Costin Bădică
Ngoc Thanh Nguyen
Marius Brezovan (Eds.)

LNAI 8083

Computational Collective Intelligence

Technologies and Applications

5th International Conference, ICCCI 2013
Craiova, Romania, September 2013
Proceedings

Springer

Costin Bădică   Ngoc Thanh Nguyen
Marius Brezovan (Eds.)

Computational Collective Intelligence

Technologies and Applications

5th International Conference, ICCCI 2013
Craiova, Romania, September 11-13, 2013
Proceedings

Springer

Volume Editors

Costin Bădică
University of Craiova
Computer and Information Technology Department
Bvd. Decebal 107
200440 Craiova, Romania
E-mail: cbadica@software.ucv.ro

Ngoc Thanh Nguyen
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
E-mail: ngoc-thanh.nguyen@pwr.edu.pl

Marius Brezovan
University of Craiova
Computer and Information Technology Department
Bvd. Decebal 107
200440 Craiova, Romania
E-mail: mbrezovan@software.ucv.ro

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-40494-8 e-ISBN 978-3-642-40495-5
DOI 10.1007/978-3-642-40495-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013945726

CR Subject Classification (1998): I.2, H.3, H.2, K.3, I.5, G.1.6, F.1, H.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

ICCCI 2013 was the fifth event of the series of international scientific conferences for research and applications in the field of computational collective intelligence. The aim of ICCCI is to provide an international forum for scientific research in the technologies and applications of computational collective intelligence. ICCCI 2013 was co-organized by University of Craiova (Romania) and Wroclaw University of Technology (Poland) and took place in Craiova (Romania) during September 11–13, 2013. The conference was run under the patronage of the IEEE SMC Technical Committee on Computational Collective Intelligence as well as of the Romanian Association of Artificial Intelligence (ARIA) and IEEE Romania (Computational Intelligence Section).

We received many papers from 24 countries all over the world. Each paper was peer reviewed by at least three members of the International Program Committee and International Reviewer Board. Only 72 papers with the highest quality were selected for oral presentation and publication in the volume of ICCCI 2013 proceedings.

The papers included in the proceedings cover the following topics: intelligent e-learning, classification and clustering methods, Web intelligence and interaction, agents and multi-agent systems, social networks, intelligent knowledge management, language processing systems, modeling and optimization techniques, evolutionary computation, intelligent and group decision making, swarm intelligence, data mining techniques and applications, cooperative problem solving, collective intelligence for text mining and innovation, collective intelligence for social understanding and mining, and soft methods in collective intelligence.

Accepted and presented papers highlight new trends and challenges of computational collective intelligence. The presenters showed how new research could lead to novel and innovative applications. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to the Honorary Chairs, Prof. Dan Claudiu Dănişor (Rector of University of Craiova, Romania), Prof. Tadeusz Wieckowski (Rector of Wroclaw University of Technology, Poland), Prof. Mircea Ivănescu, University of Craiova, Romania, and Pierre Lévy, University of Ottawa, Canada, for their support.

Our special thanks go to the Program Co-chairs, all Program and Reviewer Committee members and all the additional reviewers for their valuable efforts in the review process, which helped us to guarantee the highest quality of the selected papers for the conference. We cordially thank the organizers and chairs of special sessions who contributed to the success of the conference.

We also would like to express our thanks to the Keynote Speakers (Prof. Andrea Omicini, Prof. Sergiu Nedevschi, Prof. Jacek Koronacki and Dr. Ulle Endriss) for their interesting and informative talks of world-class standard.

We cordially thank our main sponsors, University of Craiova (Romania) and Wroclaw University of Technology (Poland). Our special thanks are due also to Springer for publishing the proceedings, and other sponsors for their kind supports.

We wish to thank the members of the Organizing Committee for their very substantial work, especially those who played essential roles: Prof. Amelia Bădică, Dr. Mihaela Colhon (Local Organizing Chairs), and the members of the Local Organizing Committee (from the Department of Computers and Information Technology, Faculty of Automation, Computers and Electronics, Department of Statistics and Business Informatics, Faculty of Economics and Business Administration, and Department of Informatics, Faculty of Exact Sciences, University of Craiova, Romania) for their excellent work and very good cooperation.

Special thanks are also due to the members of the Intelligent Distributed Systems (Sorin Ilie and Alex Muscar) research group that kindly provided technical support for running the conference.

We cordially thank all the authors for their valuable contributions and all the other participants of this conference. The conference would not have been possible without their support.

Thanks are also due to the many experts who contributed to making the event a success.

Costin Bădică
Ngoc Thanh Nguyen
Marius Brezovan

Conference Organization

Honorary Chairs

Dan Claudiu Dănişor	Rector of University of Craiova, Romania
Tadeusz Wieckowski	Rector of Wroclaw University of Technology, Poland
Mircea Ivănescu	University of Craiova, Romania
Pierre Lévy	University of Ottawa, Canada

General Chairs

Costin Bădică	University of Craiova, Romania
Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Steering Committee

Ngoc Thanh Nguyen (Chair)	Wroclaw University of Technology, Poland
Piotr Jędrzejowicz (Co-chair)	Gdynia Maritime University, Poland
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Adam Grzech	Wroclaw University of Technology, Poland
Kiem Hoang	University of Information Technology, VNU-HCM, Vietnam
Lakhmi C. Jain	University of South Australia, Australia
Geun-Sik Jo	Inha University, Korea
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Ryszard Kowalczyk	Swinburne University of Technology, Australia
Toyoaki Nishida	Kyoto University, Japan
Ryszard Tadeusiewicz	AGH-UST, Poland

Organizing Chairs

Amelia Bădică	University of Craiova, Romania
Mihaela Colhon	University of Craiova, Romania
Ileana Nicolae	University of Craiova, Romania

Program Chairs

Marius Brezovan	University of Craiova, Romania
Kazumi Nakamatsu	University of Hyogo, Japan
Piotr Jędrzejowicz	Gdynia Maritime University, Poland
Edward Szczerbicki	University of Newcastle, Australia

Special Session Chairs

Amelia Bădică	University of Craiova, Romania
Bogdan Trawiński	Wroclaw University of Technology, Poland

Liaison Chairs

Dumitru Dan Burdescu	University of Craiova, Romania
Geun-Sik Jo	Inha University, Korea

Publicity Chair

Alex Muscar	University of Craiova, Romania

Local Organizing Committee

Ion Buligiu	University of Craiova, Romania
Razvan Buse	University of Craiova, Romania
Liviu Ion Ciora	University of Craiova, Romania
Dorian Cojocaru	University of Craiova, Romania
Mihaela Colhon	University of Craiova, Romania
Nicolae Constantinescu	University of Craiova, Romania
Mirel Cosulschi	University of Craiova, Romania
Mihai Gabroveanu	University of Craiova, Romania
Eugen Ganea	University of Craiova, Romania
Ion Iancu	University of Craiova, Romania
Sorin Ilie	University of Craiova, Romania
Costel Ionascu	University of Craiova, Romania
Ilona Marogel	University of Craiova, Romania
Cristian Mihaescu	University of Craiova, Romania
Mihai Mocanu	University of Craiova, Romania
Adriana Schiopoiu	University of Craiova, Romania
Dan Selisteanu	University of Craiova, Romania
Marian Siminica	University of Craiova, Romania
Liana Stanescu	University of Craiova, Romania
Cosmin Stoica	University of Craiova, Romania
Teodor Vetezi	University of Craiova, Romania

Keynote Speakers

Andrea Omicini	University of Bologna, Italy
Sergiu Nedevschi	University of Cluj-Napoca, Romania
Ulle Endriss	University of Seoul, Korea
Jacek Koronacki	University of Amsterdam, The Netherlands

Special Sessions Organizers

1. *Special Session on Cooperative Problem Solving (CPS 2013)*

Piotr Jędrzejowicz	Gdynia Maritime University, Poland
Dariusz Barbucha	Gdynia Maritime University, Poland

2. *Special Session on Collective Intelligence in Web Systems - Web Systems Analysis (WebSys 2013)*

Kazimierz Choroś	Wrocław University of Technology, Poland
Maria Trocan	Institut Supérieur d'Électronique de Paris, France

3. *Special Session on Using Collective Intelligences and Text Mining on Innovation (UCITMI 2013)*

Chao-Fu Hong	Aletheia University, Taiwan
Tzu-Fu Chiu	Aletheia University, Taiwan

4. *Special Session on Computational Collective Intelligence for Social Understanding and Mining (CCISUM 2013)*

