 45-60). Kluwer Academic Publishers, Dordrecht.

Galindo, J., 2008. Introduction and Trends to Fuzzy Logic and Fuzzy Databases, In: Galindo J. (Ed.), *Handbook of Research on Fuzzy Information Processing in Databases* (pp. 1-33). IGI Global, London.

Gurský, P., Vaneková, V., Pribolová, J., 2008. Fuzzy User Preference Model for Top-k Search. In *IEEE World Congress on Computational Intelligence (WCCI)*. Hong Kong.

Hudec, M., 2009. An approach to fuzzy database querying, analysis and realisation. *Computer Science and Information Systems*, 6(2), 127-140.

Kacprzyk, J., Zadrożny, S., 2001. Computing with words in intelligent database querying: standalone and internet-based applications. *Information Sciences*, 134, 71–109.

Kacprzyk, J., Zadrożny, S., 1995. FQUERY for Access: Fuzzy querying for windows-based DBMS, In: Bosc, P., Kacprzyk, J. (Eds.), *Fuzziness in Database Management Systems* (pp, 415-433). Physica-Verlag, Heidelberg.

Klir, G., Yuan, B., 1995. *Fuzzy sets and fuzzy logic, theory and applications*, Prentice Hall. New Jersey.

Meier, A., Werro, N., Albrecht, M., Sarakinos, M. 2005. Using a Fuzzy Classification Query Language for Customer Relationship Management. In *Conference on Very Large Data Bases*. ACM.

Radojević, D., 2008. Interpolative realization of Boolean algebra as a consistent frame for gradation and/or fuzziness, In: Nikravesh, M., Kacprzyk, J., Zadeh, L.A. (Eds.), *Forging New Frontiers: Fuzzy Pioneers II Studies in Fuzziness and Soft Computing* (pp. 295-318). Springer-Verlag, Berlin and Heidelberg.

Siler, W., Buckley, J., 2005. *Fuzzy expert systems and fuzzy reasoning*, John Wiley & Sons. New Jersey.

Takači, A., Škrbić, S., 2008. Priority, Weight and Threshold in Fuzzy SQL Systems. *Acta Polytechnica Hungarica*, 5(1), 59-68.

Tudorie, C., 2008. Qualifying objects in classical relational database querying, In: Galindo J. (Ed.), *Handbook of Research on Fuzzy Information Processing in Databases* (pp. 218-245). IGI Global, London.

Tudorie, C., 2009. Intelligent interfaces for database fuzzy querying, *The annals of "Dunarea de Jos" University of Galati*, Fascicle III, 32(2).

Wang, T.C., Lee, H.D., Chen, C.M., 2007. Intelligent Queries based on Fuzzy Set Theory and SQL. *In Joint Conference on Information Science,* World Scientific.

Werro, N., Meier, A., Mezger, C., Schindler, G., 2005. Concept and Implementation of a Fuzzy Classification Query Language. In *International Conference on Data Mining*. CSREA Press.

Zadeh, L. A., 1965. Fuzzy sets. *Information and Control*, 8, 338-353.

Zadrożny, S., Kacprzyk, J., 2009. Issues in the practical use of the OWA operators in fuzzy querying. *Journal of Intelligent Information Systems*, 33, 307-325.

Zimmermann, H.-J., 2001. *Fuzzy Set Theory – and Its Applications*, Kluwer Academic Publishers. London.

Towards Automated Logistics Service Comparison
Decision Support for Logistics Network Management

Christopher Klinkmüller, Stefan Mutke, André Ludwig and Bogdan Franczyk
Information Systems Institute, University of Leipzig, Grimmaische Straße 12, 04109, Leipzig, Germany
{klinkmueller, mutke, ludwig, franczyk}@wifa.uni-leipzig.de

Keywords: Service Comparison, Business Process Similarity, Logistics Management.

Abstract: A recurring task when managing logistics networks in which logistics companies jointly offer services is the comparison of logistics services based on their underlying processes. The comparison is necessary for the integration of processes, the selection of logistics providers and the evaluation of a company's performance. Due to a high diversity of logistics services and their properties as well as due to the high amount of services automated logistics service comparison is needed to support this task. This paper presents basic requirements and evaluates the state of the art with regard to these requirements. In addition, an initial solution approach providing a solid base for future work is outlined.

1 INTRODUCTION

Logistics management plays a central role for most companies in manufacturing industries. It organises flows of goods and information across corporate value chains. Increased value orientation, progressive globalisation, ongoing concentration on core competencies, higher requirements towards the quality of service, and innovation in information and communication technology led to a high diversity logistics management has to deal with (Pfohl, 2004).

As a consequence, logistics companies such as warehouses or carriers start to arrange themselves in logistics networks in order to jointly offer a logistics service bundle that is able to meet customers' expectations. These logistics networks are usually managed by *Logistics Network Service Providers* (LSP) like third and fourth party logistics providers (Gudehus and Kotzab, 2009). LSPs do not necessarily have to provide own physical logistics assets such as trucks for the service delivery. Instead they need to have a wide knowledge of logistics processes and of information technology enabling them to act as the central point of contact to the customer and to coordinate logistics companies in order to flexibly configure services within the network with regard to the customers' requirements.

The main task for LSPs is therefore the network management which was introduced by (Sydow and Duschek, 2011) and which comprises four tasks: the selection of logistics companies which should be part of the network; the regulation of tasks necessary to implement the demanded logistics services; the allocation of these tasks to the companies within the network; and the evaluation of the network.

A recurring problem within those tasks is the comparison of logistics services. When selecting logistics services the LSP needs to check whether the services offered by companies fit to those required by the network. Within the allocation it needs to be examined whether those services are suitable to implement services needed by customers. Furthermore, LSPs have to find similar logistics services which indicate options to obtain economies of scale during the regulation. Finally, a central task when evaluating services is to verify that they still correspond to their initial design. As a manual comparison of logistics services can be quite cumbersome due to the high amount and diversity of logistics services and their properties the objective of this paper is to briefly outline an automated approach to the comparison of logistics services to support decision making within the management of logistics networks. In particular, the contribution of this paper is the evaluation of state of the art based on basic requirements as well as the introduction of an initial approach satisfying these requirements.

The paper is structured as follows. In section 2 the requirements towards the comparison of logistics services are outlined. Afterwards, section 3 evaluates related work with regard to these requirements. The approach is introduced in section

4. Finally, section 5 concludes the paper and gives an outlook on next steps.

2 REQUIREMENTS

This section introduces the basic requirements towards the automated service comparison within logistics network management. These requirements were determined by conducting expert interviews and case studies in the context of two research projects and in collaboration with a logistics network emphasizing the practical need for an appropriate approach. The requirements are outlined in the following and examples that are partly based on the ARIS SmartPath reference processes are used to illustrate the purpose of the requirements.

Requirement 1 (Flow semantics): The most important requirement is that the comparison of logistics services has to be based on the examination of the behaviour of the business processes which implement the services independently of which business process notation is used. Processes as sets of activities performed in coordination by a single company (Weske, 2007) and collaborations of them allow to capture the flows of goods and information which are implemented by a logistics network in order to perform the main task of logistics, namely transferring goods in space and time (Gudehus and Kotzab, 2009). The reason for explicitly looking at the behaviour of processes is that logistics processes are usually characterized by a high degree of variability. A typical example is to compare consignment processes to identify consolidation options. In order to deal with different types of goods there might be some activities whose execution depends on the type. In such a case two processes might be quite different from a structural perspective as the number of types that can potentially be handled by a company might differ from those of another company. Comparing the behaviour instead helps to determine cases which both processes can handle. The behavioural view also allows to compare the actual process execution with process templates in case of unexpected runtime variations. This would probably not be possible from a structural perspective as the variations are commonly not captured in a model. The actual behaviour instead can be reconstructed from data within information systems. Additionally, notation-independence is needed because the companies within the network usually employ different notations, e.g. BPMN, EPC etc, affecting the identification of appropriate services.

Requirement 2 (Context semantics): While the flow semantics consider how a service is delivered, it is also essential to take account of what is done. Common process notations allow to label activities using phrases like "transport goods" and "pick order". This is not sufficient in logistics where it is necessary to consider the context in which a process is executed, e.g. during regulation two transport processes can only be consolidated if their routes are close to each other or during selection it is necessary to determine if a company is able to process individual orders in a special format. Hence, the second requirement is that activities are compared under consideration of a detailed functionality description rather than simply relying on their labels.

Requirement 3 (Level of abstraction): The third requirement refers to the first two requirements. It demands that the approach must take the different levels of abstraction that services can be viewed from into consideration. For example, there might be the option to consolidate a simple transport service with a composed service which consists of a couple of services, but which depicts a similar transport. Considering the flow semantics in such a case, a simple process must be compared to a process collaboration. Furthermore, companies may provide more process details than necessary to the LSPs that are mainly interested in a coarse-grain view onto the activities and the points of interaction. In this case a few activities from an LSP's view could correspond to a complex flow of activities offered by the companies. At the context level there is also a difference between the representation of services offered by companies and of those requested by customers. While services of companies usually illustrate companies' capabilities, services demanded

Table 1: Summary of the requirements.

	Key phrase	Description
Req. 1	Flow semantics	The comparison must be based on a notation-independent analysis of the behaviour of the business processes implementing the logistics services.
Req. 2	Context semantics	Process activities have to be compared on the base of a detailed functionality description rather than relying on labels.
Req. 3	Levels of abstraction	The different levels of abstraction services can be described on need to be regarded.
Req. 4	Presentation of results	The results must enable analysts to investigate reasons for the similarity of two processes.

by LSPs are specified with regard to a certain contract. A simple example to illustrate this is a transport service. A carrier would usually name the region in which it is able to conduct transports, e.g. Central Europe etc., while in a contract there is usually a demand for a specific tour, e.g. from Hamburg to Prague. This requirement is most important when tasks are allocated to companies.

Requirement 4 (Representation of results): In order to support an LSP in decision making it is not sufficient to present the result of a comparison of two services as a single number indicating the degree of overlap or difference between the services. Such a number might indeed be useful to preselect suitable services. Unfortunately, in this case the reasons for the classification are hidden behind a single number, which makes it hard for analysts to further investigate on the most suitable solution. Thus, the fourth requirement is that an analyst must be able to examine reasons for commonalities and differences for decision making using the results.

To summarize this section Table 1 provides an overview of all four requirements.

3 RELATED WORK

After having outlined the requirements in the last section existing work is presented and assessed on the base of these requirements here. Because of the flow semantics being the central requirement and all other requirements being based on it the focus is on approaches that compare processes.

In literature a couple of equivalence notions for comparing processes can be found, e.g. bisimulation (Hidders, Dumas, van der Aalst, ter Hofstede and Verelst, 2005). Following (van Dongen, et al., 2008) those notions can be excluded from the explanations in this section for various reasons. The most important one is that they compute the equivalence of two processes, i.e. they answer the binary question if two processes are equivalent or not. As the fourth requirement states, it is important to make a statement about the degree of equivalence and to give hints for further investigation. This is clearly not satisfied by those notions. Thus, this section deals with approaches in the field of process similarity that measure the degree of equivalence.

The first approach outlined here is presented in (van der Aalst, et al., 2006). Here processes are compared on the base of finite sets of traces. These sets usually comprise a certain number of actual process executions, but can also be derived from simulations or user defined scenarios. To compare two processes using sets of traces two metrics are defined. Both are asymmetric and measure the similarity based on one of the processes. Besides counting the number of transition connections that appear in traces of the original as well as in the compared model the metrics also account for the transitions that are enabled within the traces.

In (Dijkman, et al., 2009) the Graph Edit Distance which indicates how many operations are needed to transform one process graph into another one is used to calculate the similarity. To calculate this metric, the mapping of nodes of two graphs is determined in four different ways each of them relying on activity labels.

In (van Dongen, et al., 2008) an approach is presented that relies on so called causal footprints. These footprints consist of all nodes of a process graph and two sets for each node. The first set comprises all nodes which can be executed before and the second set comprises those which can be executed

Table 2: Assessment of existing approaches.

Approach	Flow semantics	Context semantics	Levels of abstraction	Presentation of results
(Dijkman, Dumas and García-Bañuelos, 2009)	- Structure - Business process graphs	- Labels	- Not considered	- A symmetric metric
(Ehrig, Koschmider and Oberweis, 2007)	- Structure - Petri nets	- Labels	- Not considered	- A symmetric metric
(Kim and Suh, 2010)	- Structure - Special ontologies	- Context information	- Not considered	- A symmetric metric
(Lu, Sadiq and Governatori, 2009)	- Structure & behaviour - Process variant scheme	- Labels - Context information	- Not considered	- A symmetric metric
(van der Aalst, de Medeiros and Weijters, 2006)	- Behaviour - Petri nets	- Not considered	- Not considered	-Two asymmetric metrics
(van Dongen, Dijkman and Mendling, 2008)	- Behaviour - Causal footprint	- Labels	- Not considered	- A symmetric metric
(Zha, Wang, Wen, Wang and Sun, 2010)	- Behaviour - Transition adjacency relations	- Not considered	- Not considered	- A symmetric metric

after the current node. Transforming the footprints into vectors makes it possible to calculate the similarity as the cosine of the angle between these vectors. During the transformation matching activities that are based on labels and in case of EPCs also on information derived from events surrounding a function are employed. A similar approach is introduced in (Zha, et al., 2010). There a process is represented as a transition adjacency relation comprising pairs of activities of a process that can be executed directly one after the other. The similarity is defined as the ratio between the cardinality of the intersection of two transition adjacency relation sets and the cardinality of their union.

In the field of process variants an approach to determine whether a certain process variant meets a query which is a collection of features is introduced in (Lu, et al., 2009). These features can be classified as behavioural, structural or contextual features. For all classes algorithms to measure the similarity are proposed and the general similarity is then defined as the ratio of the similar features and the number of features in the query.

Some approaches rely on ontologies used to describe processes. In (Ehrig, et al., 2007) Petri net models are represented using an ontology. The similarity of two concepts from different models is the weighted sum of the syntactic, the linguistic (synonym and homonym relations) and the structural (taking related process concepts into account) similarity. The similarity of two processes is the sum of the similarities of concept pairs determined beforehand by mapping concepts from one model to those from the other one. In (Kim and Suh, 2010) five ontologies are defined to describe different views onto a process including organizational, domain, structural, resource and service aspects. Based thereon matchmaking is employed to classify the match between properties of two processes and to sum the corresponding similarity degrees.

The assessment of these approaches with regard to the requirements outlined beforehand is summarized in Table 2. As can be seen in this table, approaches exist which examine the behaviour of processes independently from a certain business process notation by relying on a representation that can be derived from such notations. While most of the approaches use labels to match activities or assume the match to be done beforehand, two approaches consider context information. However, none of the approaches fulfils both requirements. Regarding the demand for supporting different levels of abstraction it can be seen that none of the approaches addresses this requirement. Lastly, all approaches calculate a single degree of similarity but do not provide further information. The approach presented in (van der Aalst, et al., 2006) is slightly more advanced as it calculates the similarity for each of the processes being compared.

It is subject to future work and a relevant open issue to develop an approach which is designed with regard to all requirements. A first blueprint for such an approach is introduced in the next section.

4 PROPOSED APPROACH

The basic approach to the automated comparison of logistics services and the reference of each step within the approach to the requirements are presented in Figure 1.

The first step is the *transformation* of the process models into notation-independent models with activity annotations. Candidates for a meta-model are Petri nets, transition systems etc. On the base of such a meta-model different transformations have to be written in order to ensure that process models of various notations can be compared as demanded by the first requirement. Existing approaches, like (Raedts, Petkovic, Usenko, van der Werf, Groote and Somers, 2007) where BPMN models are transformed into Petri net models, can be reused.

A further important part of the first step is to annotate the models during the transformation in order to add information about the logistics functionality as necessary due to the second requirement. The annotation is based on the IOPE-model which is used within several service specification approaches like the Unified Service Description Language (Cardoso, Barros, May and Kylau, 2010). This model allows for describing activities with regard to their inputs and outputs as well as the preconditions and effects as representations of the state of the world that need to remain valid before and after activity execution. Furthermore the IOPE-model allows for applying the scheme introduced by (Hömberg, Hustadt, Jodin, Kochsiek, Nagelö and Riha, 2007). This scheme can be used to describe logistics functionality in terms of the information and goods that flow through an activity (input and output) as well as in terms of the changes made to the time and the space as well as the states of the information and goods (precondition and effect). To make these annotations interpretable for machines, different ontologies as explicit specifications of a conceptualization (Gruber, 1993) need to be employed. Regarding the third requirement the concepts of these ontologies must

reflect different levels of abstraction and must be related to each other, e.g. an ontology to describe states regarding space must enable a modeller to define regions and routes for transports. While the region is necessary to describe a company's abilities the route is needed to specify contract related requirements. This ontology should also connect the concepts route and region so that a machine is able to determine whether a company can handle routes in a certain region. The actual annotation can then be done in different ways. If the source model is already annotated in some way, these annotations also need to be transformed, e.g. if different ontologies are used, ontology matching algorithms (Euzenat and Shvaiko, 2007) will need to be integrated. In case of missing annotations they can be added manually while annotations on the base of the proposed ontologies can simply be copied.

Afterwards the second step is to *normalize* the models. This is done because of the third requirement. The goal of this step is to transform fine-grain process models into more coarse-grain ones in order to bring both models to the same level of abstraction. A simple rule could be to summarize activities that are arranged in sequence without any points of decision or interaction in between. One of the approaches supporting this step is presented in (Koliadis and Ghose, 2007) where effects of activity executions within processes are summarized supporting the summary of the overall preconditions and effects.

The third step prepares the process models for the actual comparison by *matching activities* of one process to the ones of the other process. The rationale here is to calculate the similarity of all activity pairs on the base of their annotations. Afterwards the optimal mapping is determined by an appropriate heuristic whereby optimal means that the sum of the similarity of all mapped pairs is as high as possible. This step is oriented towards the approach outlined in (Dijkman, et al., 2009) where the optimal mapping of activities is computed on the base of the syntactic and linguistic similarity of their labels. By relying on the annotations this step also accounts for the second requirement.

The fourth step is the *comparison of the models* on the base of their behaviour. As presented in the previous section there already exist approaches to compare processes from a behavioural perspective, like the one presented in (van der Aalst, et al., 2006). Nevertheless, extension is necessary to consider the fourth requirement, i.e. besides the computation of a degree of similarity the main reasons for the result must also be collected.

The last step is then to present the results to the customer using an appropriate *visualisation* that not only presents the degree of similarity but also the indicators that were collected in the previous step. As the services might rely on different notations it is important to present the results in a way that allows an analyst to investigate them although he or she is not familiar with the used process notations.

As the comparison of the original service to a set of other services is done in pairs and as there might be a lot of services that need to be compared the computation time can be high. In order to reduce it different strategies are possible. The first one is to estimate the similarity beforehand and only take those services into consideration which are believed to be similar to a certain degree, like it is done in (Yan, Dijkman and Grefen, 2010). A further option is to preselect services based on the purpose of the comparison, e.g. in the allocation and in the selection only services representing a company's capabilities are regarded. The last option mentioned here is to configure the features taken into account within the approach like it is proposed in (Lu, et al., 2009). Of course all these strategies can be commonly employed. It is also possible to proceed iteratively and refine the result set step by step.

5 CONCLUSION & NEXT STEPS

This paper motivated why it is necessary to support

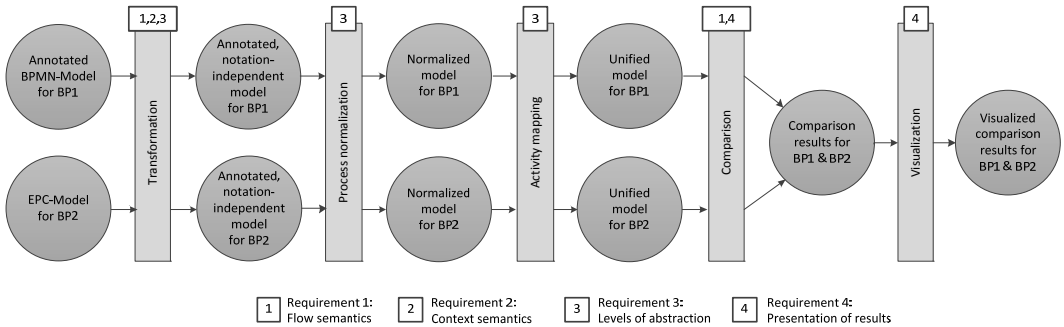


Figure 1: Basic approach for the service comparison.

the management of logistics networks with an automated approach for logistics service comparison. The central requirements determined in cooperation with a logistics network were introduced. Subsequently, the state of the art was evaluated with regard to these requirements. As a first step towards the automated comparison an approach which consists of five steps was proposed. These steps include some pre-processing in form of the transformation into a notation-independent representation as well as the normalization of the representation and the activity mapping to equalize the different levels of abstraction. Afterwards the comparison is done using the notation-independent, normalized and mapped process models. The final step is the visualization making the results interpretable for experts.

The first step to implement the basic approach is the selection of a notation-independent representation and of a basic comparison algorithm. On the one hand this represents the main functionality of the approach and on the other hand it constitutes a solid base for adding the other requirements. It is planned to evaluate the approach in each development step in order to ensure the benefit for logistics management. Hence, experts opinions and the results of the automated approach will be compared on the base of scenarios derived from logistics reference processes and from case studies conducted within the logistics network.

ACKNOWLEDGEMENTS

The work presented in this paper was partly funded by the German Federal Ministry of Education and Research under the project InterLogGrid (BMBF 01IG09010F) and by the European Regional Development Fund under the project LOGICAL (3CE396P2).

REFERENCES

Cardoso, J., Barros, A., May, N. and Kylau, U. (2010). Towards a Unified Service Description Language for the Internet of Services: Requirements and First Developments. *Proceedings of the 2010 IEEE International Conference on Services Computing*, 602-609.

Dijkman, R., Dumas, M. and García-Bañuelos, L. (2009). Graph Matching Algorithms for Business Process Model Similarity Search. *Proceedings of the 7th International Conference on Business Process Management*, 48-63.

Ehrig, M., Koschmider, A. and Oberweis, A. (2007). Measuring similarity between semantic business process models. *Proceedings of the fourth Asia-Pacific conference on Comceptual modelling - Volume 67*, 71-80.

Euzenat, J. and Shvaiko, P. (2007). *Ontology matching*. Heidelberg, Germany: Springer.

Gruber, T. R.(1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5,199-220.

Gudehus, T. and Kotzab, H. (2009). *Comprehensive Logistics*. Springer: Dordrecht.

Hidders, J., Dumas, M., van der Aalst, W.M.P., ter Hofstede, A.H.M. and Verelst, J. (2005). When are two workflows the same? *Proceedings of the 2005 Australasian symposium on Theory of computing - Volume 41*, 3-11.

Hömberg, K., Hustadt, J., Jodin, D., Kochsiek, J., Nagelö, L. and Riha, I. (2007). *Basisprozesse für die Modellierung in großen Netzen der Logistik*. Technical University Dortmund: Technical Report.

Kim, G. and Suh, Y. (2010). Ontology-based semantic matching for business process management. *ACM SIGMIS Database*, 41, 98 - 118.

Koliadis, G. and Ghose, A. (2007). Verifying Semantic Business Process Models in Inter-operation. *Proceedings of the IEEE International Conference on Services Computing SCC 2007*, 731-738.

Lu, R., Sadiq, S. and Governatori, G. (2009). On Managing Business Processes Variants. *Data & Knowledge Engineering*, 68, 642-664.

Pfohl, H.-C. (2004). *Logistikmanagement - Konzeption und Funktion*. Berlin: Springer.

Raedts, I., Petkovic, M., Usenko, Y.S., van der Werf, J.M.E.M., Groote, J.F. and Somers, L.J. (2007). Transformation of BPMN Models for Behaviour Analysis. *Proceedings of the 5th International Workshop on Modelling, Simulation, Verification and Validation of Enterprise Information Systems*, 126-137.

Sydow, J. and Duschek, S. (2011). *Management interorganisationaler Beziehungen : Netzwerke - Cluster - Allianzen*. Stuttgart: Kohlhammer.

van der Aalst, W.M.P., de Medeiros, A.K.A. and Weijters, A. J. M. M. (2006). Process Equivalence: Comparing Two Process Models Based on Observed Behavior. *Proceedings of the 4th International Conference on Business Process Management*, 129-144.

van Dongen, B., Dijkman, R. and Mendling, J. (2008). Measuring Similarity between Business Process Models. *Proceedings of the 20th international conference on Advanced Information Systems Engineering*, 450-464.

Weske, M. (2007). *Business Process Management*. Berlin, Germany: Springer.

Yan, Z., Dijkman, R. and Grefen, P. (2010). Fast business process similarity search with feature-based similarity estimation. *Proceedings of the 2010 international conference on On the move to meaningful internet systems - Volume Part I*, 60-77.

Zha, H., Wang, J., Wen, L., Wang, C. and Sun, J. (2010). A workflow net similarity measure based on transition adjacency relations. *Computers in Industry*,61, 463-471.

Application of an Artificial Immune System to Predict Electrical Energy Fraud and Theft

Mauricio Volkweis Astiazara and Dante Augusto Couto Barone
Programa de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
{*mvastiazara, barone*}*@inf.ufrgs.br*

Keywords: Artificial Immune Systems, Classifier, Pattern Recognition, Fraud Detection.

Abstract: This paper describes the application of an Artificial Immune System (AIS) to a real world problem: how to predict electricity fraud and theft. The field of Artificial Immune Systems is a recent branch of Computational Intelligence and has several possible applications, like pattern recognition, fault and anomaly detection, data analysis, agent-based systems and others. Although its potential, AIS still is not applied as much other techniques such as Artificial Neural Nets are. Various works compare AIS with other techniques using toy problems. But how much efficient is AIS when applied to a real world problem? How to model and adapt AIS to a specific domain problem? And how would be its efficiency compared to traditional algorithms? On the other hand, many companies perform activities that can be improved by Computational Intelligence, like predicting fraud. Electrical energy fraud and theft cause large financial loss to energy companies and indirectly to the whole society. This work applies AIS to predict electrical energy fraud and theft, analyzes efficiency and compares against other classifier methods. Data sample used to training and validation was provided by an electrical energy company. The results obtained showed that AIS has the best performance.

1 INTRODUCTION

The electrical energy distribution business faces a serious problem: some consumers try illegally to decrease their bills. This goal is achieved through fraud and theft. Fraud consists in handling energy company equipments aiming to decrease consumption registration. Theft is to make an unauthorized connection to the electrical energy system. In some countries, electrical energy fraud and theft cause annual losses of billions of U.S. dollars (Smith, 2004; ANEEL, 2008). Theft and fraud directly affect energy companies, but indirectly affect also honest consumers. The tampering of energy company equipments can result in poor quality energy supply to the neighbors of dishonest consumers. Also, energy taxes are increased having theft and fraud as the explanation.

To stop a fraud or theft from a dishonest consumer, energy company must perform an in locus inspection. As generally energy companies have few inspection teams, in locus inspection should be conducted in consumers more likely to be dishonest. Trying to hit dishonest consumers, energy companies use different strategies: receive anonymous tip offs about fraud and theft, make studies about consumers data and, just a few companies, apply datamining and pattern recognition techniques (Dick, 1995; Queiroga and Varejão, 2005; Monedero et al., 2006). In Brazil, CEEE-D is an energy company that still does not apply datamining and pattern recognition techniques to classify consumers as likely dishonest.

Artificial Immune System (AIS) is a relatively new branch of Computational Intelligence (CI) and is still in its infancy (Aisweb, 2009). Even though it has a wide potential application area, the algorithms and techniques of this field are not as widespread as those of Artificial Neural Nets and Genetic Algorithms. AIS can be used for pattern recognition. This work models and applies an AIS to classify CEEE-D consumers as likely dishonest aiming to analyze its efficiency. The results from AIS are compared against other well-known classification techniques.

The following sections introduce Artificial Immune Systems and discuss its application to a problem of an electricity company, including goals, data set, algorithm, experimental results, conclusions, and bibliographic references.

2 ARTIFICIAL IMMUNE SYSTEMS

The natural immune system has several properties that are interesting from a computational point of view (De Castro and Timmis, 2002), including pattern recognition, diversity, autonomy, anomaly detection, noise tolerance, resilience, learning, and memory, amongst others. Such features have inspired the development of new computational models and algorithms. AIS emerged in the 1990s as a new branch of CI (Dasgupta, 2006; Dasgupta and NIÑO, 2008). AIS are adaptive systems, inspired by theoretical immunology and observed immune functions, principles, and models, that can be applied to problem solving (De Castro and Timmis, 2002).

The scope of applications of AIS include, but are not restricted to: pattern recognition (Alexandrino et al., 2009), fault and anomaly detection (Kessentini et al., 2010), data analysis (data mining, classification etc.) (Nasir et al., 2009; Kodaz et al., 2009), agent-based systems (Hilaire et al., 2008), scheduling (Yu, 2008), machine learning, autonomous navigation and control (Zhang et al., 2009), search and optimization methods (Rodionov et al., 2011), artificial life, and security of information systems (Yu, 2011).

3 ELECTRICAL ENERGY FRAUD AND THEFT

Fraud and theft cause financial loss to energy companies in the whole World. Energy companies legally increase energy rates to compensate this kind of loss, referred to by the companies as Non-Technical Losses (NTL). In USA, estimated theft costs are between 0.5% and 3.5% of annual gross revenues (Smith, 2004). In developing countries, NTL are serious concerns for utility companies as they are about 10 to 40% of their total generation capacity (Depuru et al., 2011). In Brazil, annual NTL losses are over US$ 2 billion (ANEEL, 2008).

Basically, there are 3 situations that result in losses (Dick, 1995; Smith, 2004; Depuru et al., 2010):

1. A consumer who tampers with the meter so that it under-registers consumption; this is fraud. Figure 1 shows a tampered meter which is a kind of fraud.
2. A consumer who does not tamper with the meter, but instead creates another connection bypassing the meter. The consumer uses this illegal connection for some devices (usually devices that are large power consumers); this is theft.

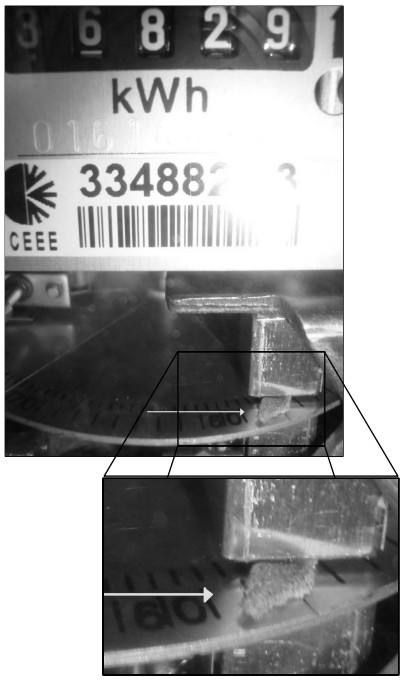


Figure 1: Picture of a tampered meter. There is a stone in the disc.

3. A non-registered consumer who makes an illegal connection. This is also theft, but this case is beyond the scope of this study, because the energy company does not have any information about these transgressors in its database.

To detect dishonest consumers, energy companies analyze consumer data and receive anonymous tip offs about dishonest consumers. Based on this information, they can determine whether a consumer is suspect. To confirm fraud or theft, an in locus inspection must be conducted. It is not, however, feasible for an energy company to inspect every consumer as the few inspection teams. Ideally in locus inspections should be conducted in consumers more likely to be dishonest, which can be ascertained through discovery of patterns in consumer data.

CEEE-D (Companhia Estadual de Distribuição de Energia Elétrica) is an energy company in southern Brazil. CEEE-D provides electricity to 72 cities and has 1,470,000 consumers (CEEE, 2011). CEEE-D is a partner in this study and provided a data set of inspected consumers to be used in the training and tests.

Table 1: Confusion matrix.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

4 GOALS AND METRICS

As described previously, the goal of this work is to analyze the effectiveness of the AIS paradigm applied to a real world problem. From this goal, three questions can be derived:

- **Question 1:** Can an AIS application learn to predict dishonest electricity consumers?
- **Question 2:** How efficient is AIS applied to this problem?
- **Question 3:** How efficient is AIS when compared to other methods?

To answer these questions, it is necessary to define metrics and how to interpret them. Thus, some concepts and metrics used in classification tasks are introduced. True Positive (TP) is the number of correctly labeled cases that belong to the positive class. In this work the positive class consists of dishonest consumers. True Negative (TN) is the number of correctly labeled cases that belong to the negative class (honest consumers). False Positive (FP) is the number of items incorrectly labeled as belonging to the positive class. Finally, False Negative (FN) is the number of items incorrectly labeled as belonging to the negative class. The four values (TP, TN, FP, and FN) constitute cells of the so-called Confusion Matrix. This matrix is created by crossing predicted values with real values. The confusion matrix is the basic output of any classifier validation as shown in Table 1.

The sum of TP and FN is the actual number of items in the positive class, whereas the sum of TN and FP is the actual number of items in the negative class. The sum of TP, TN, FP, and FN is the total number of items. From these basic values it is possible to calculate certain metrics, which are described below. Precision is defined as

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{1}$$

which means the probability of an item classified as belonging to the positive class actually to belong to the positive class.

Returning to the questions, Question 1 talks about learning. A classifier that does not learn is a random classifier. The precision of a random classifier is equal to the probability of the positive class, defined as

$$\text{Random Precision} = \frac{\text{number of positive class}}{\text{total number of items}}. \tag{2}$$

Thus, a classifier can learn if it has precision greater than that of a random classifier. Formally, this advantage of a classifier over a random classifier is called the Gain in Precision and is defined as

$$\text{Gain in Precision} = \frac{\text{Classifier Precision}}{\text{Random Precision}}. \tag{3}$$

A classifier with a Gain in Precision of 1 is no better than a random classifier. The larger the gain, the better is the classifier under consideration. Thus, the answer to Question 1 is "yes" if the Gain in Precision of the AIS is greater than 1, else it is "no".

In Question 2, it is necessary to interpret "efficient" in a business context. For the energy company, discovering dishonest consumers and stopping their fraud or theft is important because these consumers are sources of financial loss. At the same time, it is necessary an in locus inspection to confirm the fraud or theft and normally the company's inspection teams are very small. Inspecting an honest consumer is a waste of time and money. Ideally, in locus inspections should only be conducted in consumers more likely to be dishonest. Thus, Precision, which is defined in (1), is an important metric.

Another important metric is Recall (or Sensitivity), which is defined as

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{4}$$

and can be interpreted as the probability that an item of the positive class is correctly classified. Recall is an important metric too, because in a hypothetical scenario where all consumers classified as dishonest are inspected, 100% minus Recall of actual dishonest consumers remains with no inspection. This opinion that Precision and Recall are the most important metrics for this type of business is shared in (Queiroga and Varejão, 2005).

Since both metrics are important, it is necessary to use a metric that represents a balance of precision and recall. This metric is called the F-measure, and is the harmonic mean of precision and recall. The F-measure is defined as

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5}$$

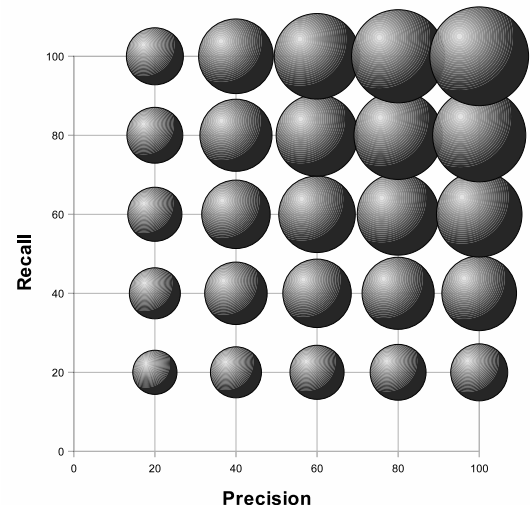


Figure 2: Representation of F-measure in bubbles.

Figure 2 illustrates F-measure as bubbles, precision as X axis and precision as Y axis. Bubbles grow as precision and recall grow. In this way, the F-measure helps to answer Question 2.

To answer Question 3, comparison of Precision, Recall, and the F-measure of an AIS with other classifier algorithms applied to the same data samples is made.

To calculate the defined metrics Leave One Out Cross Validation (Kohavi, 1995) was used. This kind of validation consists of removing one instance from the data sample to form part of the test data. The remaining instances are used as training data. The classifier is trained and tested. Then, the instances used to test are returned to the data sample and the next instance is used as test data, and so on until all instances have been used as test data. Leave One Out allows maximum utilization of all data, making the validation process less sensitive to data variations. However, this kind of validation has a high computational cost.

5 DATA SET

CEEE-D provided a data set with inspected consumers from a specific city that CEEE-D believes has a high rate of dishonest consumers. The original data set contains 4141 instances, but this includes redundant instances. After removal of redundant instances, 1249 remain. Of these instances, 440 belong to the positive class (dishonest consumers) and 854 belong to the negative class (honest consumers). In this scenario, 34% of consumers are dishonest. According to the energy company, real proportion of dishonest consumers ranges between 4 and 8%. Aiming to create a data set close to reality, the number of instances belonging to positive class was reduced to 54. It results in 5.95% proportion of dishonest consumers, a value close to the average of 4 and 8%. This proportion is shown in Figure 3.

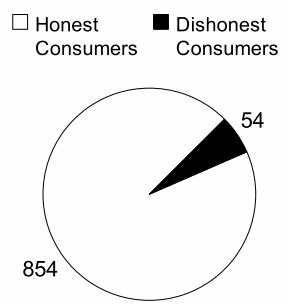


Figure 3: Proportion of dishonest consumers.

Each instance has 19 attributes involving categorical and numeric data types. These attributes were selected by an expert from the energy company based on his empirical knowledge. Attributes about energy consumption were normalized.

6 ALGORITHM

From all the algorithms based in Clonal Selection Theory (Burnet, 1959), for this analysis the Clonalg algorithm (De Castro and Timmis, 2002) was chosen because of its available documentation (Aisweb, 2009) and ease of implementation. Clonalg includes the following steps:

1. **Initialization:** create an initial random population of individuals (**P**).
2. **Antigenic Presentation:** for each antigen, do:
   (a) **Affinity Evaluation:** present it to the population **P** and determine its affinity with each element of the population **P**.
   (b) **Clonal Selection and Expansion:** select **n1** highest affinity elements of **P** and generate clones of these individuals proportionally to their affinity with the antigen: the higher the affinity, the higher the number of copies.
   (c) **Affinity Maturation:** mutate all these copies with a rate inversely proportional to their affinity: the higher the affinity, the smaller the mutation rate. Add these mutated individuals to the population **P** and reselect the best individual to be kept as the memory **m** of the antigen presented.

(d) **Metadynamics:** replace a number **n2** of individuals with low affinity by the randomly generated new ones.

3. **Cycle:** repeat Step 2 until a certain termination criterion is met.

In this work, antigens are data consumers and antibodies are data structures similar to data consumers. The antibody structure has 19 attributes, one for each consumer attribute. A hybrid representation of data was adopted keeping the original data types (categorical and real values) of each attribute.

To measure the affinity between antibodies and antigens a similarity measure based on distance is used. The smaller the distance, the higher is the similarity, and thus, the higher is the affinity. The distance between each antigen attribute and antibody attribute is calculated. The sum of all the distances is normalized by the total number of attributes, generating a value between 0 and 1. So, the value is inverted to become an affinity value. The affinity measure is defined as

$$\text{Affinity} = 1 - \frac{\sum_{i=1}^{L} D(Ag_i, Ab_i)}{L}, \qquad (6)$$

where

- Ag is the array of attributes of the antigen;
- Ab is the array of attributes of the antibody;
- L is the length of the array of attributes, in this case, 19;
- D is a function to measure distance between attributes, which depends on the data type of the attribute. The resulting value is in the range 0 and 1.

Function D depends on the data type of the attribute. For categorical attributes the Hamming distance is applied, where the result is 0 if the two values are equal, else 1. For real value attributes the following formula was applied:

$$D = \frac{|Ag_i - Ab_i|}{Max - Min}, \qquad (7)$$

where

- Ag is the array of attributes of the antigen;
- Ab is the array of attributes of the antibody;
- Max is the maximum that attribute i can assume; and
- Min is the minimum that attribute i can assume.

The size of the initial population **P** was set as 4% of the sample size. For parameters **n1** and **n2** a value of 20% of the population **P** was used. The termination criterion is that the individuals retained as memory cells reach an affinity of 0.8 or more.

Table 2: Summarized data.

Metric	Mean	Standard Deviation	Confidence Interval (level 95%)
Precision	13.97%	0.0066	[13.84%, 14.10%]
Recall	71.93%	0.0340	[71.26%, 72.59%]
F-measure	23.39%	0.0109	[23.18%, 23.61%]

This algorithm is used to generate two classifiers: one for honest consumers and the other for dishonest consumers. The classification of a new consumer is made by submitting it to both classifiers, and considering the one with the higher affinity as the label. A prototype for this AIS model was implemented in the Java programming language.

It was used the Waikato Environment for Knowledge Analysis (WEKA) software (Hall et al., 2009) to provide the other algorithms for comparison. WEKA is a workbench of machine learning that includes several algorithms. The version used was 3.6.3. All algorithms were used with default parameter values provided by WEKA except for KNN that was tested using three values for K (1, 3, and 10).

7 EXPERIMENTAL RESULTS

Precision, recall, and F-measure of 100 Leave One Out Cross Validation was calculated. Measured values have a normal distribution, so arithmetic mean was used as average. Standard deviation and confidence intervals were calculated too as shown in Table 2.

To answer the questions listed earlier, the mean of the Precision, Recall and F-measure was used. In Question 1, “Can an AIS learn to predict dishonest electricity consumers?”, it is necessary to calculate the random precision and gain in precision defined in (2) and (3), respectively:

$$\text{Random Precision} = \frac{54}{908} = 0.0595 = 5.95\%,$$

$$\text{Gain in Precision} = \frac{0.1397}{0.0595} = 2.3478.$$

The Gain in Precision of the AIS, 2.3478, is greater than 1, so the answer to Question 1 is yes,

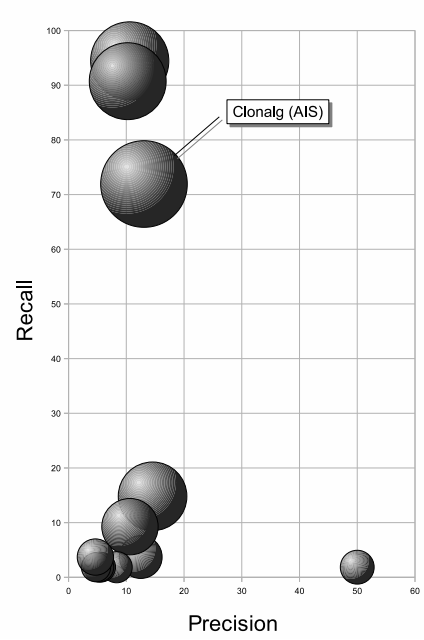


Figure 4: Comparison of results. F-measure in bubbles.

the AIS can learn to predict dishonest electricity consumers.

Question 2 is "How efficient is AIS applied to this problem?" and the answer is a consequence of the Precision, Recall and F-measure metrics; in this case, Precision = 13.07%, Recall = 71.93%, and F-measure = 23.39%.

To answer Question 3, "How efficient is AIS when compared to other methods?", Leave One Out Cross Validation was performed running several algorithms from WEKA. Table 3 shows the results ordered by F-measure. Only the top 13 algorithms are shown.

Figure 4 visually summarizes the resulting data. In terms of precision, the AIS, represented by the Clonalg algorithm, is 3rd. From obtained data can be observed that, in general, algorithms with a high precision have a low recall. In results ordered by recall, Clonalg is in third place too. Considering the balance of precision and recall through the F-measure, Clonalg is in first place. It can be concluded that, from an F-measure perspective, Clonalg achieves good performance.

Table 3: Comparison of results ordered by F-measure.

Algorithm	Precision	Recall	F-measure	#
Clonalg (AIS)	13.07%	71.93%	23.39%	1
Naive Bayes	10.60%	94.44%	19.07%	2
Voting feature intervals	10.25%	90.74%	18.42%	3
KNN (K=1)	14.55%	14.81%	14.68%	4
RandomTree	10.64%	9.26%	9.90%	5
RandomForest	12.50%	3.70%	5.71%	6
NNGE	4.65%	3.70%	4.12%	7
Fast decision tree learner	50.00%	1.85%	3.57%	8
K*	8.33%	1.85%	3.03%	9
FT Tree	5.56%	1.85%	2.78%	10
Artificial Neural Net	5.56%	1.85%	2.78%	10
KNN (K=3)	5.26%	1.85%	2.74%	11
PART decision list	4.76%	1.85%	2.67%	12

8 CONCLUSIONS

This work described how an AIS algorithm called Clonalg was applied to a real world problem: predicting electricity consumers who are sources of non-technical losses (fraud or theft) based on patterns in the data available in the energy company database. A model of antibody and antigen was shown. A distance measure was used as affinity measure. In this work was used metrics to analyze the algorithms that make sense in the electrical energy business context, different from other works that use accuracy as single metric in a simplistic way as (Brun et al., 2009; Depuru et al., 2011). Results show that the modeled AIS can learn the concept of dishonest consumers and has the best efficiency in terms of the F-measure. Thus, the AIS should be considered a potential candidate to solve pattern recognition tasks. Furthermore, it seems that there is a relation between precision and recall, where high precision is associated with low recall.

REFERENCES

Aisweb (2009). Basic immune inspired algorithms. <http://www.artificial-immune-systems.org/algorithms.shtml>. The Online Home of Artificial Immune Systems.

Alexandrino, J., Cavalcanti, G., and Filho, E. (2009). Hybrid intelligent system clonart applied to face recognition. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 102 –107.

ANEEL (2008). Nota técnica 342/2008-sre/aneel. <http://www.aneel.gov.br/cedoc/nren2008338_342.pdf>. Agência Nacional de

Energia Elétrica, Superintendência de Regulação Econômica.

Brun, A., Pinto, J., Pinto, A., Sauer, L., and Colman, E. (2009). Fraud Detection in Electric Energy Using Differential Evolution. In *Intelligent System Applications to Power Systems, 2009. ISAP '09. 15th International Conference on*, pages 1 –5.

Burnet, M. (1959). *The clonal selection theory of acquired immunity*. The Abraham Flexner Lectures. Cambridge University Press.

CEEE (2011). A ceee distribuição. <http://www.ceee.com.br/pportal/ceee/ Component/Controller.aspx?CC=1755>. Companhia Estadual de Distribuição de Energia Elétrica.

Dasgupta, D. (2006). Advances in artificial immune systems. *Computational Intelligence Magazine, IEEE*, 1(4):40 –49.

Dasgupta, D. and NIÑO, L. F. (2008). *Immunological Computation: Theory and Applications*. CRC Press, Florida, US.

De Castro, L. N. and Timmis, J. (2002). *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer, London, UK.

Depuru, S., Wang, L., and Devabhaktuni, V. (2011). Support vector machine based data classification for detection of electricity theft. In *Power Systems Conference and Exposition (PSCE), 2011 IEEE/PES*, pages 1 –8.

Depuru, S., Wang, L., Devabhaktuni, V., and Gudi, N. (2010). Measures and setbacks for controlling electricity theft. In *North American Power Symposium (NAPS), 2010*, pages 1 –8.

Dick, A. (1995). Theft of electricity-how uk electricity companies detect and deter. In *Security and Detection, 1995., European Convention on*, pages 90 –95.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18.

Hilaire, V., Koukam, A., and Rodriguez, S. (2008). An adaptative agent architecture for holonic multi-agent systems. *ACM Trans. Auton. Adapt. Syst.*, 3:2:1–2:24.

Kessentini, M., Vaucher, S., and Sahraoui, H. (2010). Deviance from perfection is a better criterion than closeness to evil when identifying risky code. In *Proceedings of the IEEE/ACM international conference on Automated software engineering*, ASE '10, pages 113–122, New York, NY, USA. ACM.

Kodaz, H., Babaoglu, I., and Iscan, H. (2009). Thyroid disease diagnosis using artificial immune recognition system (airs). In *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, ICIS '09, pages 756–761, New York, NY, USA. ACM.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Monedero, i., Biscarri, F., León, C., Biscarri, J., and Millán, R. (2006). MIDAS: Detection of Non-technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques. In Gavrilova, M., Gervasi, O., Kumar, V., Tan, C., Taniar, D., Laganà, A., Mun, Y., and Choo, H., editors, *Computational Science and Its Applications - ICCSA 2006*, volume 3984 of *Lecture Notes in Computer Science*, pages 725–734. Springer Berlin / Heidelberg.

Nasir, A. N. M., Selamat, A., and Selamat, H. (2009). An artificial immune system for recommending relevant information through political weblog. In *Proceedings of the 11th International Conference on Information Integration and Web-based Applications and Services*, iiWAS '09, pages 420–424, New York, NY, USA. ACM.

Queiroga, R. and Varejão, F. (2005). AI and GIS together on energy fraud detection. In *North American Transmission and Distribution Conference and Expo*.

Rodionov, A. S., Choo, H., and Nechunaeva, K. A. (2011). Framework for biologically inspired graph optimization. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, ICUIMC '11, pages 11:1–11:4, New York, NY, USA. ACM.

Smith, T. B. (2004). Electricity theft: a comparative analysis. *Energy Policy*, 32(18):2067 – 2076.

Yu, H. (2008). Optimizing task schedules using an artificial immune system approach. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, GECCO '08, pages 151–158, New York, NY, USA. ACM.

Yu, Y. (2011). Anomaly intrusion detection based upon an artificial immunity model. In *Proceedings of the 49th Annual Southeast Regional Conference*, ACM-SE '11, pages 121–125, New York, NY, USA. ACM.

Zhang, X.-f., Liu, J., and Ding, Y.-s. (2009). An immune coevolutionary algorithm based approach for optimization control of gas turbine. In *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, GEC '09, pages 751–756, New York, NY, USA. ACM.

A Distributed Agency Methodology applied to Complex Social Systems
Towards a Multi-dimensional Model of the Religious Affiliation Preference

Manuel Castañón–Puga[1], Carelia Gaxiola–Pacheco[1], Dora–Luz Flores[2], Ramiro Jaimes–Martínez[3] and Juan Ramón Castro[1]

[1]*Facultad de Ciencias Químicas e Ingeniería, Universidad Autónoma de Baja California, Tijuana, Mexico*

[2]*Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California, Ensenada, Mexico*

[3]*Instituto de Investigaciones Históricas, Universidad Autónoma de Baja California, Tijuana, Mexico*

{*puga, cgaxiola, dflores, rjaimes, jrcastro*}*@uabc.edu.mx*

Keywords: Fuzzy Agents, Data Mining, Social Complexity, Distributed Agency, Religion Affiliation Preference.

Abstract: The purpose of the paper is to describe a work-in-progress in the application of a distributed agency and neuro-fuzzy system methodology to a multi-dimensional model on a complex social system. This work introduces a study case focuses on decision-making modelling system on religious affiliation preferences. We use a type-2 neuro-fuzzy approach to configure cognitive rules into agent in order to built a multi-agent model for social simulation.

1 INTRODUCTION

The social systems are complex entities that represent a whole that cannot be understood by looking at its parts independently. Another characteristic is the interdependence of the parts conforming the whole: a change to one of the components in the system may potentially affect all others (Yolles, 2006).

The main goal of this part of our research is to develop a computational model of change in religious affiliation preference that incorporates available mathematical and computational theories that have not been appropriately considered in models of complex social phenomena.

Even though applications of Multi-Agent Systems (MAS) have been developed for the social sciences, MAS have been widely considered in some areas such as Artificial Intelligence (AI) (Gilbert, 2007).

1.1 Distributed Agency

The modelling of a realistic social system cannot be achieved by resorting to only one particular type of architecture or methodology. The methodology of Distributed Agency (DA) represents a research avenue with promising generalized attributes, with potentially ground-breaking applications in engineering and in the social sciences—areas in which it minimizes the natural distances between physical and sociological systems.

The methodology of DA represents a general theory of collective behaviour and structure formation, which intends to redefine agency and reflect it in multiple layers of information and interaction, as opposed to the traditional approach in which agency is only reflected in individual, atomized and isolated agents (Suarez and Castanon-Puga, 2010).

1.1.1 Modelling Complex Social System using Neuro-fuzzy and Distributed Agencies

To build the model of change of religious affiliation will follow the distributed agency methodological steps (Márquez et al., 2011):

1. Determining the levels of agency and their implicit relationships.
2. Data mining.
3. Generating a rule-set.
4. Multi-Agent Modelling (Implementation on a agent based simulation tool).
5. Validating the model.
6. A simulation and optimization experiment.
7. Analysing the outputs.

Although the methodology covers the entire lifecycle of a research process, on this paper we are describing the data mining and generating rule set steps. We are focused on the neuro-fuzzy approach in order to set up a rule set into agents.

1.1.2 Data Mining and Neuro-fuzzy System

An Interval Type-2 Fuzzy Neural Network (IT2FNN) are used for automatically generate the necessary rules. The phase of data mining using Interval Type-2 Fuzzy Logic Systems (IT2FLS) (Castillo et al., 2010; Castro et al., 2010) becomes complicated, as there are enough rules to determine which variables one should take into account. The search method of back-propagation and hybrid learning (BP+RLS) is more efficient in other methods, such as genetic algorithms (Rantala and Koivisto, 2002; Castro et al., 2008).

Since the IT2FNN method seems to produce more accurate models with fewer rules is widely used as a numerical method to minimize an objective function in a multidimensional space, find the approximate global optimal solution to a problem with N variables, which minimize the function varies smoothly (Stefanescu, 2007).

With the application of this grouping algorithm we obtain the rules, the agent receives input data from its environment and choose an action in an autonomous and flexible way to fulfill its function (Peng et al., 2008).

1.2 Religious Affiliation

When literature talks about of religious change, usually refers to the attachment or religion affiliation (Ortiz, 2006). Although some authors have argued that the concept can not be limited to this dimension, membership is one of the most important variables to study the religious phenomena (Fortuny, 1999).

The religious field is conformed by several dynamics systems. For example, we can identify some organizational entities: institutional, socio-demographic groups and individual.

Within these multiple dimensions interrelated complex processes are occurring, such changes of allegiance, change in commitment and participation, socialization and subjectivity of standards (through doctrines, values, practices), reformulation and affirming traditions. These multiple dimensions shape the religious field, and generically is known as religious change.

1.2.1 Religious Affiliation in México

In México, religious affiliation has undergone major changes since the 1950's until today. Based on population censuses, the growth rates of the evangelical population has been higher than the total Catholic population[1](Jaimes-Martínez, 2007). Baja California has one of the percentages of highest evangelical population of Northern states[2].

2 CASE OF STUDY

Tijuana is a border city located in north-western of México. Belongs to the state of Baja California, and is one of the fastest growing city in the country due to high migration rates. The population is mainly composed by migrants from southern of the country. They came to the border to further job opportunities, or looking to migrate to the United States, staying in the city long time.

2.1 Tijuana's Multi-cultural and Religious Complexity

Tijuana is an example of social and religious change. Its boundary condition has been one factor that has become a city in full development and expansion, not only by the strength of the Southern California economy, but by the early efforts to boost manufacturing by the federal government.

These factors, combined with growing internal and international migration, have transformed a town of Tijuana from a town with 12,181 inhabitants in 1930 to one with 1.2 million in 2000[3] (Alegría and Ordónez, 2002). It was so from NAFTA, Tijuana was consolidated as a major call centres maquiladora industry, with an evident increase in employment and production, but not productivity or living standards and welfare (Arias, 2008).

According to some authors, the economic balance, social and cultural development of these global processes, regional and local has had complex effects on Tijuana's society, where stands the reconfiguration of identities and new forms of social and cultural reproduction.

In this sense, the religious sphere in Tijuana has a great religious diversification as a result of different waves of migration that have shaped their society.

[1]The evangelical population has experienced rates of 8.90, between 1970 and 1980, while the total population was 3.16. Although at present growth rate 2.46 points, it is still higher than that of the population is Catholic and total population.

[2]Baja California has 7.90% and evangelical population, surpassed only by one of the first entities to which the Protestant missionaries arrived in the nineteenth century, Tamaulipas, to 8.65%. Nationally, the percentage of evangelicals is 5.20%.

[3]Tito Alegría and Gerardo Ordóñez consider the growth process of Tijuana covers from 1930 to 2000, thanks mainly to the economic expansion of Southern California.

Therefore, religious affiliation is also an indicator to study these processes of reconfiguration and realignment[4] (Jaimes-Martínez, 2007).

2.2 Preference for Religious Affiliation in Tijuana

The city has a great diversity of faiths and religious traditions. Although more numerous the Christian (Catholic, Protestant, evangelical non-biblical), there are Buddhists, Muslims, Jews and a variety of groups and beliefs generically known as New Age[5].

Considering this, we can say that every group or social stratum in Tijuana has a wide range of choice, or affinity, in the religious field in the city. Each of them is not only an expression of traditions, customs and religious practices of different groups have brought to Tijuana from their places of origin, but the dynamic formulation of these beliefs in the new environment.

3 MODELLING TIJUANA CITY

The principal difference between MAS and our proposed approach is that in our methodology the space includes transformations performed by a higher level of agency.

This upper-level agent is composed of lower-level subcomponents the may enjoy agency in their own right. It is the responsibility of this intermediate agent to present its subcomponents with individual phase-spaces that are tailored to induce the desired behaviour from the lower-level agents which inhabit it, when it chooses according to its own objective function.

Therefore, for our proposed work-in-progress case study, if we consider a municipality as an agent, this upper-level agent is composed by subcomponents, which in our case study of the city of Tijuana, Mexico, will be represented by a location set and Basic Geo-Statistic Area (AGEB in Spanish) set that compose this city. Locations is the terminology used to describe wide geographic areas of the city that are composed of AGEBs. AGEB is the terminology used to describe small geographic areas of the city that are composed of blocks.

[4]Between 1990 and 2000 Tijuana just recorded a growth rate of 8.94 evangelical population, while at the national level was 2.46.

[5]Syncretic movements oriental religions such as Buddhism, introducing ideas of self-motivation, personal growth, alternative medicine, psychology, etc.

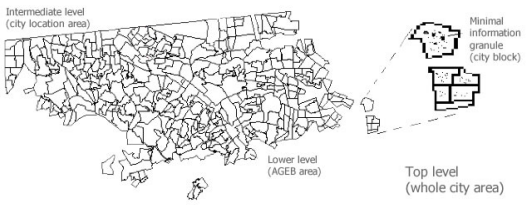


Figure 1: Levels of agents represented on the social system.

3.1 Levels of Agency

In this example we use three levels of agency: the upper-level agent is represented by the religion on the whole city, the intermediate level agents are represented by the locations and the lower level agents are the AGEB.

Using a recent census of the reunion sites distribution of the different religious organizations operating in the city, we know the exact places where they carry out activities of proselytizing. This information gives us hints of the influence of the presence of organizations in its environment and its impact on socio-demographic variables.

We are looking for relationships between demographic and economic factors (subtracted from AGEB) and distribution of meeting places of religious organizations. We believe that factors such as poverty, marginalization and other characteristics related to socio-demographic issues influencing the decision-making system of individuals in a complex and distributed way. Similarly, religious organizations act as agents who are influenced by other agencies distributed.

3.2 Data Sets

In the particular case of the city of Tijuana, the data set used came from the Instituto Nacional de Estadistica y Geografia (INEGI), the Mexican governmental organization in charge of gathering data at a federal level including aspects that are geographical, socio-demographic and economical.

The data set of the city of Tijuana is divided into 363 areas, known as AGEB [6](INEGI, 2010).

The data sets for this case study were originally compiled in an information system that is intrinsically geographical. These systems helped in the generation, classification and formatting of the required data—a fact which facilitates the edition of the different thematic layers of information, in which one can

[6]The urban AGEB encompass a part or the totality of a community with a population of 2500 inhabitants or more in sets that generally are distributed in 25 to 50 blocks.

quantify the spatial structure to visualize and interpret the areas and different spatial patterns in Tijuana.

For this paper, we going to use de following variables to exemplify the proposed approach using information from 2010 population census in México (INEGI, 2010).

- P15YMAS = Population over 15 years old.
- P15YMSE = Population over 15 years old without education.
- GRAPROES = Education.
- PEA = Working population.
- PEINAC = Non working population.
- PCATOLICA = Catholic population.
- PNCATOLICA = Non catholic population.

3.3 Neuro-fuzzy Inference System

Using the neuro-fuzzy system for the automatic generation of rules, this phase of the data extraction from the data may become complicated, as the process needs to appropriately establish the number of sufficient norms and variables that the study needs to take into account.

Using this grouping algorithm, we obtain the appropriate rule-set assigned to each agent representing an location or a AGEB of it, the agent receives inputs from its geographical environment and in turn much choose an action in an autonomous and flexible fashion (Gilbert, 2007).

The purpose of this structure without central control is to garner agents that are created with the least amount of exogenous rules and to observe the behavior of the global system through the interactions of its existing interactions, such that the system, by itself, generates an intelligent behavior that is not necessarily planned in advance or defined within the agents themselves; in other words, creating a system with truly emergent behavior.

From the 2010 census information, we create a Type-2 Fuzzy Inference System as how we could represent different agencies as a decision-making system into agents.

3.3.1 City Level Type-2 Fuzzy Inference Systems

The figure 2 shows a type-2 fuzzy inference system for Tijuana city. It depicts a set of input-output variables and a rule set. Output variables are catholic and non-catholic as a response of the system. We could use the difference between both values to make decisions into an agent as a preference decision-making system.

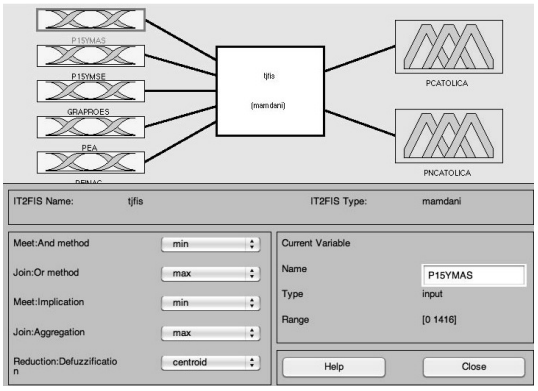


Figure 2: Fuzzy inference system for Tijuana City.

The figure 3 shows member function example for GRAPROES input variable. Type-2 fuzzy inference system allows us to introduce uncertainty into de system, that could be used to represent more dynamic changes into de Inference System because could be influenced by many real time ways.

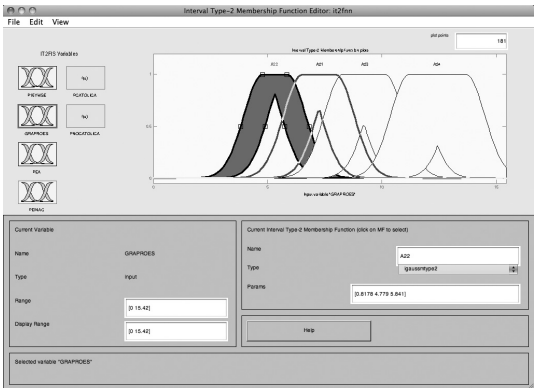


Figure 3: Fuzzy inference system input and output members function configuration for Tijuana City.

The Figure 4 depicts the resolution example of the rules by the fuzzy inference system. Different quantitative input values could be introduced and the system resolve creating different responses. Depending of the combination of inputs, we can expect different responses of the system. An agent will use this inference system as a decision-making system to show different behaviours depending of the situation.

The Figure 5 represents the response of the system to catholic preference, and the Figure 6 for non-catholic preference. We can see that there are response differences, so we can use it to make decisions.

Distributed agents do not necessarily define agents in lower-levels of description, but rather consider all levels of agency that are interconnected in a type of

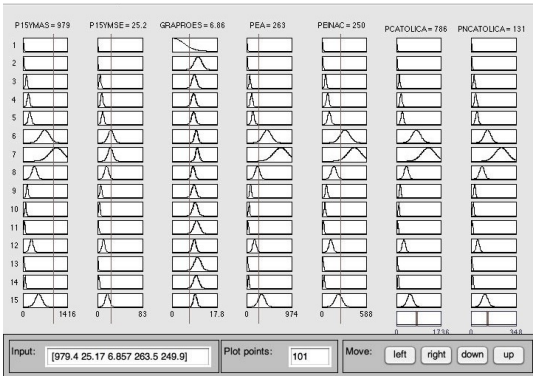


Figure 4: Fuzzy inference system rule set evaluation for Tijuana City.

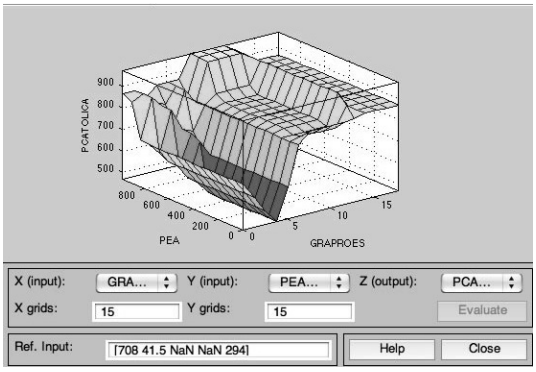


Figure 5: PEA vs. GRAPROES type-reduced surface view for Tijuana City PCATOLICA output.

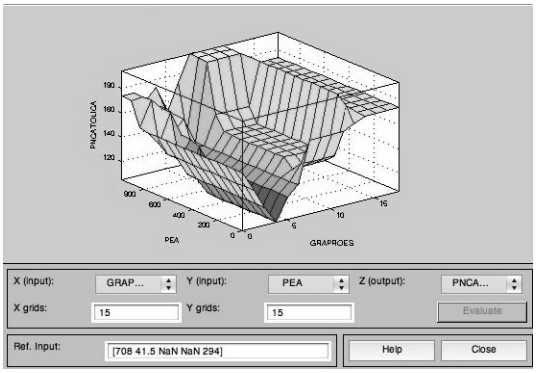


Figure 6: PEA vs. GRAPROES type-reduced surface view for Tijuana City PNCATOLICA output.

organism that spreads throughout the system.

3.3.2 Location Level Type-2 Fuzzy Inference Systems

On location layer, we can build fuzzy inference systems for agents that represents locations. Figure 7 and Figure 8 depicts the FIS response for different locations into the city. As we can see, there are differences between AGEB agents. At this level, we could be representing locations agents into a city context.

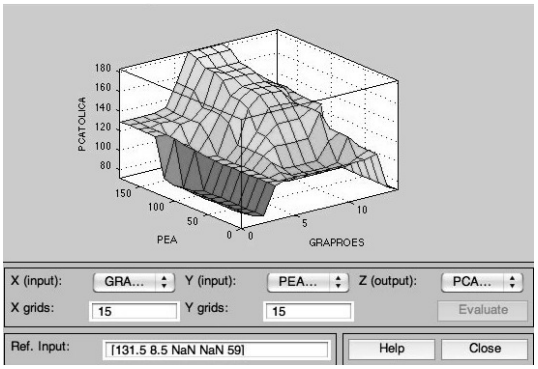


Figure 7: PEA vs. GRAPROES type-reduced surface view for location 187 PCATOLICA output.

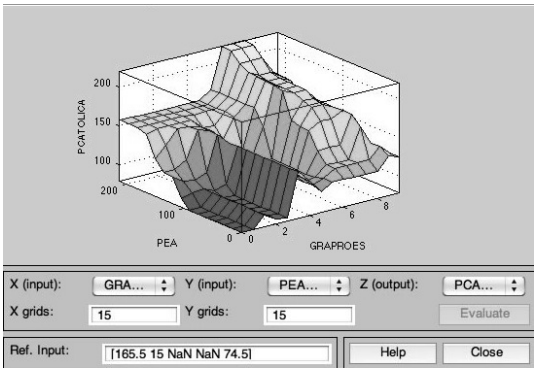


Figure 8: PEA vs. GRAPROES type-reduced surface view for location 283 PCATOLICA output.

3.3.3 AGEB Level Type-2 Fuzzy Inference Systems

On AGEB layer, we can build fuzzy inference systems for agents that represents locations. Figure 9 and Figure 10 depicts the FIS response for different AGEB into the same location. As we can see, there are differences between AGEB agents. At this level, we could be representing AGEB agents into a location context.

4 CONCLUSIONS

We use a distributed agency and neural-fuzzy system approach to develop a computational model of the decision-making system of agents in order to build a multi-agent system. We represent different levels of agency with different cognitive agents. Each agent in

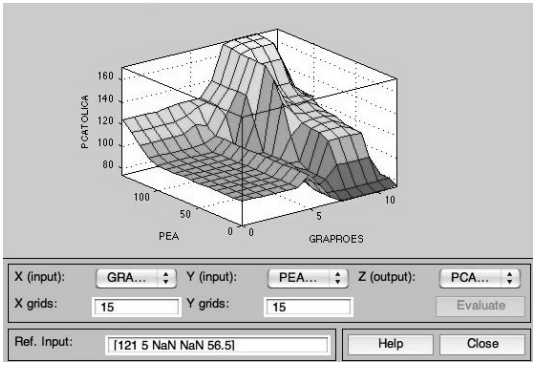


Figure 9: PEA vs. GRAPROES type-reduced surface view for AGEB 32 PCATOLICA output.

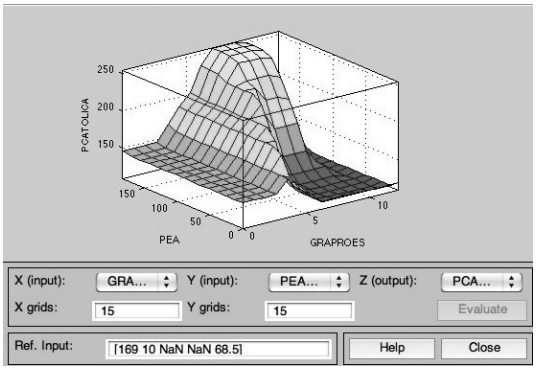


Figure 10: PEA vs. GRAPROES type-reduced surface view for AGEB 51 PCATOLICA output.

the system are a fuzzy cognitive agent that can choose religion options based on preferences.

We use the case study of the city of Tijuana, as it has an updated census of the distribution of meeting places of religious organizations in the city and their respective socio-demographic information.

The religious affiliation can be modelled with distributed agency. Establishing different layers of interaction between agents and analysing their influence on decision-making system of agents in each level, we can represent the complexity of the phenomenon of individual preference to a religious affiliation.

ACKNOWLEDGEMENTS

We would like to thank to Universidad Autónoma de Baja California for the economic support granted for this research.

REFERENCES

Alegría, T. and Ordónez, G. (2002). Regularización de la tenencia de la tierra y consolidación urbana en tijuana, b.c. Research report, El Colegio de la Frontera Norte, México.

Arias, A. L. (2008). Cambio regional del empleo y productividad manufacturera en méxico, el caso de la frontera y las grandes ciudades,1970-2004. *Frontera Norte*, 20(40):79–103.

Castillo, O., Melin, P., and Castro, J. R. (2010). Computational intelligence software for interval type–2 fuzzy logic. *Journal Computer Applications in Engineering Education*.

Castro, J. R., Castillo, O., Melin, P., Mendoza, O., and Rodríguez-Díaz, A. (2010). An interval type–2 fuzzy neural network for chaotic time series prediction with cross–validation and akaike test. *Soft Computing for Intelligent Control and Mobile Robotics*, 318:269–285.

Castro, J. R., Castillo, O., Melin, P., and Rodríguez-Díaz, A. (2008). A hybrid learning algorithm for a class of interval type–2 fuzzy neural networks. *Journal of Information Sciences*, 179(13):2175–2193.

Fortuny, P. (1999). *Creyentes y creencias en Guadalajara*. CIESAS, México.

Gilbert, N. (2007). *Computational social science: Agent-based social simulation*, pages 115–134. Bardwell, Oxford.

INEGI (2010). Censo de población y vivienda 2010. instituto nacional de estadística geografía e informática.

Jaimes-Martínez, R. (2007). *La paradoja neopentecostal. Una expresión del cambio religioso fronterizo en Tijuana, Baja California*. PhD thesis, El Colegio de la Frontera Norte, México.

Márquez, B. Y., Castañón Puga, M., Castro, J. R., and Suarez, E. D. (2011). Methodology for the Modeling of Complex Social System Using neuro-Fuzzy and Distributed Agencies. *Journal of Selected Areas in Software Engineering (JSSE)*, March:1–8.

Ortiz, O. O. (2006). Cambio religioso en la frontera norte. aportes al estudio de la migración y las relaciones transfronterizas como factores de cambio. *Frontera Norte*, 18(35):111–134.

Peng, Y., Kou, G., Shi, Y., and Chen, Z. (2008). A descriptive framework for the field of data mining and knowledge discovery. *International Journal of Information Technology and Decision Making*, 7:639–682.

Rantala, J. and Koivisto, H. (2002). Optimised subtractive clustering for neuro-fuzzy models.

Stefanescu, S. (2007). Applying nelder mead's optimization algorithm for multiple global minima. *Romanian Journal of Economic Forecasting.*, pages 97–103.

Suarez, E. D. and Castanon-Puga, M. (2010). Distributed agency, a simulation language for describing social phenomena. In *IV Edition of Epistemological Perspectives on Simulation*, Hamburg, Germany. The European Social Simulation Association.

Yolles, M. (2006). Organizations as complex systems, an introduction to knowledge cybernetics.

A Hybrid Solver for Maximizing the Profit of an Energy Company

Łukasz Domagała, Tomasz Wojdyła, Wojciech Legierski and Michał Swiderski
Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology,
16 Akademicka Str., 44-100 Gliwice, Poland
{*ldomagala, tpwojdyla, msiderski, wlegierski*}*@polsl.pl*

Keywords: Optimization, Profit, Energy, Industry.

Abstract: An energy company manages power stations, handles sales and purchases of electrical energy, CO2 emission permits and other goods. The goal of such a company is to ensure energy safety of its clients and maximize the profit. The problem is complex because of its structure and size therefore efficient automated approaches for solving it are in demand. We have generalized the problem definition to account for any structure of the power stations, market data and time scope. The definition describes a non-linear combinatorial optimization problem. We have tested a number of approaches including: constraint/ logic/ dynamic/ integer/ linear programming, local search and their hybrids using prototypes with input data from a real life process. We present a hybrid solver to produce an acceptable, near optimal solution which satisfies the requirements of an industrial application. Our research is a road-sign for development of similar software for the energy industry.

1 INTRODUCTION

The aim of the solver is to return a schedule of production, sales and purchases of all the goods for a given *time horizon* that satisfies all the *hard constraints* and optimizes the *objective function*. Hard constraints are these that cannot be violated in the solution. Satisfying them guarantees that technological and marketing requirements are met and ensures the energy safety of the company's clients. The objective function is the company's profit gained for a given time horizon. Maximizing this profit is the main goal. The optimization is performed for problem instances which consist of: technological capabilities and parameters of the power stations, market plans and estimates supplied by marketing and financial experts, trade contracts, initial states for those goods that can be accumulated.

2 PROBLEM DEFINITION

We have obtained instances of the profit maximization problem during work on a commercial project. Basing on the problem instances, we have built a generalized problem definition that accounts for any structure of the power stations, market data and time scope. To the best knowledge of the authors a definition of such a problem has never been published before.

2.1 Timing and Notation

The production, trade and constraint setup is performed for discrete time *periods*. $h \in [1,H]$ is the shortest period called, for convenience, an *hour*. Each value of h is categorized as *peak* or *off-peak*. Furthermore, consecutive values of h are grouped into periods $(H_{m-1},H_m] = M_m$ indexed by $m \in [1,M]$ where $H_0 = 0, H_{m-1} < H_m, H_M = H$, which are, for convenience, called *months*. The period of $[1,H] = y$ is, for convenience, called a *year*. An energy company handles the following *goods*: energy $\{en\}$, CO2 emission permits $\{ep_p, p \in [1,P]\}$, financial benefits $\{be_b, b \in [1,B]\}$. An energy company handles the following *objects*: power stations, production units, sales, purchases. Some object-period pairs have corresponding control variables $v(object, period)$. Numerical (unless otherwise stated) attributes $\alpha(type, object/good, period)$ are attached to objects/goods, where *period* denotes the period to which it applies. The term "volume" is used to describe the quantity of some good.

2.2 Energy Production and Trade

An *energy company* (*ec*) is divided into *power stations* $\{ps_s : s \in [1,S]\}$. Each power station ps_s is divided into *production units* $\{pu_i^s : i \in [1,U^s]\}$ which produce (electrical) *energy*. *ec* manages the volumes

$vps_{s,h}$ of energy supplied by ps_s in period h. $vps_{s,h} = fpv_{s,h}(\sum_{i=1}^{U^s} v(pu_i^s,h))$ where $fpv_{s,h}$ is a *piecewise linear function*. $fpv_{s,h}$ includes a number of components:

- The actions of a regulatory body, which may intervene with the production plans and are meant to regulate the energy market. These actions are predicted by experts as a piecewise linear function $frp_{s,h}(\sum_{i=1}^{U^s} v(pu_i^s,h))$.
- The error factor $\alpha(zo,ps_s,h)$ associated with the imperfections of the energy distribution network
- The sales of energy $\{v(se_i^s,h) : i \in [1,Se^s]\}$ managed privately by the ps_s

The energy trade consists of purchases $\{ze_i : i \in [1,Ze]\}$ and sales $\{se_i : i \in [1,Se]\}$ managed by the *ec*. The trade is further divided into *contracts* and *plans*. The contract is a signed trade agreement, whereas the trade plan is based on the expert predictions. This distinction, however, is reflected in the *variable domains* and is transparent for the solver.

2.3 CO2 Emission Permits

The CO2 emissions of the ps_s have to be covered by *permits* of $\{ep_p, p \in [1,P]\}$ types. Permits may be traded, may be granted by the government, may be *consumed*, are limited by constraints depending on the ep_p. The permit trade is managed for each ps_s separately. $\forall p \in [1,P], s \in [1,S]$ the defined sales and purchases are respectively $\{sp_i^{p,s} : i \in [1,Sp^{p,s}]\}$, $\{zp_i^{p,s} : i \in [1,Zp^{p,s}]\}$. The permits are consumed to cover emissions which are relative to energy production. Consumption volume is $v(pu_i^s,h) \cdot \alpha(co2,pu_i^s,h)$, where $\alpha(co2,pu_i^s,h)$, is the emission ratio.

2.4 Financial Benefits

Financial benefits of type $be_b, b \in [1,B]$ can be produced, purchased, sold or consumed. They are produced relatively to energy production, be_b production volume is $v(pu_i^s,h) \cdot \alpha(be_b,pu_i^s,h)$. The ratios of production are dependent on the efficiency of, and resources used by pu_i^s. Examples of financial benefit types are type for energy produced from renewable resources, type for high efficiency coal powered production, type for natural gas powered production, etc. Sales and purchases respectively, are denoted by $\{sb_i^b : i \in [1,Sb^b], b \in [1,B]\}$, $\{zb_i^b : i \in [1,Zb^b], b \in [1,B]\}$. The be_b is consumed, relatively to the volume of sold energy, to gain access to certain energy markets, be_b consumption volume is $v(se_i,h) \cdot \alpha(be_b,se_i,h)$.

2.5 Control Variables

The solution to the optimization problem is defined by values assigned to the control variables. The complete set CV of control variables (see **APPENDIX A**) is (A.1)-(A.7) where (A.1) are energy production variables, (A.2)-(A.3) are energy trade variables, (A.4)-(A.5) are financial benefits trade variables, (A.6)-(A.7) are CO2 emission permit trade variables.

2.6 Constraints

The formulas representing linear constraints are: variable unary (variable domains) (A.8), production gradient (A.9), a technological constraint of each pu_i^s, energy balance (A.10), financial benefit monthly balance (A.12), financial benefits yearly balance (A.11), CO2 permits nonnegativity (A.13), CO2 permits yearly balance (A.14). (A.12)-(A.14) are called *long period constraints*. The attribute name *ist* explicitly denotes the initial state, whenever indexation refers to element 0 e.g. $v(pu_i^s,0)$ it signifies an implicit initial state.

The formulas representing nonlinear constraints are: minimal duration for which a pu_i^s has to work after *startup* (A.15), technological constraint for the level of production (A.16), minimal number of pu_i^s turned on in ps_s (A.17), startup schedule of a pu_i^s (A.18), relation between the production levels of pu_i^s and energy provided to the *ec* by ps_s (A.19). Attribute $\alpha(startup,pu_i^s,y)$ is an ordered set of values, # is a set cardinal number.

2.7 Elements of the Objective Function

The objective function $\omega(CV)$ represents the total *ec* profit. Each control variable has a corresponding profit ratio represented by the $profit$ attribute. For production and purchases the profit ratio is negative and for sales the profit ratio is positive. The profit of control variables is linear and represented by (A.20). (A.21) and (A.22) are nonlinear elements of the cost function. The first represents the startup cost of pu_i^s i.e. the cost of turning on a disabled production unit. The latter corresponds to costs related to components of the $fpv_{s,h}$ (Section 2.2).

3 TESTED APPROACHES

The approaches have been tested on problem *instances* denoted by $inst(ec,H)$ where *ec* is the definition of objects and types, H is the time horizon. In particular $inst(ec^r,H^r)$ denotes the industrial

real life problem instance. $inst(ec^r, H^r)$ consists of 289'200 control variables, 411'838 linear non-unary constraints and 490'560 nonlinear constraints for the $H^r = 8'760$. Under the confidentiality agreement we are not allowed to disclose the structure of ec^r. For the time $cr(st)$ used to perform the optimization, the condition $cr(st) > 10min$ is called the $timeout$. $\neg timeout$ is a requirement for the solver.

We have used the following criteria to compare models: $cr(nlin)$ are nonlinearities included in the model, $cr(gopt)$ is guarantee of optimality provided, $cr(long)$ are long period constraints included in the model, $cr(teff) = cr(st)/H$ time efficiency. The approaches have been tested on personal computers with 2 x 2.2Ghz processors, 3GB of RAM and address space.

3.1 Constraint (Logic) Programming

Constraint programming (CP) (Apt, 2009; Marriott and Stuckey, 1998) is a programming paradigm with the central notion of a constraint. A constraint states relations between variable domains (allowed combinations of domain values). CP is a form of declarative programming where the program in the form of *constraint statements* is a description of the problem, rather than a path to the solution (unlike in the case of procedural programming). CP makes a distinction between 3 components required to obtain a solution: the declarative constraint statement, *constraint propagation* and *search*. Historically CP has grown out of, and has been embedded in *logic programming* and often uses the LP based backtracking search, it is however possible to embed constraint programming in a procedural language.

We have tested models written under two CP systems (Schulte, 2010; Szymanek and Kuchciński, 2010) and a CLP system (Cisco Systems, 2010) with the result of obtaining optimal solutions for problem instances with $H \in [1, \lfloor H^r/2000 \rfloor]$. The $cr(teff) \in [5s, 25s]$ depending on the CP system, constraint propagation methods and search methods. However for problem instances with $H \geq \lfloor H^r/100 \rfloor$ we have been unable to produce a solution without a timeout.

The advantages of the CP approach are that the complete problem model can be taken into account $cr(long) = cr(nlin) = true$, solution with the guarantee of optimality can be obtained $cr(gopt) = true$. The major disadvantage is that the models are impractical for problem instances near the real life problem size

3.2 Dynamic Programming

Dynamic programming (DP) (Bellman, 1953; Cormen et al., 2001) is a mathematical and computer algorithmic scheme for solving optimization problems. The method builds the final solution by expanding initial conditions step by step into more complex cases with $cr(teff)$ determined by the number of states and the complexity of the step.

By means of DP it is possible to obtain a complete solution with $cr(gopt) = true$ in polynomial time complexity by (i) calculating an initial solution for $h = 1$ and pu_1^1 and (ii) expanding the solution upon all of the *pu* and *year*. Unfortunately, the complexity of our problem generates a state-space too large for any direct approach. However, a combined approach of DP and local search can be derived if only we are able to separate a simple subproblems for DP.

We have used a DP approach to determine, for a fixed hour h, the costs of volumes $vol_h^s \in \sum_i v(pu_i^s, h)$ for all $ps_s \in \{ps_s\}$ and generate the maximized total profit pr_h for hour h. In the basic version we have determined pr_h by managing costs of production of each pu_i^s. These costs included the joint costs of maintaining the pu_i^s on $vol_{i,h}^s$ along with a few other parameters (e.g. profit and cost associated with be_b production). In the first step of the algorithm we have generated a lookup table of production volumes vol_h^s and minimal costs of achieving them for each ps_s. Assume that, for a fixed h and ps_s, $v(pu_i^s, h) \in [mn_i, mx_i]$. We denote the cost of pu_i^s in h with volume $x \in [mn_i, mx_i]$ as $ep(i,x)$. Let $m[x][z]$, where $x \in [1, U_s]$ and $z \in [\sum_{i=1}^{U_s} mn_i, \sum_{i=1}^{U_s} mx_i]$ be an optimal cost of the production units $(pu_1^s - pu_x^s)$ generating a total production equal to z. The values of $m[x][z]$ are equal to: (i) $cp(1,z)$ for $z \in [mn_1, mx_1]$, (ii) $\min(m[x-1][z-i] + cp(x,i))$ for $i \in [mn_i, mx_i]$ or (iii) ∞ in all remaining cases. This relation gave us, for each h and ps_s, an optimal configuration of $vol_{i,h}^s$ necessary to produce vol_h^s. In the second step we merged all obtained lookup tables (using DP) along with purchases ze_i modeled as an artificial *pu* with its own lookup table. In the following step we generated (using DP once again) a lookup table for hour h and for sales se_i, containing the optimal methods of selling of particular *en* volumes. In the last step of the algorithm we have compared the two obtained lookup tables and greedily chosen the best vol_h as the production volume of the *ec*.

The basic DP algorithm described above was fast ($cr(teff) = 5 - 8ms$) comparing to CP but did not include $cr(long)$ leading to a very complicated local search with poor final result for the year. Thus, we have refined the solution by introducing partial op-

timization of goods to the DP algorithm. The best configuration we have obtained by introducing only one hard constraint - CO2 emission permit, leading to $cr(teff) = 50ms$.

The main advantages of this approach are: (i) fast and always optimal solution for a fixed h and (ii) inclusion of $cr(nlin)$ without any additional time efforts. Unfortunately, the overall $cr(long)$ management is poor leading to very complex local search that need to be applied as a superior algorithm.

3.3 Linear and Integer Programming

Linear programming (LP) (Dantzig, 1963) is a natural approach to solving linear problems but does not apply directly to nonlinear problems. This drawback can be partially overcome by including relaxations of nonlinearities in the problem model, it however comes at a cost of loosing accuracy and efficiency. We have tested the LP model with a number of relaxations. First of all it has to be noted that the relaxation of (A.19) is required for the model to be of any use because it relates the production to sales and is required in the balance constraints (A.10). For any piecewise linear function fp, defined as (A.23) a convex hull relaxation (Hooker, 2006) has been used (A.24). In the initial solution of the LP model with (A.24), (A.10) are not satisfied because they contain biases caused by the relaxation. To eliminate the bias the LP model is solved again with additional constraints (A.25) that linearize the $fpv_{s,h}$ around the *steady states* obtained from the first solution. To tighten the relaxation, models of further nonlinearities can be included: startup relaxation (A.26) of (A.16), startup cost relaxation (A.27), (A.28) of (A.21) (where *vstc* is the total cost of startups) , minimum pu_i^s enabled relaxation (A.29) of (A.17). Mixed integer/linear programming (MILP) approach, can also be used to tighten the relaxation by introducing integrality constraints $integer(ar_i)$, $integer(on_{s,i,h})$.