David Camacho	Universidad Autonoma de Madrid, Spain
Pan-Koo Kim	Chosun University, Korea
Jason Jung	Yeungnam University, Korea

5. *Collective Computational Intelligence in Educational Context (ColEdu 2013)*

Danuta Zakrzewska	Lodz University of Technology, Poland
Marta Zorrilla	University of Cantabria, Spain

6. *Computational Swarm Intelligence (CSI 2013)*

Urszula Boryczka	University of Silesia, Poland

International Program Committee

Muhammad Abulaish	King Saud University, Riyadh, Kingdom of Saudi Arabia
Cesar Andres	Universidad Complutense de Madrid, Spain
Amelia Bădică	University of Craiova, Romania
Amar Balla	Ecole Supérieure d'Informatique (ESI), Algeria
Dariusz Barbucha	Gdynia Maritime University, Poland
Nick Bassiliades	Aristotle University of Thessaloniki, Greece
Maria Bielikova	Slovak University of Technology in Bratislava, Slovakia
Olivier Boissier	ENS Mines Saint-Etienne, France
Urszula Boryczka	University of Silesia, Poland
José Luís Calvo-Rolle	Universidad de La Coruña, Spain
David Camacho	Universidad Autonoma de Madrid, Spain
Tru Cao	Ho Chi Minh City University of Technology, Vietnam
Frantisek Capkovic	Slovak Academy of Sciences, Slovakia
Dariusz Ceglarek	Poznan School of Banking, Poland
Krzysztof Cetnarowicz	AGH University of Science and Technology, Poland
Tzu-Fu Chiu	Aletheia University, Taiwan
Amine Chohra	Paris-East University, France
Kazimierz Choros	Wroclaw University of Technology, Poland
Dorian Cojocaru	University of Craiova, Romania
Mihaela Colhon	University of Craiova, Romania
Tina Comes	Centre for Integrated Emergency Management, University of Agder, Norway
Phan Cong-Vinh	NTT University, Vietnam
Irek Czarnowski	Gdynia Maritime University, Poland
Paul Davidsson	Malmo University, Sweden
José Palazzo M. De Oliveira	Federal University of Rio Grande do Sul, Brazil
Roberto De Virgilio	Università degli Studi Roma Tre, Italia
Phuc Do	University of Information Technology, Vietnam
Tien Van Do	Budapest University of Technology and Economics, Hungary
Trong Hai Duong	Inha University, Korea
Atilla Elçi	Süleyman Demirel University, Turkey
Vadim Ermolayev	Zaporozhye National University, Ukraine
Rim Faiz	IHEC-University of Carthage, Tunisia
Adina Magda Florea	Politehnica University of Bucharest, Romania
Giancarlo Fortino	University of Calabria, Italy
Mauro Gaspari	University of Bologna, Italy
Dominic Greenwood	Whitestein Technologies, Switzerland
Janis Grundspenkis	Riga Technical University, Latvia

Adam Grzech	Wroclaw University of Technology, Poland
Anamika Gupta	University of Delhi, India
Quang-Thuy Ha	Vietnam National University, Hanoi, Vietnam
Le Thi Hoai An	Université de Lorraine, Metz, France
Huu-Hanh Hoang	Hue University, Vietnam
Frederik Hogenboom	Erasmus University Rotterdam, The Netherlands
Natasa Hoic-Bozic	University of Rieka, Croatia
Tzung-Pei Hong	Univesity of Kaohsiung, Taiwan
Mong-Fong Horng	Natonal Koahsiung University of Applied Sciences, Taiwan
Jingshan Huang	University of South Alabama, USA
Dosam Hwang	Yeungnam University, Korea
Lazaros Iliadis	Democritus University of Thrace, Greece
Diana Inkpen	University of Ottawa, Canada
Dan Istrate	ESIGETEL-ALLIANSTIC, France
Mirjana Ivanovic	University of Novi Sad, Serbia
Joanna Jędrzejowicz	Gdansk University, Poland
Gordan Jezic	University of Zagreb, Croatia
Joanna Jozefowska	Poznan University of Technology, Poland
Jason J. Jung	Yeungnam University, Korea
Petros Kefalas	The University of Sheffield International Faculty, CITY College, Greece
Ivan Koychev	University of Sofia "St. Kliment Ohridski", Bulgaria
Adrianna Kozierkiewicz-Hetmanska	Wroclaw University of Technology, Poland
Ondrej Krejcar	University of Hradec Kralove, Czech Republic
Piotr Kulczycki	Polish Academy of Science, Systems Research Institute, Poland
Kazuhiro Kuwabara	Ritsumeikan University, Japan
Raymond Y.K. Lau	City University of Hong Kong, China
Florin Leon	Technical University "Gheorghe Asachi" of Iasi, Romania
Xiafeng Li	Texas A&M University, USA
Joan Lu	University of Huddersfield, UK
Yannis Manolopoulos	Aristotle University of Thessaloniki, Greece
Tokuro Matsuo	Yamagata University, Japan
Adam Meissner	Poznan University of Technology, Poland
Jacek Mercik	Wroclaw University of Technology, Poland
Mihai Mocanu	University of Craiova, Romania
Grzegorz J. Nalepa	AGH University of Science and Technology, Poland
Alexandros Nanopoulos	University of Hildesheim, Germany
Filippo Neri	University of Naples Federico II, Italy
Dinh Thuan Nguyen	University of Information Technology, Vietnam

Linh Anh Nguyen	Warsaw University, Poland
Thanh Thuy Nguyen	University of Engineering and Technology, Vietnam
Alberto Nunez	Universidad Complutense de Madrid, Spain
Manuel Núñez	Universidad Complutense de Madrid, Spain
Chung-Ming Ou	Kainan University, Taiwan
Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Ramalingam Ponnusamy	Madha Engineering College, India
Radu-Emil Precup	Politehnica University of Timisoara, Romania
Paulo Quaresma	Universidade de Evora, Portugal
Ewa Ratajczak-Ropel	Gdynia Maritime University, Poland
Ioan Salomie	Technical University of Cluj-Napoca, Romania
Ali Selamat	Universiti Teknologi Malaysia
Elena Simperl	University of Southampton, UK
Liana Stanescu	University of Craiova, Romania
Stanimir Stoyanov	University of Plovdiv "Paisii Hilendarski", Bulgaria
Tadeusz Szuba	AGH University of Science and Technology, Poland
Yasufumi Takama	Tokyo Metropolitan University, Japan
Zbigniew Telec	Wroclaw University of Technology, Poland
Michel Toulouse	Oklahoma State University, USA
Trong Hieu Tran	Swinburne University of Technology, Australia
Stefan Trausan-Matu	Politehnica University of Bucharest, Romania
Bogdan Trawinski	Wroclaw University of Technology, Poland
Jan Treur	Vrije University, The Netherlands
Iza Wierzbowska	Gdynia Maritime University, Poland
Drago Zagar	University of Osijek, Croatia
Danuta Zakrzewska	Lodz University of Technology, Poland
Constantin-Bala Zamfirescu	University of Sibiu, Romania
Katerina Zdravkova	University of Sts. Cyril and Methodius, FYROM

Program Committees of Special Sessions

Special Session on Cooperative Problem Solving (CPS 2013)

Ewa Ratajczak-Ropel	Gdynia Maritime University, Poland
Ireneusz Czarnowski	Gdynia Maritime University, Poland
Mahdi Zargayouna	IFSTTAR, France
Edyta Kucharska	AGH University of Science and Technology, Poland
Iza Wierzbowska	Gdynia Maritime University, Poland
Joanna Jędrzejowicz	University of Gdansk, Poland

Special Session on Collective Intelligence in Web Systems - Web Systems Analysis (WebSys 2013)

František Čapkovič	Academy of Sciences, Bratislava, Slovakia
Ondřej Krejcar	University of Hradec Kralove, Czech Republic
Matthieu Manceny	Institut Supérieur d'Électronique de Paris, France
Alin Moldoveanu	Politehnica University of Bucharest, Romania
Tarkko Oksala	Helsinki University of Technology, Finland
Andrzej Siemiński	Wroclaw University of Technology, Poland
Behcet Ugur Toreyin	Cankaya University, Ankara, Turkey
Aleksander Zgrzywa	Wrocław University of Technology, Poland