We have tested the LP and MILP approaches, for $inst(ec^r, H^r)$ using the COIN-OR (IBM, 2010) CLP and CBC solvers. The LP model with (A.24) and (A.25) had produced the solution with $cr(teff) = 10ms$, LP model with (A.24), (A.25) and (A.29) had $cr(teff) = 40ms$, the model with additional relaxations (A.27) and (A.28) produced a solution after $cr(st) = 1853s$ with the *timeout* and $cr(teff) = 211ms$. A model with integrality constraints did not produce a solution within 1h.

To summarize, the LP relaxation approach has a number of advantages: taking into account all the long period constraints $cr(long) = true$, producing an optimal solution for the defined model $cr(gopt) = true$, a relatively short $cr(teff) = 10ms$ (LP model with (A.24) even for $inst(ec^r, H^r)$. The main disadvantage is that nonlinearities are accounted for in a very limited scope. Therefore, an additional optimization step has to be used to fully satisfy the nonlinearities.

3.4 Local Search

Local search (LS) (Aarts and Lenstra, 1997) is a meta-heuristic for solving computationally hard optimization problems. For the case of *ec* optimization the LS algorithm is organized as follows: it assumes an initial solution (step 1), repairs the violated nonlinear constraints (step 2) by applying a set of *repair_heuristics*, looks for a better solution (step 3) using a set of *improvement_heuristics* (step 3a). All value assignments $CV = vals(heur, CV)$ (step 2a,3a), are performed by choosing a neighborhood pu_j^k and applying new values to CV such that all linear constraints are satisfied and that only the variables $\{v(pu_j^k, h) : h \in [1, H]\}$ in neighborhood pu_j^k are changed from all production variables $\{v(pu_i^s, h)\}$. LS performs backtracking (*backtrack*) in the cases when constraints cannot be repaired to undo wrong heuristic choices. A solution is returned in the form of variable values $values(solution)$ and the corresponding $profit$. The details of particular heuristics shall not be discussed because they are dependent on a particular class of problem instances and are subject to customization.

1. $CV = values(init)$, $profit = -\infty$
2. for all $heur \in repair\_heuristics$:
   (a) for all $cstr \in violated(heur, CV)$: $CV = values(heur(cstr), CV)$
   (b) if $\neg \#violated(heur, CV) = 0$ then *backtrack*
3. for all $heur \in improvement\_heuristics$:
   (a) $CV = values(heur, CV)$
   (b) if $\#violated(all, CV) = 0 \wedge \omega(CV) > profit$ then $values(solution) = CV, profit = \omega(CV)$ else *backtrack*

The advantages of local search is that it can be applied to large and nonlinear problems. The disadvantage is that any LS algorithm is custom tailored for a specific problem definition and class of problem instances. It is typical that LS is highly dependent on the quality of the initial solution $CV = values(init)$ (how many constraints, and of which classes, are violated, are those difficult for LS to handle satisfied etc.)

Table 1: Execution times of the hybrid solver (CP+LP+LS) for several problem instances derived from $inst(ec^r, H^r)$.

Nr	Problem instance ($inst(ec, H)$)	CP+LP	CP+LP+LS
1	Original problem instance for a year ($inst(ec^r, H^r)$)	46.8s	161.8s
2	3 quarters of the year ($inst(ec^{r1}, H^r \cdot 3/4)$)	31.3s	111.1s
3	2 quarters of the year ($inst(ec^{r2}, H^r \cdot 2/4)$)	19.8s	74.2s
4	1 quarters of the year ($inst(ec^{r3}, H^r \cdot 1/4)$)	9.4s	41.3s
5	Higher costs of energy production ($inst(ec^{r4}, H^r)$)	12.7s	121.9s
6	Unconstrained sales and purchases ($inst(ec^{r5}, H^r)$)	183.9s	295.5s
7	Unconstrained sales and purchases, higher energy production cost ($inst(ec^{r6}, H^r)$)	448.2s	614.2s
8	Unconstrained sales and purchases, heavier constrained production gradient ($inst(ec^{r7}, H^r)$)	372.4s	490.7s

3.5 Hybrid Approach

Due to the fact that a single method approach is insufficient to time-efficiently account for all the components of the *ec* profit optimization problem a hybrid approach (solver) has been developed comprising of CP+LP+LS. The hybrid solver takes advantage of the interactions and key strengths of the methods included.

The function of CP has been reduced to performing bound propagation (Dechter and van Beek, 1997) on the constraints (A.8)-(A.14)(A.19) in order to determine *correction variables*[1] for the equality constraints in the LP model and determine infeasible constraints at the outset without time consuming proof of infeasibility by the LS.

The LP model with (A.24) and (A.25) has been used as described in Section 3.3 to obtain an initial solution for the LS. The LP model is preferred to DP because it accounts for all long period constraints (A.12)-(A.14), these constraints are "difficult" to satisfy by the LS (which leads to LS timeouts) if they are not accounted for in the initial solution. Secondly the LP model outperforms DP with respect to $cr(teff)$. The final optimization stage is LS which uses the solution provided by the the LP model as the initial variable assignment. LS is meant to account for the nonlinearities, find a feasible solution (repair phase) and optimize it (improvement phase).

The hybrid approach is supreme because it is the only one that accounts for the complete problem model and produces an acceptable solution for $inst(ec^r, H^r)$ without a timeout. The hybrid solver was tested for several problem instances, derived by modification from the industrial real life problem instance, and the results are presented in Table 1. The results show that the solver execution time is linearly dependent on the size of the problem (time horizon modifications in row 1 to 4) and is strongly dependent on the problem structure i.e. the types of constraints (tightened or relaxed) and profit ratios for control variables. This vulnerability to modification of problem structure is a feature of LP.

[1] correction variables are used to compensate rounding errors performed by the LP solver

4 CONCLUSIONS AND RELATED WORK

Hybridization is an approach to optimization problems that often yields shorter computation times than single method approaches. The relative advantage of hybrid solvers can range up to a few orders of magnitude. This means that for applications with timeouts imposed on optimization it may be the only applicable solution. Furthermore, real life problems often contain heterogeneous constraints and hybridization allows to choose techniques best suited for particular classes of constraints and let them exchange information. A survey of computational results performed by John N.Hooker in (Hooker, 2006) lists some applications of hybrid solvers and their advantages over single method approaches to problem such as: "Scheduling with earliness and tardiness cost" (Beck and Refalo, 2003) solved 5 times more problem instances, "Polyprophylene batch scheduling" (Timpe, 2002) solved previously insoluble problem in 10min, "Lesson timetabling" (Focacci et al., 1999) 2 to 50 times faster, "Min-cost multiple machine scheduling" (Jain and Grossmann, 2001) 20 to 2000 times faster, "Product configuration" (Thorsteinsson and Ottosson, 2002) 30 to 40 times faster.

For the *ec* optimization problem, on the basis of our experiments (Section 3 and Table 2), we believe that a hybrid approach (Section 3.5 and Table 2 entry 7) is the only one that can achieve performance sufficient to meet the requirements of an industrial appli-

Table 2: Comparison of experimental results for different approaches to $inst(ec^r, H^r)$.

Nr	Method	Comment	$cr(gopt)$	$cr(nlin)$	$cr(long)$	$\neg timeout$	$cr(teff)$
1	C(L)P	3 approaches tested	*true*	*true*	*true*	*false*	*na*
2	DP	basic algorithm	*true*	*true*	*false*	*true*	8*ms*
3	DP	with CO2 constraints	*true*	*true*	*false**	*true*	50*ms*
4	LP	with (A.24)(A.25)	*true*	*false**	*true*	*true*	10*ms*
5	LP	with (A.24)(A.25)(A.29)	*true*	*false**	*true*	*true*	40*ms*
6	DP + LS		*false*	*true*	*true*	*false*	*na*
7	CP+LP+LS		*false*	*true*	*true*	*true*	58*ms*

$*$ partially taken into account.

cation.

We have presented a generalized definition of the *ec* optimization problem. We have also laid out the overall structure and details of a hybrid solver developed for the generalized problem, indicating a promising area of research and leaving room for customization (especially in the LS area). We have also discussed approaches which have been discarded at an early stage of development because of their low performance, indicating areas of development which are unlikely to yield satisfactory results. To the best knowledge of the authors no other solution to the *ec* optimization problem has been reported.

The presented hybrid CP+LP+LS approach is generalizable because many complex optimization problems (other than *ec* optimization) can also be decomposed into linear and nonlinear components and then subjected to CP bound consistency, solved by LP to produce an initial solution that is next extended to a feasible solution with respect to nonlinearities by LS and improved by LS, in the same manner as described in this paper.

REFERENCES

Aarts, E. and Lenstra, J., editors (1997). *Local Search in Combinatorial Optimization*. Discrete Mathematics and Optimization. Wiley, Chichester, UK.

Apt, K. (2009). *Principles of Constraint Programming*. Cambridge University Press, New York, NY, USA, 1st edition.

Beck, J. C. and Refalo, P. (2003). A hybrid approach to scheduling with earliness and tardiness costs. *Annals of Operations Research*, 118:49–71. 10.1023/A:1021849405707.

Bellman, R. (1953). An introduction to the theory of dynamin programming. *The RAND Corporation*, Report R-245.

Cisco Systems (2010). ECLiPSe constraint logic programming system. http://www.eclipse-clp.org/.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to algorithms (2nd edition)*. MIT Press and McGraw-Hill.

Dantzig, G. B. (1963). *Linear programming and extensions*. Princeton Univ. Press, Princeton, NJ.

Dechter, R. and van Beek, P. (1997). Local and global relational consistency. *Theor. Comput. Sci.*, 173:283–308.

Focacci, F., Lodi, A., and Milano, M. (1999). Cost-based domain filtering. In *Proceedings of the 5th International Conference on Principles and Practice of Constraint Programming*, CP '99, pages 189–203, London, UK. Springer-Verlag.

Hooker, J. N. (2006). *Operations research methods in constraint programming*, pages 525–568. Handbook of Constraint Programming. Elsevier, Amsterdam.

IBM (2010). Coin-or computational infrastructure for operations research. http://www.coin-or.org/ .

Jain, V. and Grossmann, I. E. (2001). Algorithms for hybrid milp/cp models for a class of optimization problems. *INFORMS J. on Computing*, 13:258–276.

Marriott, K. and Stuckey, P. J. (1998). *Introduction to Constraint Logic Programming*. MIT Press, Cambridge, MA, USA.

Schulte, C. (2010). Gecode constraint programming system. http://www.gecode.org/.

Szymanek, R. and Kuchciński, K. (2010). Jacop constraint programming system. http://jacop.osolpro.com/.

Thorsteinsson, E. S. and Ottosson, G. (2002). Linear relaxations and reduced-cost based propagation of continuous variable subscripts. *Annals of Operations Research*, 115:15–29. 10.1023/A:1021136801775.

Timpe, C. (2002). Solving planning and scheduling problems with combined integer and constraint programming. *OR Spectrum*, 24:431–448. 10.1007/s00291-002-0107-1.

APPENDIX A

The appendix contains equations of the problem model and its relaxation. **APPENDIX A** is available at http://sun.aei.polsl.pl/∼twojdyla/opt/ec_opt_appA.pdf.

Design and Implementation of a Service-based Scheduling Component for Complex Manufacturing Systems

Lars Mönch
Department of Mathematics and Computer Science, University of Hagen, Universitätsstraße 1, 58097, Hagen, Germany
Lars.Moench@fernuni-hagen.de

Keywords: Service-oriented Computing, Scheduling, MES, Complex Manufacturing Systems.

Abstract: Scheduling is highly desirable in complex manufacturing systems. However, there is still a mismatch between academic scheduling research, the scheduling solutions offered by software vendors, and the requirements of real-world scheduling applications. In this paper, we describe the design and the development of a scheduling component prototype that is based on web services. It exploits the idea of a hierarchical decomposition of the overall scheduling problem allowing the integration of different problem-specific scheduling algorithms for sub-problems. We discuss how appropriate services can be identified and implemented and how the resulting scheduling component can be used to extend the functionality offered by manufacturing execution systems (MESs).

1 INTRODUCTION

This research is motivated by scheduling problems that are found in complex manufacturing systems, as for example, semiconductor wafer fabrication facilities (wafer fabs). Complex manufacturing systems are characterized by a diverse product mix, many machines, a large number of jobs, sequence-dependent setup times, and batching. Here, batching means that several jobs can be processed at the same time on the same machine. Scheduling is challenging in such an environment. However, it is highly desirable because of the increasing automation pressure. In contrast to previous papers (cf. Mönch and Driessel, 2005), we are not interested in proposing a new scheduling technique. Instead of this, we deal with the question of how to design scheduling components from a functional and also from a software technical point of view. It turns out that the data available in Enterprise Resource Planning (ERP) systems and Advanced Planning Systems (APS) are not fine-grained enough to allow for making detailed scheduling decisions. Furthermore, their actuality with respect to time is not appropriate. MESs are a natural carrier of scheduling functionality (McClellan, 1997; Meyer *et al.*, 2009). However, the scheduling capabilities of packaged MESs are often not appropriate because they are too generic (cf. Pfund *et al.*, 2006 for the results of a survey of the acceptance of packaged scheduling solutions in the semiconductor manufacturing industry). In this paper, we research the problem of designing a scheduling component that can be used by an MES. In a certain sense, this paper extends previous work carried out for the ERP domain (cf. Mönch and Zimmermann, 2009). The design of the component is derived taking an appropriate hierarchical decomposition of the overall scheduling problem into account. After identifying appropriate services, we implement a prototype based on web services. Such questions are rarely discussed in the literature so far (cf. Framinan and Ruiz, 2010 for a recent survey of the architecture of scheduling systems).

The paper is organized as follows. In the next section, we describe the problem and discuss related literature. We present the hierarchical decomposition of the overall scheduling problem in Section 3. Furthermore, we describe how appropriate services can be indentified. The implementation of the prototype is described in Section 4. We discuss also some limitations of the proposed approach and future research needs.

2 PROBLEM DESCRIPTION

2.1 Problem

In current MESs for complex manufacturing systems,

dispatching functionality often is offered instead of the more sophisticated scheduling functionality. Optimization kernels are typically based on genetic algorithms or on generic commercial constraint programming and mixed integer programming libraries (cf. Fordyce *et al.*, 2008). Scheduling systems are mainly developed only for parts of the manufacturing system, for example for the leading bottleneck machine group. There is only little interaction of software vendors and academic research (cf. Kellogg and Walczak, 2008). There are several reasons for this situation.

1. Scheduling of production jobs is often combined with transportation scheduling, process planning, staff scheduling, and finally advanced process control decision-making.
2. Global scheduling systems fail because humans on the shop floor are not involved in the resultant scheduling decisions. It seems that often the notion of reasonable automation is not taken into account.
3. Scheduling algorithms depend to a large extent on the objectives and constraints taken into account. That means that slight changes in the objectives and the constraints might lead to totally different algorithms. Dispatching rules strongly support this behavior by its inherent myopic view. Generic scheduling solutions have some limitations with respect to dealing with this situation. They are often not well accepted by people on the shop floor.
4. The data for scheduling decisions is located in different operative application systems. MES- and Material Control System (MCS)-related data are very important in this context.
5. Supplying appropriate data to the scheduling algorithms is important. However, scheduling algorithms that take many details into account require at the same time data that is fine-grained.

Analyzing these insights results in the conclusion that striving for a more detailed modeling is inapplicable because a more detailed consideration of constraints leads to sophisticated algorithms and also to a more difficult data supply. Therefore, it seems important to focus on the quintessence of scheduling, i.e., considering the finite capacity of the manufacturing system is more important. However, it is possible to deal with the finite capacity on a more aggregated level.

In this paper, we address the question of how a scheduling component has to be organized to take this vision into account. Therefore, the design of a service-based scheduling component is discussed that is based on an appropriate distributed hierarchical decomposition of the overall scheduling problem. The resultant design is validated by a prototypical implementation of such a component.

2.2 Related Work

A web service-based specification and implementation of ERP components is described, for example, by Brehm and Marx Gomez (2007) and Tarantilis *et al.* (2008). However, a direct application of these ideas to scheduling is not possible because of the different level of detail. A conceptual proposal for an MES based on web services that can be used in small- and medium-size enterprises in Mexico is discussed by Gaxiola *et al.* (2003). But again, no specific details of possible scheduling functionality are included in this paper. This is also true for the recent survey paper by Framinan and Ruiz (2010), where the usage of web services is only mentioned, but not further elaborated.

A service-oriented integration framework for complex manufacturing systems is presented by Qiu *et al.* (2007). A certain portion of a traditional MES, especially with respect to feedback from the shop floor, is implemented within the framework, but again, scheduling functionality is not covered.

There is some work done for the identification of services (cf. Winkler and Buhl, 2007 for the financial domain and Mönch and Zimmermann, 2009 for ERP-related services). However, to the best of our knowledge, there is no work available that addresses this question for scheduling services in complex manufacturing systems. A distributed scheduling system for complex job shops based on software agents is presented in Mönch *et al.* (2006). But in contrast to web services, software agents and multi-agent-systems are still not widely accepted in applications on the shop floor. In this paper, we will show that some of the scheduling functionality described by Mönch *et al.* (2006) can be provided using principles of service-oriented computing.

3 IDENTIFICATION OF APPROPRIATE SERVICES

3.1 Distributed Hierarchical Decomposition

As discussed in Subsection 2.1, we are interested in making manufacturing system-wide scheduling decisions without increasing the level of detail for modeling. This goal is mainly reached by an appropriate hierarchical approach.

We start by describing the assumed physical decomposition of the base system of the manufacturing system. The routing of the jobs, the dynamic entities in the system, takes place between different groups of parallel machines. Parallel machines offer the same functionality in a manufacturing system. A single group of parallel machines is called a work center. The work centers that are located in the same area of the manufacturing system are aggregated into work areas. On the highest level, we find the entire manufacturing system, i.e., the different work areas form the base system. In order to solve the overall scheduling problem, often hierarchical decomposition approaches have been applied.

In this paper, we consider a two-layer hierarchical approach. The resulting scheduling scenario is shown in Figure 1 as a Unified Modeling Language (UML) sequence diagram and will be explained in more detail below.

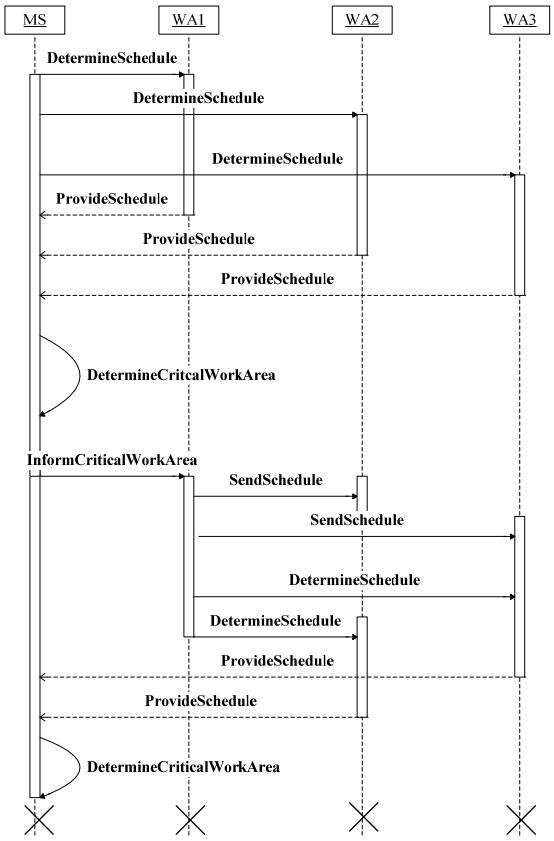


Figure 1: Sequence diagram for distributed scheduling.

In complex manufacturing systems, planned start dates and completion dates are determined with respect to a fixed work area (cf. Mönch and Driessel, 2005 and Mönch *et al.*, 2006) on the top-layer. Here, an aggregated model is used taking the capacity only on the work center level into account. Consecutive operations that are related to one work area are combined into macro operations. This results in aggregated routes that consist of these macro operations. The resulting start and completion dates can be used to set production goals for each decision-making entity, i.e., a scheduling unit, on a certain work area. This approach is called job planning to differentiate it from the more detailed scheduling. In Figure 1, we denote the decision-making entity of the entire manufacturing system by MS. The base-layer is formed by the different work area decision-making entities. We can see three work areas, denoted by WA, in Figure 1. This layer results in schedules for each single work area. Then, it determines a critical work area with respect to a criticality measure, for example, the tardiness of the jobs with respect to the work area where they are scheduled. Based on the schedule for the critical work area, updated ready dates and due dates are sent to the non-critical work areas and used to determine new schedules. This procedure is repeated for the next most critical work area in a recursive manner until the last work area is reached.

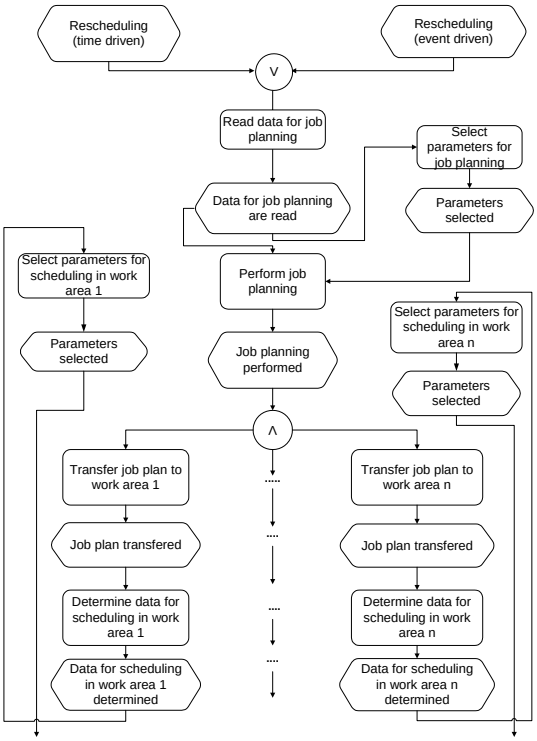


Figure 2: Functionality of the top- and base-layer (Part 1).

Figure 1 does not provide enough information from a process modeling point of view. Therefore, we provide event-driven process chains (EPC) for

the researched scheduling scenario in Figure 2 and Figure 3.

These process models can be used as a starting point for identifying appropriate services, because service-oriented architectures (SOAs) usually distinguish between process, service, and technology models (cf. Siedersleben, 2007). Figure 2 shows the job planning step and major parts of the preparation phase for the scheduling activities on the base-layer. The iterative procedure that corresponds to the decisions on the base-layer is shown by continuing the EPC from Figure 2 in Figure 3. Only a single iteration is depicted in this figure for the sake of simplicity. Note that a simple process model is provided by the EPC in Figure 2 and Figure 3.

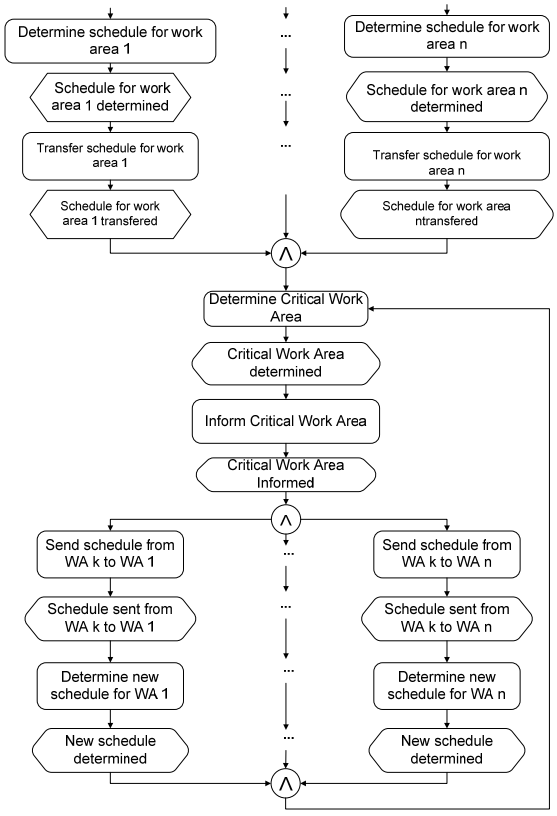


Figure 3: Functionality of the top- and base-layer (Part 2).

3.2 Identification of Services

Next, we are interested in deriving a concrete service model for the researched scenario. The basic idea for identifying services consists in determining for each function within an EPC or activity within an activity diagram, respectively whether it has to be implemented as an operation of a service or not. We use the following five criteria to make this decision:

1. **Degree of Automation (DA):** If no manual intervention is required between two consecutive functions then the two functions can be pooled together into one operation, otherwise, the two functions have to be represented by different operations.
2. **Atomicity (A):** When sub-functions of a certain function can be used within different business processes then it makes sense to implement the sub-functions using different operations. When the entire function can be re-used then a service might be implemented by orchestrating the services that correspond to the sub-functions.
3. **Modularity (M):** Functions are often carriers of an algorithm. When different algorithms are available to solve a problem, a situation-dependent selection of one algorithm is often beneficial. When such a selection is required, then an implementation of the function within a service is useful.
4. **Reusability (R):** When a function is applied in different business processes, then the function has to be implemented within a service.
5. **Number of Users (NU):** When a function is used by different humans, a certain degree of reusability is given. This function has to be implemented within a service.

Note that some of these criteria are also contained in (Kohlborn *et al.*, 2009 and Mönch and Zimmermann, 2009).

Applying these five criteria, we obtain a set of operations that can be grouped into different services. We differentiate between different types of services according to Kohlborn *et al.* (2009). Object-centric services are given by a set of operations that are used to access business objects. Task-centric services are formed by a set of operations that can be carried out without a specific business object. Hybrid services are somewhere between object- and task-centric services, because they consist of operations that are used to access data and perform tasks. Finally, process-centric services are used to support business processes using operations of other services by means of service composition. The first three service types correspond to basic operations that cannot be further decomposed. The resultant services are shown in Table 1. Here, we use the abbreviations PPC and PM for production planning and control and preventative maintenance, respectively.

Determining the necessary data for the job planning functionality is important. Therefore, we decompose the corresponding function into several sub-functions that are implemented as operations of

a service that is related to data management for job planning. Several algorithms can be used to find appropriate parameters for the algorithms used for job planning. Therefore, this function is identified as an own operation of a job planning-related service. Because different job planning algorithms are possible, we also identify the function that determines job plans as an operation of the job planning service. The function that transfers job plans to work areas is only important in the context of work area scheduling. However, within the flow control, the job plans are available. Therefore, this function is not identified as an operation of the job planning service.

Table 1: Applying the criteria to the scheduling functions.

Criterion	DA	A	M	R	NU
Data Gathering Job Planning	0	X	0	X	None
Parameter Setting Job Planning	X	0	X	X	PPC
Calculate Job Plans	X	0	X	X	PPC
Transfer Job Plans to Work Areas	0	X	0	0	None
Data Gathering for Scheduling a Work Area	0	X	0	X	None
Parameter Setting Job Planning	X	0	X	X	PC
Determine Schedule for Work Area	X	0	X	X	PC, PM
Transfer Schedule to Top-Layer	0	X	0	0	None
Determine Critical Work Area	X	0	X	X	PPC
Inform Critical Work Area	0	X	0	0	None

The data management service for work area-related scheduling is justified in a similar manner as the data management service for job planning. Again, a refinement into several sub-functions is performed. Due to different possible algorithms for parameter setting within work area scheduling algorithms, we identify a corresponding parameter setting operation for the work area scheduling service. Different scheduling algorithms, e.g., decomposition- or local-search-based approaches, can be used to determine schedules for jobs in each single work area. Therefore, this function is represented by an operation of the work area-related scheduling service. Because the work area schedules are available within the flow control, we do not identify a separate operation to transfer schedules to the top-layer of the hierarchy. Several methods are available to determine a critical work area. Therefore, this function is identified as an operation of the job planning-related service. It is reasonable to inform the decision-making entity of a critical work area. Therefore, we include this functionality into the operation that determines the critical work area.

Table 2: Identified services for the scheduling process.

Service	Operation
Object-centric	
DataManagement-JobPlanning	ReadCapacityMS()/ ChangeCapacityMS()
	ReadJobsMS()/ AddJobsMS()
	ReadAggregatedRoutes-MS()/AddAggregated RoutesMS()
DataManagement-SchedulingWork-Area	ReadMachinesWA()/ AddMachinesWA()
	ReadJobsWA()/ AddJobsWA()
	ReadPartialRouteWA()/ AddPartialRoute()
hybrid	
JobPlanning	InitializeJP()
	SetParametersJP()
	ReadDataJP()
	DetermineJP()
	EvaluateJP()
	DetermineCriticalWA()
SchedulingWA	InitializeSWA()
	SetParameterSWA()
	ReadDataSWA()
	DetermineScheduleWA()
	EvaluateScheduleWA()
	IsWACritical()
process-centric	
Data Manage-mentScheduling	DetermineAggregatedCapa-cities()
AutomatedSche-dulingManufacturing System	StartProcess()

After the identification of operations, they have to be grouped into appropriate services as discussed before. We show the results of this second step in Table 2.

Note that two process-centric services are described in Table 2. The first service is related to data management issues of an automated scheduling process. This service is based on operations of the DataManagementSchedulingWorkArea service. For example, the single machine capacities are determined using the ReadMachinesWA() operation. They are aggregated to capacities of certain work centers that are used within the job planning approach. The second process-centric service represents the automated hierarchical scheduling process. It contains only a single operation to launch this process. Note that the SchedulingWA service allows for the integration of problem specific scheduling approaches for different work areas.

4 IMPLEMENTATION OF THE PROTOTYPE

4.1 Implementation Issues

In this section, we will discuss our technology model. A prototype is designed and coded to check the feasibility of the proposed component. A client application implemented in the C# programming language provides a graphical user interface for calling the services and therefore, serves as a test driver. Web services, implemented in C#, are run on an ASP .NET development server. They implement the services described in Section 3.

We do not use a real MES for our prototype because we do not have access to an MES. Instead of that, the MES is simulated by a so called blackboard service, basically a SQL Server database. Within the prototype, we implement the AutomatedSchedulingManufacturingSystem service and the DataManagementScheduling service.

However, we are not interested in evaluating concrete scheduling algorithms for the scheduling scenario in this paper because this was done in previous research (cf. Mönch and Driessel, 2005 and Mönch *et al.*, 2006). Our main interest is related to the process flow. The architecture of the prototype is depicted in Figure 4.

4.2 Orchestration of the Services

An orchestration is required to implement the process-centric services. Orchestration of the services is performed using the BPEL engine of the Oracle BPEL process manager. The Oracle BPEL process manager runs on a J2EE application server. The decision to use Oracle BPEL is based on the fact that we used this tool for several previous prototypes.

Because web services are stateless, we compose them into a BPEL process that owns state variables that can be used to manage the state of the system. We can see from Figure 4 that the AutomatedSchedulingManufacturingSystem and the DataManagementScheduling services are implemented as BPEL processes.

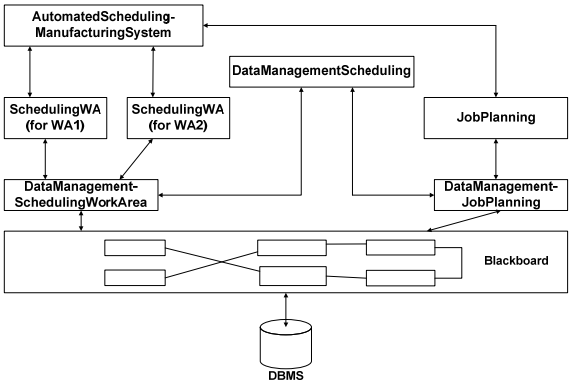


Figure 4: Architecture of the Prototype.

SOAP messages are used to exchange data between the different web services. This approach requires that appropriate XML data structures for routes, aggregated routes, job plans, and work area schedules are defined because SOAP is XML-based. Because we also have to write data, we use SOAP- and not REST-based web services.

4.3 Limitations of the Approach

There are some limitations of the proposed approach. Task objects, goals, and solution methods are the main building blocks of an enterprise task (cf. Ferstl and Sinz, 2006). Business processes work on task objects and change the attributes of these objects. Therefore, we can conclude that business processes are always related to persistency issues. Many object-centric services are the consequence (see Table 2). A tight coupling using joint master data is typical for scheduling and more generally for the production planning and control domain.

This may lead to undesirable side effects of services, such as problems with the global state consistency of the distributed application. The state of the art master data management in MESs has to be extended to fulfill the SOA requirements. We can see this requirement from Figure 4, where the blackboard service is used to represent the master data from the MES.

The second limitation is also related to data management in the service-based prototype. A certain

number of XML-based SOAP messages have to be exchanged between the different operations of the services that form the scheduling component. A loss of performance might be the result of this situation.

5 CONCLUSIONS

In this paper, we described a scheduling component for complex manufacturing systems that is based on a hierarchical decomposition of the overall scheduling problem. We discussed the identification of appropriate services. The orchestration of these services is shown. The implementation of a prototype for such a component based on web services is also discussed.

There are some directions for future work. While it is possible to design and implement such a component, there is still much more effort required to integrate the resultant component into a real-world MES. A rigor assessment of the performance of the overall application, especially with respect to computing time, is also necessary.

Furthermore, it seems fruitful to combine software agents with service-oriented computing techniques as proposed by Huhns (2008). It is highly desirable to enrich the multi-agent-system FABMAS (cf. Mönch *et al.*, 2006) that implements a similar hierarchical scheduling approach as described in the present paper by web services. It is differentiated between decision-making and staff agents in FABMAS. Staff agents support the decision-making agents. It seems possible to replace the staff agents by web services as discussed in the present paper. The decision-making agents can be used to carry out a more sophisticated orchestration of these services.

ACKNOWLEDGEMENTS

The author would like to thank Daniel Kaiser for implementing the prototype and for interesting discussions on the topic of this paper.

REFERENCES

Brehm, N., Marx Gomez, J., 2007. Web service-based specification and implementation of functional components in federated ERP-systems. In *Proceedings Business Information Systems (BIS) 2007, LNCS 4439*, 133-146.

Ferstl, O., Sinz, E. J., 2006. *Grundlagen der Wirtschaftsinformatik*. Oldenbourg, München, Wien, 5th Edition.

Fordyce, K., Bixby, R., Burda, R., 2008. Technology that upsets the social order - a paradigm shift in assigning lots to tools in a wafer fabricator - the transition from rules to optimization. In *Proceedings of the 2008 Winter Simulation Conference*, 2277-2285.

Framinan, J., Ruiz, R., 2010. Architecture of manufacturing scheduling systems: literature review and an integrated proposal. *European Journal of Operational Research*, 295, 237-246.

Gaxiola, L., Ramirez, M. D. J., Jimenez, G., Molina, A., 2003. Proposal of holonic manufacturing execution systems based on web service technologies for Mexican SMEs. In *Proceedings HoloMAS 2003*, LNAI 2744, 156-166.

Huhns, M. H., 2008. Services must become more agent-like. *WIRTSCHAFTSINFORMATIK*, 50(1), 74-75.

Kellogg, D., Walczak, S. 2008. Nurse scheduling: from academia to implementation or not? *INTERFACES*, 37, 355-369.

Kohlborn, T., Korthaus, A., Chan, T., Rosemann, T., 2009. Identification and analysis of business and software services – a consolidated approach. *IEEE Transactions on Services Computing*, 2(1), 50-64.

McClellan, M., 1997. *Applying Manufacturing Execution Systems*. St. Lucie Press, Boca Raton.

Meyer, H., Fuchs, F., Thiel, K., 2009. *Manufacturing Execution Systems: Optimal Design, Planning, and Deployment*. McGraw-Hill.

Mönch, L., Driessel, R., 2005. A distributed shifting bottleneck heuristic for complex job shops. *Computers & Industrial Engineering*, 49, 673-680.

Mönch, L., Zimmermann, J., 2009. Providing production planning and control functionality by web services: state of the art and experiences with prototypes. In *Proceedings of the 5th Annual IEEE Conference on Automation Science and Engineering*, 495-500.

Mönch, L., Stehli, M., Zimmermann, J., Habenicht, I., 2006. The FABMAS multi-agent-system prototype for production control of wafer fabs: design, implementation, and performance assessment. *Production Planning & Control*, 17(7), 701-716.

Pfund, M., Mason, S. J., Fowler, J. W., 2006. Semiconductor manufacturing scheduling and dispatching. Chapter 9 in *Handbook of Production Scheduling*, Herrmann, J.W. (ed.), Springer, New York, 213-241.

Qiu, R. G., 2007. A service-oriented integration framework for semiconductor manufacturing systems. *International Journal Manufacturing Technology and Management*, 10(2-3), 177-191.

Siedersleben, J., 2007. SOA revisited: Komponentenorientierung bei Systemlandschaften. *WIRTSCHAFTSINFORMATIK*, 49, 110-117.

Tarantilis, C. D., Kiranoudis, C. T., Theodoadoakopoulos, N. D., 2008. A web-based ERP system for business services and supply chain management: application to real-world process scheduling. *European Journal of Operational Research*, 187, 1310-1326.

Winkler, V., Buhl, H. U., 2007. Identifikation und Gestaltung von Services: Vorgehen und beispielhafte Anwendung im Finanzdienstleistungsbereich. *WIRTSCHAFTSINFORMATIK*, 49, 257-266.

Node Positioning
Application for Wireless Networks Industrial Plants

Pedro H. G. Coelho, Jorge L. M. do Amaral and José F. do Amaral
Rio de Janeiro State University, UERJ, Department of Electronics and Telecommunications,
Rua São Francisco Xavier, 524 – Sala 5027E, 20550-900, RJ – Rio de Janeiro, Brazil
{phcoelho, jamaral, franco}@uerj.br

Keywords: Wireless Networks, Node Positioning, Artificial Immune Systems.

Abstract: This article discusses the positioning of the nodes of a wireless network of an industrial plant for the network to meet the application requirements, particularly with respect to coverage characteristics and reliability. Issues involving these two parameters are investigated and it is intended to submit proposals using the concepts of computational intelligence to solve the problem.

1 INTRODUCTION

The possibility of using wireless network has been widely discussed in the areas of industrial automation, environmental monitoring, and location of road vehicles among others. The great advantage of not using cable for data transmission is the ease of network installation in all environments, including those where it is not possible to lay cables, be for the difficulty of access, or for being a dangerous area or not allowed access. Another advantage is the ease of maintenance of equipment. In the listed applications, it is of paramount importance the safety, reliability, availability, robustness, and network performance in carrying out the monitoring and process control. That is, the network cannot be sensitive to interference or stop its operation because of an equipment failure, nor can have high latency in data transmission and ensure that the information is not lost (Zheng and Myung, 2006), (Santos, 2007).

A network of wireless smart sensors is responsible for conducting the monitoring of a process or an environment, process the collected information and send it to other sensors or routers closer to the gateway. The sensors are powered by batteries and positioned according to the process to be monitored (Gomes, 2008).

Data transmission in a wireless network in today's industrial automation is faced with the problem of interference generated by other equipment and obstacles. In an attempt to minimize these effects, various methods of intermediate nodes positioning are used. The intermediate nodes or routers are responsible for making the routing of data, generated by sensors in the network to the gateway through hops, directly or indirectly. Such devices are responsible for meeting the safety, reliability and robustness criteria of the network and are of paramount importance in directing data transmission. However, they can leave all or a great part of the network dead, if they have any fault (Hoffert et al., 2005).

Most of the presented solutions to this problem use optimization algorithms to find the smallest number of routers needed to make the network meet the criteria related to the decrease of energy consumption of each node. Moreover, monitoring the total area, simplifying the network with the lowest cost, meeting the traffic demand not evenly distributed, reducing the latency of the data, the reduction of computational complexity, minimizing the burden placed on the nodes and maximizing the number of nodes that can communicate with the gateway are also key issues in the problem (Youssef and Younis, 2007), (Molina et al., 2008). These solutions typically face problems of scalability and changes of the network configuration the over the years. Thus, for each network configuration it would be necessary to develop a specific solution in accordance with new obstacles, positioning of the sensor nodes and, consequently, new positions for the routers.

This article is divided into four sections. This section is an introduction to the problem of positioning nodes in wireless networks. Section two discusses the importance of studying node

positioning. In section three possible solutions to the problem using computational intelligence (CI) are discussed. The forth section closes the paper presenting research directions and preliminary conclusions.

2 NODE POSITIONING ISSUES

This section briefly discusses the positioning of nodes characteristics and its main constraints.

Why study the positioning of the nodes? For the network to meet the application requirements, especially regarding reliability and coverage issues. Positioning of the nodes can cause a dramatic impact on the efficient operation of networks.

The positioning of the nodes can be static, which is done before the network operation, or dynamic, where the repositioning of nodes continues on the network in operation.

Static positioning of the nodes depends on the method of nodes distribution, for instance controlled or randomized. It also varies with the optimization objective which may include the coverage area, connectivity, longevity or data fidelity. The role of node in the wireless network should also be taken into account in the positioning process e. g. the node can act as a sensor repeater, a base station or cluster-head node. Questions such as where and when to relocate nodes naturally arise and the characteristics of the network also play an important role in the whole process.

One has to define the network coverage or in other words the accessibility to the gateway. The critical nodes are those for which its load is overwhelming and the sensitivity of the network to a loss of such a node is high. Fault tolerance is also a key issue and it is of paramount importance to know what happens if a node fails. In such a case it is important to know if an alternative path exists for those paths having that faulty node. Determining the number and the positioning of repeater nodes is also an information one should have.

For industrial wireless networks every node must be able to communicate with the gateway, either directly or through other nodes, so that the targeted coverage should be equal to 100 %.

As far as the used criteria are concerned for each of such networks, every node must have a certain number of neighbors in order to increase the availability of alternative paths. That means also that a number of network nodes must be in direct connection with the gateway. The number of hops for which a message reaches a node to the gateway has to be closely monitored once the increase in the number of hops raises the message latency. Depending on the refresh rate of the measurements, in case of wireless sensor applications, this can be an important issue. Also the number of retransmissions from other nodes necessary for reaching the gateway should be carefully considered as increasing the number of retransmissions may shorten the battery life.

In terms of fault tolerance it is important to know what percentage of the network that is still working if a particular node fails. What is the most critical node on the network in relation to this criterion?

Those issues are to be considered in the building of the node positioning for industrial wireless networks.

3 CI BASED PROPOSALS

This section presents proposals for solving the problem of positioning nodes for wireless industrial networks employing computational intelligence techniques. Such techniques are very promising for application to the problem at hand because they allow consideration of heuristic optimization issues related to the theme.

The problem of positioning nodes is an NP-Hard (Molina, 2008) and in view of this, it is usual to use heuristics and stochastic optimization schemes. Thus potential techniques applicable to the solution of the problem involve those related to computational intelligence such as genetic algorithms, collective intelligence, such as ant colony optimization, artificial immune systems and others. Before proceeding it is necessary to emphasize an important feature of the problem of positioning nodes in wireless networks in industrial environments. Note that in this case, the nodes in the network are positioned at locations to be instrumented and connectivity with the central node, usually located in the control room, it is absolutely necessary otherwise the consequences can be devastating. This situation is distinct from a network of wireless computers only for INTERNET access in which connectivity can be lost and then resumed without major losses to the user, since the greatest interest is to achieve a high throughput.

3.1 Node Positioning using Artificial Immune Systems

The immune system is one of most important ones for the survival of humans and animals. It has the

task of fighting the invaders, which cause diseases through complex mechanisms. Such mechanisms are complementary and fit to perform the recognition of pathogens (viruses, bacteria, foreign molecules etc.) and inhibit its action in the body of the individual and are divided into (Amaral, 2006) (Castro, 2001):

1. Recognition of pathogen - is accomplished by lymphocytes, i.e. B and T cells that have receptors for the purpose of joining the pathogen to subsequently eliminate it;

2. Affinity maturation of lymphocyte receptors and pathogen - there will be hypermutation receptors so that they are able to fit "perfectly" to the antigen;

3. Cloning of the antibody with higher affinity - cloning of lymphocytes that are better suited to the pathogen;

4. Distinction between self and non-self - this mechanism is of paramount importance for the individual able to survive without any autoimmune disease that destroys the cells and proteins of the organism itself. It will make the distinction between body proteins and the invaders;

5. Immunological memory - is a database stored in the memory immune receptors, which act more quickly and effectively against the next infection caused by the same pathogen.