Special Session on Using Collective Intelligences and Text Mining on Innovation (UCITMI 2013)

Fang-Cheng Hsu	Aletheia University, Taiwan
Pen-Choug Sun	Aletheia University, Taiwan
Chia-Ling Hsu	Tamkang University, Taiwan
Ai-Ling Wang	Tamkang University, Taiwan
Hung-Ming Wu	Aletheia University, Taiwan
Feng-Sueng Yang	Aletheia University, Taiwan
Min-Huei Lin	Aletheia University, Taiwan
Ai-Yun Su	Tamkang University, Taiwan
Ya-Fung Chang	Tamkang University, Taiwan
Leuo-Hong Wang	Aletheia University, Taiwan
Kuan-Shiu Chiu	Aletheia University, Taiwan
Peng-Wen Chen	Oriental Institute of Technology, Taiwan
Ming-Chien Yang	Aletheia University, Taiwan

Special Session on Computational Collective Intelligence for Social Understanding and Mining (CCISUM 2013)

David F. Barrero	Universidad de Alcala, Spain
John Breslin	DERI, Ireland (To be confirmed)
Manuel Cebrian	NICTA, University of Melbourne, Australia
Tutut Herawan	University of Malaya, Malaysia
Sachio Hirokawa	Kyushu University, Japan
Sang-Wook Kim	Hanyang Unversity, Korea
Shintaro Okazaki	Universidad Autonoma de Madrid, Spain
Seung-Bo Park	Kyung Hee University, Korea
Maria Dolores Rodriguez-Moreno	Universidad de Alcala, Spain
David Garcia	ETH Zurich, Switzerland
Hideaki Takeda	NII, Japan (To be confirmed)

I-Hsien Ting	National University of Kaohsiung, Taiwan
Martin K.J. Waiguny	Auckland University of Technology, New Zealand

Collective Computational Intelligence in Educational Context (ColEdu 2013)

José Luis Crespo Fidalgo	University of Cantabria, Spain
Dalia Kriksciuniene	Vilnius University, Lithuania
Ernestina Menasalvas Ruiz	Universidad Politécnica de Madrid, Spain
Adam Niewiadomski	Lodz University of Technology, Poland
Cristóbal Romero Morales	University of Córdoba, Spain
Marek Rudnicki	Lodz University of Technology, Poland
Virgilijus Sakalauskas	Vilnius University, Lithuania

Computational Swarm Intelligence (CSI 2013)

Rafał Skinderowicz	University of Silesia, Poland
Mariusz Boryczka	University of Silesia, Poland
Wojciech Wieczorek	University of Silesia, Poland
Rafael Parpinelli	State University of Santa Catarina, Brazil
Ajith Abraham	Machine Intelligence Research Labs (MIR Labs)

Additional Reviewers

Ion Iancu	University of Craiova, Romania
Nicolae Constantinescu	University of Craiova, Romania
Andreea Urzica	Politehnica University of Bucharest, Romania

Table of Contents

Invited Paper

Social Networks

Intelligent Decision Making

Web Intelligence and Interaction

Agent and Multi-agent Systems

Intelligent Knowledge Management

Language Processing Systems

Modeling and Optimization Techniques

Evolutionary Computation

Clustering, Classification and Data Mining

Cooperative Problem Solving

Collective Intelligence in Web Systems – Web Systems Analysis

Collective Intelligences and Text Mining on Innovation

Computational Collective Intelligence for Social Understanding and Mining

Collective Computational Intelligence in Educational Context

Computational Swarm Intelligence

Complexity and Interaction: Blurring Borders between Physical, Computational, and Social Systems

Preliminary Notes

Andrea Omicini and Pierluigi Contucci

Alma Mater Studiorum – Università di Bologna
{andrea.omicini,pierluigi.contucci}@unibo.it

Abstract. Complex systems of any kind are characterised by autonomous components interacting with each other in a non-trivial way. In this paper we discuss how the views on complexity are evolving in fields like physics, social sciences, and computer science, and – most significantly – how they are converging.

In particular, we focus on the role of *interaction* as the foremost dimension for modelling complexity, and discuss first how *coordination* via mediated interaction could determine the general dynamics of complex software system, then how this applies to *complex socio-technical systems* like social networks.

Keywords: Complex systems, interaction, interacting systems, statistical mechanics, coordination models, socio-technical systems.

1 Complexity and Interaction: An Introduction

The notion of complexity is definitely a multi-disciplinary one, ranging from physics to biology, from economics to sociology and organisation sciences. Most interestingly, systems that are said "complex" are both natural and artificial ones: so, for instance, we observe and model complex physical systems, and at the same time we design and build complex computational systems.

Along this line, moving from the pioneering work of Simon [1] on complex artificial systems – whose acceptation of complexity and complex system is the one implicitly adopted here –, it is nowadays widely recognised that there exist some "laws of complexity" that characterise any complex system, independently of its specific nature [2]. No matter whether we are modelling the behaviour of a human organisation, the life of an intricate ecosystem, or the dynamics of a huge market-place, we can anyway expect to find some repeated patterns, some shared schema, some common laws that make all such systems look similar—of course, when they are observed at the right level of abstraction. However, the exact source of what all complex systems share, the precise nature of such common factors to which all complex systems might be reducible, is still unknown in essence.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 1–10, 2013.

In this paper we argue that *interaction* – its nature, structure, dynamics – is the key to understand some fundamental properties of complex systems of any kind. Accordingly, in this preliminary notes we first elaborate on the role of interaction in complex systems, then we provide some perspectives on how the evolving views on complexity coming from physics and computer science could be seen as converging, by adopting complex socio-technical systems such as social networks as a key case study.

2 Complexity and Interaction in Computational Systems

> ... by a complex system I mean one made up of a large number of parts that interact in a non simple way [1]

Complexity is nowadays one of the most relevant traits of the systems of interest in many scientific fields. In particular, interaction is recognised as a fundamental dimension for modelling and engineering complex computational systems [3]: in a world where software systems are made of an always-increasing amount of objects, components, processes, or agents, and where the Internet – with billions of interacting clients and servers – represents the most widespread application environment, it is quite apparent that interaction is today the most relevant source of complexity for software systems of any sort.

In [4] the study of interaction as a first-class subject of research is shown to be at the core of a number of diverse scientific areas dealing with complex systems, whose results are "bridged" towards computer science, devising out a linear conceptual path:

Interaction — Complex systems cannot be described, understood, or built by merely dealing with the nature and behaviour of their individual components. Instead, dealing with *interaction* as first-class subject of study is a key issue: this calls for special interaction-oriented paradigms, models, technologies, and methodologies aimed at modelling and engineering complex systems.

Environment — Individual components of a system cannot be understood separately from the *environment* where they live and interact. Studying the environment where a system is situated, its nature and dynamics, as well as its interaction with the system components, is a fundamental pre-condition for understanding the essence and evolution over time of complex systems of any sort.

Mediated interaction — Interaction is always *mediated*, and the nature of mediators affects interaction. The notion of *mediator*, along with its structure and behaviour, is essential for modelling and engineering the space of interaction within complex systems.

Infrastructure — In order to govern the interactions among participants of large, complex systems, a suitable *infrastructure* is required, which could enforce collective laws and norms to rule the interaction among individual components, as well as between the system and its environment—essentially, by enacting laws through a coherent apparatus of mediators.

It is quite easy to see that the above interaction-related notions – along with the conceptual path they implicitly sketch – do not pertain to complex computational systems only: instead, as discussed in [4], they apply – in diverse ways – to either physical, biological, social, or computational systems.

When applied to computational systems, the same notions basically draw the foremost lines of evolution of contemporary computational systems: *(i) interaction* has become an essential and independent dimension of computational systems, orthogonal to mere computation [5,3]; *(ii) environment* is nowadays conceived as a first-class abstraction in the modelling and engineering of complex computational systems, such as pervasive, adaptive, and multi-agent systems [6]; *(iii)* environment-based *mediation* [7] is the key to designing and shaping the interaction space within complex software systems, in particular socio-technical ones [8]; finally, *(iv)* middleware and software infrastructure provide complex socio-technical systems with the mediating abstractions required to rule and govern social and environment interaction [9].

The nature, structure, and behaviour of such mediators are better discussed in Subsection 4.1, where the notion of *coordination medium* is introduced and reviewed.