The artificial immune systems exploit mechanisms found in natural immune systems to develop techniques for solving problems. The natural immune systems provide protection against numerous pathogens such as viruses, bacteria and others.

Some basic concepts of natural immune systems will be described so that we can develop the application in node positioning. Antigens are substances that are not recognized by the immune system as the body itself. There are two types of immune systems the innate and the adaptive. The first is the first line of defense of the living organism and reacts similarly to different pathogens such as the skin. Note that some pathogens cannot be fought by the innate immune system. The adaptive system fight against specific pathogens. Its main components are B cells which produce antibodies and T cells that attack the abnormal cells. The response of the innate immune system remains constant, the adaptive gives immunity against re-infection of the same infectious agent. Pathogens or molecules present antigens that are recognized by B cells. Note that the marriage is not always perfect. Since the antigen recognized, the B cell begins to produce antibodies. Each B cell produces only one type of antibody. For example, antibody to influenza virus is different from that for pneumonia. The more efficient antibodies are cloned.

Now an algorithm using artificial immune system techniques will be described.
The algorithm is as follows:

1 - Initialization: Original placement or pattern of antibodies.

2 - Training: Presentation of antigens for the iterative network of antibodies against antigens and antibodies.

3 - Competition: winners antibodies in accordance with an affinity function

4 - Cloning; reproducing the efficient antibodies.

5 - Convergence: each antibody is associated with an antigen and each antigen antibody should have a winner within a minimum defined distance.

6 - Pruning: After all training unrelated antibody with any antigen is removed.

Preliminary tests indicate that the above proposal is satisfactory.

4 CONCLUSIONS

The artificial immune systems are algorithms inspired by the functioning of the human immune system to solve optimization problems, pattern recognition and others. The most widely used algorithms in solving the problems mentioned above are the immune network algorithms, clonal selection and negative selection (Castro and Timmis, 2002). The artificial immune networks are algorithms that mimic the functioning of the immune network in combating human infectious diseases in slaughter. This network provides human immune B cells capable of recognizing and to recombine in the absence of the pathogenic agent, thereby forming a network capable of eliminating the invaders. They are formed in accordance with the degree of affinity between B cells. If the affinity between them is high, then the cell B is joined to the network, otherwise it will be repelled away from the network. This action of union or inhibition of B cells occur until the network stabilizes and so could fight off diseases. The purpose of this paper is to solve the problem of positioning nodes in wireless industrial networks using artificial immune, based on the human immune system. The algorithms based on immune networks have very desirable characteristics in solving this problem, among which we mention: scalability, self-organization, learning ability and continuous treatment of noisy data. It is intended to build positioning algorithms based on models of artificial immune networks (Castro 2001), aiming to

get the best settings for a wireless network industry, positioning the router nodes in the network, so that all devices to communicate with gateway without loss of information.

REFERENCES

Zheng, J. and Lee, M. J. (2006) *A Comprehensive Performance Study of IEEE 802.15.4*, *In: Sensor Network Operations, IEEE Press*, Wiley InterScience, Chapter 4, pp. 218-237.

Santos, S. T. dos (2007) Wireless Sensor in Monitoring and Control: MSc Dissertation, *Federal University of Rio de Janeiro*, COPPE/UFRJ. In Portuguese.

Gomes, S. P., Carvalho, S. V. de, and Rodrigues, R. C. M. (2008) Optimization of Energy Consumption in Wireless Sensor Networks by Decision Markov Processes, *In: 7th Brazilian Conference on Dynamics Control and Applications*. In Portuguese.

Hoffert, J., Klues, K. and Orjih, O. (2005) "Configuring the IEEE 802.15.4 MAC Layer for Single-sink Wireless Sensor Network Applications", In: Technical Report,http://www.dre.vanderbilt.edu/~jhoffert/802_15_4_Eval_Report.pdf .

Youssef, W., and Younis, M. (2007) "Intelligent Gateways Placement for Reduced Data Latency in Wireless Sensor Networks", *In: ICC'07 International Conference on Communications*, Glasgow, pp. 3805-3810.

Molina, G., Alba, E., and Talbi, E. G. (2008) "Optimal Sensor Network Layout Using MultiObjective Metaheuristics", *In: Journal of Universal Computer Science*, Vol. 14, No. 15, pp. 2549-2565.

Castro, L. N. (2001) Immune Engineering: Development and Application of computational Tools inspired in Artificial Immune Systems, *In: Ph.D Thesis, UNICAMP, Campinas*, pp. 287. In Portuguese.

Amaral, J. L. M, (2006) Artificial Immune Systems Applied to Fault Detection, *In: Ph.D. Thesis, PUC-RJ, Rio de Janeiro*, RJ, pp. 121. In Portuguese.

Castro, L. N. and Timmis, J. (2002) Artificial Immune System: A New Computational Intelligence Approach. *Springer-Verlag*, pp. 357.

POSTERS

An Ontological Knowledge-base System for Composing Project Team Members

Yu-Liang Chi
Dept. of Information Management, Chung Yuan Christian University, 200 Chung-Pei Rd., Chung-Li, Taiwan 320, Republic of China
maxchi@cycu.edu.tw

Keywords: Team Member Composition, Knowledge-based System, Ontology, Semantic Rules.

Abstract: In teamwork-based projects, human play a critical role in achieving project success. This study utilizes ontological approaches to build project teaming models into ontology. It helps to develop a set of logic rules for identifying semantic relationships between individuals. By following a knowledge base creation process, the factual data of project, workers, and teaming factors can be inserted into ontological knowledge base. Based on knowledge inference, reliable knowledge bases are established for selecting project team members in runtime. A case study is presented to demonstrate the effectiveness of the proposed design.

1 INTRODUCTION

Collaboration is a major feature of teamwork-based projects, which are frequently implemented by high performance project teams. Effective project teaming thus has become essential in human-side project management. In project management, project teaming refers to managing a project team with assignment of project tasks and roles (Beranek *et al.*, 2005), and the appropriate composition of the development or workplace team that performs ad-hoc project tasks. Industry experts and academic researchers continue to work on identifying factors and composition approaches for effective project teaming. While current methods and considerations are presented mostly as predefined and syntactic criteria, further consideration of the effect of derived semantic relations and facts should also be carried out. The characteristics typically considered in composing a quality team include team size, personal commitment, current workload of the individual, leadership, skill competence, years of experience, communications skill, and so on (Chen and Lin, 2004). A need for cross-functional composition with regard to the skill backgrounds of team members is recognized in projects and is multi-disciplinary in nature. Configurational and task-oriented approaches to project teaming require the composition of a team to depend on tasks of the project work (Coates *et al.*, 2007). Such tasks contribute to the technical and explicit foundations of a software project team.

A solid technical foundation alone does not guarantee a quality composition of the project team. For example, Krishnan (1998) reported that the effects of three team-related measures include not only the domain and language experience of the team, but also the capabilities of the team personnel with regard to information system product costs and quality. This is particularly true when it is recognized that culture and human or "soft" factors, for example differences in individual characteristics of preferences, also contribute to team success (Gorla and Lam, 2004). Regarding personality, the Myers-Briggs Type Indicator has been widely employed to assess software engineer personality types (Stewart *et al.*, 2005), as well as to assess the influence of team member personality namely Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness to experience, also known as the "big-five personality factors" on individual role, social role and task accomplishment. Thus, the human-side of project management should be integrated into technological project management methods and tools.

2 DEVELOPING TEAM MEMBER COMPOSITION MODEL

Project team knowledge has numerous sources and different aspects. To build the knowledge model of

project teaming, his study refers to the discipline of knowledge engineering (KE) proposed by Guarino (1995). Typical KE includes expertise gathering, knowledge model building, and knowledge representation. Detailed steps are described as the followings:

Expertise Gathering: Expertise gathering is the focus of identifying critical elements of project teaming. Observation of project teaming events identifies three primary subjects including *Project, Worker* and *Teaming factors*. Further expertise gathering are also implemented, including elicitation, analysis, and transformation are implemented.

- Property elicitation. Over 100 properties are gathered during this stage. For example, the collected properties of projects are basic features such as project's budget and skill needs.
- Analysis. A total of 35 properties are identified. New or subordinate subjects are generated by either separating existing subjects or assembling relevant subjects. For example, the subjects *Project_Type*, *Title*, *Skills* and *Personality* are subordinate to the *Teaming_Factors* subject.
- Transformation. This step transforms subjects and their dependent properties into an ontological representation. The representation is formed by a formatted expression written as *{**Concept**: Property list}*. Examples are presented below.

  {**Project**: *Project_Description, PM, Budget, Number_of_HR, Skill_Need, PM_Skill_Need, Length_of_Time, has_Type, Qualified_Basic_Worker, Watch_List, …* }

  {**Worker**: *Age, Gender, Seniority, Salary, has_Title, has_Skill, member_of, Hostile_to, has_Experience, …*}

Building Knowledge Model: A knowledge model is based on an abstract view of the task domain, and can be used as an intermediary between the real world and information systems. Two type of relations including "*is-a*" and "*has-a*" are developed. The "*is-a*" denotes an inheritance relationship between two concepts. For example, the *Teaming_factors* concept has sub-concepts such as *Title, Skills and Personality*. The "*has-a*" denotes the "*part-of*" relation between two concepts. These properties stipulate a schema for describing the concepts. Users can employ these properties to contribute their factual knowledge to the knowledge base or to obtain implicit knowledge via inference mechanisms.

Furthermore, two types of property's content including "Asserted property" and "Inferred property" need to be defined during model building. The asserted properties provide the basis of inference engine to deduce new knowledge. On the other hand, the content of inferred property is implicit, but can be obtained by inferring factual knowledge via a reasoned. The inferred property plays a critical role for rule-based reasoning.

Knowledge Representation: This study employs Web Ontology Language (OWL) as the notation and formalism for representing the knowledge to be stored in ontology. After constructing the team composition model, OWL is utilized for knowledge representation. OWL is highly appropriate for representing structured knowledge using classes and properties organized in taxonomies.

3 CREATING RELATIONSHIPS USING SEMANTIC RULES

Horrocks and Patel-Schneider (2004) have reported several limitations and issues of OWL in syntax and computation, particularly in relationships between roles chains using rules, causing indeductibility, logical undecidability, by embedding the word problem in inferences. The rules apply the syntax "*Antecedent* ➔ *Consequent*". Both antecedent and consequent are conjunctions of atoms of the form $atom_1 \wedge .. \wedge atom_n$, where a variable is indicated by a question mark (e.g., *?x*). The semantic rules are used to extend the power of the ontological approach to identify semantic relationships between instances.

This study utilizes the *Project_Type* concept to manage characteristics of typical historical projects as best practices. Several properties used in rules development are detailed below. The *PT_Skill_Need* and *PT_PM_Skill_Need* properties are used to indicate the skills needed by workers and project managers, respectively. Furthermore, the *PT_Personality_Need* property describes preferred personality types for performing the project. These properties can be used to develop rules to connect other concepts such as *Worker* to obtain candidate members for a project team. The following five rules are examples developed in this study.

Rule-1 is used to identify qualified team members with reference to best practice. Rule-2 helps identify candidate project manager(s) based on the qualified workers with reference to best practices. Rule-3 is applied to group senior workers as candidate team members. Rule-4 adds the *PT_Personality_Need* property to deduce whether

the qualified workers possess the preferred personalities. Rule-5 examines whether a qualified worker is hostile to someone then both of them will be inserted into the *Watch_List* property of the project.

Rule-1: Project(?x) ∧ has_Type(?x, ?y) ∧ PT_Skill_Need(?y, ?z) ∧ Same_Skill_Worker(?z, ?a) ➔ Qualified_Basic_Worker(?x, ?a)

Rule-2: Project(?x) ∧ has_Type(?x, ?y) ∧ PT_PM_Skill_Need(?y, ?z) ∧ Same_Skill_Worker(?z, ?a) ∧ has_Title(?a, Project_Manager) ➔ PM(?x, ?a)

Rule-3: Project(?x) ∧ has_Type(?x, ?y) ∧ PT_Skill_Need(?y, ?z) ∧ Same_Skill_Worker(?z, ?a) ∧ Seniority(?a, ?b) ∧ swrlb:greaterThan(?b, 5) ➔ Qualified_Advanced_Worker(?x, ?a)

Rule-4: Project(?x) ∧ has_Type(?x, ?y) ∧ PT_Skill_Need(?y, ?z) ∧ Same_Skill_Worker(?z, ?a) ∧ PT_Personality_Need(?y, ?b) ∧ has_Personality(?a, ?b) ➔ Quality_Intensive_Worker(?x, ?a)

Rule-5: Project(?x) ∧ has_Type(?x, ?y) ∧ PT_Skill_Need(?y, ?z) ∧ Same_Skill_Worker(?z, ?a) ∧ Hostile_to(?a, ?b) ➔ Watch_List(?x, ?b)

4 CASE STUDY

Before implementing this case study, known facts (instances) of concepts must be identified. For example, instances about workers, including *age*, *salary*, and *skill*, must be given into the asserted properties. Some instances regarding the example scenario are detailed below. Table 1 lists known instances of the *Project_Type* concept. Each instance involves three known property values, such as skills required of the project manager, skills required of workers, and the personalities preferred by the project. These instances are considered to represent the best practices for future projects.

Table 1: Instance samples of the *Project_Type*.

Type	PM Skills*	Member Skills*	Personalities**
BPM	PC; PMC; PP	BM; CM; SAD; DP	High_A; High_E
ERP	PC; PMC; PP	BM; CUT	High_C; High_E
GCM	PP	CM; UAT;DP	Low_N; High_C
HRM	PP	QA; TR	Low_N; High_C
MES	PMC; PP;	RD; TR; SAD	High_A; High_E
PLM	PP; CM	BM; CM; CUT	High_O; High_C

*Skills. BM: business modeling; CM: configuration management; CUT: coding and unit test; DP: deployment; PC: project closure; PMC: project monitor and control; PP: project planning; and etc.

**Personalities. A: Agreeableness; C: Conscientiousness; E: Extraversion; N: Neuroticism; O: Openness

Table 2 lists partial instances of the *Worker* concept. The row headings indicate the property names for each instance. A worker comprises eight known property such as title and gender. The *has_Skill* property presents a list of skill items. The symbol '×' indicates that a worker has this corresponding skill. In the *Personality* property, the symbols *N, A, C, E, and O* denote Neuroticism, Agreeableness, Conscientiousness, Extraversion and Openness respectively. Furthermore, the symbol '+' represents positive psychological power, while the '-' indicates negative psychological power. Total 17 persons are identified for the following experiments.

Table 2: Instance samples of the *Worker*.

Name	Allen	Alvin	Cindy	Eric	Leon	Mavis	Phil	Stan	Ted
Title	IC_L.	IC	BC	PM	PM	IC_L.	PM	BC	IC
Gender	M	M	F	M	M	F	M	M	M
Salary	48k	52k	40k	70k	160k	50k	120k	67k	70k
Seniority	7	3	2	2	3	8	5	8	7
has_skills									
BM			×					×	
CM		×							
UAT			×		×				
Hostile_to	-	Ian	-	Flying	-	Jeff	Stan Ted	Phil	Phil Eva
Personality									
N	+		+			+	+		
O				−		−	−		

The first case uses instances of *Project_Type* as a reference for best project practices. For example, when a project is newly created, the decision makers identify the project has having typical features like the *BPM*. As shown in Figure 1, a user selects the *BPM* as a known fact in the *has_Type* property. This property value is initially the only factual knowledge associated with the new project. After firing the JESS rule engine, the project obtained five inference results as presented within inferred properties. The rules engine utilizes known facts of the *BPM* to provide for the computational needs of Rule-1 to 5. For example, Rule-1 is applied to identify qualified workers using the instances in the *PT_Skill_Need* property of the *BPM*, including *BM, CM, SAD* and *DP*. A total of nine workers were inferred into the *Qualified_Basic_Worker* property. Rule-2 deduced two qualified project managers such as *Eric* and *Leon* for this new project. Rule-3 deduced four candidate workers for *Qualified_Advanced_Worker* property. These workers are all highly qualified and each had over 5 years of working experience. Rule-4 treats preferred personalities as noted criteria in the property of the *BPM*. A total two workers were

recommended inside the *Quality_Intensive_Worker* property. These workers have at least one personality item conforming to *Agreeableness, Extraversion*, or both. Finally, Rule-5 contributed five workers to the *Watch_List* property. These workers may have interpersonal relationship issues based on the record of the *Hostile_to* property of the *BPM*.

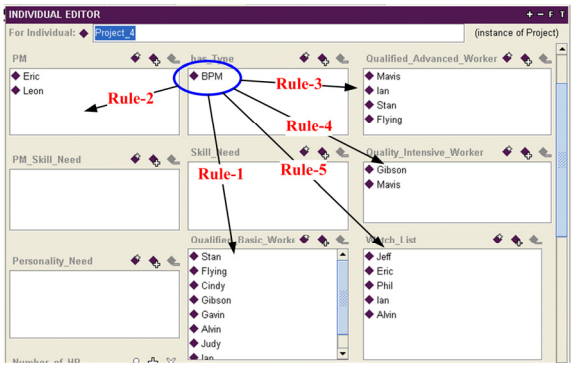


Figure 1: Using the instance of *Project_Type* as a reference to infer candidate team members.

5 DISCUSSION AND CONCLUSIONS

This study employs OWL as the notation for representing knowledge to be stored in the ontology. SWRL rules are applied to infer semantic relationships of instances. Once ontology and rules are used for knowledge representation, it is possible to stipulate practical facts as factual knowledge. The experimental results demonstrate that the proposed design can support the system for identifying appropriate project teams. Additionally, the proposed design stresses that the system can be continually maintained by factual knowledge providers rather than system developers. The inference mechanism then helps establish a new and complete knowledge base for maintaining system reliability. Consequently, the combination of semantic rules and ontologies can manage intricate information such as the project teaming problem mentioned in this study.

ACKNOWLEDGEMENTS

The author would like to thank the National Science Council of the ROC for financially supporting this research under Contract No. NSC 99-2410-H-033 - 028 -MY3.

REFERENCES

Beranek, G., Zuser, W., and Grechenig, T. (2005). Functional group roles in software engineering teams. *ACM SIGSOFT Software Engineering Notes*, *30*(4), 1-7.

Chen, S. J., and Lin, L. (2004). Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering. *IEEE Transactions on Engineering Management*, *51*(2), 111-124.

Coates, G., Duffy, A. H. B., Hills, W., and Whitfield, R. I. (2007). A preliminary approach for modeling and planning the composition of engineering project teams. *Proceedings of the Institution of Mechanical Engineers*, *221*(7), 1255-1265.

Gorla, N., and Lam, Y. W. (2004). Who should work with whom: Building effective software project teams. *Communications of the ACM*, *47*(6), 79-82.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, *43*(5-6), 625-640.

Horrocks, I., and Patel-Schneider, P. F. (2004). Reducing OWL entailment to description logic satisfiability. *Journal of Web Semantics*, *1*(4), 345-357.

Krishnan, M. S. (1998). The role of team factors in software cost and quality: An empirical analysis. *Information Technology and People*, *11*(1), 20-35.

Stewart, G. L. Fulmer, I. S., and Barrick, M. R. (2005). An exploration of member roles as a multilevel linking mechanism for individual traits and team outcomes. *Personnel Psychology*, *58*(2), 343-365.

Study on Task Decomposition in Emergency Logistics based on System Dynamics

Jun Su[1] and Li-jun Cao[2]
[1]School of Management, Jinan University, Guangdong, Guangzhou, 510632, P.R.China
[2]School of International Business, Jinan University, Guangdong, Zhuhai, 519070, P.R.China

Keywords: System Dynamics, Dynamic Alliance of Logistics, Task Decomposition, Emergency.

Abstract: It analyzed several key factors by system dynamics that the task decomposition in emergency logistics impact on dynamic alliance of logistics. These factors included inter-constraints, quality of cooperation, collaboration time, ability to adapt with each other, core capabilities of logistics. To establish diagram and system dynamics model, it can forecast and analysis disadvantages of task decomposition in emergency logistics. It can for the government emergency management to provide strategic adjustment decision support. On this basis, it simulated the task decomposition of system dynamics model on dynamic alliance of logistics by EXCEL, tested and verified this way was a feasible approach.

1 INTRODUCTION

When unexpected events occurred, government organized the dynamic alliance of logistics quickly for transporting emergency supplies. Their primary job is to break down the missions into several sub-tasks, and then look for federates of dynamic alliance of logistics for each sub-tasks. In the process, the government should consider which way is the best of task decomposition.

The extent of task decomposition determines the number of federates in logistics dynamic alliance adapt to the emergency incident, the different running status of logistics dynamic alliance, and the success or failure to the emergency task ultimately. But the extent of task decomposition influence by many indicators, such as the mandate of the total stipulated completion time, each sub-task stipulated completion time, the working ability in core part of task, and ability to adapt to each other. It should be used to the co-ordinate system for ensuring access to the optimal task decomposition scheme.

When unexpected event occurred, the management system of emergency logistics is a non-stable, non-equilibrium dynamics of the process system. It should not be used the way as solve stable systems to resolving such issues. The system dynamics is to study the behavior of complex feedback systems in the computer simulation method, it can start from the system as a whole, find and study of related factors within the system. It also can focus on the dynamics of process and causality in logistics system, and to solve complex problems in a non-complete non-state analysis of information (Hu et al., 2006).

Currently, it has a lot of studies in task decomposition, particularly in large enterprises and multi-enterprise collaboration between departments in manufacturing. Pi (2006) studied and explored the significance and role in task decomposition of aerospace; Chen (1998) focused on analysis of task decomposition in the Boeing Commercial Aircraft Manufacturing Engineering System; Hu et al. (2005a) proposed the optimization of the virtual enterprise partner selection model based on the task decomposition, and the same year, she proposed process of building a virtual enterprise framework based on task decomposition (Hu et al., 2005b); Zhang et al. (2007) addressed a multi-level projects across the enterprise network planning method based on task decomposition, to solve multi-level program consistency problem in cross-enterprise projects.

This article built a dynamic alliance of logistics simulation model of task decomposition, with the impact of the relationship between the relevant indicators based on system dynamics theory .Finally, it discussed the model simulation results and applications.

2 THE ANALYSIS OF THE CAUSAL RELATIONSHIP IN TASK DECOMPOSITION OF EMERGENCY LOGISTICS

Use of system dynamics to build the flow of causal relationship diagram, It can effectively express the relationship of system feedback, and identify the location of the proper task decomposition. In the process of task decomposition of logistics dynamic alliance, the quality of cooperation, collaboration time and so on, can be affected by many factors, such as logistics facilities and equipment damage, road conditions after expected events, government policies, as well as the impact from different quality of the federate and so on. According to the causal relationship between the determinants, and being combined with other variables in the decomposition of the project, it used VENSIM to build system flow diagram, shown in Figure1.

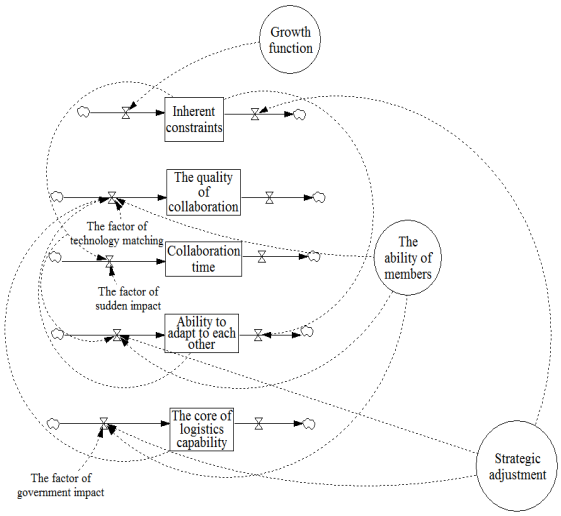


Figure 1: The flow chart of causal elements in emergency logistics task decomposition.

It is usually that the more federates the more bindings. To be targeted strategic adjustments timely, such as replace members of union, can improve the collaboration efficiency and adapt of ability, and prevent the growth trend of inherent constraints in logistics dynamic alliance. But at the same time, the completion time of task will be extended, and emergency supplies cannot be delivered on time. It can bring many dangers to the people in disaster areas and the losses of economic.

3 CONSTRUCTION OF SYSTEM DYNAMICS MODEL

Logistics dynamic alliance is an organization which be needed to maintain close between logistics enterprises. It requires federates matched hardware and software resources. With the increasing number of federates, the demanding of breakpoints have become increasingly in the supply chain. It expressed as the increasing of intrinsic constraints.

Based on this consideration, this paper modified the Pearl curve model, the formula is expressed as:

$$y = \frac{K}{1 + ae^{-b(n-1)}} \tag{1}$$

In here, “n” is expressed that the required total number of enterprises in a particular task of emergency logistics alliance, and it takes a positive real number. The “k” is the limit of the “y”, and it takes “50”. “a” and “b” are the model parameters, and it takes that “a” is “1”and “b” is “1”.

In addition, it should be noted that the followings, such as: if it has only one company in alliance (n = 1), at this time, that means “y” is “25”, this is the minimum constrains; but with the increasing number of federates, the increasing “y” was, and the “50” is assumed maximum constrains of “y”.

It uses DYNAMO language to the identity (Zhao, 2010). “K” is the current moment and “J” is the last moment. “DT” said that the steps between “K” and “J”. It makes “DT” is “1” initially, and you can adjust it in the actual simulation process.

Conveniently, it will use the letters to replace each variable, as shown in Table 1:

Table 1: The alphabet of variable corresponding.

Letters	Variable
A	Inherent constraints
B	The quality of collaboration
C	Collaboration time
D	Ability to adapt to each other
E	The core of logistics capability
M	Strategic adjustment
N	The ability of members
SF	The factor of sudden impact
GF	The factor of government impact
TF	The factor of technology matching

With the causal relationship in Figure 1, the alliance model is expressed as:

$$pearl(n) = \frac{K}{1 + ae^{-b(n-1)}} \quad (2)$$

With the increasing number of federates, the inherent constraints is growth. To reduce the inherent constraints, government can make strategic adjustments to the members of logistics dynamic alliance. Shown as:

$$A.K=A.J+PEARL\ (n)\ -M \quad (3)$$

The growth of the quality of collaboration will be affected by ability to adapt to each other, the core of logistics capability, and the factor of technology matching. Shown as:

$$B.K=B.J+DT*\ (D.J+E.J+N+TF) \quad (4)$$

With the increasing of inherent constraints, the collaboration time is increased. In addition, some unexpected event will lead to collaboration time changes. Shown as:

$$C.K=C.J+DT*\ (A.J+SF) \quad (5)$$

Ability to adapt to each other is mainly affected by the ability of members and the factor of technology matching. In addition, it also can be influenced by the government and inherent constraints. Shown as:

$$D.K=D.J+DT*\ (N+TF+M-A.J) \quad (6)$$

The core logistics capabilities can be influenced by the decision of government. in addition, it can be also affected by the strategic adjustment, and the ability of members. Shown as:

$$E.K=E.J+DT*\ (M+N+GF) \quad (7)$$

4 SIMULATIONS

4.1 Realization of the Simulation

It used EXCEL to achieve the model simulation.

The initial value of variable represents the initial state of the system. According to the actual input, these variables will be simulated by iterative changes to future operating conditions of dynamic alliance of logistics. The flow diagram of system dynamics, which used in the characteristic parameters of the reaction system, should be depended on specific characteristics of dynamic alliance of logistics in the simulation.

After the simulation running, firstly, it was input the initial value of variables. Then, it can be set parameters to simulate the actual situation according to the special. By view of output value and table, future running of the dynamic alliance of logistics can be mastered. After the model data generated, the data generated will be out of the curve form of visual representation.

4.2 Example

It selected the representative data form one particular Union in the task decomposition stage, shown in Table 2, and selected the other parameters for model to do the initial value of iteration. Here, "n" is "10".

Table 2: The table of initial value of each variable in Table.

Variable	A	B	C	D	E
initial value	0	20	50	10	20

The data listed in Table 2 is designed to verify the validity of the model of artificial data. In practice, the representative from government and dynamic alliance of logistics enterprises provided the data and input to the program according to the actual. By the simulation of the data in Table 2, the output of the model can express the relative value of each factor trends. It verified that system dynamics model created whether show the effectiveness of impact of relationships between the task decomposition and the evaluation factors or not. And it also verified whether can achieve the best solution by application of this model. Put the data in Table 2 into EXCEL and get changes in each index. Shown as Figure 2.

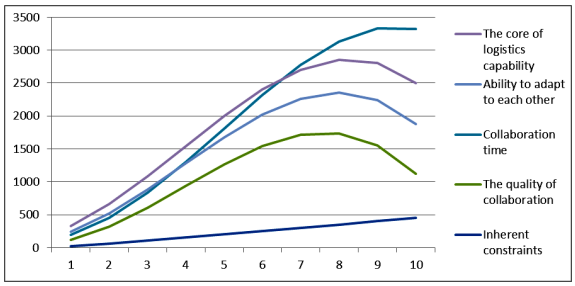


Figure 2: The changes map in each index to model.

The simulation results of the analysis of the figure:

(1) Adding a degree of inherent constraints will bring the improving of the quality of collaboration. However, when the alliance members to a certain amount of time, the quality of collaboration will deteriorate. The ability to adapt to each other also will be bad.

(2) Internal constraints may not necessarily bring about the extension of time collaboration. As

members of the co-ordinated, but to a certain extent, increasing of the number of members will cut down the collaboration time.

By validated, it can illustrate the feasibility of this model. This model can provide some reference value of information to the government in the extent of the emergency task decomposition.

4.3 Significance

Mainly reflected in two aspects:

(1) Some person from emergency management department of government and dynamic alliance of logistics are in charge of discussing to achieve the initial values of variable. Next to simulation, it can be given the optimal extent of decomposition. Then informed of the operational status of the future trend of alliance, it can help the government have more in-depth grasp of the dynamic alliance of logistics.

(2) When the emergency of task decomposition cannot be changed, it can design a set of strategies of different intensity adjustment programs and put it simulate together with the current of initial value from logistics alliance. This has been a different result set. From these results, select a few good according to the logistics alliance needs and the strategic adjustment of the corresponding intensity is the best solution. It can provide a basis for strategic adjustment in emergency management for government.

5 CONCLUSIONS

This text used the method of system dynamics to construct modeling and simulation studies in task decomposition of dynamic alliance of logistics in the supply chain of emergency. Based on analysis of the key elements of the task decomposition of causality in dynamic alliance of logistics, it established a causal flow diagram and system dynamics model. And furthermore, it used random data to achieve the simulation in EXCEL. As can be conclude from the simulation results, the design of the model can express the causal relationship between the key factors of the task decomposition in emergency logistics effectively. The conclusion is in line with the operation of conventional dynamic alliance of logistics.

REFERENCES

Bin Hu et al., 2006. Modeling and Simulation of Corporate Lifecycle Using System Dynamics. *Chinese Journal of Management Science*, 3, 142-147.

Ya-feng Pi., 2006. Study on Spaceflight Model Work Breakdown Structure (WBS). *Journal of North China Institute of Astronautic Engineering*, 16(3), 1-3.

Gang Chen, 1998. Application of Work BreakdownStructure in Boeing. *Aviation Engineerging & Mainienance*, 2, 33-35.

Xin-yue Hu et al., 2005. Partner optimal selection of Virtual Enterprises Based on Work Breakdown Structures. *Manufacturing Automation*, 27(1), 24-27.

Xin-yue Hu.et al., 2005b. Configuration Process of Virtual Enterprises Based on Work Breakdown Structures. *Industrial Engineering Journal*, 8(6), 15-20.

Hong-guo Zhang et al., 2007. Hierarchical network planning method based on WBS for cross-enterprise collaborative project. *Computer Integrated Manufacturing Systems*, 13(3), 513-519.

Ming-guang Zhao, 2010. Analysis of the granularity of task decomposition in enterprise alliance based on system dynamics. *Machinery Design & Manufacture*, 9, 254-256.

An Efficient Technique for Detecting Time-dependent Tactics in Agent Negotiations

Jakub Brzostowski[1] and Ryszard Kowalczyk[2]
[1]*Institute of Mathematics, Silesian University of Technologz, ul. Kaszubska 23, Gliwice, Poland*
[2]*Faculty of Information and Communication Technologies, Swinburne University of Technology, John St, Hawthorn, Australia*
jakub.brzostowski@polsl.pl, rkowalczyk@swin.edu.au

Keywords: Negotiation, Negotiation Tactic, Tactic Detection.

Abstract: The paper proposes an efficient technique for detecting a negotiation strategy used by an opponent during the encounter. It is based on simple transforms that transform the series of offers into a series of values determining the shape of the observed concession curve. It allows for detecting whether the partner is using a time-dependent tactic and what is the specific tactic through determination of the beta parameter used on the side of the negotiation partner. Such information can be further used in choosing a negotiation strategy that can cope with a particular type of the opponent behaviour, and thus improving the negotiation outcomes.

1 INTRODUCTION

Negotiation is process of exchanging offers and counter-offers between parties with conflicting interests that aims at finding a solution satisfying the interests of parties taking part in this interaction (Jennings et al., 2001). There are variety of approaches of learning and reasoning during the negotiation process.

Zeng and Sycara (Zeng and Sycara, 1996) propose a learning approach based on Bayesian updating of beliefs about the environment and negotiation partner. Li and Tesauro (Li and Tesauro, 2003) proposed an approach based on approximate optimization of expected utility using depth-limited combinatorial search and Bayesian updating. The work of Hindriks and Tykhonov (Hindriks and Tykhonov, 2008) proposes to employ Bayesian learning to learn the preference of the negotiation partner assuming a very specific form of the preferences. However, such approaches focus on learning the preferences' structure but not the negotiation strategy. Oliveira and Rocha (Oliveira and Rocha, 2000) propose a framework for multi-issue negotiation between agents where the bid formation is supported by reinforcement learning. This approach is used for both learning from previous interaction and learning from the current encounter.

Work by Nastase (Nastase, 2006) presents a concession curve analysis which is used to predict the negotiation outcomes based on the features of the concession curve. Our approach is suitable for automated negotiation and aims at extracting just one parameter describing the shape of concession curve and its nature is different from the nature of parameters exctracted in the work of Nastase that are suitable in the analysis of negotiations conducted by humans.

Works such as (Oliver, 1997)(Matos et al., 1998)(Gerding and Somefun, 2006) employ evolutionary computing to determine the optimal profile of negotiation strategies. The works by Hou (Hou, 2004) Ren and Zhang (Ren and Zhang, 2007) and the work (Brzostowski, 2007) propose to predict the concession curve using regression analysis. In this work we propose simpler method of prediction based on concession curve transforms which is computationally very cheap. The proposed approach overcomes the problem of wrong estimation of parameters encountered sometimes by regression analysis, and it gives high level of certainty that the time-dependent tactic is used when it is actually used.

The paper is structured as follows. In the second section we recall the concept of decision functions. The third section presents the transforms used to transform the series of partner's offers that is further used to determine the parameter corresponding to the shape of concession curve. The fourth section presents an evaluating experiment allowing for validation of the proposed technique. The fifth section presents conclusions.

2 DECISION FUNCTIONS

In this work we will consider the acceptance region in a form of interval containing real numbers. Multiple attributes will be consider in further work. The negotiation agent can use a decision function for generating offers that it is going to propose. The decision function is a function mapping a time point into the value of offer. The time point corresponds to the current negotiation moment. The decision function may be dependent on different types of parameters and values. Among them there can be negotiation deadline, the borders of acceptance range and other formal descriptions of agent preferences. There are variety of ways of implementing negotiation strategy in a form of decision function. Faratin (Faratin et al., 1998) proposed different types of tactics which can be used to generate negotiation behaviour. Three types of tactics were proposed in his approach, namely: the time-dependent tactics, the behaviour-dependent tactics and resource-dependent tactics. This three types of tactics are called pure tactics. In this work we are only interested in the prediction of time-dependent tactic.

The objective of an agent is to reach an agreement with the negotiation partner in the time range $[0, t_{max}]$ (t_{max} - deadline). The tactic allows for generation of offer in each time point of this range. If before proposing the offer x an agent receives from the counterpart an offer y exceeding the value of x (in terms of utility value) then the agent accepts y. The time-dependent tactic is constructed in such a way that during the whole encounter an agent will concede up to the reservation value when it meets deadline. If an agent a using that type of tactic wants to propose an offer $x^t_{a\to b}$ for the issue j at time t ($0 \le t \le t_{max}$) then that offer can be generated in the following way (Faratin et al., 1998):

$$x^t_{b\to a}[j] = \begin{cases} \min^a_j + \alpha^a_j(t)(\max^a_j - \min^a_j) & \text{if } U^a_j \text{ is decreasing} \\ \min^a_j + (1-\alpha^a_j(t))(\max^a_j - \min^a_j) & \text{if } U^a_j \text{ is increasing} \end{cases}$$

where min^a_j and max^a_j are the boundaries of the acceptance range of the issue j of the agent a. The function $\alpha^a_j(t)$ is a function defined over time giving values in the interval $[0,1]$ ($0 \le \alpha^a_j(t) \le 1$) that can be further rescaled to fit the space in which the agent is conceding. Faratin (Faratin et al., 1998) proposed two families of functions used to implement the time-dependent tactic, namely, polynomial and exponential as follows:

- **Polynomial:** $\alpha^a_j(t) = k^a_j + (1-k^a_j)(\frac{min(t,t^a_{max})}{t^a_{max}})^{\frac{1}{\beta}}$
- **Exponential:** $\alpha^a_j(t) = e^{(1-\frac{min(t,t^a_{max})}{t^a_{max}})^{\beta} ln k^a_j}$

where k^a_j is the initial concession of agent a and β specifies the way of conceding (shape of concession curve).

3 TIME-DEPENDENT TACTICS TRANSFORMS

Let us consider a function transform of the following form:

$$F^1_\beta(f)(x) = \frac{Log(\frac{f(x)-f(0)}{f(t_e)-f(0)})}{Log(x)-Log(t_e)} \quad (1)$$

where $t_e \in (0,x)$. We will prove that this transform can transform the polynomial decision function into very simple function which is constant and equals to β value.

Theorem 1. *Let the function f be defined in the form of polynomial decision function:*

$$f(t) = min^a_j + (k^a_j + (1-k^a_j)(\frac{min(t,t^a_{max})}{t^a_{max}})^{\frac{1}{\beta}})(max^a_j - min^a_j)$$

The transform F^1_β transforms the function f into constant function which equals $\frac{1}{\beta}$ for all values of the domain ($x \in [0, t^a_{max}]$).

Proof. $f(0) = \min^a_j + k^a_j(\max^a_j - \min^a_j)$

$$\begin{aligned} F^1_\beta(f)(x) &= \frac{Log(\frac{(1-k^a_j)(\frac{min(x,t^a_{max})}{t^a_{max}})^{\frac{1}{\beta}})(\max^a_j - \min^a_j)}{(1-k^a_j)(\frac{min(t_e,t^a_{max})}{t^a_{max}})^{\frac{1}{\beta}})(\max^a_j - \min^a_j)})}{Log(x)-Log(t_e)} = \\ &= \frac{Log(\frac{(\frac{min(x,t^a_{max})}{t^a_{max}})^{\frac{1}{\beta}})}{(\frac{min(t_e,t^a_{max})}{t^a_{max}})^{\frac{1}{\beta}})})}{Log(x)-Log(t_e)} \end{aligned}$$

We make a simplifying assumption that x does not exceed t^a_{max}, then:

$$\begin{aligned} F^1_\beta(f)(x) &= \frac{\frac{1}{\beta}Log(\frac{x}{t_e})}{Log(x)-Log(t_e)} \\ &= \frac{1}{\beta} \end{aligned}$$

□

Let us now consider a transform of the following form:

$$F^2_\beta(f)(x) = \frac{\frac{x\partial f(x)-f(0)}{\partial x}}{f(x)-f(0)} \quad (2)$$

we will prove that this transform acts similarly to the transform F^1_β.

Theorem 2. *Let the function f be defined in the form of polynomial decision function:*

$$f(t) = min_j^a + (k_j^a + (1-k_j^a)(\frac{min(t,t_{max}^a)}{t_{max}^a})^{\frac{1}{\beta}})(max_j^a - min_j^a)$$

The transform F_β^2 transforms the function f into constant function which equals $\frac{1}{\beta}$ for all values of the domain ($x \in [0, t_{max}^a]$).

Proof.

$$F_\beta^2(f)(x) = \frac{x\frac{\partial(1-k_j^a)(\frac{min(x,t_{max}^a)}{t_{max}^a})^{\frac{1}{\beta}}}{\partial x}}{(1-k_j^a)(\frac{min(x,t_{max}^a)}{t_{max}^a})^{\frac{1}{\beta}}}$$

Let us further assume that $x \in [0, t_{max}^a]$ then

$$F_\beta^2(f)(x) = \frac{\frac{1}{\beta}\frac{x}{t_{max}^a}(\frac{x}{t_{max}^a})^{\frac{1}{\beta}-1}}{(\frac{x}{t_{max}^a})^{\frac{1}{\beta}}} = \frac{1}{\beta}$$

□

As we have shown in the case of time-dependent tactic there exist transforms that allow to determine the value of β parameter when the agent is using time-dependent tactic generated with the use of polynomial decision function. However, the transforms work for continuous functions. Therefore, we form linearly interpolated function $g(x)$ from the concession curve and then we transform the obtained function into functions h^1 and h^2 using two transformation methods. In the next step we sample the transforms in hypothetical points densly selected from the domain of transforms. The obtained series h_i^1 and h_i^2 are averaged to approxiamate the value of β and the standard deviations for series are computed. The deviations are used to determine the level of certainty that the polynomial decision function was used by the predicted partner.

4 EVALUATING EXPERIMENT

We run 25 negotiations for different values of β parameters for both parties using the polynomial decision function. For a fixed value of deadline (common for both parties), aspiration levels and reservation values we run negotiations for differing values of β parameter (five possible values for both parties). We set up the experiment in the following way: For the first party (party a with one issue):

$$min^a = 15 \quad max^a = 25 \quad t_{max}^a = 20$$

For the second party (party b with one issue):

$$min^b = 10 \quad max^a = 20 \quad t_{max}^b = 20$$

As shown in the Table 1 the estimations of the value

Table 1: The values of β estimated from the view point of the first party for the second party using the mean value of series h_i^1.

$\frac{1}{\beta}(a)$ / $\frac{1}{\beta}(b)$	0.1	0.5	1	2	10
0.1	0.0986	0.0996	0.0997	0.0998	0.0998
0.5	0.4990	0.4993	0.4994	0.4995	0.4995
1	1	1	1	1	1
2	2.004	2.003	2.003	2.003	2.002
10	10.1274	10.1274	10.1274	10.1159	10.1159

Table 2: The values of standard deviations for series h_i^1 for different negotiation scenarios analogous to the first Table.

$\frac{1}{\beta}(a)$ / $\frac{1}{\beta}(b)$	0.1	0.5	1	2	10
0.1	0.0006	0.0005	0.0004	0.0003	0.0003
0.2	0.0014	0.0012	0.0011	0.0010	0.0009
1	0	0	0	0	0
5	0.0088	0.0086	0.0083	0.0081	0.0079
10	0.311	0.311	0.311	0.307	0.307

β parameter determined with the transform approach are quite precise. This means that it is possible to determine the strategy that agent b used using simple method of sequence of offers transformations. The next Table (2) presents the standard deviations (namely how the estimated value of β deviates from the mean value of β over the negotiation scenario). For all β values the standard deviations are close to zero. Low standard deviations means that we have high degree of confidence that the estimated values of β are close to the actual values of β. One exception is the value of β equal to 10 where the standard deviation is around 0.311. Therefore, the certainty that the β value 10 was used is lower. The reason for lower certainty in this case is the shape of concession curve which is quite flat up to the solving negotiation round. The Table 3 presents the estimations of β value for the second type of transform in analogous way as the first Table. As we can see the results are similar; the estimations are close to the actual value of β used by the counterpart. However, as we can see in

Table 3: The values of β estimated from the view point of the first party for the second party using the mean value of series h_i^2.

$\frac{1}{\beta}(a)$ / $\frac{1}{\beta}(b)$	0.1	0.5	1	2	10
0.1	0.0996	0.0995	0.0995	0.0996	0.0999
0.5	0.4987	0.4988	0.4989	0.4989	0.49986
1	1	1	1	1	1
2	2.0082	2.0080	2.0078	2.0073	1.9919
10	10.2397	10.2397	10.2397	9.97882	9.97882

Table 4: The values of standard deviations for series h_t^2 for different negotiation scenarios analogous to the third Table.

$\frac{1}{\beta}(b)$ \ $\frac{1}{\beta}(a)$	0.1	0.5	1	2	10
0.1	0.0094	0.0054	0.0045	0.0040	0.0038
0.2	0.0152	0.01266	0.01179	0.0110	0.0107
1	0	0	0	0	0
2	0.0093	0.0906	0.880	0.0858	0.09393
10	4.6394	4.6394	4.6394	4.6499	4.6499

Table 5: The values of estimated β parameters by the use of nonlinear regression analysis.

$\frac{1}{\beta}(b)$ \ $\frac{1}{\beta}(a)$	0.1	0.5	1	2	10
0.1	0.388	0.0297	0.0331	0.3524	0.03727
0.2	0.4987	0.4989	0.4990	0.4991	0.4992
1	1	1	1	1	1
2	2	2	2	2	2
10	10	10	10	10	10

Table 4 the standard deviations for the β value equal to 10 are quite high (around 4.6394). Similarly, as in the case of first transform the reason for that is the flatness of the concession curve generated using the β value equal to 10.

As we can see in the Table 5 the values of β estimated with the use of non-linear regression analysis are very precise except for small values of $\frac{1}{\beta}$. The reason for this is that the regression algorithm gets stucked in the local minimum while estimating β value. That may happen for sharp values of β parameters such as 0.1 That is were the method based on transforms outperforms the regression-based approach. Low number of data causes the regression algorithm to obtain wrong estimations. As we can see in Table 6 the values of estimated variance are very close to 0 for all estimated values of β which means the result of regression analysis may be quite misleading when the algorithm gets stucked in local minimum. Such a result is obtained in the first row (when estimating the value 0.1). The value of estimated variance indicates how certain we are that the polynomial time-dependent tactic was used. The method based on transforms manages to estimate the value of β quite precisely even if the certainty (standard deviation) that the polynomial time-dependent tactic was used is not very high.

Table 6: The values of estimated variance (approximations) obtained by the regression algorithm when estimating the values of β.

$\frac{1}{\beta}(b)$ \ $\frac{1}{\beta}(a)$	0.1	0.5	1	2	10
0.1	0	0.000158	0.000252	0.000311	0.000377
0.2	0	0	0	0	0
1	0	0	0	0	0
2	0	0	0	0	0
10	0	0	0	0	0

5 CONCLUSIONS

We proposed a novel approach for detecting the time-dependent tactic used by the negotiation partner. We use simple transforms to transform the series of offers into a series of values indicating what value of β parameter is used on the side of the negotiation partner. Using this method we are able to determine if the partner is using time-dependent tactics. Moreover, we are able to determine the β parameter used by partner. Such an approach may be further used to choose a negotiation strategy that can cope with a particular type of behaviour.

REFERENCES

Brzostowski, J. (2007). *Predictive decision-making mechanisms based on off-line and on-line reasoning*. PhD thesis, Swinburne University of Technology.

Faratin, P., Sierra, C., and Jennings, N. R. (1998). *Negotiation among groups of autonomous computational agents*. University of Londond.

Gerding, E. H. and Somefun, D. J. A. (2006). Multi-attribute bilateral bargaining in one-to-many setting. In *Agent Mediated Electronic Commerce VI*.

Hindriks, K. and Tykhonov, D. (2008). Opponent modelling in automated multi-issue negotiation using bayesian learning. In *AAMAS '08: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, pages 331–338, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Hou, C. (2004). Modelling agents behaviour in automated negotiation. Technical Report KMI-TR-144, Knowledge Media Institute, The open University, Milton Keynes, UK.

Jennings, N. R., Faratin, P., Lomuscio, A., Parson, S., and Wooldridge, C. S. M. (2001). Automated negotiation: Prospects, methods and challenges. *International Journal of Group Decision and Negotiation*, 10(2):199–215.

Li, C. and Tesauro, G. (2003). A strategic decision model for multi-attribute bilateral negotiation withalternating offers. In *Proceedings*. ACM.

Matos, N., Sierra, C., and Jennings, N. (1998). Determining successful negotiation strategies: An evolutionary approach. In *ICMAS '98: Proceedings of the 3rd International Conference on Multi AgentSystems*, page 182, Washington, DC, USA. IEEE Computer Society.

Nastase, V. (2006). Concession curve analysis for inspire negotiations. *Group Decision and Negotiation*, 15:185–193.

Oliveira, E. and Rocha, A. P. (2000). Agents advances features for negotiation in electronic commerce and virtualorganisations formation process. In Dignum, F. and Sierra, C., editors, *Agent Mediated Electronic Commerce, the European AgentLink Perspective.*, volume 1991, pages 77–96. Springer-Verlag.

Oliver, J. (1997). A machine learning approach to automated negotiation and prospects for electroniccommerce.

Ren, F. and Zhang, M. (2007). Prediction of partners' behaviors in agent negotiation under open and dynamic environments. In *Proceedings of the 2007 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology - Workshops*, pages 379–382, Washington, DC, USA. IEEE Computer Society.

Zeng, D. and Sycara, K. (1996). Bayesian learning in negotiation. In Sen, S., editor, *Working Notes for the AAAI Symposium on Adaptation, Co-evolution and Learningin Multiagent Systems*, pages 99–104, Stanford University, CA, USA.

Fuzzy Classifier for Church Cyrillic Handwritten Characters

Cveta Martinovska[1], Igor Nedelkovski[2], Mimoza Klekovska[2] and Dragan Kaevski[3]
[1]*Computer Science Faculty, University Goce Delcev, Tosho Arsov 14, Stip, R. Macedonia*
[2]*Faculty of Technical Sciences, University St Kliment Ohridski, Ivo Ribar Lola bb, Bitola, R. Macedonia*
[3]*Faculty of Electrical Engineering and Information Technologies, University St Cyril and Methodius, Rugjer Boshkovik bb Skopje, R. Macedonia*
cveta.martinovska@ugd.edu.mk, igor.nedelkovski@uklo.edu.mk, mimiklek@yahoo.com, d.kaevski@gmail.com

Keywords: Handwritten Character Recognition, Historical Manuscripts Recognition, Fuzzy Decision Techniques, Feature Extraction, Recognition Accuracy and Precision.

Abstract: This paper presents a fuzzy methodology for classification of Old Slavic Cyrillic handwritten characters. The main idea is that the most discriminative features are extracted from the outer character segments defined by intersections. Prototype classes are formed using fuzzy aggregation techniques applied over the fuzzy rules that constitute the descriptions of the characters. Recognition methods use features like number and position of spots in outer segments, compactness, symmetry, beams and columns to assign a pattern to a prototype class. The accuracy and precision of the fuzzy classifier are evaluated experimentally. This fuzzy recognition system is applicable to a large collection of Old Church Slavic Cyrillic manuscripts.

1 INTRODUCTION

Recognition of handwritten characters has been a subject of intensive research in the last 20 years (Arica and Yarman-Vural, 2001); (Vinciarelli, 2002). Different approaches for developing handwritten character recognition systems are proposed, like Fuzzy Logic (Malaviya and Peters, 2000); (Ranawana et al., 2004), Neural Networks (Zhang, 2000) and Genetic Algorithms (Kim and Kim, 2000).

This paper describes a character recognition system developed for digitalization of a large Old Cyrillic manuscripts collection found in Macedonian churches and monasteries. This process cannot be performed using the existing computer software due to the specific properties of Old Slavic characters.

A novel classification methodology based on the fuzzy descriptions of characters is proposed. Number and position of spots, beams and columns that appear in the outer segments of the topological character map are considered as significant features. This character recognition system is applicable to a large historical collection of manuscripts that originate from various periods and locations. The manuscripts used for church liturgical purposes are unaffected by style changes. They are written in Constitutional Script. This Script looks like printed text, where character contour lines can be easily extracted.

2 CHARACTER ANALYSIS AND FEATURE EXTRACTION

Manuscripts are converted to black and white bitmaps. The first step of processing is extracting the characters using contour following function (Fig. 1). Visual prototype of a normalized character is analyzed to determine character features and their membership functions. Several features are examined, such as compactness, x-y symmetry, presence of beams and columns in three horizontal and vertical segments and number of spots in outer segments.

According to visual features, the characters of the Church Slavic alphabet can be grouped in several subsets. There is a subset whose members are Г, В and Б that have emphasized vertical lines on the left-side or left column. Another subset contains characters such as П and Ш that have a right-side and left-side column. The third subset consists of characters like П, Г and Б that have noticeable horizontal line in the upper segment (upper beam). The fourth subset consisting of characters as Ш and

ꙋ has horizontal line in the bottom segment (bottom beam).

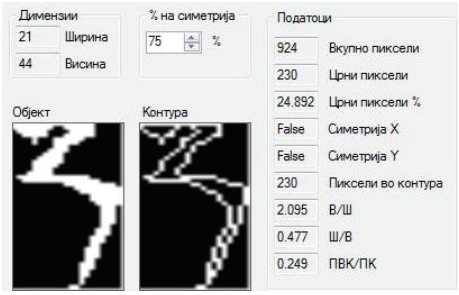


Figure 1: Extracting a character contour.

These features are illustrated in Fig. 2. Particular character can be a member of several of these subsets.

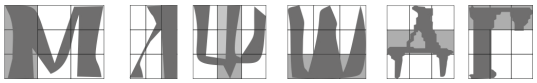


Figure 2: Vertical lines and horizontal lines of characters.

The character can be intersected in such a way that 6 segments are formed. Four outer segments provide useful information for the proposed character recognition system (Fig. 3).

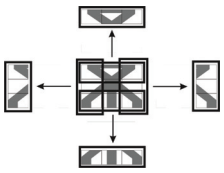


Figure 3: Intersections of the characters that form the upper, down, left and right segments.

Visual prototype of a character is formed applying fuzzy intersection and fuzzy union operators over a set of character samples (Fig. 4).

a) Fuzzy Intersection b) Fuzzy Union

Figure 4: a) Fuzzy intersection and b) fuzzy union.

3 FUZZY CLASSIFIER

Fuzzy classifier for Church Slavic characters is based on character prototypes created in the form of fuzzy linguistic rules. Fuzziness emerges from the fact that texts are written by individuals with different ways of writing and manuscripts originate from different historical periods characterized with certain styles of writing.

3.1 Operators for Fuzzy Aggregation

The precision of the character recognition system to a certain extent depends on the proper selection of features. This is done by calculating the overall measure for the features applying the fuzzy aggregation techniques. The general fuzzy membership function μ_G that combines the fuzzy information $(\mu_1, \mu_2, \dots, \mu_N)$ for the character features can be represented as:

$$\mu_G = Agg(\mu_1, \mu_2, \dots, \mu_N) \quad (1)$$

where Agg is a fuzzy aggregation operator.

This approach uses operators defined by Yager (Yager, 1990) for the union and for the calculation of weighted median aggregation.

Let $w_1, w_2, \dots, w_N$ represent weights associated with fuzzy sets $A_1, A_2, \dots, A_N$. Yager defines the union using the following formula:

$$U(a_1, a_2, \dots, a_N) = \min\left\{1, \left(\sum_{i=1}^{N} (a_i)^{\alpha}\right)^{\frac{1}{\alpha}}\right\} \quad (2)$$

where α is a real non-zero number and the value that can be obtained as a result of the union ranges between 1 and $\min(a_1, a_2, \dots, a_N)$.

The weighted median aggregation is defined by the following formula (Malaviya and Peters, 1995):

$$Med(a_1, \dots, a_N, w_1, \dots, w_N) = \left(\sum_{i=1}^{N} (w_i a_i)^{\alpha}\right)^{\frac{1}{\alpha}} \quad (3)$$

where $\sum_{i=1}^{N} w_i = 1$ and α is a real non-zero number with values between $\max(a_1, a_2, \dots, a_N)$ and $\min(a_1, a_2, \dots, a_N)$.

3.2 Fuzzy Descriptions of Characters

The Church Slavic character recognition system operates in two working regimes: building the prototypes and character recognition. The first regime creates a matrix of combined characteristics. Using this matrix, fuzzy rules are generated in the form of linguistic descriptions of the characters.

The rules contain only the features that are relevant for the character classification and identification. For example, the character “B“ is described by the following combination of features: two vertical holes, x symmetry, left column, one spot in the left segment, one spot in the upper segment, two spots in the right segment, and one

spot in the lower segment. In the fuzzy description of this character less significant features are upper beam and lower beam.

Let the number of significant features for the particular character is S and the number of segments is C. The importance of every feature in the aggregation process is represented by a certain weight. The input values in the system are the features of the character segments for which the fuzzy values are calculated.

Generally, the input matrix with the character features has dimension C x G, where G is the total number of features that can be of structural nature (symmetry, compactness, number of spots) or to denote position:

$$I = \{i_{cg} | c = [1, C], g = [1, G]\} \quad (4)$$

From the elements of the above matrix, a K_j matrix with dimensions p x C can be formed for each segment, where p is a number of significant features for the segment and $j = 1, \dots, C;\ p \leq G$.

$$K_j = \{\bar{k}_{p_1 j}, \bar{k}_{p_2 j}, \dots, \bar{k}_{pcj}\} \quad (5)$$

Using the weighted median aggregation operator, the feature vector for a particular character is calculated

$$\mu_j = \mathrm{Med}(a_1, \dots, a_N, w_1, \dots, w_N) \quad (6)$$

Then, using the union operator, the most significant features from the list of features obtained in the previous step, are selected:

$$\mu_p = \min\left\{1, (\sum \mu_{pj})\right\} \quad (7)$$

Weight matrices implicitly represent fuzzy rules that describe the character prototypes. The weights are obtained by statistical calculations from the training samples. The number of appearances of a particular feature is measured for every character. Higher frequency of a feature decreases its recognition importance. Smaller weights are assigned to more frequent and hence less important features.

Weight matrices are used to reduce the number of features that are considered for each character. Different features are considered at each step in the recognition phase and character prototypes that possess these features are activated. Finally, only the most similar character prototype is winner.

4 CHARACTER RECOGNITION

Procedure for character recognition consists of several steps: 1. Determining the membership functions for the global features of the unknown symbol. 2. Calculating the membership functions of an unknown symbol for all the prototypes according to formula:

$$\mu_n = \frac{\sum_{c=1}^{C} w_c . \mu_c}{C} \quad n = 1, .. N \quad (8)$$

3. Selecting the possible prototypes that are most similar to the unknown character, following the formula:

$$\mu_A = \bigcup_{n=1}^{N} \mu_n \quad (9)$$

The result of the classification process is a list of prototypes that have the most similar features with the features of the unknown character.

5 EXPERIMENTAL RESULTS

Several experiments are performed to test the performance of the proposed fuzzy classifier. Table 1 shows the recall and the precision measures for each character. Recall (R) is computed as a fraction of the number of retrieved correct characters divided by the total number of relevant characters:

$$R = TP/(TP + FN) \quad (10)$$

Precision (P) is computed as a fraction of the number of retrieved correct characters, divided with the number of retrieved characters:

$$P = TP/(TP + FP) \quad (11)$$

In formulas (10) and (11), the TP (True Positive) is the number of correctly predicted examples, FP (False Positive) is the number of negative examples wrongly predicted as positive, and FN (False Negative) is the number of positive examples wrongly predicted as negative. The sum of precision and recall i.e. F1 metric is computed as

$$F1 = 2RP/(R + P) \quad (12)$$

The proposed fuzzy classifier recognizes the characters with an average recall of 0.69, average precision of 0.72 and an overall average measure of precision and recall F1 of 0.70.

6 CONCLUSIONS

In this paper a novel methodology for recognition of Old Slavic Cyrillic handwritten characters based on fuzzy prototypes is described. Fuzzy descriptions of

the characters are represented as fuzzy rules. Fuzzy aggregation techniques are used to combine different character features, such as number and position of spots in outer segments, compactness, symmetry, beams and columns.

Table 1: Precision and recall of the fuzzy classifier.

	Number of characters	recall	precision
Aa	10	0.4	1
b	7	0.67	0.67
v	6	0.83	1
g	10	1	0.58
d	12	0.75	0.56
e	7	0.43	1
/	5	0.4	1
\	6	0.67	0.36
z	4	0.25	0.5
J	12	0.83	1
i	5	0.4	1
k	1	1	o.5
l	10	0.9	0.75
m	4	0.25	1
n	11	1	0.73
o	8	1	0.61
p	4	0.5	0.67
r	5	0.2	1
s	9	0.89	0.73
t	7	1	0.78
U	4	0.75	1
f	7	1	0.87
H	6	0.67	0.8
h	9	0.28	0.67
w	5	0	0
]	7	1	1
c	11	0.91	0.58
;	10	0.8	0.89
[	7	0.86	0.75
q	14	0.86	0.63
Q	9	0.67	0.67
2	9	1	0.9
`	5	0.2	1
1	1	1	1
5	2	1	0.67
3	5	0	0
u	3	1	0.43
Total	257	0.69	0.72

Higher weights are assigned to features that are more discriminative. For example, three spots right/left/up or down and two holes are the most indicative for the recognition process.

The accuracy and precision of the proposed fuzzy classifier are acceptable and motivational for future work and improvement.

Presented experimental results of this visual methodology are comparable to the recognition of the human visual system. Characters that are misclassified are also unrecognizable for the humans. Besides the fuzzy classifier a decision tree classifier is designed. The recognition results of the two classifiers are comparable. Both classifiers use the same set of discriminative features.

For future work a combination of these two classifiers is planned to achieve more accurate and precise recognition of the Old Slavic Cyrillic characters.

REFERENCES

Arica, N., Yarman-Vural, F. T., 2001. An Overview of Character Recognition Focused on Off-Line Handwriting. *IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications and Rev.*, vol. 31, no. 2, pp. 216-233.

Kim G., Kim S., 2000. Feature Selection using Genetic Algorithms for Handwritten Character Recognition, *In: L.R.B. Schomaker and L.G. Vuurpijl (Eds.), Proc. of the 7th Int. Workshop on Frontiers in Handwriting Recognition,* pp 103-112.

Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On Combining Classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3.

Malaviya A., Peters L., 1995. Extracting Meaningful Handwriting Features with Fuzzy Aggregation Method, *Proc. of the 3rd Int. Conf. on Document Analysis and Recognition*, Montreal, pp. 841-844

Malaviya A., Peters L., 2000. Fuzzy Handwritten Description Language: FOHDEL, *Pattern Recognition*, 33, pp. 119-131.

Ranawana, R., Palade V., Bandara, GEMDC, 2004. An efficient Fuzzy method for Handwritten Character Recognition, *In M.Gh. Negoita et al. (eds.), KES 2004, LNAI 3214*, Springer-Verlag, pp.698-707.

Vinciarelli, A., A Survey on Off-line Cursive Word Recognition, 2002. *Pattern Recognition* 35, pp.1433-1446.

Yager, R., 1990. On the Representation of Multi-Agent Aggregation using Fuzzy Logic, *Cybernetics and Systems* 21, pp.575-590.

Zhang, G. P., 2000. Neural Networks for Classification: A Survey. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 30 no.4, pp.451-462.

Multiagent Model of Stabilizing of Petroleum Products Market

Leonid Galchinsky
Department of Management and Marketing, National Technical Universiry of Ukrain, Peremohy 37 Street, Kyiv, Ukrain
hleonid@gmail.com

Keywords: Multi-Agent Models, Oligopolistic Market, Asymmetry in Prices, Tacit Collusion, Petroleum Products, Price Stabilization.

Abstract: The problem of developing of multi-agent models of stability in market prices of petroleum products is presented. The problem of price stabilization occurs due to external factors: the result of sudden changes of crude oil on world markets or exchange rate changes. In addition the market price dynamics is also influenced by internal factors such tacit collusion sellers. Shown the theoretical possibility to reduce of asymmetry in prices through the use stabilization fund of petroleum products, which the public body can use at the moment when there is a of price jump through the sale of petroleum products by stable prices.

1 INTRODUCTION

The behavior of fuel prices has a global impact on the whole economy of a particular country. Increases in fuel prices automatically lead to higher prices of commodities with high demand and transportation services. The result is the decrease in purchasing power and reduction in profitability of companies, especially those with energy-intensive production.

The expenses for petroleum products are involved in the consumer market prices; transportation costs also affect the prices of all goods of consumer market.

This question is particularly important in emerging economies, especially in such countries as Ukraine, where practically immediate reaction of all industries to changing prices occurs. This factor affects not only the economy, but also the social situation of the general public and political processes in it.

The modern market of oil products in Ukraine is characterized by the large number of economic entities, acting alone or co-operating, in conditions dissimilar to classical equilibrium markets. In this market the main sources of equilibrium disturbance are external factors, primarily world prices of crude oil and exchange rates. Due to non-stationarity of these factors and cooperative actions of market agents, prices of petroleum products, including retail gasoline prices, are changing daily.

Retail gasoline prices in Ukraine depend on many factors; the main ones are the national currency fluctuations, changes in world oil prices, the activities of oil producing and refining companies, oil traders, government policy etc. Thus, the problem of finding the mechanism for stabilizing oil prices arises. The ways of price stabilization – from direct administrative methods to the market-based approaches – have long been known. This paper deals with the mechanism for smoothing oil price shocks through targeted interventions of oil products, provided by the state, in moments of disturbance in fuel prices threatening to destabilize the market.

2 RELATED WORKS

The intensive research of price dynamics in the oil market as well as research of multi-agent approach to modeling price competition in oligopolistic markets was held over the last 20 years. The asymmetry of prices for petroleum markets in different countries was studied in (Bacon, 1991), (Borenstein et al., 1992); (Matt Lewis, 2003), (Veremenko and Galchinsky, 2010); (García, 2010).

In (Kephart et al., 2000); (Tsvesovat and Carley, 2002); (Happenstall et al., 2004); (Levin et al., 2009); (Ramezani et al., 2011) the possibilities and properties of applying multi-agent approach to modeling the competition in oligopolistic markets were explored.

3 MODEL

In oligopolistic markets, the decisions of each firm don't only affect their own profit but also the profit of their competitors. Therefore, firms react to the actions of their competitors and in every decision the companies consider not only the direct impact on their income, but also the reaction effects of competitors. This so-called oligopolistic interdependence lays the foundation in modeling the market behavior as a multi-agent system. There are several reasons for choosing the multi-agent approach, although the game theory was about to be chosen as the theoretical basis. However, for games with more than two players the results of the game theory approach are far from building a constructive design scheme. Even in games with no coalitions there is no exact algorithm for finding equilibrium in general, because it is very difficult to consider the real constraints on the strategy of all players analytically. For coalition games claim the existence of equilibrium was not even proven, so we will find the solution of the problem in another way, with the agent modeling method.

Let us determine the following factors in the model:

- Consumer - a vehicle with the driver. It is characterized by the type of fuel being used and fuel tanks capacity, the use of fuel per 100 km, the frequency and range of travel, the propensity to traveling and saving money.
- Gas Station - a gas station that provides services to consumers and the companies, which buy fuel for their vehicles. It is characterized by the volume of containers for storage, type of fuel, its availability, and geographical location.
- Refinery station, which is characterized by type of fuel it produces, volumes of containers for storage, fuel prices.
- Country is an agent that displays activity of the state and sets a number of rules for the market functioning and import-export operations.
- Trader is a mediator between refineries and gas stations. Sells fuel in bulk, making transportation to the appropriate object. Characterized by means of transportation and storage facilities for fuel.

The environment also holds information about the concentration and location of agents in the country, the transport grid, grid with railroad connections.

Each agent has its own program behavior based on finite-state machines, which describes its condition and the conditions of transition from one state to another.

Each agent can communicate with any other agent through the messaging mechanism. Thus the «consumer», that is within visibility range of certain agent of a «station» will be able to receive notice of the price on its fuel. Similarly «station» agents will be able to receive data available in the region traders and their prices. Also, each agent has a specific set of actions with which he manipulates the state of the environment. For example, for the «consumer» agents they are: go (move around the environment), refuel and wait. In case of failure of any agent to act in the market (the agent goes bankrupt) he is removed from the model. Similarly, agents may also enter the model. Inputs for the model are:

$$\{M\ ,\ PZ\ ,\ PN\ ,\ LOC\ ,\ S\ \},$$

where

PZ_t^m - For purchases of fuel by network S in t time;

PN^m - The original retail price of network m;

LOC_k^m - The location of station k of network m;

$S_{i,j}$ - Number of consumers of fuel in the square with coordinates (i,j);

M - The number of retail networks;

The main mechanism for the distribution of fuel consumed is the function of demand, taking into account not only for a particular network, but also the maximum possible demand.

$$D = \begin{cases} D_{\max} \cdot N_{AZS}^i, \\ A - B p_i + C \sum_{j \neq i} p_j \end{cases}$$

The model of agents' behavior relies on rule-based algorithm, proposed in [1]. Variables and logical conditions were added in the implemented algorithm to model collusion between the agents. The collusion is valid until significant changes happen in the agent's input parameters.
In account of this it is possible to make an algorithm for the agent:

1. Set the price specified in the preceding period
2. Collect data for neighbors
3. Get prices for fuel
4. Get on the environment of consumers for the current period
5. Determine the cost of 1 liter fuel, taking the fixed costs into account
6. Forecast fuel demand, given the cost of fuel, the current price and the price of neighboring agents to forecast demand for fuel

7. Check messages from neighboring agents for available collusion suggestions.
8. Decide on pricing, using a set of rules.
9. Put the price set in the next period.

The printed form should be completed and signed by one author on behalf of all the other authors, and sent on to the secretariat either by normal mail, e-mail or fax.

4 MARKET SIMULATION

The basis of the algorithm is the set of rules for changing prices, which also contains rules for checking the usefulness of the collusion. The main indicator, appearing in the rules is $P_{\text{int}} = \frac{\sum_{i=1}^{n} \frac{p_i}{l_i}}{\sum_{j=1}^{n} \frac{1}{l_j}}$, where p_i is price of agent i in the neighborhood, and l_i is the distance between this agent and the i-th agent.

The numerical constants for price change rates were determined basing on real data in the studied region and on the characteristics of prices asymmetry. For this purpose the initial values based on expert judgments were taken and then specified through minimizing the residual function with the help of Nelder-Mead method on a set of historical data in Kyiv region for the period 2010-2011.

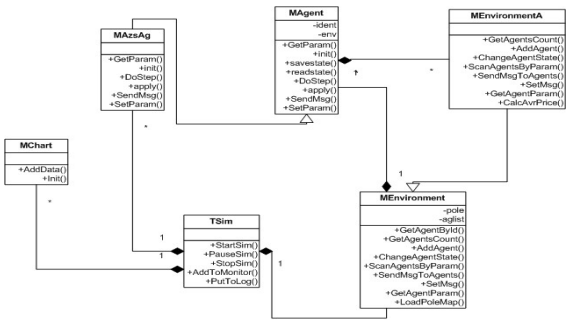


Figure 1: The class diagram in UML notation.

The diagram of classes shows, that the main class which provides the entire program is the class TSim. It is a kind of the experimental abstraction, and it includes instances of the agent model classes. Agent model is represented by two classes: MAzsAg and MEnvironment. According to the paradigm of agent modeling, MAzsAg is a software agent which can receive messages, react to the environment changes and interact with other agents through the environment. MEnvironment class is the agents' environment which provides their identification, messaging and performs a mechanism for interaction between agents and between agent and environment.

5 EXPERIMENTAL RESULTS

Since the agent-based model relies on the interaction between retailing petroleum products networks, it is firstly needed to consider the opportunity for the state to intervene in the retail market in order to prevent collusions between the agents. Thus, the state petroleum retail network can be considered as such regulator. Taking into account, that the oligopolists have significant market shares, the state-owned market share, sufficient for the desired effect on the market, must be determined.

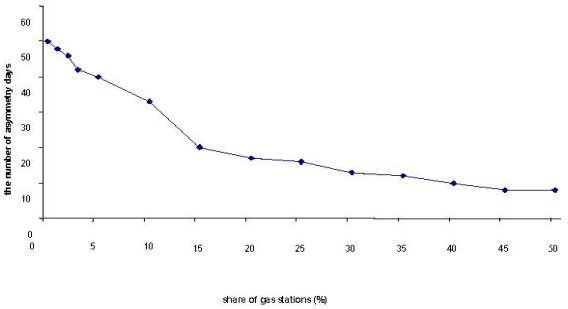


Figure 2: Dependence of the length of the return prices to normal levels of the market share of the state regulator.

As you can see, the effect is noticeable when the market share exceeds 15-20%. Further increase in market share slightly increases this effect. In respect that that the cost of a public network can be quite high (the cost of building a gas station is estimated at 0.5 million.), the regulator may be too expensive. Analysts estimate the total costs could reach up to $2 billion. These costs are currently estimated as too high in order to implement.

Due to the fact that it is difficult to enact the above-mentioned type of controller, government can bring such regulator to the wholesale market. Given that during the jump in prices some retailers do not have enough fuel, the state can sell their stocks to reduce the effects of the shock. Thus, signing contracts with the network stations and having their margin on the sale of petroleum products restricted, is a way to indirectly affect the price situation in the market.

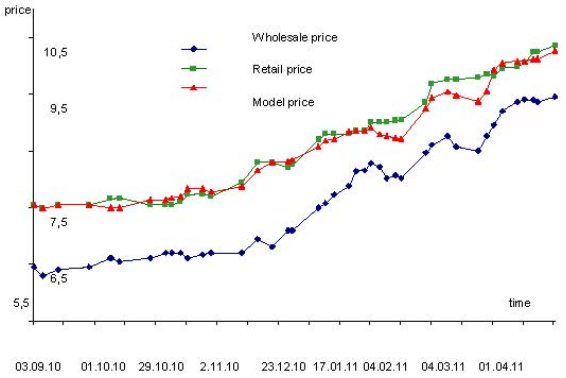


Figure 3: The behavior of the gasoline prices with regulation and without.

So the state can enter the wholesale market with stabilization reserve during the prices' jumps and sell fuel under contracts to station networks, which have demand for fuel. The need for profitability of such fund should be taken into account. To evaluate the effectiveness of control, the scheme, rearranged in Figure 3, can be used. The comparison of price without regulator and with the presence of the regulator clearly indicates the effect of stabilization.

6 CONCLUSIONS

The results indicate that basing on the proposed multi-agent model, the implementation of the regulator, which can effectively reduce the level of asymmetry in oil prices, is possible in principle. The state agency, acting not administratively, but through market-based control methods, might play a role of such regulator. Further research in this area should be aimed at clarifying the mechanism of influence on prices by the state regulator.

REFERENCES

Bacon, R. W., 1991, Rockets and Feathers; the Asymmetrical Speed of Adjustment of UK Retail Gasoline Prices to Cost Changes, *Energy Econ.* 1, pp. 211 – 218.

Borenstein S, Cameron A and Gilbert R, "Do Gasoline Prices Respond Asymmetrically To Crude Oil Price Changes?" *National Bureau of Economic Research*, 1992, Working Paper No. 4138.

Matt Lewis. "Asymmetric Price Adjustment and Consumer Search: An Examination of the Retail Gasoline Market" University of California, Berkeley Department of Economics November 13, 2003.

Veremenko I, Galchinsky L., Modeling the dynamics of retail prices for oil products market of Ukraine, *"Business Inform"* № 1, 2010, pp. 20-26.

Perdiguero García, Jordi, 2010. "Dynamic pricing in the spanish gasoline market: A tacit collusion equilibrium," *Energy Policy,* Elsevier, vol. 38(4), pages 1931-1937, April.

Alison Heppenstall, Andrew Evans and Mark Birkin Using Hybrid Agent-Based Systems to Model Spatially-Influenced Retail Markets *Journal of Artificial Societies and Social Simulation* vol. 9, no. 3.

Tsvetovat, M. and Carley, K., 2002, Emergent Specialisation in a Commodity Market: A Multi-Agent Model, *Computational and Mathematical Organisation Theory*, 8, pp. 221 – 234.

Dynamic pricing by software agents Jeffrey O. Kephart, James E. Hanson, Amy R. Greenwald. *Computer Networks 32 (2000) 731-752.*

S. Ramezani, P.A.N. Bosman, J.A. La Poutré. Adaptive Strategies for Dynamic Pricing Agents. *Proceedings of the 23rd Benelux Conference on Artificial Intelligence*, 423–424, 2011.

Y. Levin, J. McGill, and M. Nediak. Dynamic pricing in the presence of strategic consumers and oligopolistic competition. *Management Science*, 55(1):32–46, 2009.

An Intelligent Transportation System for Accident Risk Index Quantification

Andreas Gregoriades[1], Kyriacos Mouskos[2] and Harris Michail[3]
[1] *Department of Computer Science and Engineering, European University Cyprus, Nicosia, Cyprus*
[2] *Cyprus Transport and Logistics Ltd, Nicosia, Cyprus*
[3] *Department of Electrical Engineering and Information Technology, Cyprus University of Technology, Limassol, Cyprus*
A.gregoriades@euc.ac.cy, mouskosks@gmail.com, harris.michail@cut.ac.cy

Keywords: Bayesian Networks, Dynamic Traffic Assignment, Road Safety.

Abstract: Traffic phenomena are characterized by complexity and uncertainty, hence require sophisticated information management to identify patterns relevant to safety and reliability. Traffic information systems have emerged with the aim to ease traffic congestion and improve road safety. However, assessment of traffic safety and congestion requires significant amount of data which in most cases is not available. This work illustrates an approach that aims to alleviate this problem through the integration of two mature technologies namely, simulation-based Dynamic Traffic Assignment (DTA) and Bayesian Networks (BN). The former generates traffic flow data, utilised by a BN model that quantifies accident risk. Traffic flow data is used to assess the accident risk index per road section and hence, escape from the limitation of traditional approaches that use only accident frequencies to quantify accident risk. The development of the BN model combines historical accident records obtained from the Cyprus police and domain knowledge from road safety.

1 INTRODUCTION

Road safety constitutes a problem of paramount importance worldwide (Bartley, 2008). To deal with this problem, intelligent transportation systems (ITS) have emerged. ITS are also used in the following areas: congestion control, mobility enhancement, delivering environmental benefits, and boosting productivity and expanding economic and employment growth. The work presented herein describes a novel approach and tool for assessing the accident risk index of road networks. This prerequisites the assessment of accident risk. According to (Zheng, 2009), accident risk models are divided into two categories: social risk models, that measure probabilistic (frequentist) collective damage, and individual risk models, that measure probabilistic (frequentist) individual damage. These are categorized into aggregate and disaggregate methods. The former, use global statistics and the former specific events (Bartley 2008). However, predicting accident risk requires not only frequencies of crashes per road section but also traffic flow data. However, in most cases traffic flow and accident data cannot be found together. To that end authorities perform safety analysis using only crash data which is an approximate approach to accident risk estimation. This paper aims to address this problem through the development of a novel Intelligent Traffic Information System (ITIS) that leverages the capabilities of two mature methodologies namely simulation-based Dynamic Traffic Assignment (DTA) embedded in the VISTA simulator (Ziliaskopoulos et al., 1996) and Bayesian Belief Networks (BN). The former is widely used in transportation planning and operations to predict drivers' decisions (where and when to travel on the road network), and in work was used to estimate traffic flow conditions for each road section. The latter is a powerful uncertainty modelling technique used for the quantification of accident risk under varying conditions.

The paper is organised as follows. Next section describes the methodology. Subsequent sections concentrate on data pre-processing and BN model development. The integration of VISTA with the BN along with the results that emerge from the amalgamation of the two technologies in an ITS, is described next. The paper finishes with conclusions.

2 METHODOLOGY

The Road Safety Assessor, ITS system proposed herein is the amalgamation of probabilistic risk assessment with a mesoscopic traffic simulation, namely VISTA. The need for this integration boils down to the limitations of traditional traffic information systems that mainly concentrate of data warehousing. The methodology proposed utilises data marts to generate projections of future system behaviour. To that end, intelligent information management techniques are employed to distil knowledge used to develop models that enable the prospective system behaviour. The two models that emerged from this process are the accident risk assessment model and the traffic simulation model. The accident risk assessment employed is causality-based and uses BN. In BN each node is used to represent a random variable that has been identified to have a causal influence on accident risk. Each directed edge represents an immediate dependence or direct influence between parent and child variables (Jensen, 2001). Evidence is entered in the model through instantiation of leaf node on the model. Inference is achieved by belief propagation through the models topology. BN technology is used to model how traffic and infrastructural factors influence accident risk. The second component of the approach is a road traffic simulator based on DTA. The DTA model is used in VISTA through the Dynamic User Equilibrium (DUE) model (Peeta et al., 2000). The use of DTA model enhances the limitations of existing practices by providing a consistent way of producing estimates of traffic flow conditions of road networks using limited information from traffic flow detectors. Moreover, it produces timely and complete traffic volume estimates for all sections of a road network and hence, can be used to assess accident risk using time varying conditions. The integration of BN with VISTA in the proposed traffic information system enables the dynamic assessment of accident risk using simulated traffic conditions and prior knowledge embedded in the BN. A pilot study conducted with the system aimed to assess the safety performance of the Nicosia road network in Cyprus and to investigate how it will behave under different scenarios.

Initially the road traffic model of Nicosia was specified, implemented, verified and validated in VISTA. Models in VISTA are represented by nodes connected by unidirectional links that represent flow of traffic in one direction. It is possible to have more than one link between two nodes to indicate separate lanes and lane direction. The completed VISTA simulation model was integrated with an accident risk assessor implemented in Java. The simulator provided the risk assessor with the traffic volumes of all road sections of the network for every 15 min interval. Traffic volumes along with infrastructural properties of the network were used by the BN to assess accident risk on a simulation step basis. For the development of the BN topology and the parameterization of its prior knowledge, historical road accident data were utilized.

3 ARCHITECTURE OF THE ITS

The Road Safety Assessor tool emerged from the integration of VISTA with BN technologies. The main components of the tool are: the BN engine, the accident risk assessor, the VISTA simulator, the data pre-processor that incorporates the scenario generator, the results analyzer and the visualizer. The tool was developed using a component-based software engineering methodology. With the initial specification of the system requirements captured, we proceeded in the identification of suitable software components that matched the initial system requirements. These components were subsequently integrated to implement parts of the system's functionality. In particular the Bayesian inference engine and the visualization components were selected after thorough investigation. The glue-code that enabled components integration was implemented in Java. The risk assessor quantifies accident risk using a Bayesian inference engine that utilizes the probabilistic model of accident risks. Input to the BN assessor is categorized into static and dynamic. The former is obtained from the VISTA database and the latter is the output of the VISTA simulation.

Input to the accident risk assessor is organized in the form of scenarios. An input scenario to the BN assessor is defined by the static and dynamic properties of each road section. Static information is obtained from the VISTA database and in combination with the dynamic input from the simulator. This provides the baseline for generating a number of plausible test scenario variations for each road section. Generated scenarios are executed by the risk assessor to quantify the probability of accident. The scenario generator is responsible for generating plausible scenario variations to stress-test the safety performance of each road section. The visualizer processes the results and depicts these to the user graphically. Input scenarios are executed by

the BN model. Each scenario evidence is propagated down the BN topology to produce the posterior probability of accident risk per scenario.

The integration of the VISTA with the BN model was realized through asynchronous data interchange. To establish communication between VISTA and the risk assessor it was imperative to pre-process VISTA's output data prior to being utilized by the BN in the risk assessor. Specifically, VISTA variables are continuous by nature, hence, had to be converted into categorical/discrete to be processed by the BN model, since it uses only discrete nodes. Hence, it was necessary to discretize the output from VISTA prior to instantiating the BN model. For the discretization process it was necessary to refer to domain experts that specified the cut-off values for each variable. Specifically, for traffic volume three states were defined, namely, low, average and high. The first corresponding to less than 100 vehicles per 15 time interval, the second to less than 350 and the last to greater than 350.

4 BN MODEL DEVELOPMENT

Development of BNs requires the specification of the topology and the conditional probability tables. To that end historical accident records were obtained from the traffic safety department of the Cyprus Police. Preliminary compilation of the data was performed with the SPSS statistical package to reduce the dimensionality of the data. The accident dataset covered all accidents occurred in the Nicosia area from 2002 until 2008 and comprised over 9000 records. Each record consisted of 43 (six continuous and 37 categorical) input parameters covering global, local, temporal, accident, driver and car characteristics collected at the site of the accident by the police officers, eye witnesses and the involved parties. Each record was associated with a single categorical output parameter pertaining to accident severity, namely light, severe and fatal, as evaluated by the police officer at the site of the accident.

However, for the development of the BN model topology it was imperative to enhance the dataset with additional information regarding the traffic conditions of each accident record from VISTA simulation. Therefore each accident record was mapped on a geospatial GIS platform and subsequently import on VISTA to obtain the dynamic information of each accident location at different time intervals. This yielded an enhanced dataset of accident records.

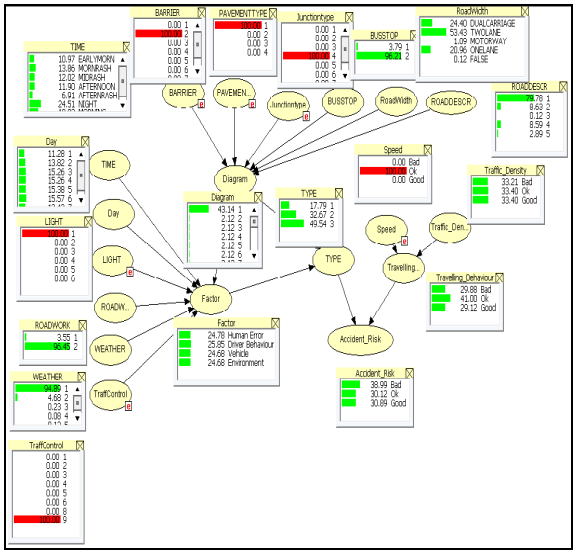


Figure 1: Data inferred BN topology at execution time.

A preliminary analysis of the dataset provided a generic indication of the influence of each variable to road accident risk. Data pre-processing was performed in two steps (a) replacement of missing and erroneous parameter values by the mean value, and (b) grouping related values of multi-valued categorical parameters so as to have a manageable number of states per parameter. Next, to reduce the dimensionality of the dataset, Principal Component Analysis (PCA) was used. This helped to identify the core variables of the model. Results from the dimensionality reduction using PCA, yielded 19 variables for the BN topology. The topology depicted in Figure 1, was learned from processed dataset using the Expectation Maximisation algorithm (Jensen, 2001). Figure 1 also shows an instantiation of the BN model in Hugin researcher tool. The developed ITS utilises the Hugin engine using its API. Each variable in this figure is accompanied by a monitor window that shows its states. The input evidence is showed as a solid bar in the monitor window of each variable. Collectively all variable instantiations correspond to one scenario variation that is provided by the scenario generator component of the tool that uses input from VISTA. In each scenario variation variables that are not instantiated using input from VISTA is varied systematically to produce additional scenario permutations that instantiates the BN model.

To estimate the accuracy of the developed BN model, validation was performed using the accident dataset obtained from the police. The dataset was utilised to identify locations on the network with high accident frequency. These are the networks

black spots. These points were used to validate the model after it was implemented. Specifically, a subset of the accident dataset was used to validate the system. Black-spots that were identified using the dataset, were used to test the BN accuracy under varying conditions of traffic flow data.