3 Complexity and Interaction in Statistical Mechanics

Whereas the concepts illustrated in the previous section generally apply to many sorts of systems, quite a different view over complexity comes from physics, and in particular from *statistical mechanics*. Statistical mechanics is the branch of theoretical physics born after a set of ideas and methods, originally introduced by Boltzmann [10], used to de-axiomatise thermodynamics—i.e., to derive its laws from those of mechanics by means of probability methods. There, the key point is to relate the *macroscopic* observables quantities – like pressure, temperature, etc. – to suitable *averages* of *microscopic* observables—like particle speed, kinetic energy, etc. Since the method works based on the laws of large numbers, it turns out to be effective for those systems made of a large number of particles / basic components.

In a similar way as computer science, statistical mechanics has expanded beyond its origins: first, into many directions within physics; then, in the last decades, towards fields as diverse as biology [11], economics [12,13], and computer science itself [14,15], while its relevance in social sciences is growing fast as well. Cross-fertilisation with all those assorted fields concerns many different aspects, but focuses in particular on the notion of *complexity* as it emerges from statistical mechanics. In order to introduce its main features, it is useful first of all to review the basic properties of statistical mechanics systems, starting from the elementary classical cases up to the more recent and refined ones.

3.1 Interaction in Statistical Mechanics

Historically, the ideal gas behaviour is probably the first example of a thermodynamic system successfully understood by adopting statistical mechanics method.

The laws governing its behaviour may indeed be deduced, with elementary methods of probability, from physical laws like conservation of energy and momentum for the particles the gas is composed of, with the crucial identification of the mean kinetic energy with the temperature of the system.

The key point here is that the mathematical model of the ideal gas contains the assumption of *mutual independence* among particles, that is, the behaviour of the particles, as well as the resulting overall behaviour of the system, is not affected by their mutual *interaction*: the factorisation law that mathematically encodes the property of independence is related to the microscopic fact that the ideal gas particles do not interact with each other. Under this assumption, the probability distribution of the whole system is the product of those of each particle that composes it—in computer science terms, the properties of the system can be *compositionally* derived by the properties of the single components [3]. The foremost consequence of the above assumption is that the system as a whole does not display any sort of macroscopic sudden shift, or abrupt change: in physics, such a sort of system is technically said to have no *phase transitions*.

The introduction of an interaction among particles structurally changes the macroscopic properties and, correspondingly, the mathematical ones. The probability distribution of the system does not factorise anymore – in computer science terms, the system is no longer compositional [3] –, and the study of its properties becomes much more challenging. *Interacting systems* – that is, systems where particles do not behave independently of each other – are the only suitable candidates to describe real cases beyond the idealised ones. The versatility of interacting systems in the modelling of physical systems is especially proven by the fact that they are the only ones capable to explain phase transitions – like liquid-gas transition – and much more, such as collective emerging effects. While a system made of independent parts can be represented by isolated single nodes, an interacting system is better described by nodes connected by lines or higher-dimensional objects. From the point of view of information and communication theories, an ideal non-interacting gas is a system of non-communicating nodes, whereas an interacting system is made of nodes connected by channels.

3.2 Complexity in Statistical Mechanics

Interaction is a necessary ingredient for complexity in statistical mechanics but definitely not a sufficient one. The simplest standard prototype of an interacting system is the one made of magnetic particles. There, individual particles can behave according to a magnetic field which leaves their probabilistic independence undisturbed. At the same time, two magnetic particles interact with each other, and the strength of their interaction is a crucial tuning parameter to observe a phase transition. If the interaction is weak, the effect of a magnetic field is smooth on the system; instead, if the interaction is strong – in particular, higher than a threshold – even a negligible magnetic field can cause a powerful cooperative effect on the system. The system can be in one of two equilibrium states: the up and the down phase.

Complexity arises when the possible equilibrium states of a system grow very quickly with the number of particles, regardless of the simplicity of the laws governing each particle and their mutual interaction—in other terms, roughly speaking, complexity is much more related to size in number, rather than to complexity of the laws ruling interaction. Such view of complexity in statistical mechanics emerged in a fundamental achievement of theoretical physics, that is, the solution to the so-called mean field *theory of spin glasses* [16]. Beside the physical origins of that theory, the mathematical model describing those systems – see [17] for a rigorous account – accounts not just merely for interaction between particles: instead, it also features the property of being alternatively either *imitative* or *anti-imitative* with the same probability. There, prototypical cooperation and competition effects are both present, and the resulting emerging collective effect is totally new. The equilibrium state space that was identified is endowed with a hierarchical structure: configurations are organised in families, families in superfamilies, and so on. From the geometrical point of view, those spaces are sometimes called *ultrametric*, or tree-like.

3.3 From Statistical Mechanics to Social Systems

In order to illustrate the growing levels of complexity, along with the increasing relevance of interaction, a parallel with social systems is surely of some use here. A group of isolated individuals neither knowing nor communicating with each other is the typical example of a *compositional* social system—that is, a system whose global behaviour results from the independent "sum" of the behaviour of its single components. No sudden shifts are expected in this case at the collective level, unless it is caused by strong external exogenous causes.

In order to obtain a collective behaviour displaying genuinely endogenous phenomena, the individual *agents* should be in a state of exchange of information—i.e., they have to meaningfully interact with each other. The foremost issue here is that *the nature of the interaction determines the nature of the collective behaviour* at the aggregate level. For instance, a simple imitative interaction is capable to cause strong polarisation effects even in presence of extremely small external inputs. Nevertheless, in order to obtain the real complexity of a social system, with clusters iteratively organised in sub-clusters, the interaction ought to be not only *imitative* (cooperative) but also *counter-imitative* (competitive), randomly extracted with equal probability.

There are clear indications that the complex behaviour of many observed socio-economic systems – and in particular the crisis events [18] –, can be approached with the quantitative tools developed within those statistical mechanics ideas. Among the issues for future investigations, *structural stability* of modern society is likely to be key. In such a context, research has to face the challenge of devising out the opportunities to take and the dangers to avoid in relation to the fast changes underwent by social connectivity [19,20] that, in the last decades, has witnessed an enormous increase.

4 Perspectives: Coordination and Socio-technical Systems

In order to draw some consequences from the above notes about complexity and interaction in computational, physical, and social systems as well, two are the possible starting points.

First of all, while physical systems are to be observed, understood, and possibly modelled, computational systems are to be designed and built. In particular, whereas for physical systems the laws of interaction, and their role for complexity, are to be taken as given, to be possibly formalised mathematically by physicists, in the case of software systems the laws of interaction have first to be defined through amenable abstractions and computational models by computer scientists, then exploited by computer engineers in order to build computational systems.

Secondly, a particularly-relevant class of social systems, nowadays, is represented by socio-technical systems – for instance, social platforms like FaceBook [21] and LiquidFeedback [22] –, where active components are mostly represented by humans, whereas interaction is almost-totally regulated by the software infrastructure.

Accordingly, in the remainder of this section we first suggest how coordination models and languages could be used to define the laws of interaction in complex computational systems. Then, we derive a novel perspective over socio-technical systems, where scientific and technical tools from both computer science and physics could be exploited to assess new macroscopic properties of complex systems.

4.1 Coordination Media for Ruling Interaction

Defining the abstractions and the computational model for ruling the interaction space in computational systems basically means to define their *coordination model* [5,23,24]. Coordination models are typically enforced via suitable *coordination middleware*, providing *coordination media* – like Linda tuple spaces [25], or Reo channels [26] – as the mediators for the interaction among *coordinables* – that is, the components of the system representing the coordinated entities –, enforcing the *coordination laws* that rule interaction among coordinables [23]. According to [27], coordination laws shape the interaction space by *constraining* the admissible interactions among components, and between components and the environment—that is, by defining the acceptable perceptions and actions by coordinables, as well as their mutual coordination, independently by their inner structure and behaviour.

Roughly speaking, this means that the global properties of complex coordinated systems that depend on interaction can be enforced through an appropriate choice of the coordination model, essentially based on its expressiveness—that is, the ability to capture and inject suitably-expressive laws of coordination governing system interaction [28,29]. For instance, tuple-based coordination models have been shown to be expressive enough to support self-organising coordination patterns for nature-inspired distributed systems [30].

Along this line, it is quite natural to draw a parallel with physical systems, where the nature of interaction among components (particles) changes structurally the macroscopic properties of systems. In particular, as discussed in Section 3, interacting systems are the only ones capable to model phase transitions, and, more generally, collective emerging effects.

So, in the same way as the study of stigmergy coordination in social insect colonies [31] has proven how coordination models can be used to support stigmergy in computational systems, and also to develop new coordination patterns such as *cognitive stigmergy* [32,8], the study of physical systems could possibly lead to the definition of new sorts of global, macroscopic properties for computational systems inspired by physical ones. For instance, an interesting line of research could involve trying to understand *(i)* whether notions such as *phase*, *phase transition*, or any other macroscopic system property, could be transferred from statistical mechanics to computer science; *(ii)* what such notions would imply for computational systems; and *(iii)* which sort of coordination model could, if any, support such notions.

4.2 A Twofold View of Socio-technical Systems

A particularly-interesting sort of complex system nowadays are the so-called socio-technical systems—that is, artificial systems in which human interaction plays a central role. The so-called Web 2.0 [33] and social platforms like FaceBook or LiquidFeedback are outstanding examples of such a kind of systems, whose interest here comes from their twofold nature of both social systems and computational systems [34,8]

As complex social systems, their complex behaviour is in principle amenable of mathematical modelling and prediction through notions and tools from statistical mechanics, as discussed in Subsection 3.3. As complex computational systems, they are designed and built around some (either implicit or explicit) notion of coordination, ruling the interaction within components of any sort—be them either software or human ones.

Altogether, socio-technical systems are sorts of social systems whose macroscopic properties can be described by exploiting the conceptual tools from physics, and at the same time be enforced by the computational abstractions made available by coordination models. In other terms, social platforms – and complex socio-technical systems in general – precisely represent those kinds of systems where the acceptation of complexity as developed by statistical mechanics (and subsequently expanded to physics and social systems), along with the corresponding mathematical tools for behaviour modelling and prediction, could finally meet the abstractions and tools from computer science (in particular, coordination models and languages) that make it possible to suitably shape the interaction space within complex computational systems. As a result, we envision complex socio-technical systems whose implementation is based on suitable coordination middleware, and whose macroscopic properties can be modelled and predicted by means of mathematical tools from statistical physics.

5 Conclusion

In this paper we elaborate on the notion of complexity as it emerges from fields like physics and computer science, and foresee some possible lines of convergence. In particular, we focus on interaction as the key issue for complex systems, and discuss its role in physics and computer science, with a perspective on social systems. Then, we elaborate on coordination models and middleware as the possible sources of abstractions and technology for enforcing global properties in complex computational systems, which could then be modelled as physical systems, and engineered as computational ones.

In particular, we suggest that socio-technical systems such as large social platforms could represent a perfect case study for the convergence of the ideas and tools from statistical mechanics and computer science, being both social and computational systems at the same time. To this end, in the near future we plan to experiment with social platforms like FaceBook and LiquidFeedback, by exploiting coordination technologies for setting macroscopic system properties, and statistical mechanics tools for predicting global system behaviour.

Acknowledgments. The authors would like to thank the organisers of ICCCI 2013 – and in particular Costin Bădică – for inviting our contribution as both a keynote speech and a written contribution in the proceedings.

This work has been partially supported by the EU-FP7-FET Proactive project SAPERE – Self-aware Pervasive Service Ecosystems, under contract no. 256874.

References

1. Simon, H.A.: The architecture of complexity. Proceedings of the American Philosophical Society 106(6), 467–482 (1962)
2. Kauffman, S.A.: Investigations. Oxford University Press (January 2003)
3. Wegner, P.: Why interaction is more powerful than algorithms. Communications of the ACM 40(5), 80–91 (1997)
4. Omicini, A., Ricci, A., Viroli, M.: The multidisciplinary patterns of interaction from sciences to Computer Science. In: Goldin, D.Q., Smolka, S.A., Wegner, P. (eds.) Interactive Computation: The New Paradigm, pp. 395–414. Springer (September 2006)
5. Gelernter, D., Carriero, N.: Coordination languages and their significance. Communications of the ACM 35(2), 97–107 (1992)
6. Weyns, D., Omicini, A., Odell, J.J.: Environment as a first-class abstraction in multi-agent systems. Autonomous Agents and Multi-Agent Systems 14(1), 5–30 (2007); Special Issue on Environments for Multi-agent Systems
7. Ricci, A., Viroli, M.: Coordination artifacts: A unifying abstraction for engineering environment-mediated coordination in MAS. Informatica 29(4), 433–443 (2005)

8. Omicini, A.: Agents writing on walls: Cognitive stigmergy and beyond. In: Paglieri, F., Tummolini, L., Falcone, R., Miceli, M. (eds.) The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi. Tributes, vol. 20, pp. 543–556. College Publications, London (2012)
9. Viroli, M., Holvoet, T., Ricci, A., Schelfthout, K., Zambonelli, F.: Infrastructures for the environment of multiagent systems. Autonomous Agents and Multi-Agent Systems 14(1), 49–60 (2007) (Special Issue: Environment for Multi-Agent Systems)
10. Boltzmann, L.: Lectures on Gas Theory. University of California Press (1964)
11. Kauffman, S.A.: The Origins of Order: Self-organization and Selection in Evolution. Oxford University Press (January 1993)
12. Bouchaud, J.P., Potters, M.: Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management, 2nd edn. Cambridge University Press, Cambridge (2003)
13. Mantegna, R.N., Stanley, H.E.: Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge (1999)
14. Mézard, M., Montanari, A.: Information, Physics, and Computation. Oxford University Press, Oxford (2009)
15. Nishimori, H.: Statistical Physics of Spin Glasses and Information Processing: An Introduction. International Series of Monographs on Physics, vol. 111. Clarendon Press, Oxford (2001)
16. Mézard, M., Parisi, G., Virasoro, M.A.: Spin Glass Theory and Beyond. An Introduction to the Replica Method and Its Applications. World Scientific Lecture Notes in Physics, vol. 9. World Scientific, Singapore (1986)
17. Contucci, P., Giardinà, C.: Perspectives on Spin Glasses. Cambridge University Press, Cambridge (2012)
18. Stanley, H.E.: Econophysics and the current economic turmoil. American Physical Society News 17(11), 8 (2008); The Back Page
19. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
20. Granovetter, M.S.: The strength of weak ties. American Journal of Sociology 78(6), 1360–1380 (1973)
21. FaceBook: Home page, `http://www.facebook.com`
22. LiquidFeedback: Home page, `http://liquidfeedback.org`
23. Ciancarini, P.: Coordination models and languages as software integrators. ACM Computing Surveys 28(2), 300–302 (1996)
24. Ciancarini, P., Omicini, A., Zambonelli, F.: Coordination technologies for Internet agents. Nordic Journal of Computing 6(3), 215–240 (1999)
25. Gelernter, D.: Generative communication in Linda. ACM Transactions on Programming Languages and Systems 7(1), 80–112 (1985)
26. Arbab, F.: Reo: A channel-based coordination model for component composition. Mathematical Structures in Computer Science 14, 329–366 (2004)
27. Wegner, P.: Coordination as constrained interaction (extended abstract). In: Hankin, C., Ciancarini, P. (eds.) COORDINATION 1996. LNCS, vol. 1061, pp. 15–17. Springer, Heidelberg (1996)
28. Zavattaro, G.: On the incomparability of Gamma and Linda. Technical Report SEN-R9827, CWI, Amsterdam, The Netherlands (October 1998)
29. Denti, E., Natali, A., Omicini, A.: On the expressive power of a language for programming coordination media. In: 1998 ACM Symposium on Applied Computing, SAC 1998, Atlanta, GA, USA, February 27-March 1, pp. 169–177. ACM (1998); Special Track on Coordination Models, Languages and Applications

30. Omicini, A.: Nature-inspired coordination for complex distributed systems. In: Fortino, G., Badica, C., Malgeri, M., Unland, R. (eds.) Intelligent Distributed Computing VI. SCI, vol. 446, pp. 1–6. Springer, Heidelberg (2012)
31. Theraulaz, G., Bonabeau, E.: A brief history of stigmergy. Artificial Life 5(2), 97–116 (1999)
32. Ricci, A., Omicini, A., Viroli, M., Gardelli, L., Oliva, E.: Cognitive stigmergy: Towards a framework based on agents and artifacts. In: Weyns, D., Van Dyke Parunak, H., Michel, F. (eds.) E4MAS 2006. LNCS (LNAI), vol. 4389, pp. 124–140. Springer, Heidelberg (2007)
33. O'Reilly, T.: What is Web 2.0: Design patterns and business models for the next generation of software. Communications & Strategies 65(1st Quarter), 17–37 (2007)
34. Verhagen, H., Noriega, P., Balke, T., de Vos, M. (eds.): Social Coordination: Principles, Artefacts and Theories (SOCIAL.PATH), AISB Convention 2013, University of Exeter, UK, April 3-5. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour (2013)