5 RESULTS

Results from the accident risk assessor were used to calculate the accident risk index (ARI) of each road section. BN scenarios for each road segment were labeled accident prone if the BN accident risk probability was above a pre-specified threshold value. BN scenarios that fell below the threshold value were ignored. Scenarios were defined on the fly by the scenario generator component. Each segment is evaluated against scenarios that describe traffic condition at different time intervals and driver profiles. To assess the ARI it was imperative to normalize the number of accidents that were predicted by the BN with the traffic volume per time interval, for each road section. To that end, the developed system uses a systematic approach that utilizes the traffic volume estimates from the VISTA simulation and the accidents predicted using the BN risk assessor. Traffic volume acts as a normalizing factor for the number of accidents predicted using the BN risk assessor. In this study, the ARI is defined as:

> Accident Risk Index (ARI) = Number of accidents predicted by the BN/estimated traffic flow rate per time period of the day, from DTA

ARI results gave rise to road sections that inherently have safety issues. These are the network's black spots. An illustration of the preliminary results produced by the method is depicted in Figure 2. This figure illustrates a subset of the results and indicates that sections with IDs, 3, 21 and 47 have the highest ARI.

6 CONCLUSIONS

The ITS system described herein illustrates a novel approach to quantifying road safety using probabilistic inference expressed in causal relationships between factors leading to accidents with DTA simulation. The method escapes from the problem of traffic data shortage through the use of DTA simulation. VISTA provides complete traffic volume data estimates for all road sections of the network on a 24 hour basis. This constitutes advancement over existing methods that base their analysis on limited data obtained from a scarce number of traffic sensors on the network.

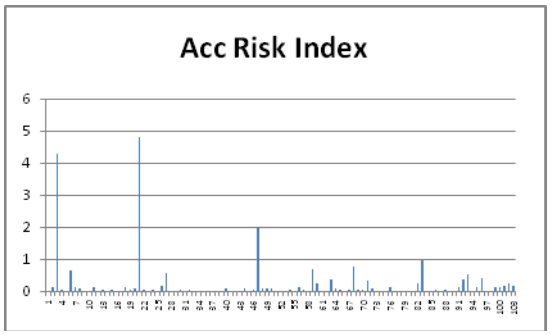


Figure 2: All road section with their ARI values (Y axis).

REFERENCES

Bartley, P., 2008. Traffic Accidents: Causes and Outcomes. *Nova*.

Florian, M., Mahut, M., Tremblay, N., 2008. Application of a simulation-based dynamic traffic assignment model. *European Journal of Operational Research*, 189, 1381–1392.

Jensen, F., 2001. Bayesian Networks and Decision Graphs. *Springer*.

Peeta, S., Ziliaskopoulos, A., 2001. Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Networks and Spatial Economics*, 1 (3/4), 233-65.

Zheng, X., Liu, M., 2009. An overview of accident forecasting methodologies. *Journal of Loss Prevention in the Process Industries*, 22(4), 484-491.

Ziliaskopoulos, A., Lee, S., 1996. A Cell Transmission Based Assignment-simulation Model for Integrated Freeway/Surface Street Systems. *Proc., 75th Transportation Research Board*, Annual Meeting, Washington, DC.

Stakeholders Analysis for Utility Relocation in Construction Project

Ying-Mei Cheng and Chi-Hsien Hou
Department of Civil Engineering and Hazard Mitigation Design, China University of Technology,
56 Hsing-Lung Road, Section 3, Taipei, 116, Taiwan
yingmei.cheng@msa.hinet.net, hou26306208@yahoo.com.tw

Keywords: Stakeholders, K-prototypes, Utility Relocation.

Abstract: Communication is a complicated task while executing utility relocation projects in developed city. The main reason behind this is the number of stakeholders involved. This research tries to identify and classify the stakeholders during the utility relocation projects through interviews with the experts, questionnaire and clustering approach. First, 25 stakeholders and the 6 attributes, Power, Profit, Influence, Impact, Legitimacy, and Urgency are identified from interviews with the experienced engineers. The questionnaire is then developed based on the 6 attributes. The k-prototypes approach is adopted to analyze the results of the questionnaires and classify these stakeholders. The project managers can customize their communication techniques or choose suitable timing to involve the stakeholders with similar characteristics for each group in order to promote communication efficiency, and reach the anticipated objective.

1 INTRODUCTION

Many researches have discussed stakeholders management, analysis or mapping in recent years (Smith et al., 2004; Bourne, 2005; Newcombe, 2010; Jeffrey et al., 2010; Jing et al., 2011; etc.). All of these researches emphasize that the stakeholders' management influences the success of a project. Construction projects are full of uncertainties and risks because of the on-site condition, especially when the construction project includes the relocation of utility lines. Most of the utilities are buried underground in Taipei, Taiwan, which entail the water system, electricity, gas, sewage and so on, and they each under a different jurisdiction with different specialization. Communication among the different stakeholders is complex and difficult. For this reason, recognizing the stakeholders in utility relocation project to improve communication among them and ensure project success is the objective of this research. This research tries to identify the stakeholders and their attributes during the utility relocation projects through interviews with the experts. The attributes become the basis for the questionnaires. The k-prototypes approach is then applied to analyze the results of the questionnaires and classify the stakeholders during utility relocation.

2 STAKEHOLDERS

The concept of stakeholders was first raised by Freeman in 1984. Freeman defines the stakeholder as any group or individual who can affect or is affected by the achievement of the organization's objectives (Freeman, 1984). According to "A Guide to the Project Management Body of Knowledge" (PMI, 2008), project managers spend the majority of their time communicating with team members and other project stakeholders, whether they are internal or external to the organization. PMI (2008) also states that project stakeholders are individuals and organizations that are actively involved in the project or whose interests may be affected as a result of project execution or project completion.

2.1 Stakeholders Recognition

Taipei is a fully developed city in Taiwan with crowded population. Most of the infrastructures, such as the electricity system, and gas utilities were built decades ago. Thus, when a property owner wants to build new infrastructure, communication becomes a major issue for utility relocation. This research will analyze the stakeholders during utility relocation of the MRT (Mass Rapid Transit) construction project, a classic example for utility relocation in Taiwan. Generally, utilities involved in

such project include Street Light, Sewage System, Water System, Gas, Electricity, Telecommunication, Signalization, Military Information and Storm Drainage. Many jurisdictions and agencies are involved. Different types of utilities also require different expertise. This research utilizes the engineers' practical experiences to identify 25 stakeholders during the utility relocation project. Table 1 shows how Taiwanese engineers typically categorize the stakeholders:

Table 1: Stakeholders' classification using practical experiences.

Group 1 - Cable Pipeline Units	
1	Taipower Power Supply Station
2	Taipower District Office
3	Chunghwa Telecom District Office
4	The Parks and Street Lights Office, Taipei City Government
5	Traffic Engineering Office, Taipei City Government
6	Network Transmission Squad of the Signal Group, Army Corps
7	Fixed Line Companies
8	Telecommunication Companies
9	Cable Companies
Group 2 - Fluid Pipeline Units	
1	Storm Drainage Section of the Hydraulic Engineering Office, Public Works Department, Taipei City Government
2	Sewage Systems Office, Public Works Department, Taipei City Government
3	Engineering Division, Taipei Water Department
4	Taipei City Fire Department
5	Natural Gas Companies
Group 3 - Client	
1	Client (Department of Rapid Transit Department, TCG)
Group 4 - Contractors	
1	Material Suppliers
2	Utility Contractors
Group 5 - Elected Representative & Law Enforcement	
1	Local Traffic Police
2	Local Police
3	Local Borough Office
4	Local Representatives and Council Members
Group 6 - User	
1	Local Community Management Center
2	Local Financial Sector
3	Local Businesses
4	Local Residents

2.2 Stakeholders Classification

Ronald (1997) identified 3 attributes: Power, Legitimacy, and Urgency and use them to classify the stakeholders into 7 groups – Dormant, Discretionary, Demanding, Dominant, Dangerous, Dependent, and Definitive. In his research, Power means the ability of those who possess power to bring about the outcomes they desire (Salancik and Pfeifer, 1974). Legitimacy is a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions (Suchman, 1995). Ronald argued that Urgency is based on the time sensitivity and the criticality, so they define urgency as the degree to which stakeholder claims call for immediate attention (Ronald, 1997). Newcombe (2003) included the property developer, British Rail, design practice, insurance company, general public, contractor, users, and local authority as the key stakeholders in the Swindon redevelopment project. He applied the power/predictability matrix and the power/interest matrix to classify the stakeholders and analyze the stakeholders' influence. Bourne (2005) used the Stakeholder Circle methodology to classify and prioritize stakeholders, develop strategies and monitor effectiveness. Different from the above mentioned researches which used qualitative method or analysis software to classify the stakeholders, this research tries to classify the stakeholders by using the quantitative attributes or characteristics of stakeholders.

3 METHODOLOGY

This research identifies the stakeholders of utility relocation projects through interviews with the experts. 25 stakeholders are first identified from the interviews, and then 7 attributes, Power, Interest, Influence, Impact, Legitimacy, Urgency, and Public/Private sector are adopted to set up the questionnaires. The 6 former attributes are numeric data type. Power, Legitimacy, and Urgency are defined in section 2.2. Interest refers to the stakeholders' level or concern regarding the project outcomes. Influence is the stakeholders' active involvement in the project. Impact means the stakeholders' ability to affect changes to the project's planning or execution (PMI, 2008). The last data is categorical data type, which represents whether the stakeholders belong to the public or private sector. Because the k-prototypes approach

can be applied toward mixed data type, it is adopted in this research to analyze the results of the questionnaires and classify the stakeholders during utility relocation.

3.1 Questionnaire

The questionnaire is designed to get the attribute value for each stakeholder. The questionnaire assigns a seven-point Likert scale for the 6 attributes of each of the 25 stakeholders, which will be discussed later on. The "7" in the scale means the highest, "6" means higher, "5" means high, "4" means average, "3" means low, "2" means lower, and "1" means the lowest. In order to achieve objectivity and professional result, the members need to have utility relocation related experience. In addition, the recipients of the questionnaires are from both public and private sectors, for example, Department of Rapid Transit Systems, Sewage Systems Office, Chunghwa Telecom, Water Department, gas companies, contractors and so on. There are 37 participants for this questionnaire, including engineering staff or officials who have participated in utility relocation related projects, among which 14 has over 20 years of experience, 13 with 10 to 20 years of experience, 5 with 5 to 10 years, and 5 with 1 to 5 years.

3.2 K-prototypes Algorithm

The clustering algorithms have numerous scientific and practical applications, such as in artificial intelligence, pattern recognition, and medical research. In general, it can be divided into various categories based upon their principles and algorithms. The traditional clustering methods include the following: 1) Partitioning methods; 2) Hierarchical methods; 3) Density-based methods; and 4) Grid-based methods. The k-prototypes algorithm is a type of Partitioning methods proposed by Huang (1998). This algorithm provides a straightforward approach to integrate the k-means and k-modes algorithms to cluster mixed-data-type objects. The objective function is defined as follows:

$$P(W,Q)=\sum_{l=1}^{k}\left(P_l^r+P_l^c\right) \tag{1}$$

in Equation (1),

$$P_l^r=\sum_{i=1}^{n}w_{i,l}\sum_{j=1}^{p}\left(x_{i,j}-q_{l,j}\right)^2 \tag{2}$$

and

$$P_l^c=\gamma\sum_{i=1}^{n}w_{i,l}\sum_{j=p+1}^{m}\delta\left(x_{i,j},q_{l,j}\right) \tag{3}$$

where W is an $n\times k$ partition matrix, $Q = \{Q_1, Q_2, \ldots, Q_k\}$ is a set of objects in the same object domain, where $1 \le l \le k$. Let $X = \{X_1, X_2, \ldots, X_n\}$ be a set of n objects. Object X_i is represented as $[x_{i,1}, x_{i,2},\ldots, x_{i,m}]$, where $1 \le i \le n$, $1 \le j \le m$. Equation (2) is the squared Euclidean distance measure of the numeric attributes and Equation (3) is the simple matching dissimilarity measure of the categorical attributes. The weight γ is used to maintain a balance between both data types. Interested readers are encouraged to refer to Huang's paper for details on this algorithm (Huang, 1998).

4 ANALYSIS RESULT AND DISCUSSION

This research used the questionnaires to assign values to stakeholders' 6 attributes with the addition of whether the stakeholders are from public or private sector, and the k-prototypes approach to analyze the stakeholders. Based on Roland's classification result (Ronald, 1997), this research uses 7 as the initial number of groups. The numbers of each group is shown in Figure 1. Only 1 stakeholder (local community management center) is classified under Group 1, so the researchers consider that most of its attribute values are close to the means of Group 6, it means the characteristics of local community management center are similar to those of Group 6, so it was combined with Group 6 into the new Group.

After the adjustment, table 2 shows the means of each attribute in each group. Group 1 now includes the Storm Drainage Section of Hydraulic Engineering Office of Public Works Department and Sewage Systems Office in Taipei City Government, Sewage Systems Office of Public Works Department in Taipei City Government, and client (Department of Rapid Transit Department, TCG). Group 2 includes the Parks and Street Lights Office and Traffic Engineering Office of Taipei City Government, Network Transmission Squad of the Army Corps Signal Group, and Taipei City Fire Department. Group 3 includes Taipower Power Supply Station, Taipower District Offices, Chunghwa Telecom District Offices, Engineering Division of Taipei Water Department, and natural gas companies. Group 4 includes the fixed line companies, telecommunication companies, cable

companies, and utility subcontractors. Group 5 includes local traffic police, local police stations, local borough offices, and local representatives and council members. Finally, Group 6 includes material suppliers, local financial sector, local businesses, local residents, and local community management center. The classification results indicates that Group 3 has the highest values in power, profit, influence, legitimacy, and urgency while Group 1 has the second highest values in power, profit, influence, legitimacy, urgency, and the highest value in impact. Project managers need to pay more attention to the stakeholders within these 2 groups.

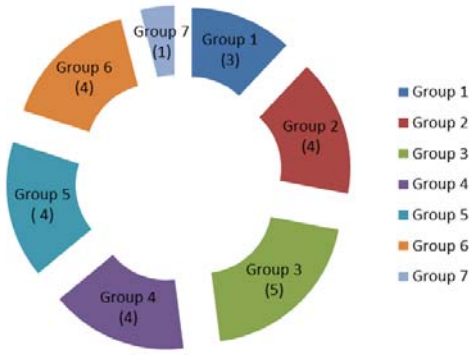


Figure 1: The numbers of each group (Before adjustment).

Table 2: The means of attribute in each group.

Attribute \ Group	1	2	3	4	5	6
Power	5.4234	4.7432	5.4541	4.7365	4.0676	3.6270
Interest	5.0180	4.5203	5.3568	4.8378	3.9662	4.3946
Influence	4.9459	4.6149	5.7027	4.5203	3.9122	3.8865
Impact	5.4955	4.1081	5.3459	4.1554	3.6149	3.6865
Legitimacy	5.3604	4.9730	5.4703	4.4122	3.6014	3.4865
Urgency	5.0090	4.8108	6.0595	4.5203	3.1081	3.3297

5 CONCLUSIONS

This research utilized questionnaire and k-prototypes clustering approaches to classify the stakeholders for utility relocation projects. Comparing with the traditional classification method, which depends on the engineers' subjective opinions, this method proposed objective and quantitative classification. The authors first interviewed experienced engineers to identify a list of 25 stakeholders, who are then classified into 6 groups. Stakeholders in each group are with similar characteristics. According to this information, project managers can plan for communication accordingly. For example, the project team can seek advices from the group with the highest attribute values early in the processes. In conclusion, project managers/team can customize their communication strategy for each group.

REFERENCES

A Guide to the Project Management Body of Knowledge, *Project Management Institute*, PMI, 2008.

Bourne, L., 2005. Project Relationship Management and the Stakeholder Circle^{TM}, *PhD thesis*, *RMIT University*, Australia.

Freeman, R. E., 1984. Strategic management: A stakeholder approach. Boston: Pitman.

Huang, Z., 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery* 2, 283–304.

Jeffrey S. Harrison, Douglas A. Bosse, and Robert A. Phillips, 2010. "Managing for Stakeholders, Stakeholder Utility Functions, and Competitive Advantage", *Strategic Management Journal*, 31: 58–74.

Jing Yang, Geoffrey Qiping Shen, Lynda Bourne, Christabel Man–Fong Ho, and Xiaolong Xue, 2011. A Typology of Operational Approaches for Stakeholder Analysis and Engagement, *Construction Management and Economics*, 29, 145–162.

Newcombe, R., 2003. From Client to Project Stakeholders: A Stakeholder Mapping Approach, *Construction Management and Economics*, 21, 841–848.

Ronald K. Mitchell, 1997. Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts, *Academy of Management Review*, 22(4), 853-886.

Salancik, G. R., Pfeifer, J., 1974. The bases and use oí power in organizational decision making: The case of universities. *Administrative Science Quarterly*, 19, 453—473.

Smith, J., Love, P., E., D., 2004. Stakeholder Management during Project Inception: Strategic Needs Analysis, *Journal of Architectural Engineering*, vol. 10, No. 1.

Suchman. M. C. 1995. Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, 20, 571-610.

Assessment and Choice of Software Solution with the Analytic Network Process Method

Jerzy Michnik[1] and Krzysztof Kania[2]
[1]*Operations Research Department, University of Economics in Katowice, ul. 1 Maja 50, Katowice, Poland*
[2]*Knowledge Engineering Department, University of Economics in Katowice, ul. 1 Maja 50, Katowice, Poland*
{*jerzy.michnik, krzysztof.kania*}*@ue.katowice.pl*

Keywords: Analytic Network Process (ANP), Software Assessment.

Abstract: Assessment and choice of software solution belong of the most difficult tasks facing business and IT experts. Multiple criteria decision methods can help in making that kind of a decision. However most of the methods is based on the assumption of criteria independence which is rarely fulfilled in practice. We applied the Analytic Network Process (ANP) as an aid for a choice of software solution and compare its results with the Analytic Hierarchy Process (AHP) method which is more often used in such a task.

1 INTRODUCTION

Under the pressure of rapid development of information and communication technologies (ICT) and the growing importance of information systems in business, the organizations often face the problem of matching the available ICT to business needs. A choice and successful implementation of the versatile ICT system is a serious challenge for an organization. Such a project lasts a long time, needs substantial investment and re-organization of most of the business procedures. A complete ICT system is a compound structure with many specialized components and an exchange of information between its components.

The aim of our work was to develop the effective decision aid for an assessment and choice of ICT solution characterized by a wide range of attributes by applying the Analytic Network Process (ANP) method for ranking the decision alternatives. The ANP seems to be more suitable then the other methods as it has the ability to handle the complicated decision model with many criteria and dependencies among them.

The article is structured as follows. Sec. 2 defines a research problem, Sec. 3 presents the ANP method. Sec. 4 contains the model of Business Intelligence (BI) systems assessment. Sec. 5 concludes the article.

2 THE PROBLEM OF CHOOSING THE ICT SYSTEM

Selecting a specific software is a stage of the whole decision process, where requirements formulated in the sphere of business and ICT meet together. Managers want the software to give them the greatest possible business opportunities and focus mainly on the software functionalities, while the ICT professionals have to take into account many technological limitations, existing and legacy systems, the possibility of performing additional tasks (administration, support and safety) and many others.

Selecting a software for the large-scale systems is a strategic decision because it determines the operating environment for a long time and bounds an organization to a particular vendor. It can be implemented in different ways (Woitsch et al., 2009). The most common approaches to that task can be described as heuristic approach as they are based on knowledge of experts involved in the process of software assessment. Hence, a quality of the decisions depends primarily on a quality of knowledge and an experience of experts. Yet, the serious difficulty encumber the heuristic procedure in practice. It is evoked by the enormous quantity of information that have to be processed on the way to final decision. ICT is a highly compound system of several components with many various sub-elements that are characterized by a large number of qualitative and quantitative features.

Risk of making the wrong choice is high because

of the variety of offers, the high cost of software implementation that excludes rapid changes of environment, very long time needed for software deployments and because of technological and hardware linkages established during the project. All these risks can be reduced by a precise projection of business needs and technological constraints during the procedure of software selecting. Unfortunately, the desired properties of the system are often dependent on each other and they create a net rather then a simple hierarchy. So, it is necessary to use a method that gives the possibility of more than just a simple imposition of a set of weights.

Some number of formal methods that support software selection has been reported. The authors of Computer Science Technical Report mentioned, among others, 4 papers that use the linear weighted attribute method (simple additive weighting) (Fritz and Carter, 1994). The other similar multiple criteria decision methods, like SMART (Valiris et al., 2005) or ELECTRE II (Stamelos et al., 2000), also have been tested. Lai et al. report the results of a case study where the AHP method was employed to support the selection of multimedia authoring system (Lai et al., 2002). Selecting the best software product among the alternatives for each module in the development of modular software systems has also been done with the aid of AHP (Jung and Choi, 1999).

All methods mentioned above base on an assumption that the criteria, considered in the evaluation of alternatives, are independent. Yet, the ICT system is compounded from the interfering modules and it leads to some dependencies between criteria. Use of a method that can involve dependencies in the analyzed system may substantially improve the results.

Wu proposed a hybrid model that combines the Decision Making Trial and Evaluation Laboratory (DEMATEL) with the ANP and the zero-one goal programming (ZOGP) to get an effective solution that considers both financial and non-financial factors (Wu, 2008). Recently, ANP has been used to select most suitable simulation software (Ayağ, Zeki, 2011) and ERP system (Wieszała et al., 2011).

Almost all of the publications cited above concern the assessment of single, specialized software or consider (Wu, 2008) the series of the mutually non-excluded IT projects. Our evaluation deals with more complex implementation of a whole, multi-modular ITC system in an enterprise when only one alternative is to be selected.

3 THE ANALYTIC NETWORK PROCESS

The Analytic Network Process (ANP) (Saaty, 2005) is defined as a multiple criteria method that derives priority scales of absolute numbers from individual judgments. The numbers come out from the pairwise comparisons of elements of the studied system. One provides the judgment by answering two kinds of questions: 'Which of the two elements is more dominant with respect to a criterion?' or 'Which of the two elements influences the third element more with respect to a criterion?'

The ANP procedure can be summarized in the following steps:

1. Set up: a) the control criterion representing the decision problem, b) the main groups of criteria (named components or clusters) characterizing the decision problem, c) the criteria that belong to each cluster, d) the decision alternatives, e) the relations between elements of the decision model (criteria and alternatives).
2. Make all pairwise comparisons for relations in the model using the two kinds of questions mentioned above.
3. Perform the following operations: a) calculate priority vectors for supermatrix and cluster matrix, b) build the unweighted supermatrix, c) weight the unweighted supermatrix with the cluster matrix, d) calculate the limit supermatrix.
4. Read out the overall priorities for alternatives from the limit supermatrix. Discuss the results. If needed, make the suitable modifications of the model and repeat the procedure.

All steps besides Step 3, are the tasks that need to be made by people engaged in the decision process (decision maker(s) and/or analyst). Step 3 has a computational character and can be automatized with a suitable software (in this work, like in many others, the specialized software "Superdecisions" has been used). The short description of the operations of Step 3 is presented below.

A priority vector is derived from paired comparisons matrix by normalizing its columns and taking the geometric mean form rows (in the same way as in the AHP). Let's assume that we need to compare p elements of the model with respect to some control criterion. So, the pairwise comparison matrix C will be the square matrix of size $p \times p$. Saaty (Saaty, 2005) suggests to use the following scale to translate the verbal comparisons (easier to obtain from decision makers) into numbers: equal importance = 1; moderate

importance = 3, strong importance = 5, very strong importance = 7, extreme importance = 9. The even numbers 2, 4, 6, 8 are used for an assessment lying between the above main points of scale.

Each priority vector becomes a column of matrix $W_{ij} = [w_{kl}]_{n_i \times n_j}$, where n_i (n_j) is a number of elements in cluster i (j). Let's assume that there is a system of N clusters. Then the supermatrix will be constructed from $N \times N$ blocks, i.e. $W = [W_{ij}]_{N \times N}$. W_{ij} represents the influence of the elements from cluster i on the elements from cluster j. The supermatrix W represents the influence priority of the element on the left of the matrix on the element at the top of the matrix.

In the next step, the supermatrix is transformed into a weighted supermatrix, i.e. to the matrix, whose columns sums to unity. Initially the supermatrix columns are made up of several eigenvectors which, in normalized form, sum to one and hence that column sums to the number of nonzero eigenvectors. The weighted supermatrix can be obtained by weighting the initial supermatrix with the cluster matrix. The cluster matrix contains eigenvectors representing the priorities of clusters with respect to the general control criterion (in most cases it will be a main objective).

In the end the limit matrix is derived by raising the weighted matrix to an arbitrarily high power. This procedure sums up the influences along paths of different length in the underlying network and determines the overall priorities.

4 THE ANP MODEL FOR BI SYSTEM SELECTION

The need for analysis and evaluation of BI environments results from the fact that none of the largest vendors of integrated analytical platforms offers full functionality required in management practice of business or public organizations. Moreover, an open source software often has functionality similar to commercial tools or even enhances specific business analytic modules. Flexibility and ease of adaptation to the particular needs are the advantages of an open source software.

Our example follows and supplements BI environments evaluation presented in (Dudała et al., 2010). The evaluation was conducted in five modules (according to classical architecture of BI environments): Database/Data Warehouse server, ETL tools, OLAP, Data Mining and Reporting tools. In each module a set of criteria was proposed. This modules take a role of criteria clusters in the ANP model. They are complemented by two other clusters: Main objective and Alternatives. The structure of the model, generated by Superdecisions software, is presented in fig. 1.

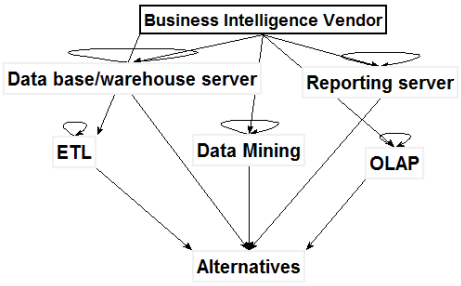


Figure 1: Clusters and their interrelations in the ANP model.

The cluster numbering and priorities of criteria clusters are as follows: 1. Alternatives; 2. Business Intelligence Vendor (main objective); 3. Data Base/Data warehouse server (0.042); 4. Data Mining (0.193); 5. ETL (0.229); 6. OLAP (0.418); 7. Reporting server (0.116).

The limited space prevent us to present the complete input data. Below, there are given only some parts of the unweighted supermatrix. Table 1 contains the example of criteria and their initial priorities.

Table 1: Criteria in cluster 6 OLAP.

No.	Criterion	Priority
61	Diff. data sources OLAP	0.139376
62	Graph. interf. OLAP	0.163747
63	Lic/fin cond. OLAP	0.162196
64	MDX language	0.139376
65	MS Office int. OLAP	0.139376
66	Security OLAP	0.127965
67	User supp. OLAP	0.127965

The priorities of alternatives with respect to the selected criteria are presented in Tab. 2 (sample for the cluster OLAP).

Table 2: Selected alternatives' priorities with respect to criteria from cluster 6 OLAP.

Alternative	Diff. data sources OLAP	User supp. OLAP
V1	0.084863	0.187670
V2	0.459105	0.187670
V3	0.154541	0.363056
V4	0.154541	0.199955
V5	0.146949	0.061648

A number of dependencies between criteria have to be considered. They regard criteria belonging to the same cluster and are represented by 'inner dependence loops' in fig. 1. The dependence of 'Lic/fin

cond.' on other criteria in cluster 'Performance' on 'Scalability', 'Paralell computing' and in cluster 'OLAP' is a good example (see Tab. 3).

Table 3: The dependence of 'Lic/fin cond.' in cluster 6 OLAP.

No.	Criterion	Priority
62	Graph. interf. OLAP	0.113512
64	MDX language	0.539254
65	MS Office int. OLAP	0.244404
67	User supp. OLAP	0.102830

Calculations for the ANP and AHP models have been done with Superdecisions software (ANP Team, 2012). Fig. 2 shows the final priorities of alternatives given by ANP. For the sake of comparison the results of AHP are also presented.

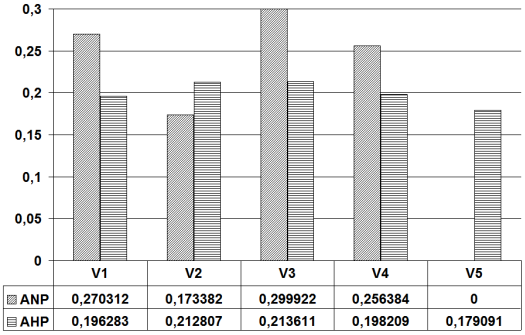


Figure 2: Final priorities of alternatives given by ANP and AHP models.

As it can be seen at Fig 2, the assessment with the ANP not only has changed the overall ranking but also has differentiated vendors much more then AHP. Hence, the ANP gives a better base for the final decision.

5 CONCLUSIONS

A software selection is a task that have to take into account multiple, often interdependent factors. This article shows how this task can be done with the ANP method. In comparison with the other methods, ANP allows better modeling of the needs of users as it allows for the relationships between elements of the modeled system. In fact, our example has demonstrated that an inclusion of the interrelations among factors may lead to different results in comparison to methods with independence principle (represented here by AHP).

We have built and solved the ANP model for an extended problem of BI system selection, in which all main modules has been considered: Data Base/Data Warehouse Server, Data Mining, ETL, OLAP and Reporting Server. Each of these modules, in turn, contained 7-14 criteria. Altogether the problem embodied 49 criteria, and additionally there were some dependencies between them. Inevitably, requiring the hundreds of comparisons, the procedure became really labor intensive and high demanding. The only excuse of this inconvenience is that several hours spent on pair-wise comparisons may be assumed as not so high cost in comparison with the overall time and money expense of such a big and important project. This also suggests a potential direction of the future study towards the methods with similar capabilities but less laborious and less demanding.

REFERENCES

ANP Team (2012). www.superdecisions.com.

Ayağ, Zeki (2011). Evaluating simulation software alternatives through ANP.

Dudała, J., Gołuchowski, J., Kajfosz, K., Kania, K., and Staś, T. (2010). Business intelligence environment comparison (in polish). Technical report, University of Economics in Katowice, Katowice.

Fritz, C. and Carter, B. (1994). A classification and summary of software evaluation and selection methodologies. Technical Report 940823, Mississippi State University.

Jung, H. and Choi, B. (1999). Optimization models for quality and cost of modular software systems. *European Journal of Operational Research*, 112(3):613–619.

Lai, V., Wong, B., and Cheung, W. (2002). Group decision making in a multiple criteria environment: A case using the AHP in software selection. *European Journal of Operational Research*, 137(1):134–144.

Saaty, T. L. (2005). *Theory and Applications of the Analytic Network Process. Decision Making with Benefits, Opportunities, Costs and Risks*. RWS Publications, Pittsburgh.

Stamelos, I., Vlahavas, I., Refanidis, I., and Tsoukiàs, A. (2000). Knowledge based evaluation of software systems: a case study. *Information and Software Technology*, 42(5):333–345.

Valiris, G., Chytas, P., and Glykas, M. (2005). Making decisions using the balanced scorecard and the simple multi-attribute rating technique. *Performance Measurement and Metrics*, 6(3):159–171.

Wieszała, P., Trzaskalik, T., and Targiel, K. (2011). Analytic network process in ERP selection. In *Multicriteria Decision Making'10-11*, pages 261–286. University of Economics in Katowice, Katowice.

Wu, W. (2008). A hybrid approach to IT project selection. *WSEAS Transactions on business and Economics*, 5(6):361–371.

Substations Optimization
Foundations of a Decision Making System

Luiz Biondi Neto[1], Pedro H. G. Coelho[1], Francisco Soeiro[1], Osvaldo Cruz[1] and David Targueta[2]
[1]*State University of Rio de Janeiro, Av. Maracanã, 524, Rio de Janeiro, RJ, Brazil*
[2]*São Simão M. S. Ltda, Mal Camara 160, Sala 1808, Rio de Janeiro, RJ, Brazil*
{lbiondi, phcoelho, soeiro}@uerj.br, ocruzconsultoria@gmail.com, dtsilva@ssmao.com.br

Keywords: Decision Support System, Substations Optimization.

Abstract: The optimization of building processes for a power substation is based on the adopted configuration structure and includes a simulation of the methods for the mechanical, civil and electrical processes. Thus it is necessary to know the scope of the service area, the substation load and its connected transmission lines, the terrain topography, and the environmental impact, issues that will be only known after the choice of the area and the project details. The purpose of this work was to bring the foundations of a decision support system regarding the reduction of the structure weight and its concrete volume. A laboratory reduction model validated the work.

1 INTRODUCTION

For an electric utility, changes in legislation and the growth of energy use require the need for new tools and techniques, to achieve the highest level of quality of power supply to the consumer at the lowest cost and always preserving the environment. For this reason, the research carried out, combines mechanical civil and electrical engineering, and therefore makes use of different methodologies, depending on the area in which one seeks to optimize envisioning a decision support system (D'Ajuz, 1985).

The main objectives of this research aimed to develop possible solutions for optimization of construction of substations and consisted of:

1. Model and simulate the investigated metallic or composite structures by estimating their weights, aiming their reduction in the optimized Electrical System (ES).
2. Shape the foundations of the pillars concerned to the investigated metal or composite structures in order to .reduce the concrete volume.
3. Model, simulate and test the Electrical System on a reduced scale.

2 METHODOLOGY

2.1 Reducing the Weight Structure

Studies were undertaken in order to minimize the weight of the structure with three different situations, from the most traditional to the most innovative on the market with technical characteristics that meet the preliminary optimized substation. Three structures were investigated:

1. Lattice-like structures (traditionally used).
2. Tubular structures (used in our proposal).
3. Centrifuged Concrete Structures (steel and concrete).

These structures must be sized appropriately in order to resist traction forces, self weight, weight of equipment and wind acting on them. For the calculations is necessary to know precisely the topography of the region adjacent to the land and own land for the construction, the angular distribution lines related to the substation, and the climatic characteristics of the region, especially in relation to the wind, which at this stage project are not yet defined.

The computational tool (Bhati, 2005) used to model, simulate, analyze and estimate parameters in the three cases studied, was the Finite Element Method (FEM).

2.1.1 Lattice-like Structures

Lattice Systems are those consisting of undeformable elements joined together by hinges, considered perfect, and subject only to loads applied to the joints or nodes. Thus the elements or bars are only subject to normal efforts, traction or compression. In the plane lattice, the set of construction elements, e.g. round bars, flat or angles, are interconnected under triangular form geometry, by pins, welding, rivets or bolts, designed to form a rigid structure in order to withstand only the normal efforts. Figure 1 shows part of the plant that was used for calculating the unilateral drift due to crosswind for a 138 kV Electrical Substation.

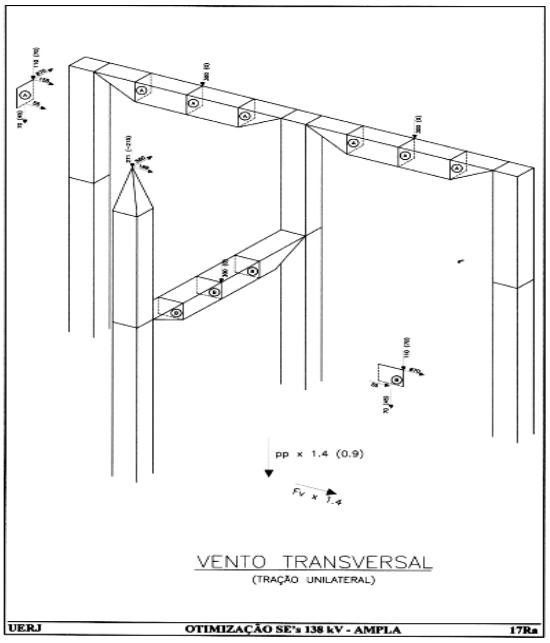


Figure 1: Unilateral drift due to crosswind for a 138 kV SE.

2.1.2 Tubular Structures

Tubular profiles can have three different geometries: circular, rectangular and square. The geometry of these profiles is their main advantage, because its closed section allows a significant increase in resistance. Besides, the effective reduction of the foundations structure yields huge savings for these buildings, and shows good integration to the environment.

The circular profiles provide a better distribution of stresses on the tube due to their geometry, in which all cross-sectional points are equidistant and therefore were investigated in our research. Figure 2 shows the FEM used in the 138 kV ES having tubular structure.

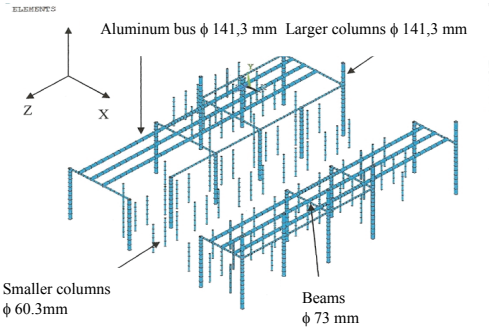


Figure 2: FEM modelling used for tubular structure.

2.1.3 Centrifuged Concrete Structures

The excellent visual integration with the urban environment, given the texture of the concrete and the elegance of the structure, allows the installation of centrifuged reinforced concrete in any area, minimizing the impact on the environment and landscape. Throughout this work simulations using FEM were performed indicating that the weight of the centrifuged reinforced concrete structure is much larger than that of the tubular steel and the same occurred with the lattice one. Consequently, the amount spent on concrete foundations using centrifuged concrete is much larger than the structures used in tubular steel and the same occurred with the lattice structure. Figure 3 shows a brief view of a 138 kV Electrical Substation with centrifuged concrete.

Tables I and II summarize some specifications, technical and economic characteristics and important peculiarities in these types of structures.

Table 1: Structures operational characteristics.

kV Class	Estimated Weight (kgf)			Av. Cost (US$/kgf)			Corrosion & Fire Resist.			Installation and execution of work			Visual Pollution		
	LAT	TUB	CENT	LAT	TUB	CENT	LAT	TUB	CENT	LAT	TUB	CENT	LAT	TUB	CENT
138	50260	12000	145000	>1.9 T	>4.5 NT	<0.6 T	1	3	4	E S	Ex S	F E S	4	2	0
69	38760	5980	200000												
34.5	14590	2250	40000												

LEGEND:
LAT – Lattice
TUB – Tubular
CENT – Centrifuged Concrete

0 – Very Low E- Easy
1 – Low F – Fast
2 – Medium Ex – Expensive
3 – High S – Safe
4 - Very High

T – Broadly tested in ESs
NT – Not Broadly tested in ESs

Table 2: Costs of the investigated structures.

kV Class	Structure Cost (US $)		
	Lattice	Tubular	Centrifuged
138	94935	54000	80556
69	72214	26910	111112
34.5	27559	10125	22223

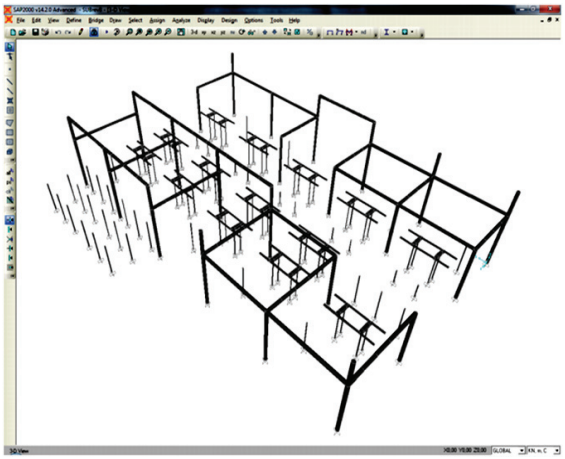


Figure 3: Centrifuged concrete structure for a 138 kV ES.

As far as the cost is concerned, the tubular structure is also very attractive, as evidenced in Table 2.

2.2 Simulation of the Concrete Volume

Generally, the foundations of substations can be classified as shown in Table 3

Table 3: Foundations of substations types.

Foundations						
Shallow			Pile			
Block	Shoe	Concrete Slab	Pre-cast Piles Steel, Concrete and Centrifuged Concrete	Piles moulded on site with a coating tube		
				Strauss	Franki	Helical	Pillar

To accurately estimate the type and volume of concrete foundations it is basically required to know the loads to be transferred to the foundations of the investigated structures, and evaluate the reports of the land survey for the construction of the ES. It is undeniable that there is an inevitable link between the geological conditions and the design of the foundations (Groenewald, 2009). Mentioned below are some needs which must be fully met during the detailed design of the project.

1. Definition of the loads to be transferred to foundations;
2. Important developments in geomorphology;
3. Geotechnical local site;
4. Data on slopes and hillsides on the ground;
5. Data on erosion, occurrence of soft soil on the surface;
6. Need to make cuts and embankments on the ground;
7. Compressibility and resistance in the survey;
8. The level of groundwater;
9. Executive feasibility;
10. Economic viability.

It can be seen, through simulation, that the tubular steel frame weighs less than 10% of the centrifuged concrete structure and less than 25% of the lattice structure, fact which would lead directly to its choice. The lighter the structure, the lower the concrete volume to be used, resulting in lower cost, as shown in Table 4.

Table 4: Cost of concrete.

kV Class	Concrete Cost (US $)					
	Lattice		Tubular		Centrifuged	
	30 MPa	50 MPa	30 MPa	50 MPa	30 MPa	50 MPa
138	30000	50000	15000	25000	30000	50000
69	18334	30556	1667	2778	11667	19445
34.5	13334	22223	1000	1667	7500	12500

2.3 Reduced Model Testing

The choice of the reduction coefficient of the reduced model was based taking into account not only the physical limitations found in the Laboratory of Structures and Materials (LEM) at PUC-Rio, where tests were performed, but also the equipment and instrumentation required to the tests, which followed a high technical accuracy required in these experiments and available on the LEM.

For the tests of the prototype scale model were considered reductions in the dimensions of the parts, taking into account the equivalence of physical resistance to the tubes easily available for purchase on the market. The height of the prototype is decisive for the calculation of the reduction coefficient under the penalty of exceeding the limits permitted in the laboratory tests, which led to the ratio of 1:6 (one to six).

The calculations of the reduced model were based on the original study design. The values of the geometric properties of the prototype, such as length, width and height of the structure, and external diameter and wall thickness of tubular profiles were taken from the design of the 69 kV ES performed using the structural analysis program SAP2000. Figure 4 shows the model with the actual dimensions.

All profiles are circular tubes with the following dimensions:

Columns

- - outside diameter of 219.1 mm and 12.7 mm thickness;
- - crossbeams - outside diameter of 219.1 mm and 12.7 mm thickness.

Longitudinal Beams

- - outside diameter of 101.6 mm and thickness 5.7 mm.

As explained earlier, the reduced model was constructed using a reduction factor of 1:6. Figure 5 shows a schematic drawing of the dimensions of the reduced model.

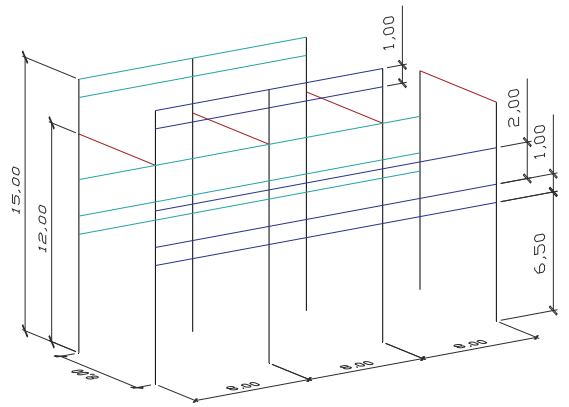


Figure 4: Actual dimensions of the 69 kV ES in meters.