Trend Based Vertex Similarity for Academic Collaboration Recommendation

Tin Huynh, Kiem Hoang, and Dao Lam

University of Information Technology - Vietnam,
Km 20, Hanoi Highway, Linh Trung Ward, Thu Duc District, HCMC
{tinhn,kiemhv}@uit.edu.vn, daolavi@gmail.com

Abstract. In this paper, we propose a new method used for collaboration recommendation in the academic domain. The proposed method is based on combination of probability theory and graph theory for modeling and analysing co-author network. In the co-author network, similar vertices are explored as potential candidates for collaboration recommendation. Taking the trend information into considering similarity of vertices in the network is the main contribution of this research. We did experiments with co-author networks extracted from the DBLP. Co-authorship that will occur in future used to evaluate accuracy of collaboration recommendation methods. We used metadata of publications from 2001 to 2005 for building the training network. The testing networks were built with publications from 2006-2008 (the testing network 1 for the near future prediction) and 2009-2011 (the testing network 2 for the far future prediction). The experimental results show that the proposed method, called TBRSS (Trend Based Relation Strength Similarity), outperforms other existing methods.

Keywords: Academic Social Network, Vertex Similarity, Collaborative Trend, Collaboration Recommendation.

1 Introduction

The explosive growth and complexity of information that is added to the Web daily challenges all search engines. One solution that can help users deal with flood of information returned by search engines is recommendation. Recommender systems have attracted attention of research community. Recommender systems identify user's interests through various methods and provide specific information for users based on their needs. Rather than requiring users to search for information, recommender systems actively suggest related information to users.

Adomavicius and Tuzhilin provide a survey of the state-of-the-art and possible extensions for recommender systems [3]. Traditional recommender systems are usually divided into three categories: (1) content-based filtering; (2) collaborative filtering and (3) hybrid recommendation systems [3]. Content-based approaches compare the contents of the item to the contents of items in which the user has

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 11–20, 2013.

previously shown interest. Collaborative Filtering (CF) determines similarity based on collective user-item interactions, rather than on any explicit content of the items. These traditional approaches do not mention relationships which can effect to behaviours and interests of individuals. Combining the social network analysis approach with traditional approaches can help us dealing with these disadvantages.

In section 2, we briefly present the overview of research related to recommender systems, applying social network analysis for recommender systems, link prediction in social networks. Especially, we focus on scientific collaboration networks. Section 3 presents vertex similarity measures and our proposed method that analyses the co-author network based on research trend for collaboration recommendation. The experiments, evaluation and discussion will be introduced in section 4. We conclude the paper and suggest future works in section 5.

2 Related Work

Recommender systems are widely used nowadays, especially in E-Commerce. Park et al. collected and classified articles on recommender systems from 46 journals published between 2001 and 2010 to understand the trend of recommender system research and provide practitioners and researchers with insight and future direction on recommender systems [16]. Their statistical numbers showed that recommender systems have attracted the attention of academics and practitioners. The majority of those research papers relates to movie (53 out of 210 research papers, or 25.2%) and shopping (42 out of 210 research papers, or 20.0%) [16]. In another research, Li et al. said that the utilization of recommender system in academic research itself has not received enough attention [12].

The basic idea of the traditional recommendation approaches is to discover users with similar interests or items with similar characteristics or the combination of these. The traditional approaches do not mention relationships which can effect to behaviours and interests of individuals. Therefore, developing the recommendation system research using social network analysis will be an interesting area further research [16].

Jianming He et al. presented a social network-based recommender system (SNRS) which makes recommendations by considering a user's own preference, an item's general acceptance and influence from friends [9]. Yunhong Xu et al. presented using social network analysis as a strategy for E-Commerce Recommendation [18]. Walter Carrer-Neto et al. presented a hybrid recommender system based on knowledge and social networks. Their experiments in the movie domain shown promising results compared to traditional methods [4].

Recently, it has emerged some researches applied social network analysis in the academic area such as building a social network system for analysing publication activities of researchers [1], research paper recommendation [10][15][12][8], collaboration recommendation [6][13], publication venue recommendation [14][17].

In work related to publication venues recommendation, Pham et al. proposed a clustering approach based on the social information of users to derive the recommendations [17]. They studied the application of the clustering approach in two scenarios: academic venue recommendation based on collaboration information and trust-based recommendation. Luong et al. proposed a new social network-based approach that automatically finds appropriate publication venues for authors's research paper by exploring their network of related co-authors and other researchers in the same field [14].

For collaboration recommendation, Chen et al. introduces CollabSeer, an open system to recommend potential research collaborators for scholars and scientists [6]. CollabSeer considers both the structure of a co-author network and an author's research interests for collaborator recommendation. It suggests a different list of collaborators to different users by considering their position in the co-author network structure. In another research, Lopes et al. considered both the semantic issues involving the relationship between the researchers in research areas and the structural issues by the analysis of the existent relationships among researchers [13]. They proposed an innovative approach to recommend new collaborators and to intensify existing collaborations.

In summary, the traditional recommendation methods do not consider social relationships that can effect to behaviours or interests of researchers. In order to overcome the weakness of the traditional methods, the new methods based on analysing academic social network have been studied. To the best of our knowledge, there are not any existing methods taken the trend factor into account. The next section presents our proposed method taking the trend factor into analysing co-author networks for the academic collaboration recommendation.

3 Vertex Similarity Measures

Vertex similarity computation is a crucial step of discovering the missing links. Vertex similarity score is a value which presents the similarity between vertices. There are two main approaches: local structure based measures and global topology based measures.

3.1 Popular Vertex Similarity Measures

Local structure based measures, such as Jaccard [6], Cosine [6], Adamic-Adar [2], use local neighbourhood information to compute the similarity of two vertices. These measures share the same intuition that two vertices are more similar if they share more common neighbours. Therefore, only neighbours are considered as factors affected to the similarity of two vertices, while others are not taken into the computation.

Jaccard: Jaccard similarity coefficient of two vertices is the number of common neighbours divided by the number of vertices that are neighbours of at least one of the two vertices being considered. It is defined in equation:

$$Sim_{Jaccard}(X,Y) = \frac{|n_X \bigcap n_Y|}{|n_X \bigcup n_Y|}, \text{ where}$$

- n_X: is the set of neighbours of vertex X.
- $|n_X \bigcap n_Y|$: number of common sharing neighbours of X and Y.
- $|n_X \bigcup n_Y|$: number of neighbours of X and Y.
- $|n_X|$: number of neighbours of vertex X.

Cosine: Another measure base on local structure is Cosine similarity which shares the same intuition with Jaccard similarity. Cosine similarity is defined as follows:

$$Sim_{Cosine}(X,Y) = \frac{|n_X \bigcap n_Y|}{\sqrt{|n_X|.|n_Y|}}$$

Adamic-Adar: Adamic-Adar also based on local structure. It can be expressed in follows: two vertices are more similar if their common neighbours have less neighbours besides these two. It is defined in the following equation:

$$Sim_{Adamic-Adar}(X,Y) = \sum_{Z \in (n_X \bigcup n_Y)} \frac{1}{\log |n_Z|}$$

Instead of only using local neighbourhood information, global topology based measures can be used for vertex similarity calculation, such as SimRank [11], P-Rank[19]. These measures based on the same intuition: two vertices are more similar if their immediate neighbours in the network are themselves similar. So, the computation of vertex similarity using SimRank, P-Rank are all recursive process. Moreover, a small structure change, such as adding a new vertex or a new edge, will eventually propagate the effect to the whole network. Therefore, it is not feasible to apply these algorithms to a large scale dynamic network [6].

3.2 Trend Based Relation Strength Similarity (TBRSS)

Most of popular vertex similarity measures that mentioned above can be applied for un-weighted networks. None of them are asymmetric. In other research papers, Chen et al. proposed a measure called the relation strength similarity (RSS) [6]. The RSS is base on the idea of relation strength, which defines how close two adjacent vertices are. It is asymmetric and can be applied for weighted networks. Chen et al. applied this measure for discovering missing links in science networks [5][7]. This measure does not mention to the research trend factor that can effect to the similarity of researchers, vertices in the science networks. Our proposed method is an improved version of the RSS. We took the trend factor into account and we proposed a new method called TBRSS. It is a combination of graph theory and probability theory, includes considering the trend or time factor. TBRSS uses vertices' s information including trend factor to define how the vertices are similar. The trend-based similarity score of vertex X and vertex Y is calculated as follows:

$$Direct\_Sim_{TBRSS}(X,Y) = \begin{cases} \dfrac{f(trend)_{XY}}{\sum_{\forall Z \in N(X)} f(trend)_{XZ}} & \text{if X and Y are adjacent} \\ 0 & \text{otherwise,} \end{cases}$$

- $f(trend)_{XY}$: is a function depended on the relation trend factor.
- $N(X)$: is the set of neighbor vertices of X.

In case of that is X and Y are not adjacent, if a network has a simple path that is $Z_1, Z_2, ..., Z_k$ from X to Y (Z_1 is vertex X, Z_k is vertex Y), the $WeightOf\_DirectPath_{TBRSS}$ measure of X and Y is defined as follows:

$$WeightOf\_DirectPath_{TBRSS}(X,Y) = \prod_{i=1}^{k} Direct\_Sim_{TBRSS}(Z_i, Z_{i+1})$$

In large scale networks, computation of the $WeightOf\_DirectPath_{TBRSS}$ (X,Y) may be overloaded. In addition, similarity of any two vertices could be meaningless and the value of $WeightOf\_DirectPath_{TBRSS}$ can approximately reach zero if they reach each other through a long simple path including many intermediate vertices. So, we use *a threshold parameter* r as *a heuristic parameter* to control the performance of system. Thus, we only look for simple paths at most r hops away while others are not considered for the $WeightOf\_DirectPath_{TBRSS}(X,Y)$ as follows:

$$WeightOf\_DirectPath_{TBRSS}(X,Y) = \begin{cases} \prod_{i=1}^{k} Direct\_Sim_{TBRSS}(Z_i, Z_{i+1}) & if k \leq r \\ 0 & \text{otherwise,} \end{cases}$$

Moreover, if network has m simple paths $p_1, p_2, ..., p_m$ then the $Indirect\_Sim_{TBRSS}$ of X and Y can be calculated as following:

$$Indirect\_Sim_{TBRSS}(X,Y) = \sum_{m=1}^{M} WeightOf\_DirectPath_{TBRSS}(X,Y)$$

Finally, the Trend Based Relation Strength Similarity of any two vertices X and Y can be calculated as following:

$$Sim_{TBRSS}(X,Y) = Direct\_Sim_{TBRSS}(X,Y) + Indirect\_Sim_{TBRSS}(X,Y)$$

3.3 TBRSS in the Case of Co-author Networks

For a specified author in the co-author network, co-author relationships in recent years are more effective for the collaboration trend than co-author relationships in a long time ago. So, we define *coefficient k* and *parameter t* to evaluate influence of co-author relationships in recent years as a trend factor. For any two authors X and Y, if we consider the direction of the collaboration from X to Y then the function presented collaboration trend can be calculated as following:

$$f(trend)_{XY} = f(t)_{XY} = k.n_1(t) + (1-k).n_2(t), \text{ where}$$

- k: is a coefficient that effects to the collaboration trend ($0 < k < 1$).
- n_1: is a function returned number of times that X has co-author with Y in recent t years.
- n_2: is a function returned number of times that X has co-author with Y before t years ago.

Depending on specific problem, the coefficient k and parameter t can be adapted to archive the best results.

4 Experiments and Evaluation

In order to evaluate the performance of the TBRSS (our proposed method) in collaboration recommendation, we used the accuracy of link prediction in co-author network. Our experiments were done with a co-author network extract and built from the DBLP.

4.1 Dataset

We use the DBLP dataset downloaded in Oct. 2012 to build co-author network and study the performance of different measures in terms of their ability to predict link. Specifically, we use 11 years of dataset, from 2001 to 2011. First 5 years were used as the training process. So, the authors who have the publications published between 2001 and 2005 are used to build a training co-author network G_0. The degree distribution of G_0 is shown in table 1 and figure 1.

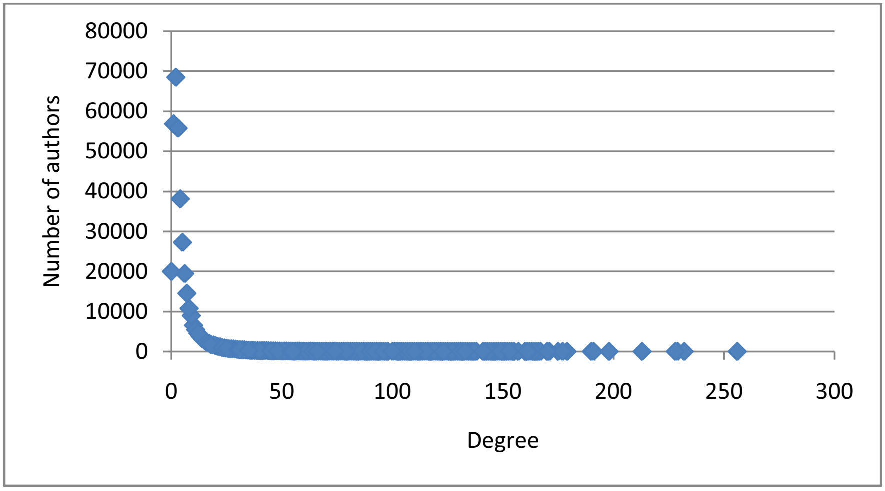


Fig. 1. Degree distribution of vertices in the training network G_0

Table 1. Degree Distribution of vertices in the training network G_0

The training network G_0	No. of Vertices	Lowest Degree	Highest Degree
High Degree Authors Group	162	117	256
Mid Degree Authors Group	1150	58	116
Low Degree Authors Group	366755	0	57

The authors who have publications in interval [2006, 2008] and also in interval [2001, 2005] are used to build the testing network, G_1. We repeat the same procedure to produce one more testing network in interval [2009,2011], G_2. So, we have three co-author networks: G_0 as the training network, G_1 as the first testing network for the near future prediction and G_2 as the second testing network for the far future prediction. The statistics of G_0, G_1 and G_2 are shown in the table 2.

Table 2. The statistics for the training and testing networks

Networks	Number of authors	Number of papers	Average degree
G_0	369704	453980	5.4562
G_1	158551	389944	5.1662
G_2	134703	445369	5.2334

4.2 Experiments for Collaboration Recommendation

In order to make a thorough evaluation, the authors in the training network are classified into 3 different groups of degree distribution: high, mid, and low degree authors. High degree authors are those with degree number in the top 1/3 of all the degrees; low degree authors are those with degree number in the bottom 1/3 of all the degrees; and mid degree authors are all the remaining. For each type, we randomly pick up 100 authors to do experiments.

For each authors who were picked up, we apply the vertex similarity measures on the training network to get the similarity score of them with other authors in training network and then we recommend top-n authors whom they have the highest similarity score with. In our experiments, we applied the threshold parameter $r = 3$ to control finding a path between any two vertices in G_0. We did different experiments with k=0.6, 0.7, 0.8, 0.9 and t=1 ([2005]), t=2 ([2004-2005]), t=3 ([2003-2005]), t=4([2002-2005]). We archived the best result with k=0.9 and t=1 ([2005]). It means that the function based on the relation trend in this case can be calculated as following:

$$f(trend)_{XY} = f(1)_{XY} = 0.9.n_1(1) + (1 - 0.9).n_2(1), \text{ where}$$

- $n_1(1)$: number of co-authorship that X has with Y in the last year (2005).
- $n_2(1)$: number of co-authorship that X has with Y more than one year in the past (before 2005).

4.3 Evaluation and Discussion

In order to evaluate the performance of collaboration recommendation, we checked if the testing networks contain links from an input author to top similar authors returned by different methods.