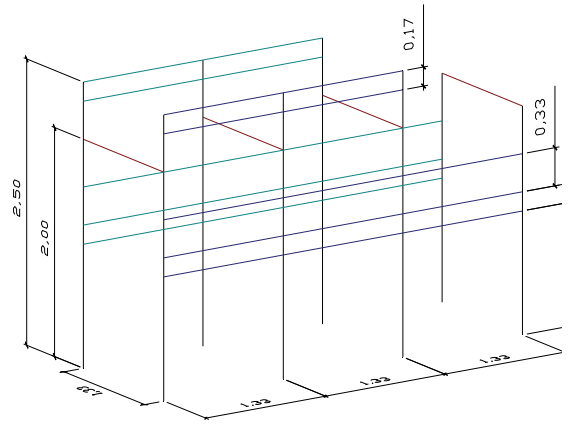


Figure 5: Reduced Model Dimensions of the 69 kV ES

3 CONCLUSIONS

Under the specific viewpoint of optimization of power substations, object of this research, the obtained results seem very promising. For future work it is intended to give more depth to the tubular steel structures and their respective founding, simulating more cases using finite element software, in addition to those already made in this research.

The test results of the reduced model indicate that the integrity of the structure was confirmed, considering the details of the boundary conditions of the investigated structures, loading and material, where there was no need for any reinforcement or modification of the original structure.

Finally, it was found that the optimized SEs were actually efficient from the studied viewpoint.

REFERENCES

D'Ajuz, A., 1985. *Electrical Equipments – Specifications and Applications in High Voltage Substations,* Furnas/UFF Ed., in Portuguese.

Bayliss, C. and Hardy, B., 2007. *Transmission and Distribution Electrical Engineering,* Newness Ed.

Bhatti, M. A., 2005. *Fundamental Finite Element Analysis and Applications with Mathematica and Matlab Computations,* John Wiley & Sons.

Groenewald, A., J., S., 2009. The Use of Tubular Conductors in the Design of High Voltage Substations. In *CIBRE 6th S. Africa Regional Conf.,* 2009.

Carneiro, F., L., 1996. *Dimensional Analysis and Similarity and Physical Models Theory*, UFRJ Ed., in Portuguese.

Evaluating a Petroleum Exploration Opportunity trough Data Mining

Marcos Affonso, Kate Revoredo and Leila Andrade
Centro de Ciências Exatas e Tecnologia, Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, Brazil
{marcos.affonso, katerevoredo, leila}@uniriotec.br

Keywords: Data Mining, Decision Support System, Bayesian Network.

Abstract: A petroleum exploration opportunity (EO) is defined as a mapped region with potential for possessing a sufficient petroleum accumulation that may justify an exploration project .This article proposes to build a predictive model that it is able to economically evaluate a petroleum exploration opportunity through data mining techniques.

1 INTRODUCTION

Although, in the last decades, it has been growing the search for alternative sources of energy, such as Alcohol, the consumption of petroleum derivatives keeps growing in Brazil as shown the analysis of the Ministry of Mines and Energy (MME) (Figure 1).

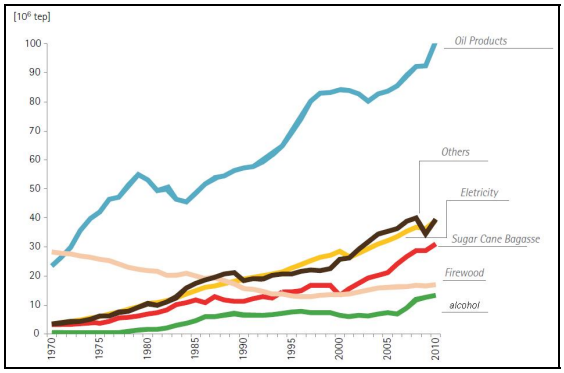


Figure 1: Consumption of energy in Brazil by source - (Source: MME - 2011).

Any geographic area with clues of petroleum is called Exploration Opportunity (EO). Given an EO, engineers evaluate the viability of an exploration project to confirm those expectations. Geological researches aim to identify necessary geological conditions, meanwhile economic studies focus on economic viability of the project.

The modality of economic evaluation usually applied in these cases is Net Present Value (NPV) (Newendorp et al., 2009). NPV is based on cash flow and production curve. The evaluation task deals with uncertain information like oil price, oil quality and depth of the accumulation.

The world success index in petroleum activity is about 20%, that is, 20 out of 100 exploration wells succeed in finding oil in quantity and quality sufficient to turn an EO into an Oil Field. Part of this low index is due to errors in economic evaluation.

On the other hand, data mining techniques (Witten et al., 2011) are been used to learn models for depicting a dataset. In the literature, one can find examples of data mining application as in Healthcare (Canlas, 2009).

This work argues that data mining techniques can be used to improve an economic evaluation of an EO.

The article has the following structure. Section 2 explains the purpose of an economic evaluation and how it is made. In Section 3 Data Mining techniques are reviewed. Section 4 presents our proposal. Section 5 describes the outcomes from our experiments applied over historical dataset provided by a petroleum company. Section 6 presents related works and, finally, Section 7 concludes with final observations.

2 NET PRESENT VALUE (NPV)

In the petroleum industry, people mostly use NPV to economically evaluate an EO. The first step to evaluate an opportunity is to elaborate an oil production curve, where a production estimate, year by year is reflected.

The next step is to do the calculation of the cash

flow related to that production curve. The cash flow represents the net return (revenue minus expenditure), year by year.

After gathering all this information, it's possible to calculate the NPV, that it is the sum of all cash flow discounted.

As one can see, this methodology uses very conventional analytic approaches and applies a linear method to calculate the NPV.

3 DATA MINING

Data Mining (DM) is part of a bigger process called Knowledge Discovery in Databases (KDD) (Witten et al., 2011). Usually the terms KDD and Data Mining are used indiscriminately. However, DM is just a step in the KDD process.

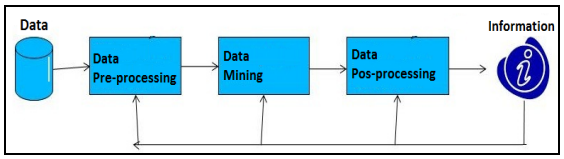


Figure 2: Knowledge discovery in database process.

Figure 2 shows the sequence of steps in order to achieve new and useful Knowledge. First, techniques of pre-processing like cleaning, missing values treatment, detection of noise data and outliers are applied. It's also in this step that continuous attributes are normalized and discretized if necessary. Normalization avoids that attributes with great ranges of values are deprecated over others and discretization aims to reduce the number of attribute values through the division in intervals.

Next step is data mining, where machine learning (Mitchell, 1997) algorithms are applied to learn a model that reflects the dataset, expliciting hidden patterns.

The learning of the model is called supervised whenever the example class is known and taken in consideration during the learning process. The learned model can be a classifier or a predictor depending on the type of the class variable.

Finally, the learned model passes through a post-processing step, where the interesting patterns are filtered, visually presented and interpreted. KDD is an iterative process. Therefore, one can return to previous step whenever necessary. The process is finished when knowledge is found.

Due to the high degree of uncertainty involving the exploration activities, this work looked for a model able to represent this uncertainty. Bayesian Networks (Koller et al., 2009) is an example of such a model.

3.1 Bayesian Network

Bayesian Network (BN) is a model that combines Graph Theory and Probability Theory with strong theoretical basement. BN represents probability distributions in a concise manner and uses graphs to express dependences among domain variables (Koller et al., 2009). That is, BN is a directed acyclic graph where the nodes represent domain variables and the edges represent dependences between these variables. Each variable is associated to a Conditional Probability Distribution (CPD) that indicates the degree of influence among variables.

With the knowledge modeled one can make some probability inferences.

There are two kinds of inference algorithms: exact and approximate algorithms. The former are the most precise, but they consume much time and machine resources. Sometimes they are not appropriate when the BN has many variables and a complex network topology. The latter consumes fewer resources and sometimes is the only alternative to get the inference done. They do not present exact result, but are useful when the error rate is acceptable not compromising the result.

Learning a BN means to build the graph representing the dependencies among the variables of the domain and the conditional probability distribution for each variable. A BN can be learned from a specialist. However, in domains with a lot of variables and complex dependences among them, this process can be costly and prone of errors. Therefore, it is interesting to use methods for learning a BN automatically from data. Examples of machine learning algorithms for learning BN are K2, Hill-Climbing (Koller et al., 2009) and Simulated Annealing.

4 NPV CALCULATION AND BAYESIAN NETWORKS

In this article, it's proposed to use historical dataset for learning a model able to both predict the NPV and depict the exploration domain. Since, there is uncertainty concerning the analyses of an Exploration Opportunity the learned model is represented through a Bayesian Network. Therefore, the following methodology is used: (i) Elicit with specialists the relevant variables concerning

economic analysis of an EO, (ii) Gather historical data about these variables, (iii) Learn a BN using the collected data, (iv) Validate the learned BN with the specialists.

The step (iii) consists of executing the KDD process described in Section 3. In the data mining step, it was used Bayesian algorithms to learn the BN. Any toolbox for data mining, as Weka (Seewald et al, 2010), can be used for this task. During the learning process validation metrics as Accuracy, Correlation Coefficient and ROC Curve (Witten et al, 2011) are used. Moreover, to avoid overfitting, Cross Validation is considered. Therefore, the BN can be used for predicting the NPV value of an EO.

Besides the data analysis in step (iii), to verify if the model has good generalization, the step (iv) has the objective to validate the model with the specialists. In this evaluation it's possible to verify if the final model describes in a concise way the domain in study.

5 EXPERIMENTAL ANALYSES

An Experimental Analyses was conducted using data from a petroleum company. After consulting the specialists, the relevant variables were defined (Table 1). And it was collected 700 examples. All variables are continuous-valued except Basin and Fluid were nominal.

Table 1: Relevant variables to calculate NPV.

Variable	Description
Volume	Estimate volume of oil (meter)
Water_Depth	Depth of ocean (meter)
Profundity	Depth of the EO (meter)
Oil_Quality	Measure of the quality of oil (API)
Taxes	Brazilian taxes over oil production
TMA	Minimum attractive rate (return expected by the company)
Oil Price	Price of the oil in the market (USD)
Fluid	Type of fluid (oil or gas)
distance	Distance from EO to shore (meter)
Basin	Sedimentary Basin where is located the EO
Rock Variables	
Area	Area of rock that contains oil (Km2)
Thickness	Thickness of the rock (meter)
Porosity	Porosity of the rock (%)
Permeability	Permeability of rock (mD)
Saturation	Saturation of the rock (%)

5.1 Pre-processing

Some pre-processing techniques were performed. First, outliers were identified in the dataset. Algorithms based on quartile values were applied for this task. Next, normalization algorithms were applied to continuous attributes, converting the values to a range between zero and one [0;1].

Finally, all numeric attributes were discretized using bins with equal width. The number of bins was optimized for each attribute using a proper algorithm.

5.2 Data Mining

After the pre-processing step Data Mining was applied. Algorithm K2 was the learning algorithm used. One restriction imposed to the learning algorithms was to consider a maximum of 3 parents for each variable. The learned model presented an accuracy value of 70%. Another way of analyze the performance of a model is through the Confusion Matrix. This matrix shows the performance of the model in more details. The Confusion Matrix shown that the model learned presented low TP rate for some classes. Besides that, some classes have no representative, leading to conclusion that the dataset was unbalanced.

In an attempt to reduce the class imbalance, it was applied the SMOTE algorithm (Chawla et al., 2002). It produced a resamping using the Synthetic Minority Oversampling technique. This technique creates synthetic examples of the classes less represented. After the application of SMOTE it was noted a gain of 0.48 in TP rate in average. The accuracy was also improved, from 70 to 78%.

With the objective of finding a more precise model, it was applied others BN learning algorithms: Hill Climber (without restriction on variable order), Tabu (similar to Hill climber, but extends the search a little more, even when it finds a supposed optimum point). Besides that, it was tested some parameter options as Local/Global Scope and Maximum number of parent nodes (2 or 3).

Naïve Bayes algorithm was used as baseline for comparison. The results are shown in Figure 3. Each BN model was built using different algorithms, metric scope and number of parent nodes.

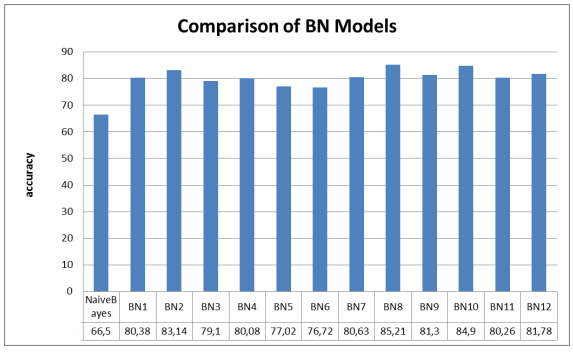


Figure 3: BN model comparison.

Figure 3 shows that all models were better than Naive Bayes baseline, and the best of all in performance had options Global Scope, K2 algorithm and a limit of 3 parents (BN8 model). Judging by the metrics, our BN model has achieved good performance.

5.3 Post-processing

After the data mining process, the learned Bayesian Network was presented to the specialists in order to validate it. They approved, but asked for some minus changes in order to better represent the domain. For example, they knew that the variables Area, Thickness and Porosity strongly influence the Volume, so the graph was rearranged to express this knowledge and the CPDs was relearned.

6 RELATED WORKS

In (Schoeninger, 2003), a specialist system based on Fuzzy Logic that estimates the risks of exploration activities was proposed. Moreover, it calculates the probability of a geologic success of an EO. Some experiments with BN were performed, but the author gave it up claiming having troubles building the CPDs, since they were built manually, collecting information from the specialists. Different from our proposal, Schoeninger did not consider economic aspects of an EO.

In (Junior, 2003), a Neural Network to predict the optimum value of a bid at an auction of exploration areas was defined. The focus of our work concerns a posterior period, when the auction was finished, the concession granted to a company, and there is no more competition for petroleum areas.

7 CONCLUSIONS

This article exposed the problems that involve an economic evaluation of a Petroleum Exploration Opportunity and how it intends to contribute to solve these problems.

Ours preliminary experiments indicate that is possible to build a model able to predict the Net Present Value related to an Exploration Opportunity. The dataset used in ours experiments contains only information about areas located in Brazil and explored by a Brazilian company. The model learned can be used by other companies to evaluate their Exploration Opportunity.

The authors have submitted a similar article at IADIS 2011, but this time our work is more complete and with new experiments.

REFERENCES

Junior, Repsol, (2003). "*A Competição e a Cooperação na Exploração e Produção de Petróleo*" – COPPE/UFRJ – Master thesis. Pag 62-63; 171. Energetic Planning.

Schoeninger, C., 2003. "*Tratamento de Informações Imperfeitas na Análise de Risco de Prospectos em Exploração Petrolífera*" – Federal University of Santa Catarina (UFSC) – Master thesis.

Newendorp, P., Schuyler, J., (2009). "Decision Analysis for Petroleum Exploration", 2nd Edition, *Planning Press*, Pag 24.

Witten, I., Frank, E., (2011). "Data Mining: Practical Machine Learning Tools and Techniques", *3rd ed. Elsevier*. Pag 5; 9; 278-279.

Canlas, R., (2009). "*Data Mining in Healthcare: Current Applications and Issues*", article to fulfill requirements for the Master of Science in Information Technology, Carnegie Mellon University - Australia

Chawla, N., Bowyer, K., Hall, Kegelmeyer, W., (2002). "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, Morgan Kaufmann Publishers.

Mitchell, T., (1997). "*Machine Learning*", McGraw Hill.

Seewald, A., Scuse, D., (2010). "Weka Manual", University of Waikato http://www.cs.waikato.ac.nz/ml/weka/.

Koller, D., Friedman, N., (2009). "*Probabilistic Graphical Models*", The MIT Press.

PSO-based Linear SVM Classifier Selection for Credit Risk Evaluation Modeling Process

Paulius Danenas and Gintautas Garsva
Department of Informatics, Kaunas Faculty, Vilnius University, Muitines St. 8, LT- 44280 Kaunas, Lithuania
{paulius.danenas, gintautas.garsva}@khf.vu.lt

Keywords: Support Vector Machines, Linear SVM, Particle Swarm Optimization, Credit Risk, Evaluation, Bankruptcy, Machine Learning

Abstract: A research on credit risk evaluation modelling using linear Support Vector Machines (SVM) classifiers is proposed in this paper. The classifier selection is automated using Particle Swarm Optimization technique. Sliding window approach is applied for testing classifier performance, together with other techniques such as discriminant analysis based scoring for evaluation of financial instances and correlation-based feature selection. The developed classifier is applied and tested on real bankruptcy data showing promising results.

1 INTRODUCTION

Credit risk evaluation is defined as one of the most important domains in financial sector as it shows the ability to regenerate income by lending money; yet, calculation of the possibility to get back the money invested is the most critical problem. Machine learning and artificial intelligence techniques are novel and state-of-the-art methods which help to develop tools for this problem by overcoming the drawbacks of statistical tools and deriving more robust and accurate solutions.

Discriminant analysis was one of the first techniques applied in credit evaluation (Altman, 1968). Support Vector Machines (SVM) classifiers gained a lot of attention as they showed abilities to get classification results comparable to Neural Networks but avoiding their main difficulties such as local minimas. Selection of hyperparameters is a sophisticated task thus various metaheuristic and evolutionary techniques have been adopted for solving this task including swarm intelligence techniques such as Ant colony Optimization (Zhou et al, 2007). Particle Swarm Optimization (abbr. PSO) has previously been applied for SVM optimization in credit risk domain – personal credit scoring (Xuchuan et. al, 2007), financial distress prediction (Chen et al., 2010; Wang, 2010), consumer credit scoring analysis (Yun et al., 2011). Linear SVM (LIBLINEAR) has also been tested to show competitive results to original C-SVC classifier (Danenas et. al, 2010; Danenas et al, 2011), which proved that they can be a good alternative in terms of both complexity and speed. According to these aspects, linear SVM and PSO are selected for model development. The research presented in this paper proposes a hybrid method based on linear Support Vector Machines classification and Particle Swarm Optimization. The proposed method is also tested in "sliding window" approach manner, which means that it can be useful to identify more general trends. Moreover, proposed approach might be useful while trying to improve the performance of these methods by identifying the most relevant financial attributes and developing a new classifier based on that particular technique.

2 USED METHODS

Support Vector Machines (SVM). SVM solves following quadratic minimization problem:

$$\min -\sum_{i=1}^{\ell}\alpha_i + \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\textit{subject to } \sum_{i=1}^{\ell} y_i \alpha_i = 0,\ \forall i : 0 \le \alpha_i \le C$$

where the number of training examples is denoted by l, training vectors $X_i \in R, i = 1,..,l$ and a vector $y \in R^l$ such as $y_i \in [-1;1]$. α is avector of l values where each component α_i corresponds to a training example (x_i, y_i). If training vectors x_i are not linearly

separable, they are mapped into a higher (maybe infinite) dimensional space by the kernel function $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$.

Fan et al. (Fan et al., 2008) proposed a family of linear SVM and logistic regression classifiers for large-scale SVM classification which do not use kernel functions for transformation into other dimensional space; although with less flexibility, it can perform effectively especially using large amounts of data. The formulations of the algorithms by are given in the paper of Fan et al.; four of them (L2-regularized L1-loss SVC, L2-regularized L2-loss SVC, L2-regularized logistic regression, L1-regularized L2-loss SVC) are used in the experiment. These classifiers are formulated as minimization problems, but they all share the concept of cost parameter C and bias usage. This proposes a possibility for heuristic selection of classifiers themselves.

Particle Swarm Optimization (PSO). The PSO algorithm, introduced by Kennedy and Eberhart, is based on behavior of flock of birds which search for food randomly in some area, knowing only the distance from the food. Thus all the particles have one fitness PSO is expressed in terms of particles (birds) and searched target described by fitness value; the location of each particle is determined by velocity describing its flying direction and distance. Two extreme values are tracked by each particle - the optimal solution found by the particle itself (*pbest*), and the optimal solution found by the whole swarm (*gbest*). Unmodified PSO algorithm is used in this research, thus its details are not presented in this paper, but can be found in other sources, such as (Kennedy et al., 2001).

2.1 PSO Approach for Linear SVM Optimization

A classification technique based on Particle Swarm Optimization and linear SVM combination, namely PSO-LinSVM is proposed in this paper. Each particle P = <p_1;p_2;p_3> is represented as follows:

p_1 – integer value, that represents the algorithm used for classification:

0 - L2-regularized logistic regression
1 – L2-regularized L2-loss SVC
2 – L1-regularized L2-loss SVC
3 - L2-regularized L1-loss SVC

p_2 – real value, cost parameter C

p_3 –real value, which represents bias term

The fitness function is defined as maximization of sum of TPR values:

$$f(fitness) = \sum_{N_C}^{1} TPR_i ,$$

where N_C is the number of classes. Most of the authors (Wang, Chin et al.) choose accuracy for fitness evaluation; however, in case of imbalanced learning, accuracy is not the best option, so sum of TP rate values is selected for this case, which allows selection of classifier that balances between identification of both "majority" and "minority" classes. These evaluations are obtained by performing k-fold cross-validation training; k is considered to be quite small (k = 5 is used for the experiment), considering the amount of data used in research. The optimal solution can be obtained only in case of perfect classification; as this happens very rarely, the main goal is to find best satisfactory solution.

2.2 Sliding Window Testing Approach

This research adopts techniques used earlier by Danenas et al. (Danenas et al., 2010; Danenas et al., 2011), extending it with PSO application for classifier optimization step. Thus the modified algorithm is defined as follows:

1. Evaluate each financial entry manually or by using expert techniques to compute bankruptcy classes (discriminant models used in banking are sued in this research).
2. Apply data preprocessing steps – elimination of unevaluated instances, data imputation and standardization.
3. Perform the following steps for each $m \in [1, n-k]$, where n is the total number of periods, k is the number of periods are used for forecasting:
   a. Feature selection;
   b. Classifier and parameter selection, using Particle Swarm Optimization;
   c. Train classifier using data from first m periods.
   d. Apply hold-out testing using data from period p, $p \in [m+1, m+k]$; $p \in \mathrm{N}$.

Note, that feature selection step is important for 2 reasons:

1. Quality and complexity - data dimensionality reduction;

2. Ratio importance - a new classifier based on other evaluator but using a set of statistically significant attributes obtained from the data is developed.

The output of each iteration in experimental stage is the trained classifier and the list of selected

attributes for each period.

3 EXPERIMENT RESULTS

3.1 Data used in the Experiment

The dataset that was applied for the experiment consists of entries from 785 USA Transportation, Communications, Electric, Gas, And Sanitary Services companies with their 1999-2008 yearly financial records (balance and income statement) from financial EDGAR database.

Each instance has 51 financial attributes (indices used in financial analysis). "Risky" and "Non-Risky" classes were formed using Zmijewski's scoring technique widely used in banking.

Table 1: Main characteristics of datasets used in experiments.

Year	Entries labeled as		Total entries	No of selected attributes	Bankrupt 1 years after	Bankrupt >1 year after
	Risky (R)	Not risky (NR)				
1999	376	166	542	11	-	-
2000	423	192	615	8	0	0
2001	383	226	609	13	2	1
2002	376	239	615	11	1	0
2003	417	220	637	9	0	0
2004	460	194	654	9	1	1
2005	478	173	651	8	1	4
2006	375	118	493	8	0	1
2007	367	112	479	11	0	6
2008	38	12	50	8	-	-
Total	**3693**	**1652**	**5345**		**5**	**13**

Note that ratios in original Zmijewski were not used in order to avoid linear dependence between variables. Main characteristics of the datasets formed for the experiment are presented in Table 1. It also shows financial ratios which were considered relevant by feature selection procedure; the number of such features is larger than the ones which are considered in original evaluator.

3.2 Computational Results

Correlation-based feature subset selection (Hall, 2001) algorithm with Tabu search for search in attribute subsets was applied for feature selection.

The search space for PSO was set to $C \in [0;50]$, $bias \in [0;1]$, as well as the number of run iterations was set to 10. PSO was configured to run with 20 particles and inertia rate of 0.8. Velocity for p_2 was set to 3, for p_3 was set to 0.2.

Table 2 presents the results obtained by PSO-LinSVM classifier: classifier parameters, obtained by PSO, classification accuracy together with True Positive and F-Measure rates for each class. It is clear that classification accuracy did not show stable increase while providing the classifier with more data each year. While performing testing procedure with first year data, accuracy decreased to 80% in 2004 although next year it returned to 83.8% was relatively stable, and later in fell to 82%; similar trends might be identified while analyzing testing results obtained with Year 2 and Year 3 data. It is important to note that instances marked as "risky" were identified better.

Table 2: Experimental classification results.

Training period			2000	2001	2002	2003	2004	2005	2006	2007
Linear classifier			L1-SVM (dual)	L2-SVM (dual)	L2-SVM (dual)	L2-RLR	L2-SVM (primal)	L2-SVM (dual)	L2-SVM (dual)	L2-SVM (primal)
C			15,3157	47,8343	24,7346	29,0490	22,3727	38,0860	6,5322	48,0734
Bias			1,000	0,196	0,749	0,797	0,873	0,838	0,436	0,508
Year 1	Accuracy		77,941	78,409	80,220	83,689	80,640	83,806	82,887	82,000
	TP	R	0,969	0,952	0,981	0,987	0,952	0,957	0,970	0,974
		NR	0,461	0,521	0,464	0,482	0,412	0,462	0,385	0,333
	F-Measure	R	0,846	0,843	0,867	0,895	0,878	0,900	0,896	0,892
		NR	0,609	0,653	0,618	0,637	0,535	0,579	0,520	0,471
Year 2	Accuracy		80,032	77,080	84,146	83,232	83,806	84,742	82,000	-
	TP	R	0,979	0,947	0,985	0,990	0,957	0,959	0,974	-
		NR	0,521	0,436	0,503	0,407	0,462	0,496	0,333	-
	F-Measure	R	0,857	0,844	0,897	0,896	0,900	0,905	0,892	-
		NR	0,670	0,568	0,653	0,567	0,579	0,611	0,471	-
Year 3	Accuracy		77,237	80,488	83,384	86,032	84,124	84,000	-	-
	TP	R	0,966	0,952	0,987	0,987	0,967	0,974	-	-
		NR	0,405	0,456	0,418	0,462	0,444	0,417	-	-
	F-Measure	R	0,848	0,873	0,897	0,915	0,902	0,902	-	-
		NR	0,551	0,582	0,576	0,615	0,575	0,556	-	-
Average testing accuracy			**78,403**	**78,660**	**82,583**	**84,318**	**82,857**	**84,183**	**82,444**	**82**

4 CONCLUSIONS AND FUTURE DEVELOPMENT

This paper presents an approach for credit risk evaluation using linear SVM classifiers, selected and optimized by Particle Swarm Optimization, combined with sliding window testing technique and feature selection using correlation analysis. Linear SVM classifiers perform well when applied to large scale problems; this is one of the main reasons why they were selected as classification technique. The developed classifiers were applied for real-world dataset, combined with widely applied Zmijewski technique as an evaluator and basis for output formation. Analysis of experimental results shows that the performance still needs to be improved to be more stable and reliable. Particle Swarm Optimization topology has not been investigated in this research, thus further steps will involve more detailed investigation into PSO performance. Imbalanced learning is another field where significant improvements might lead to increase in overall performance; this procedure is especially important if labelling is done automatically (as, in our case, using Zmijewski's model), as this might lead to highly imbalanced datasets. Notably, misidentification of bankrupt company might cost more to the creditor than the misdentification of "healthy" one, thus this problem is especially important if there are much less bankrupt companies or companies with high risk than companies which belong to another classes.

REFERENCES

Altman, E., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. In *The Journal of Finance*, Vol. 23 (4), pp.589–-609. American Finance Association; Blackwell Publishing

Chen, C.-Y., Chen, M.-Y., Hsieh C.-H., 2010. A Financial Distress Prediction System Construction based on Particles Swarm Optimization and Support Vector Machines. In *Proceedings of 2010 International Conference on E-business, Management and Economics* (IPEDR), Vol.3, IACSIT Press, Hong Kong pp. 165-169.

Danenas P., Garsva G., 2010. Credit risk evaluation using SVM-based classifier. In *Lecture notes in Business Information Processing*, Heidelberg: Springer-Verlag Vol. 57, Part 1, 2010, pp. 7-12, Springer.

Danenas P., Garsva G., Gudas S., 2011. Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. In *Proceedings of the International Conference on Computational Science (ICCS 2011*), Vol. 4, pp. 1699-1707, Procedia Computer Science.

Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C., 2008. LIBLINEAR: A library for large linear classification, In *The Journal of Machine Learning Research*, Vol. 9, pp.1871–4.

Hall, M.A. *Correlation-based Feature Subset Selection for Machine Learning*. Hamilton, New Zealand (1998)

JSwarm-PSO: Swarm optimization package, http://jswarm-pso.sourceforge.net/

Kennedy, J., Eberhart, R. C., Shi, Y. *Swarm intelligence*. Morgan Kaufmann Publishers, 2001.

Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

Wang X., 2010. Corporate Financial Warning Model Based on PSO and SVM. In *2010 2nd International Conference on Information Engineering and Computer Science (ICIECS)*, Wuhan, pp.1-5.

Xuchuan, J.M.Y., 2007. Construction and Application of PSO-SVM Model for Personal Credit Scoring. *Proc. of the 7th international conference on Computational Science, ICCS '07, Part IV*, pp. 158-161.

Yun, L., Cao Q.-Y.; Zhang H., 2011. Application of the PSO-SVM Model for Credit Scoring. Proceedings of *Seventh International Conference on Computational Intelligence and Security (CIS),* pp.47-51.

Zhou J., Zhang A., Bai T., 2008. Client Classification on Credit Risk Using Rough Set Theory and ACO-Based Support Vector Machine. In: *Proceedings of Wireless Communications, Networking and Mobile Computing (WiCOM '08)*, pp.1-4

Zmijewski, M,, 1984. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. In *Journal of Accounting Research*, Vol. 22, pp. 59–82.

Polymorphic Random Building Block Operator for Genetic Algorithms

Ghodrat Moghadampour
VAMK, University of Applied Sciences, Technology and Communication, Wolffintie 30, 65200, Vaasa, Finland
mg@puv.fi

Keywords: Evolutionary Algorithm, Genetic Algorithm, Function Optimization, Mutation Operator, Multipoint Mutation Operator, Polymorphic Random Building Block Operator, Fitness Evaluation and Analysis.

Abstract: Boosting the evolutionary process of genetic algorithms by generating better individuals, avoiding stagnation at local optima and refreshing population in a desirable way is a challenging task. Typically operators are used to achieve these objectives. On the other hand using operators can become a challenging task in itself if applying them requires setting many parameters through human intervention. Therefore, developing operators, which do not require human intervention and at the same time are capable of assisting the evolutionary process, is highly desirable. Most typical genetic operators are mutation and crossover. However, experience has proved that these operators in their classical form are not capable of refining the population efficiently enough. In this work a new dynamic mutation operator called polymorphic random building block operator with variable mutation rate is proposed. This operator does not require any pre-fixed parameter. It randomly selects a section from the binary presentation of the individual, then generates a random bit-string of the same length as the selected section and applies bitwise logical AND, OR and XOR operators between the randomly generated bit-string and the selected section from the individual. In the next step all three newly generated offspring will go through selection procedure and will replace a possibly worse individual in the population. Experimentation with 33 test functions and 11550 test runs proved the superiority of the proposed dynamic mutation operator over single-point mutation operator with 1%, 5% and 8% mutation rates and the multipoint mutation operator with 5%, 8% and 15% mutation rates.

1 INTRODUCTION

Most often genetic algorithms (GAs) have at least the following elements in common: populations of chromosomes, selection according to fitness, crossover to produce new offspring, and random mutation of new offspring.

A simple GA works as follows: 1) A population of n l-bit strings (chromosomes) is randomly generated, 2) the fitness $f(x)$ of each chromosome x in the population is calculated, 3) chromosomes are selected to go through crossover and mutation operators with p_c and p_m probabilities respectively, 4) the old population is replaced by the new one, 5) the process is continued until the termination conditions are met.

However, more sophisticated genetic algorithms typically include other intelligent operators, which apply to the specific problem. In addition, the whole algorithm is normally implemented in a novel way with user-defined features while for instance measuring and controlling parameters, which affect the behaviour of the algorithm.

1.1 Genetic Operators

For any evolutionary computation an appropriate representation (encoding) of problem variables must be chosen along with the appropriate evolutionary computation operators. Data might be represented in different formats: binary strings, real-valued vectors, permutations, finite-state machines, parse trees and so on.

Decision on what genetic operators to use greatly depends on the encoding strategy of the GA. For each representation, several operators might be employed (Michalewicz, 2000). The most commonly used genetic operators are crossover and mutation.

1.1.1 Crossover

The simplest form of crossover is single-point: a single crossover position is chosen randomly and the parts of the two parents after the crossover position are exchanged to form two new individuals (offspring). The idea is to recombine building blocks (schemas) on different strings.

In two-point crossover, two positions are chosen at random and the segments between them are exchanged. Two-point crossover reduces positional bias and endpoint effect, it is less likely to disrupt schemas with large defining lengths, and it can combine more schemas than single-point crossover (Mitchell, 1998). Two-point crossover has also its own shortcomings; it cannot combine all schemas.

Multipoint-crossover has also been implemented, e.g. in one method, the number of crossover points for each parent is chosen from a Poisson distribution whose mean is a function of the length of the chromosome. Another method of implementing multipoint-crossover is the "parameterized uniform crossover" in which each bit is exchanged with probability p, typically $0.5 \le p \le 0.8$ (Mitchell, 1998).

In parameterized uniform crossover, any schemas contained at different positions in the parents can potentially be recombined in the offspring; there is no positional bias. This implies that uniform crossover can be highly disruptive of any schema and may prevent coadapted alleles from ever forming in the population (Mitchell, 1998).

The one-point and uniform crossover methods have been combined by some researchers through extending a chromosomal representation by an additional bit. There has also been some experimentation with other crossovers: segmented crossover and shuffle crossover (Eshelman et al., 1991; Michalewicz, 1996).

Segmented crossover, a variant of the multipoint, allows the number of crossover points to vary. The fixed number of crossover points and segments (obtained after dividing a chromosome into pieces on crossover points) are replaced by a segment switch rate, which specifies the probability that a segment will end at any point in the string.

The shuffle crossover is an auxiliary mechanism, which is independent of the number of the crossover points. It 1) randomly shuffles the bit positions of the two strings in tandem, 2) exchanges segments between crossover points, and 3) unshuffles the string (Michalewicz, 1996). In gene pool recombination, genes are randomly picked from the gene pool defined by the selected parents.

1.1.2 Mutation

The common mutation operator used in canonical genetic algorithms to manipulate binary strings $a = (a_1,...a_\ell) \in I = \{0,1\}^\ell$ of fixed length ℓ was originally introduced by Holland (Holland, 1975) for general finite individual spaces $I = A_1 \times ... A_\ell$, where $A_i = \{\alpha_{i_1},...,\alpha_{i_{k_l}}\}$. By this definition, the mutation operator proceeds by:

i. determining the position $i_1,...,i_h (i_j \in \{1,...,l\})$ to undergo mutation by a uniform random choice, where each position has the same small probability p_m of undergoing mutation, independently of what happens at other position
ii. forming the new vector $a'_i = (a_1,...,a_{i_1-1}, a'_{i_1}, a_{i_1+1},...,a_{i_h-1}, a'_{i_h}, a_{i_h+1},...a_\ell)$, where $a'_i \in A_i$ is drawn uniformly at random from the set of admissible values at position i.

The original value a_i at a position undergoing mutation is not excluded from the random choice of $a'_i \in A_i$. This implies that although the position is chosen for mutation, the corresponding value might not change at all (Bäck et al., 2000).

Mutation rate is usually very small, like 0.001 (Mitchell, 1998). A good starting point for the bit-flip mutation operation in binary encoding is $P_m = 1/L$, where L is the length of the chromosome (Mühlenbein, 1992). Since $1/L$ corresponds to flipping one bit per genome on average, it is used as a lower bound for mutation rate. A mutation rate of range $P_m \in [0.005, 0.01]$ is recommended for binary encoding (Ursem, 2003). For real-value encoding the mutation rate is usually $P_m \in [0.6, 0.9]$ and the crossover rate is $P_m \in [0.7, 1.0]$ (Ursem, 2003).

While recombination involves more than one parent, mutation generally refers to the creation of a new solution from one and only one parent. Given a real-valued representation where each element in a population is an n-dimensional vector $x \in \Re^n$, there are many methods for creating new offspring using mutation. The general form of mutation can be written as:

$$x' = m(x) \quad (1)$$

where x is the parent vector, m is the mutation function and x' is the resulting offspring vector. The more common form of mutation generated offspring vector:

$$x' = x + M \quad (2)$$

where the mutation M is a random variable. M has often zero mean such that

$$E(x') = x \quad (3)$$

the expected difference between the real values of a parent and its offspring is zero (Bäck et al., 2000).

Some forms of evolutionary algorithms apply mutation operators to a population of strings without using recombination, while other algorithms may combine the use of mutation with recombination. Any form of mutation applied to a permutation must yield a string, which also presents a permutation. Most mutation operators for permutations are related to operators, which have also been used in neighbourhood local search strategies (Whitley, 2000). Some other variations of the mutation operator for more specific problems have been introduced in (Bäck et al., 2000). Some new methods and techniques for applying crossover and mutation operators have also been presented in (Moghadampour, 2006).

1.1.3 Other Operators and Mating Strategies

In addition to common crossover and mutation some other operators are used in GAs including inversion, gene doubling and other operators for preserving diversity in the population. For instance, a "crowding" operator has been used in (De Jong, 1975; Mitchell, 1998) to prevent too many similar individuals ("crowds") from being in the population at the same time. This operator replaces an existing individual by a newly formed and most similar offspring.

In (Mengshoel et al., 2008) a probabilistic crowding niching algorithm in which subpopulations are maintained reliably, is presented. It is argued that like the closely related deterministic crowding approach, probabilistic crowding is fast, simple, and requires no parameters beyond those of classical genetic algorithms.

Diversity in the population can also be promoted by putting restrictions on mating. For instance, distinct "species" tend to be formed if only sufficiently similar individuals are allowed to mate (Mitchell, 1998). Another attempt to keep the entire population as diverse as possible is disallowing mating between too similar individuals, "incest" (Eshelman et al., 1991; Mitchell, 1998).

Another solution is to use a "sexual selection" procedure; allowing mating only between individuals having the same "mating tags" (parts of the chromosome that identify prospective mates to one another). These tags, in principle, would also evolve to implement appropriate restrictions on new prospective mates (Holland, 1975).

Another solution is to restrict mating spatially. The population evolves on a spatial lattice, and individuals are likely to mate only with individuals in their spatial neighborhoods. Such a scheme would help preserve diversity by maintaining spatially isolated species, with innovations largely occurring at the boundaries between species (Mitchell, 1998).

The efficiency of genetic algorithms has also been tried by imposing adaptively, where the algorithm operators are controlled dynamically during runtime (Eiben et al. 2008). These methods can be categorized as deterministic, adaptive, and self-adaptive methods (Eiben & Smitt, 2007; Eiben et al. 2008). Adaptive methods adjust the parameters' values during runtime based on feedback from the algorithm (Eiben et al. 2008), which are mostly based on the quality of the solutions or speed of the algorithm (Smit et al., 2009).

2 THE POLYMORPHIC RANDOM BUILDING BLOCK OPERATOR

The *polymorphic random building block* (PRBB) operator is a new self-adaptive operator proposed here. The random building block (*RBB*) operator was originally presented in (Moghadampour, 2006; Moghadampour, 2011; Moghadampour, 2012), where promising results were also reported.

In this paper we modify the original idea of the operator by applying multiple logical bitwise operators, namely AND, OR and XOR during mutation process in order to produce new offspring. During the classical crossover operation, building blocks of two or more individuals of the population are exchanged in the hope that a better building block from one individual will replace a worse building block in the other individual and improve the individual's fitness value. However, the polymorphic random building block operator involves only one individual.

The polymorphic random building block operator resembles more the multipoint mutation operator, but it lacks the frustrating complexity of such an operator. The reason for this is that the random building block operator does not require any pre-defined parameter value and it automatically

takes into account the length (number of bits) of the individual at hand. In practice, the polymorphic random building block operator selects a section (s_1) of random length (l_s) from the binary presentation of the individual at hand. In the next step the operator produces randomly a binary string (s_2) of the same size (l_s) and then applies AND, OR and XOR bitwise operators between s_1 and s_2 in turn in order to produce three new offspring. In the next step these newly generated offspring go through selection procedure one by one to be either selected or discarded.

This operator can help breaking the possible deadlock when the classic crossover operator fails to improve individuals. It can also refresh the population by injecting better building blocks into individuals, which are not currently found from the population. Figure 1 describes the random building block operator.

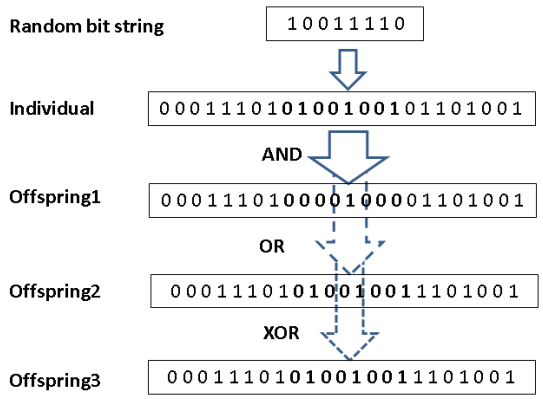


Figure 1: The polymorphic random building block operator. A random building block is generated and is combined with the individual through AND, OR and XOR operators to generate three new offspring.

This operation is implemented in the following order: 1) for each individual *ind* of binary length l in the population a section length l_s proportionate to the number of variables in the problem is randomly generated so that $1 \le l_s \le \frac{l}{number\_of\_variables}$, 2) two crossover points cp_1 and cp_2 are randomly selected so that $l_s = |cp_2 - cp_1|$, 3) a random bit string *bstr* of length l_s is generated, 4) bits between the crossover points on the individual *ind* go through bitwise AND, OR and XOR logical operators with the bits on the bit string *bstr* to generate three new offspring, and 5) each newly generated offspring go through the selection procedure.

2.1 Survivor Selection

After each operator application, new offspring are evaluated and compared to the population individuals. Newly generated offspring will replace the worst individual in the population if they are better than the worst individual. Therefore, the algorithm is a steady state genetic algorithm.

3 EXPERIMENTATION

The random building block operator, three versions of single-point mutation operator (with 1%, 5% and 8% mutation rates) and three versions of multipoint mutation operator (with 5%, 8% and 15% mutation rates) were implemented as prat of a genetic algorithm to solve the following demanding minimization problems: Ackley's ($\forall x_i : -32.768 \le x_i \le 32.768$), Colville's ($\forall x_i : -10 \le x_i \le 10$), Griewank's F1 ($\forall x_i : -600 \le x_i \le 600$), Rastrigin's ($\forall x_i : -5.12 \le x_i \le 5.12$), Rosenbrock's ($\forall x_i : -100 \le x_i \le 100$) and Schaffer's F6 ($\forall x_i : -100 \le x_i \le 100$). Some of these functions have a fixed number of variables and some others are multidimensional in which the number of variables could be determined by the user. For multidimensional problems with an optional number of dimensions (n), the algorithm was tested for $n = 1, 2, 3, 5, 10, 30, 50, 100$. The exception to this was the Rosenbrock's function for which the minimum number of variables is 2. The efficiency of each of the operators in generating better fitness values was studied.

During experimentation only one operator was tested at each time. To simplify the situation and clarify interpretation of experimentation results the operators were not combined with other operators, like crossover.

Single-point mutation operator was implemented so that the total number of mutation points ($total\_mut\_points$) was calculated by multiplying the mutation rate (m_rate) by the binary length of the individual (ind_bin_length) and the population size (pop_size):

$$total\_mut\_points = m\_rate \times ind\_bin\_length \times pop\_size \quad (4)$$

Then during each generation for the total number of mutation points one gene was randomly selected from an individual in the population and mutated. Multipoint mutation operator was implemented so that during each generation for the total number of

mutation points ($total\_mut\_points$) a random number of mutation points (sub_mut_points) from a random number of individuals in the population was selected and mutated. This process was continued until the total number of mutation points was consumed:

$$total\_mut\_points = \sum_{i=1}^{n} sub\_mut\_points_i \quad (5)$$

For each test case the steady-state algorithm was run for 50 times. The population size was set to 9 and the maximum number of function evaluations for each run was set to 10000. The exception to this was the Rosenbrock's function for which the number of function evaluations was set to 100000 in order to get some reasonable results.

The mapping between binary strings into floating-point numbers and vice versa was implemented according to the following well-known steps:

1. The distance between the upper and the lower bounds of variables is divided according to the required precisions, $precision$ (e.g. the precision for 6 digits after the decimal point is $1000000_{(10)}$) in the following way:

$$(upperbound - lowerbound) \times precision \quad (6)$$

2. Then an integer number l is found so that:

$$(upperbound - lowerbound) \times precision \leq 2^l \quad (7)$$

Thus, l determines the length of binary representation, which implies that each chromosome in the population is l bits long. Therefore, if we have a binary string x' of length l, in order to convert it to a real value x, we first convert the binary string to its corresponding integer value in base 10, $x'_{(10)}$ and then calculate the corresponding floating-point value x according to the following formula:

$$x = lowerbound + x'_{(10)} \times \frac{upperbound - lowerbound}{2^l - 1} \quad (8)$$

The variable and solution precisions set for different problems were slightly different, but the same variable and solution precisions were set the same for all operators. During each run the best fitness value achieved during each generation was recorded. This made it possible to figure out when the best fitness value of the run was actually found. Later at the end of 50 runs for each test case the average of the best fitness values and the required function evaluations were calculated for comparison. In the following, test results for comparing the efficiency of polymorphic random building block operator with different versions of mutation operator are reported.

Experimentation results indicated that the polymorphic random building block operator had produced much better results than different versions of the single-point mutation operator in all test cases. The difference in performance seemed to be significant for Colville's function and Ackley's and Griewank's functions when the number of variables increases.

Very low p-values for T-test and F-test indicated that the performance values achieved by Polymorphic Random Building Block operator were significantly smaller than the ones achieved by other operators.

The performance of the polymorphic random building block operator against the single-point mutation operator was also tested on Rastrigin's, Rosenbrock's and Schaffer's F6 functions.

Studying results proved that the polymorphic random building block operator has been able to produce significantly better results in more than 87% of test cases. The results indicated that for Rosenbrock50 and Rosenbrock 100 the polymorphic random building block had on average produced worse results than the single mutation point operator. However, studying the results showed that there are huge differences between the median values (in parentheses) of test results for the benefit of the polymorphic random building block. While the median values for polymorphic random building block operator were less than the average values, the situation was vice versa in all cases for different versions of single point mutation operators. For Rosenbrock50 in 58% of test cases the fitness value achieved by polymorphic random building block operator was less than 351, which is the average of fitness values achieved by single mutation operator with 8% mutation rate. This means that in 58% of test cases polymorphic random building block had a better performance in finding the best fitness value for Rosenbrock's function with 50 variables.

For Rosenbrock100 in 60% of test cases the fitness value achieved by polymorphic random building block operator was less than 342, which is the average of fitness values achieved by single mutation operator with 8% mutation rate. This means that in 60% of test cases polymorphic random building block had a better performance over mutation operator with 1% and 5% mutation rates in finding the best fitness value for Rosenbrock's function with 100 variables.

Very low p-values for T-test and F-test indicated that the performance values achieved by polymorphic random building block operator are significantly smaller than the ones achieved by other operators.

Analysis showed that the differences between average fitness values achieved with different operators wer not significant for Rosenbrock's function with 50 and 100 variables. The superiority of polymorphic random building block operator becomes clear if we recall that it produced in most cases better results than the average values achieved by other operators.

The performance of the polymorphic random building block operator was also compared against the multipoint mutation operator in which several points of the individual were mutated during each mutation operator. As it was earlier mentioned the number of points to be mutated during each mutation operation was randomly determined. Mutation cycles were repeated until total mutation points were utilized. Clearly, the total number of mutation points was determined by the mutation rate, which was 5%, 8% and 15% for different experimentations.

Comparing results proved that the fitness values achieved by the building block operator were better than the ones achieved by different versions of multipoint mutation operator in all cases. Differences between the average fitness values achieved for Ackley's and Griewank's functions with 30, 50 and 100 variables by the polymorphic random building block and different versions of multipoint mutation operator were even more substantial.

Very low p-values for T-test and F-test indicated that the performance values achieved by polymorphic random building block operator were significantly smaller than the ones achieved by other operators.

The performance of the polymorphic random building block operator against the multipoint mutation operator was also tested on Rastrigin's, Rosenbrock's and Schaffer's F6 functions.

Experimentation showed that the polymorphic random building block operator had also outperformed multipoint mutation operator with 5%, 8% and 15% mutation rates. In most cases differences in performance were huge in favour of polymorphic random building block operator.

A small p-value for T-test and very low p-value for F-test indicate that the performance values achieved by polymorphic random building block operator were significantly smaller than the ones achieved by other operators.

4 CONCLUSIONS

In this paper a dynamic mutation operator; polymorphic random building block operator for genetic algorithms was proposed. The operator was tested against single-point mutation operator with 1%, 5% and 8% mutation rates and multipoint mutation operator with 5%, 8% and 15% mutation rates.

Comparing test results revealed that the polymorphic random building block operator was capable of achieving better fitness values within less function evaluations compared to different versions of single-point and multipoint mutation operators. The fascinating feature of polymorphic random building block is that it is dynamic and therefore does not require any pre-set parameter.

However, for mutation operators the mutation rate and the number of mutation points should be set in advance. The polymorphic random building block can be used straight off the shelf without needing to know its best recommended rate. Hence, it lacks frustrating complexity, which is typical for different versions of the mutation operator.

Therefore, it can be claimed that the polymorphic random building block is superior to the mutation operator and capable of improving individuals in the population more efficiently.

4.1 Future Research

The proposed operator can be combined with other operators and applied to new problems and its efficiency in helping the search process can be evaluated more thoroughly with new functions. Moreover, the polymorphic random building block operator can be adopted as part of the genetic algorithm to compete with other state-of-the-art algorithms on solving more problems.

REFERENCES

Bäck, Thomas, David B. Fogel, Darrell Whitely & Peter J. Angeline, 2000. Mutation operators. In: *Evolutionary Computation 1, Basic Algorithms and Operators*. Eds T. Bäck, D.B. Fogel & Z. Michalewicz. United Kingdom: Institute of Physics Publishing Ltd, Bristol and Philadelphia. ISBN 0750306645.

De Jong, K. A., 1975. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*. Ph.D. thesis, University of Michigan. Michigan: Ann Arbor.

Eiben, A. and J. Smith, 2007. *Introduction to Evolutionary Computing*. Natural Computing Series. Springer, 2nd edition.

Eiben, G. and M. C. Schut, 2008. *New Ways To Calibrate Evolutionary Algorithms*. In *Advances in Metaheuristics for Hard Optimization*, pages 153–177.

Eshelman, L. J. & J. D. Schaffer, 1991. Preventing premature convergence in genetic algorithms by preventing incest. In *Proceedings of the Fourth International Conference on Genetic Algorithms*. Eds. R. K. Belew & L. B. Booker. San Mateo, CA : Morgan Kaufmann Publishers.

Holland, J. H., 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor: MI: University of Michigan Press.

Mengshoel, Ole J. & Goldberg, David E., 2008. *The crowding approach to niching in genetic algorithms*. Evolutionary Computation, Volume 16 , Issue 3 (Fall 2008). ISSN:1063-6560.

Michalewicz, Zbigniew (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Third, Revised and Extended Edition. USA: Springer. ISBN 3-540-60676-9.

Michalewicz, Zbigniew, 2000. Introduction to search operators. In *Evolutionary Computation 1, Basic Algorithms and Operators*. Eds T. Bäck, D.B. Fogel & Z. Michalewicz. United Kingdom: Institute of Physics Publishing Ltd, Bristol and Philadelphia. ISBN 0750306645.

Mitchell, Melanie, 1998. *An Introducton to Genetic Algorithms*. United States of America: A Bradford Book. First MIT Press Paperback Edition.

Moghadampour, Ghodrat (2006). *Genetic Algorithms, Parameter Control and Function Optimization: A New Approach*. PhD dissertation. ACTA WASAENSIA 160, Vaasa, Finland. ISBN 952-476-140-8.

Moghadampour, Ghodrat (2011). *Random Building Block Operator for Genetic Algorithms*. 13th International Conference on Enterprise Information Systems (ICEIS 2011), 08 - 11 June 2011Beijing – China.

Moghadampour, Ghodrat (2012). *Outperforming Mutation Operator with Random Building Block Operator in Genetic Algorithms*. In *Enterprise Information Systems International Conference, ICEIS 2011* Beijing, China, June 8-11, 2011 Revised Selected Papers. Eds. Runtong Zhang, Zhenji Zhang, Juliang Zhang, Joaquim Filipe and José Cordeiro. Springer-Verlag LNBIP Series book.

Mühlenbein, H., 1992. How genetic algorithms really work: 1. mutation and hill-climbing. In: *Parallel Problem Solving from Nature 2*. Eds R. Männer & B. Manderick. North-Holland.

Smit, S. K. and Eiben, A. E., 2009. *Comparing Parameter Tuning Methods for Evolutionary Algorithms*. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 399–406, May 2009.

Spears, W. M., 1993. Crossover or mutation? In: *Foundations of Genetic Algorithms 2*. Ed. L. D. Whitely. Morgan Kaufmann.

Ursem, Rasmus K., 2003. *Models for Evolutionary Algorithms and Their Applications in System Identification and Control Optimization (PhD Dissertation)*. A Dissertation Presented to the Faculty of Science of the University of Aarhus in Partial Fulfillment of the Requirements for the PhD Degree. Department of Computer Science, University of Aarhus, Denmark.

Whitley, Darrell, 2000. Permutations. In *Evolutionary Computation 1, Basic Algorithms and Operators*. Eds T. Bäck, D.B. Fogel & Z. Michalewicz. United Kingdom: Institute of Physics Publishing Ltd, Bristol and Philadelphia. ISBN 0750306645.

Indoor Location Estimation in Sensor Networks using AI Algorithm

József Dániel Dombi
Department of Software Engineering, University of Szeged, Árpad tér 2., Szeged, Hungary
dombijd@inf.u-szeged.hu

Keywords: Indoor Location and Tracking, Sensor Network, Fingerprinting, Machine Learning.

Abstract: To determine the indoor location of a person or object, we can use a suitable wireless network. There are different kinds of wireless networks available for this. Independent of the type of the network, using RSSI it is possible to find the position of the moving person close by. Here, we present Wireless Sensor Network and apply it in a real environment. We will mainly concentrate on locating a person using standard artificial intelligence methods. In our system we define nodes (the fingerprint), and supervised learning algorithms that should predict these nodes. In addition, we test whether we can get nice results if we change the granularity of the nodes. Real simulation demonstrates that this system can supply the current position of the moving person with good accuracy.

1 INTRODUCTION

Here, a locating system is used for tracking and defining the current position of a person or object. The most important distinguishing feature of such a system is the type of wireless communication used, and the application information presented to the user. The granularity of the position can vary from one application to another. For example, finding out whether a person is in a room requires less information, while locating a person who is sitting in front of a desk requires more accurate information.

Therefore, many different systems and technologies have been proposed. GPS devices are available for everyday use in modern outdoor applications (Enge and Misra, 1999). The GPS system has a limited accuracy, and can be used where satellites are ”visible”, because buildings block the GPS transmissions. The earliest investigation for indoor positioning was done by Bahl et al. who observed that an RF signal source exhibits spatial variation, but is consistent in time. They created a system called Radar (Bahl and Padmanabhan, 2000). They used four 802.11 access points to locate a laptop at its true position to an accuracy of 2-3 meters. Since then, there have been a lot of improvements in Radar’s fingerprint matching algorithms (Agrawala and Shankar, 2003) (Haeberlen et al., 2004) (Ladd et al., 2005).

These studies showed that the Received Signal Strength Indicator (RSSI) has a larger variation because it is subject to the detrimental effects of fading and shadowing.

Other techniques, such as Active Badge (Hopper et al., 1993) and a commercial system like Versus (Versus, 2012), use infrared emitters and detectors to achieve an accuracy of 5-10m. Active Bet (Harter et al., 1999) (Ward and Jones, 1997) and Cricket (Priyantha et al., 2000) combine the RF and ultrasound signal to estimate the distance. These systems have accuracies ranging from a few meters to a few centimeters. In a commercial system (Ubisense, 2012), ultra-wideband emitters and receivers have been used to realize indoor locations.

In this study we use a wireless sensor network. If a large number of sensors are deployed, the network can monitor large areas. We can apply a sensor network in a variety of situations like those for monitoring the environment. Sensor nodes can measure temperature, a heartbeat, humidity and so on. However, collecting a large amount of data leads to an increase in traffic and in the energy consumption of sensors. Moreover, increasing the data collection time has a negative impact on the location data collection method. In a wireless sensor network it is vital to keep the energy consumption low. Our Sensor Network protocol is similar to the ZigBee (ZigBee, 2012) protocol, which includes IEEE 802.15.4 for MAC and PHY. Here, we implemented a positional estimation technique based on standard artificial intelligence methods using RSSI in a sensor network and evaluated its position-estimation ability. The remainder of this paper is organized as follows. Sec-

tion 2 outlines the standard AI methods, then Section 3 describes the experimental setup. After, Section 4 presents the result of our experiments. In the last section, we summarize our findings and draw some pertinent conclusions.

2 AI METHODS

In our system the signal strengths are got by a router. Currently, different routers send the RSSI to the PC. More than three RSSI values are used to determine the position of the node inside the building. First, we have to investigate the relationship between the distance and signal strength from a given router point. If one knows the distances from a node to at least three different routers, one can calculate the position of the node in the system.

In a real environment the power received is a very complex function of distance. Even if a good model is available to determine the position of the node, it still requires a lengthy calculation. Hence, the location of the RSSI is more complicated and it is harder to solve. In our model, we simplify the system. We do not worry about calculating the exact position of the object. For us, it is sufficient to determine the nearest node (fingerprint).

Standard artificial intelligence methods offer a good solution for estimating the location and reducing the distance error. Here, we implemented the decision tree model and neural network model. An exact knowledge of the position is not required by either method. We can train and use the methods without asking for it. Both methods have good classification capabilities and are suitable for our purpose, where we wish to determine the location that best matches the observed signal strength data.

Using a decision tree means we have to generate all possible decision trees that correctly classify the training set and then choose the simplest one. The number of such trees is finite, but very large. One of the most widely used decision tree method is ID3 (Quinlan, 1986). It constructs the simple decision tree, but this approach cannot guarantee that better trees have not been overlooked. The basic structure of ID3 is iterative. The window, which is a subset of a training set, is chosen at random and a decision tree is formed from it. ID3 examines all candidate attributes and chooses attribute A to maximize the gain. This tree correctly classifies all objects in the window. All other objects are then classified using the tree. If the tree returns the correct answer for all these objects, it is then correct for the entire training set and the process terminates. If not, a selection of the incorrectly classified objects is added to the window and the process continues. Recent articles (Yim, 2008) have examined how a decision tree works in a location system based on a fingerprint, and it is found that the accuracy of the decision tree is no worse than a Neural Network or Bayesian system.

A neural network is capable of representing the relationship between the inputs (signal strengths) and outputs (nodes). The learning strategy should calculate the free parameters of the model (also called the ”weights” of the network). Here, the standard multilayer perceptron (MLP) is implemented. The architecture of MLP is organized as follows: the signals flow sequentially through the different layers from the input to the output layer. For each neuron, it first calculates a scalar product between a vector of weights and the vector given by the output of the previous layer. A transfer function is then applied to the result to produce the input for the next layer. A commonly applied transfer function is the sigmoid function. In a single hidden layer, if the number of hidden layers is sufficiently large then any continuous function can be approximated to some desired accuracy. Roberto Battiti et al. (Battiti et al., 2002) examined how a neural network might be used to locate an object. They found that with MLP it is possible to determine the position of the person within 1.82 meters.

In our study, we compare the performance of both methods to see how well they determine the location of an object in a sensor network environment.

3 EXPERIMENTAL SETUP

Our experimental testbed is located on the first floor of a 2-storey building. We define nodes (position of the fingerprint), and the distances between the nodes are equal, namely a distance less than 2 m. Part of the layout of the floor and position of fingerprint are shown in Figure 1. In the tests, we employed a special type of sensor network called RTLS (RTLS, 2012). We placed four routers per room and two on the corridor at the locations indicated in Figure 2.

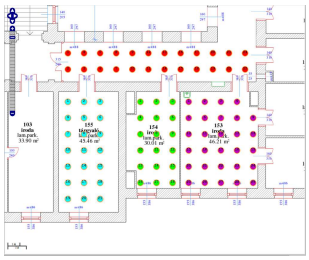


Figure 1: Map of the floor and position of each fingerprint.

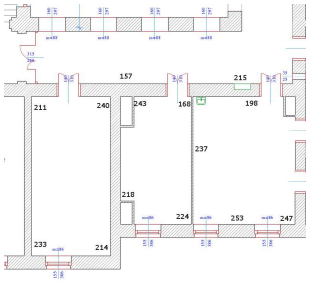


Figure 2: The position of routers inside the building.

In this figure, routers are represented by numbers (i.e.: 211, 240, 243, 168, etc.) and we see that the position of each router is usually in the corner of a room. Users wear a transmitter (also called a tag) device on their wrist as a watch (see Figure 3, which has a unique ID called the address.

Figure 3: Two different kinds of watch. Both of them function as a sensor.

The tag can measure and transmit the temperature and battery level; and, of course, the routers can measure the RSSI value. It is also possible to send audio data through this sensor network.

By default, the tag will send a broadcast message every 4 seconds. When a router gets a message it can transmit this data to the coordinator (zero in Figure 4, a special router). The packet received by the coordinator contains the address of the measured tag, the RSSI value measured by the router, and any other data measured by the tag. In this network there is a time delay in the routers. The routers wait for a while to receive RSSI values, then they aggregate them and transmit this data to the coordinator as a single packet. There is a size limit of the packet so in this way the router should be able to send a packet to coordinator every 400 milliseconds. The aggregated package contains only the latest RSSI value received from the tag. As we mentioned above, the coordinator can receive packets from different routers and it forwards them to a PC. The program running on this PC can collect the RSSI values. The primary task of the program is to determine which measurement belongs to the given tag at any one time. Our network is self-organized. This means that a tag can communicate with the routers and these routers send the received information on to the next router, which is closer to the coordinator and is connected to the PC. A tag tries to reach the nearest router, and if it cannot communicate with this router then it will search for another router. Figure 4 shows how communication is established and maintained. In this figure we can see that there is coordinator (C1), which is connected to the PC and there are six routers and a Tag (E1).

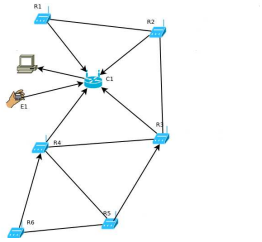


Figure 4: Communication between routers and tags.

We generated a fingerprint by performing calibration measurements. For each node, we measured over 20 values in a second and stored the RSSI values obtained by the router. These were the reference values that were used for testing the system.

4 RESULTS

As mentioned previously, we tested the AI methods on a first floor of a building. We positioned the routers and coordinator. First we had to collect samples and then we used the cross-validation method. This method partitions a given data sample into complementary subsets. Then we performed an analysis on one subset (called the train set), and validated our analysis on the other subset (called the test set). Multiple rounds of cross-validation were performed using different partitions, and the average over the rounds was the result of the validation.

It should be noted that the objective of our training algorithm was to build a model with good generalization capabilities when it was tested with values not present in the train set. The number of parameters and the length of the train phase determined the goodness of the generalization.

In a real environment it may happen that the given tag cannot reach the router (missing value). In that case, we define the worst RSSI value. In addition we define a new attribute that contains this information. When the value is one, the router receives a signal, and when the value is zero, the router doesn't receive an RSSI value of the given time. The maximum value of RSSI that we measured was -54dBm and minimum value was -90 dBm. For each measurement, out of 14 routers 6 on average send a message saying that they

receive an RSSI value, and only 3 routers on average can measure valuable RSSI values - which means that they can measure values better than -85 dBm. We created different kinds of tests which varied the granularity of the nodes: single position, triple position, and the room. Single position means that we would like predict the current position of the object. Triple position means that we aggregated 3 nearest node values into one, and we tried to predict this new position. In this case, we were only interesed in locating the object in a certain part of the room. Room position means that we merged all the node values in the room into a single node in order to locate the object. The results are shown in the following table.

Table 1: The results of the methods.

Granularity of nodes	Decision tree	Neural network
Single Position	38%	40%
Triple position	65%	53%
Room	91%	89%

As we see, the two methods have a similar performance in most cases. The percentage value tells us the degree of certainty of location an object. We tried different kinds of parameter input for the two learning methods and we obtained similar results. In the decision tree, we get the whole tree and examine the decisions. The decision tree has an average size of 250 and an average number of leaves around 125. The time needed for the learning method and the evaluation of the values is less for a tree than that for a neural network.

5 CONCLUSIONS

Many indoor positioning methods have been published that can be used in a variety of situations. For any kind of wireless network, the fingerprint method is the most commonly used approach. Previous studies showed that AI algorithms can perform well in locating an object. These studies used different types of networks. In this paper we compared two different kinds of AI method in a wireless sensor environment, which is similar to the ZigBee network. The bandwidth of this network is very low, but it can transmit audio and data measurements in real time with just one radio chip. We carried out different kinds of tests using this wireless sensor network, and we discovered that in most cases the decision tree and neural network approaches have a similar performance. When we increase the granualty of the nodes, we get much better results in terms of accuracy.

ACKNOWLEDGEMENTS

The study presented here was supported by the Hungarian national grant GOP-1.1.1-07. I would like to thank Ákos Kiss for his valuable advice, and also Péter Kenderesi, Péter Molnár and Balázs Szabó for providing position data.

REFERENCES

Agrawala, A. K. and Shankar, A. U. (2003). WLAN location determination via clustering and probability distributions. In *PerCom*.

Bahl, P. and Padmanabhan, V. N. (2000). RADAR: An in-building RF-based user location and tracking system. In *INFOCOM*.

Battiti, R., Villani, A., Villani, R., and Nhat, T. L. (2002). Neural network models for intelligent networks: Deriving the location from signal patterns. In *Citeseer*.

Enge, P. and Misra, P. (1999). Special issue on gps: The global positioning system. In *Proceedings of the of the IEEE*.

Haeberlen, A., Flannery, E., Ladd, A. M., Rudys, A., Wallach, D. S., and Kavraki, L. E. (2004). Practical robust localization over large-scale 802.11 wireless networks. In *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking*.

Harter, A., Hopper, A., Steggles, P., Ward, A., and Webster, P. (1999). The anatomy of a context-aware application. In *Proceedings of the Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom-99)*.

Hopper, A., Harter, A., and Blackie, T. (1993). The active badge system. In *INTERCHI'93 Conference on Human Factors in Computing Systems*.

Ladd, A. M., Bekris, K. E., Rudys, A., Kavraki, L. E., and Wallach, D. S. (2005). Robotics-based location sensing using wireless ethernet. *Wireless Networks*.

Priyantha, N. B., Chakraborty, A., and Balakrishnan, H. (2000). The cricket location-support system. In *MOBICOM*.

Quinlan, R. (1986). Induction of decision trees. *Machine Learning*, 1.

RTLS (2012). Rtls. http://www.rtls.eu.

Ubisense (2012). Ubisense. http://www.ubisense.net.

Versus (2012). Versus. http://www.versustech.com.

Ward, A. and Jones, A. (1997). A new location technique for the active office. In *IEEE Personal Communications*.

Yim, J. (2008). Introducing a decision tree-based indoor positioning technique. *Expert Syst. Appl*, 34(2).

ZigBee (2012). Zigbee. http://www.zigbee.org/Specifications.aspx.

An Impact of Model Parameter Uncertainty on Scheduling Algorithms

Radosław Rudek[1], Agnieszka Rudek[2], Andrzej Kozik[3] and Piotr Skworcow[4]
[1]*Institute of Business Informatics, Wrocław University of Economics, Wrocław, Poland*
[2]*Department of Systems and Computer Networks, Wrocław University of Technology, Wrocław, Poland*
[3]*Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wrocław, Poland*
[4]*De Montfort University, Water Software Systems, Leicester, U.K.*
radoslaw.rudek@ue.wroc.pl, {agnieszka.wielgus, andrzej.kozik}@pwr.wroc.pl, pskworcow@dmu.ac.uk

Keywords: Scheduling, Flowshop, Learning, Heuristic, Robustness.

Abstract: This short paper presents a preliminary analysis of the impact of model parameter uncertainty on the accuracy of solution algorithms for the scheduling problems with the learning effect. We consider the maximum completion time minimization flowshop problem with job processing times described by the power functions dependent on the number of processed jobs. To solve the considered scheduling problem we propose heuristic (NEH based) and metaheuristic (simulated annealing) algorithms. The numerical experiments show that NEH and simulated annealing are robust for this problem with respect to model parameter uncertainty.

1 INTRODUCTION

Classical flowshop scheduling problems are perceived to be more interesting in a theoretical context than as a practical research (Gupta and Stafford, 2006). It follows from observations that algorithms constructed on the basis of the classical models usually provide unsatisfactory (unstable) solutions for real-life flowshop problems, since these models do not take into consideration additional factors such as the learning effect that is significant in practice (Biskup, 2008), (Lee and Wu, 2004), (Rudek, 2011).

A schedule for a real-life problem (e.g., in manufacturing or computer systems) is calculated on the basis of a model and values of its parameters. Due to the possible differences between estimated and real values of problem parameters (e.g., shape of the learning curve, job processing times), the algorithms that are efficient for the modelled problem do not have to be accurate for the real problem. Therefore, it is crucial to evaluate how values of parameters (uncertainty) affect the quality of solutions provided by such algorithms, thereby determine their usefulness.

Thus, in this paper, we will analyse the impact of values of parameters on the accuracy of solution algorithms for the scheduling problems with the learning effect. In particular, we will consider the maximum completion time minimization flowshop problem with job processing times described by the power functions dependent on the number of processed jobs.

This paper is organized as follows. Next section contains the problem formulation. Approximation algorithms with the analysis of their efficiency are given subsequently. The last section concludes the paper.

2 PROBLEM FORMULATION

There are given a set $J=\{1,\ldots,n\}$ of n jobs and m machines, namely $M=\{M_1,\ldots,M_m\}$. Each job j consists of a set $O=\{O_{1,j},\ldots,O_{m,j}\}$ of m operations. Each operation $O_{z,j}$ has to be processed on machine M_z ($z=1,\ldots,m$). Moreover operation $O_{z+1,j}$ may start only if $O_{z,j}$ is completed. It is assumed that machines have to process jobs in the same order, i.e., a permutation flowshop, and each machine can process one operation at a time. There are no precedence constraints between jobs, operations are non-preemptive and are available for processing at time 0 on M_1. Further, instead of operation $O_{z,j}$, we say job j on machine M_z.

Due to the learning effect the processing time $\tilde{p}_j^{(z)}(v)$ of job j processed as the vth in a sequence on machine M_z is described by a non-increasing positive function dependent on the number of previously processed operations $(v-1)$, i.e., on its position v in a sequence. The function $\tilde{p}_j^{(z)}(v)$ of the job processing time that models the learning effect is called the learning curve. Moreover, each job j is characterized by its normal processing $\tilde{p}_j^{(z)}$ time on machine M_z that is de-

fined as the time required to perform a job if the machine is not affected by learning, i.e., $p_j^{(z)} \triangleq \tilde{p}_j^{(z)}(1)$.

Following (Mosheiov and Sidney, 2003), in this paper, we focus on a problem, where the processing time of job j processed as the vth on machine M_i is described by:

$$\tilde{p}_j^{(z)}(v) = p_j^{(z)} v^{\alpha_j^{(z)}}, \quad (1)$$

where $p_j^{(z)}$ and $\alpha_j^{(z)}$ are the normal processing time and the learning index, respectively, of job j on machine M_z. Moreover, we will analyse the problem with the special cases of (1), where $\alpha_j^{(z)} = \alpha$ for $j = 1,\ldots,n$ and $z = 1,\ldots,m$.

For the m-machine *permutation* flowshop problems the schedule of jobs on the machines can be unambiguously defined by their sequence (permutation). Let $\pi = \langle \pi(1),...,\pi(i),...,\pi(n) \rangle$ denote the sequence (permutation) of the n jobs where $\pi(i)$ is the job in position i of π. Also, let Π be the set of all job permutations. Thus, for each job $\pi(i)$, i.e., scheduled in the ith position in π, we can determine its completion time $C_{\pi(i)}^{(z)}$ on machine M_z as follows:

$$C_{\pi(i)}^{(z)} = \max\left\{C_{\pi(i)}^{(z-1)}, C_{\pi(i-1)}^{(z)}\right\} + \tilde{p}_{\pi(i)}^{(z)}(i), \quad (2)$$

where $C_{\pi(1)}^{(0)} = C_{\pi(0)}^{(z)} = 0$ for $z = 1,\ldots,m$ and $C_{\pi(i)}^{(1)} = \sum_{l=1}^{i} \tilde{p}_{\pi(l)}^{(1)}(l)$ is the completion time of a job placed in position i in the permutation π on M_1. On this basis, the maximum completion time (makespan) for the given π can be defined as $C_{\max}(\pi) = C_{\pi(n)}^{(m)}$.

The objective is to find such a schedule π^* of jobs on the machines that minimizes the maximum completion time (makespan): $\pi^* \triangleq \arg\min_{\pi\in\Pi}\left\{C_{\max}(\pi)\right\}$. For convenience, the problem according to the three field notation scheme $X \mid Y \mid Z$ will be denoted as $Fm|\tilde{p}_j(v) = p_j v^{\alpha_j}|C_{\max}$ and its special case ($\alpha_j^{(z)} = \alpha$) as $Fm|\tilde{p}_j(v) = p_j v^{\alpha}|C_{\max}$.

3 ALGORITHMS

In this section, we will briefly describe the algorithms that are analysed in the further part of this paper. Namely, we present the extensive search algorithm (ESA), the random schedule algorithm (RND), the shortest processing time (SPT) rule, NEH (Nawaz et al., 1983) and simulated annealing (SA) (Kirkpatrick et al., 1983). Note that the problem $Fm|\tilde{p}_j(v) = p_j v^{\alpha_j}|C_{\max}$ is strongly NP-hard even without the learning effect for $m \geq 3$, and it seems to be strongly NP-hard for $m = 2$ with the learning effect.

The extensive search algorithm (ESA) is an exact method that searches the total solution space, which size is $O(n!)$.

The random schedule algorithm (RND) provides a random schedule (permutation) as a solution; its complexity is $O(n)$.

The shortest processing time (SPT) rule constructs the solution according to the non-decreasing order of the normal processing times of jobs on machine M_1, i.e., $p_j^{(1)}$; its computational complexity is $O(n \log n)$.

The NEH algorithm (Algorithm 1) is based on the method introduced by (Nawaz et al., 1983). It starts with an initial solution $\pi_{initial}$ that determines the order of jobs that are subsequently inserted into the resulting solution π^* such that the criterion value $C_{\max}(\pi^*)$ is minimized. The computational complexity of this algorithm is $O(mn^3)$.

Algorithm 1: NEH.

```
1: Determine the initial sequence of jobs
   in π_initial and set π* := ∅
2:   Get the first job j from π_initial
3:   Insert j in such a position in π*
     for which C_max(π*) is minimal
4:   Remove j from π_initial
5:   If π_initial ≠ ∅ Then go to Step 2
6: The permutation π* is the given solution
```

Algorithm 2: SA.

```
1: Determine initial solution π_init
   and π = π* = π_init, T = T_0
2: For i = 1 to Iterations
3:   Choose π' by a random interchange of
     two jobs in π
4:   Assign π = π' with probability
     P(T,π',π) = min{1, exp(-(C_max(π')-C_max(π))/T)}
5:   If C_max(π) < C_max(π*) Then π* = π
6:   T = T/(1+λT)
7: The permutation π* is the given solution
```

The presented simulated annealing (SA) algorithm (Algorithm 2), that is based on (Kirkpatrick et al., 1983), starts with an initial solution $\pi_{initial}$ and generates in each iteration a new permutation π' based on the current solution π by interchanging of two randomly chosen jobs in π. This new solution π' replaces π (i.e., $\pi = \pi'$) with the following probability $P(T,\pi',\pi) = \min\left\{1, exp\left(-\frac{C_{\max}(\pi')-C_{\max}(\pi)}{T}\right)\right\}$, where T is the temperature that decreases in a logarithmical manner $T = \frac{T}{1+\lambda T}$, and the values of the initial temperature T_0 and of the parameter λ are chosen empirically. The solution π^* with the minimal found cri-

terion value $C_{\max}(\pi^*)$ is also stored. The algorithm stops after *Iterations* steps, thus, its overall computational complexity is $O(Iterations \cdot mn)$.

4 NUMERICAL ANALYSIS

In practice, a schedule for a real-life problem (e.g., in manufacturing systems) is calculated on the basis of a model and values of its parameters. Due to the possible differences between estimated and real values of problem parameters (e.g., shape of the learning curve, job processing times), the algorithms that are efficient for the modelled problem do not have to be accurate for the real problem. Therefore, it is crucial to evaluate how uncertain values of parameters affect the quality of solutions provided by such algorithms. Some of the analysed algorithms were described in (Rudek, 2011).

Let REAL denote the flowshop problem $Fm|\tilde{p}_j(v) = p_j v^{\alpha_j}|C_{\max}$, where job processing times are described by (1) and the values of the job parameters ($p_j^{(z)}$, $\alpha_j^{(z)}$) are precise. However, in practice it is usually difficult to obtain such accurate values and solution methods are based on uncertain (estimated) values. Following this, let ESTIM denote the flowshop scheduling problem, where the exact values of job parameters are unknown. In this case, job parameters are estimated, and we assume that job processing times are described by $\widehat{\tilde{p}}_j^{(z)}(v) = \widehat{p}_j^{(z)} v^{\widehat{\alpha}}$, where $\widehat{p}_j^{(z)}$ and $\widehat{\alpha}$ are the estimated values of $p_j^{(z)}$ and $\alpha_j^{(z)}$, respectively.

In the further part of this section, we provide the numerical analysis of the presented algorithms concerning the impact of the imprecise model on their efficiency. It is done according to the following steps. First, we draw values of job parameters for the problem REAL. Next, we solve the problem REAL using an algorithm A, which find a schedule π with criterion value $C_{\max}(\pi)$. Based on the parameters of the problem REAL, we draw or determine values of parameters for the problem ESTIM ($Fm|\tilde{p}_j(v) = p_j v^{\alpha}|C_{\max}$), which simulates their estimation. Next, we use the algorithm A to calculate a schedule $\widehat{\pi}$ for the problem ESTIM. For this schedule, we calculate the corresponding criterion value $C_{\max}(\widehat{\pi})$ for the problem REAL. The difference $C_{\max}(\widehat{\pi}) - C_{\max}(\pi)$ informs about the usefulness of the algorithm A in case of imprecise values of job parameters. Algorithms with smaller differences are more stable (robust), than those with greater.

The values of parameters for the problem REAL are generated as follows. For each pair of $n \in \{10, 25, 50\}$ and $m \in \{2, 3\}$, 100 random instances are generated from the uniform distribution in the following ranges of parameters: $p_j^{(z)} \in [1, 10]$, $\alpha_j^{(z)} \in [-0.51, -0.15]$ for $j = 1, \ldots, n$ and $z = 1, \ldots, m$. In all experiments in this paper, p_j are integers and α_j are rational values with accuracy of two decimal place, e.g., for $\alpha_j^{(z)} \in [-0.51, -0.15]$ it is $\alpha_j^{(z)} \in \{-0.51, -0.50, -0.49, \ldots, -0.15\}$. The values of $\alpha_j^{(z)} \in [-0.51, -0.15]$ corresponds to the learning curves in range between 70% and 90%, which are most common in practice (Biskup, 2008).

The values of the normal processing times for ESTIM are $\widehat{p}^{(z)} = p_j^{(z)}(1+\Delta_p)$, where Δ_p is the estimation error, which allows us to control precision of parameters for the analysis; it simulates the estimation process. The values of Δ_p and $\widehat{\alpha}$ are provided for particular experiments in Table 1.

Let $A_R = \{\text{ESA}, \text{ESA}_{\max}, \text{RND}, \text{SPT}, \text{NEH}, \text{SA}\}$ denote the algorithms that calculate the schedule for the problem REAL, where $\text{ESA}_{\max}$ is the algorithm that calculates the schedule with the maximum possible criterion value (opposite to ESA). ESA and $\text{ESA}_{\max}$ clearly show the place of the errors provided by the analysed algorithms in reference to the optimum and the worst criterion values. Note that the algorithms RND provide the same solution (schedule) for REAL and ESTIM. On the other hand, let $A_E = \{\widehat{\text{ESA}}, \widehat{\text{SPT}}, \widehat{\text{NEH}}, \widehat{\text{SA}}\}$ denote the corresponding algorithms from A_R that calculate the schedule for the problem ESTIM.

The initial solution for NEH and SA is a random permutation (in this case natural) and values of the parameters of SA were chosen empirically as follows: $Iterations = 1000000$, $T_0 = 1000000$ and $\lambda = 0.01$.[1]

The algorithms are evaluated, for each instance I, according to the relative error $\delta_A(I) = \left(\frac{C_{\max}(\pi_I^A)}{C_{\max}(\pi_I^*)} - 1\right) \cdot 100\%$, where $C_{\max}(\pi_I^A)$ denotes the criterion value provided by algorithm $A \in \{A_R, A_E\}$ for instance I and $C_{\max}(\pi_I^*)$ is the optimal solution of instance I (if $n = 10$) or the best found solution of instance I (if $n \geq 25$) provided by the considered algorithms. The optimal solution is provided by ESA for the problem REAL. The results concerning the percentage values of mean, minimum and maximum relative errors and mean criterion values $\overline{C}_{\max}$ (rounded to integer) provided by the analysed algorithms are presented in Table 1.

First, we discuss the results provided by the heuristic and metaheuristic algorithms for the prob-

[1] All algorithms were coded in C++ and simulations were run on PC, Intel® Core™i7–2600K Processor and 8GB RAM.

Table 1: The impact of model parameter uncertainty on the errors of the algorithms for $p_j^{(z)} \in [1,10]$, $\alpha_j^{(z)} \in [-0.51,-0.15]$, $\Delta_p \in [-0.25,0.25]$, $\widehat{\alpha} = -0.322$.

n	m	Algorithms	$\bar{C}_{max}$	Errors		
				Mean	Min	Max
10	2	ESA	36	0.00	0.00	0.00
		ESA_{max}	54	44.35	21.37	74.72
		RND	44	19.58	4.05	46.09
		SPT	39	5.20	0.00	18.63
		NEH	37	1.60	0.00	8.09
		SA	36	0.00	0.00	0.00
		$\widehat{ESA}$	38	3.09	0.00	16.41
		$\widehat{SPT}$	39	5.60	0.15	21.73
		$\widehat{NEH}$	38	3.45	0.00	17.11
		$\widehat{SA}$	38	3.05	0.00	16.41
	3	ESA	41	0.00	0.00	0.00
		ESA_{max}	62	49.64	28.56	80.89
		RND	52	21.85	6.16	48.14
		SPT	45	10.31	0.54	34.50
		NEH	43	2.20	0.00	10.15
		SA	41	0.01	0.00	0.44
		$\widehat{ESA}$	43	4.31	0.00	16.21
		$\widehat{SPT}$	46	10.93	0.25	30.58
		$\widehat{NEH}$	43	5.21	0.57	19.90
		$\widehat{SA}$	43	4.18	0.00	16.21
25	2	RND	82	19.57	7.95	32.41
		SPT	75	6.65	0.10	19.38
		NEH	72	2.34	0.12	5.82
		SA	70	0.00	0.00	0.00
		$\widehat{SPT}$	75	6.74	0.32	18.47
		$\widehat{NEH}$	73	4.93	0.55	12.76
		$\widehat{SA}$	73	3.90	0.26	13.73
	3	RND	84	22.71	3.30	41.19
		SPT	77	12.77	4.96	29.46
		NEH	71	3.51	0.51	7.61
		SA	70	0.00	0.00	0.00
		$\widehat{SPT}$	78	12.90	5.05	28.37
		$\widehat{NEH}$	75	7.53	2.11	19.88
		$\widehat{SA}$	75	6.10	1.12	18.19
50	2	RND	129	19.32	9.61	31.39
		SPT	119	9.35	0.20	20.08
		NEH	113	3.09	0.12	7.18
		SA	109	0.00	0.00	0.00
		$\widehat{SPT}$	118	9.44	0.43	20.15
		$\widehat{NEH}$	116	6.47	0.51	13.04
		$\widehat{SA}$	113	4.19	0.42	11.86
	3	RND	141	20.41	10.33	31.11
		SPT	133	14.21	5.87	28.94
		NEH	122	3.78	1.02	7.41
		SA	117	0.00	0.00	0.00
		$\widehat{SPT}$	134	13.93	6.46	30.50
		$\widehat{NEH}$	125	7.66	2.95	15.01
		$\widehat{SA}$	125	5.96	1.40	14.29

lem REAL, for which the exact values of model parameters are known. It can be seen in Table 1 that SA finds solutions with criterion values close to the optimum. On the other hand, the differences between the mean relative errors provided by SA and NEH is about 3.5% and for the maximum errors 10%; for SPT it is about 14% for mean and 35% for maximum errors. The random solution is usually equally between the optimal and the worst case ($n = 10$) and provides mean and maximum errors (in reference to SA) about 20% and 45%, respectively.

However, if the applied algorithms are based on uncertain values of model parameters (solve the problem ESTIM), then their accuracy decreases in reference to the criterion value found by the algorithms, which are based on exact values (solve the problem REAL). It can be seen in Table 1 (for $n = 10$), that SA is more robust with respect to model parameter uncertainty than ESA. Namely, solutions obtained for ESTIM by $\widehat{SA}$ have lower criterion values (in reference to REAL) than provided by $\widehat{ESA}$. Note that the mean relative errors of NEH and SA increase about 3-5% if model parameters are uncertain. The exception is SPT, which is robust to the analysed model parameter uncertainty, however, it provides solutions with relative errors greater than $\widehat{NEH}$ and $\widehat{SA}$. Note that the considered algorithms calculate schedules that are significantly lower than a random solution (RND).

From the numerical analysis follows that NEH and SA can be efficiently applied to solve the considered real-life problem even if the model parameters are uncertain.

5 CONCLUSIONS

In this paper, we analysed the impact of model parameter uncertainty on the accuracy of solution algorithms for the makespan minimization flowshop scheduling problem with job processing times described by the power functions dependent on the number of processed jobs. We showed that the considered algorithms are efficient even if the values of problem parameters are not precisely identified.

REFERENCES

Biskup, D. (2008). A state-of-the-art review on scheduling with learning effects. *European Journal of Operational Research*, 188:315–329.

Gupta, J. N. D. and Stafford, J. E. F. (2006). Flowshop scheduling research after five decades. *European Journal of Operational Research*, 169:699–711.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Lee, W.-C. and Wu, C.-C. (2004). Minimizing total completion time in a two-machine flowshop with a learning effect. *International Journal of Production Economics*, 88:85–93.

Mosheiov, G. and Sidney, J. B. (2003). Scheduling with general job–dependent learning curves. *European Journal of Operational Research*, 147:665–670.

Nawaz, M., Enscore, J. E. E., and Ham, I. A. (1983). A heuristic algorithm for *m*-machine, *n*-jobs Flow-shop sequencing problem. *Omega*, 11:91–95.

Rudek, R. (2011). Computational complexity and solution algorithms for flowshop scheduling problems with the learning effect. *Computers & Industrial Engineering*, 61:20–31.

AUTHOR INDEX

AUTHOR INDEX (CONT.)

Proceedings will be submitted for indexation by:

Proceedings of ICEIS 2012
14th International Conference on Enterprise Information Systems
ISBN: 978-989-8565-10-5
http://www.iceis.org