Table 3 and figure 2 compares the accuracy of our proposed method with the others. For local structure measures, Adamic-Adar outperforms Cosine, Jaccard. RSS, an measure proposed by Chen et al., outperforms all local structure measures. RSS archived 72.33% top 1, 65.61% top 2, 61.50% top 3, 57.80% top 4 and 54.91% top 5 for the testing with G_1 (near future prediction). And TBRSS, proposed in this paper, is an improved version of RSS based on combination of graph theory and probability theory plus the relation trend factor. TBRSS archived 77.67% top 1, 70.95% top 2, 66.85% top 3, 63.59% top 4 and 60.36%

Table 3. Percent of all similar authors in the top who have a link to input authors

Methods	Testing Net 1 (2006-2008)					Testing Net 2 (2009-2011)				
	top 1	top 2	top 3	top 4	top 5	top 1	top 2	top 3	top 4	top 5
Jaccard	52.17	44.15	40.72	38.57	36.59	29.77	23.91	22.15	20.84	19.81
Cosin	47.16	41.64	39.26	36.55	34.70	27.76	22.07	20.58	18.82	17.92
Adamic-Adar	61.20	54.52	51.57	47.90	44.41	41.81	34.78	30.31	27.23	25.07
RSS	72.33	65.61	61.50	57.80	54.91	44.00	36.89	34.49	31.71	29.74
TBRSS	**77.67**	**70.95**	**66.85**	**63.59**	**60.36**	**49.67**	**41.90**	**38.39**	**35.23**	**32.77**

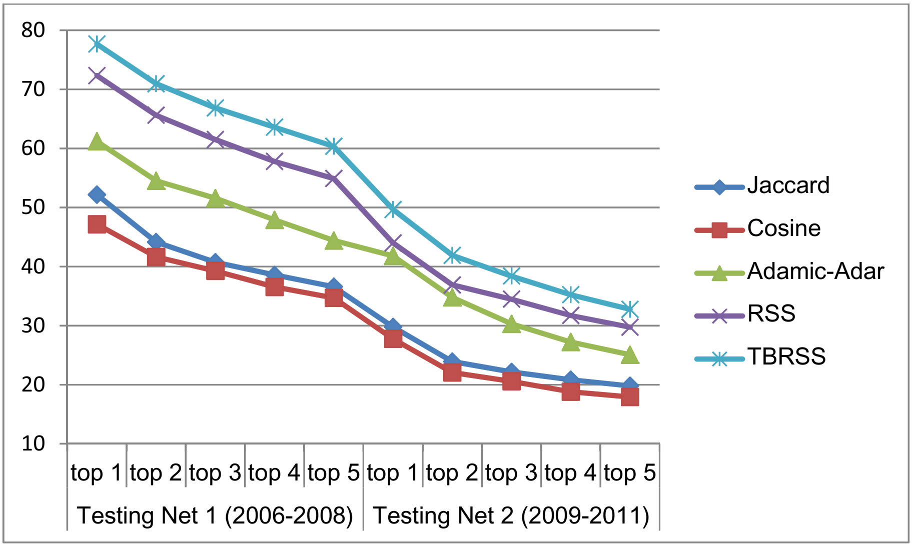


Fig. 2. Evaluation based on the accuracy of link prediction

top 5 for the testing with G_1. So, we can said that our proposed method work better than the others. For the testing network 2 (far future prediction), the accuracy decrease compare with the testing network 1 for all methods. That said that prediction for near future is easier than far future prediction (figure 2).

5 Conclusion and Future Work

The objective of this research is to implement and evaluate a new method based on combining probability theory and graph theory, in addiction with the relation trend factor for analysing co-author networks, to recommend collaboration for computer science researchers. Our proposed method was empirically tested using a dataset of publications extracted from the DBLP in 11 years (from 2001-2011). We compared our proposed method with other popular vertex similarity measuring methods. The experimental results show that our method outperform all other methods. The highest accuracy evaluated base on link prediction for top 1 similar vertex archived 77.67% in the G_1(near future prediction) and 49.67% in the G_1(far future prediction). For other tops of similar vertices, our method also work better than the others.

Our key tasks in the future are continuing to enhance collaboration recommendation methods by combining link-based approach and content-based approach. We also consider to take into account other factors which effect to collaboration trend of researchers. We will also develop recommender modules for publication search engines.

References

1. Abbasi, A., Altmann, J.: A social network system for analyzing publication activities of researchers. TEMEP Discussion Papers 201058, Seoul National University; Technology Management, Economics, and Policy Program, TEMEP (2010)
2. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Social Networks 25(3), 211–230 (2003)
3. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. on Knowl. and Data Eng. 17, 734–749 (2005)
4. Carrer-Neto, W., Hernández-Alcaraz, M.L., Valencia-García, R., García-Sánchez, F.: Social knowledge-based recommender system. application to the movies domain. Expert Systems with Applications 39(12), 10990–11000 (2012)
5. Chen, H.H., Gou, L., Zhang, X., Giles, C.L.: Capturing missing edges in social networks using vertex similarity. In: Proceedings of the Sixth International Conference on Knowledge Capture, K-CAP 2011, pp. 195–196. ACM, New York (2011)
6. Chen, H.H., Gou, L., Zhang, X., Giles, C.L.: Collabseer: a search engine for collaboration discovery. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, JCDL 2011, pp. 231–240. ACM, New York (2011)
7. Chen, H.H., Gou, L., Zhang, X.L., Giles, C.L.: Discovering missing links in networks using vertex similarity measures. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC 2012, pp. 138–143. ACM, New York (2012)

8. Ekstrand, M.D., Kannan, P., Stemper, J.A., Butler, J.T., Konstan, J.A., Riedl, J.T.: Automatically building research reading lists. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, pp. 159–166. ACM, New York (2010)
9. He, J.: A social network-based recommender system. Ph.D. thesis, Los Angeles, CA, USA, aAI3437557 (2010)
10. Huynh, T., Luong, H., Hoang, K., Gauch, S., Do, L., Tran, H.: Scientific publication recommendations based on collaborative citation networks. In: Proceedings of the 3rd International Workshop on Adaptive Collaboration (AC 2012) as part of The 2012 International Conference on Collaboration Technologies and Systems (CTS 2012), May 21-25, pp. 316–321. Denver, Colorado (2012)
11. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 538–543. ACM, New York (2002)
12. Li, C.P.W.: Research paper recommendation with topic analysis. In: 2010 International Conference on Computer Design and Applications (ICCDA), pp. 264–268. IEEE (2010)
13. Lopes, G.R., Moro, M.M., Wives, L.K., de Oliveira, J.P.M.: Collaboration recommendation on academic social networks. In: Trujillo, J., et al. (eds.) ER 2010. LNCS, vol. 6413, pp. 190–199. Springer, Heidelberg (2010)
14. Luong, H., Huynh, T., Gauch, S., Do, L., Hoang, K.: Publication venue recommendation using author network's publication history. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part III. LNCS, vol. 7198, pp. 426–435. Springer, Heidelberg (2012)
15. Ohta, M., Hachiki, T.T.A.: Related paper recommendation to support online-browsing of research papers. In: 2011 Fourth International Conference on Applications of Digital Information and Web Technologies (ICADIWT), pp. 130–136 (2011)
16. Park, D.H., Kim, H.K., Choi, I.Y., Kim, J.K.: A literature review and classification of recommender systems research. Expert Syst. Appl. 39(11), 10059–10072 (2012)
17. Pham, M.C., Cao, Y., Klamma, R., Jarke, M.: A clustering approach for collaborative filtering recommendation using social network analysis. J. UCS 17(4), 583–604 (2011)
18. Xu, Y., Ma, J., Sun, Y.H., Hao, J., Sun, Y., Zhao, Y.: Using social network analysis as a strategy for e-commerce recommendation. In: PACIS, p. 106. AISeL (2009)
19. Zhao, P., Han, J., Sun, Y.: P-rank: a comprehensive structural similarity measure over information networks. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 553–562. ACM, New York (2009)

Application of Graph Cellular Automata in Social Network Based Recommender System

Krzysztof Małecki, Jarosław Jankowski, and Mateusz Rokita

Faculty of Computer Science, West Pomeranian University of Technology, Szczecin, Poland
{kmalecki,jjankowski,mrokita}@wi.zut.edu.pl

Abstract. Recommending systems are used in various areas of electronic commerce. Social platforms make it possible to design recommender systems based on social network analysis and connections between users. This paper presents an alternative approach, which uses graph cellular automata. Empirical research was based on datasets from social platforms that confirmed the effectiveness of the proposed solution and is a motivation for extended research in this area.

Keywords: social networks, recommender systems, graph-based cellular automata.

1 Introduction

Over the past few years, online social platforms have become the most popular websites on the Internet. These take the form of blogs and microblogs systems oriented on publication of text messages [26], social networks and communities [4], multimedia systems [2] up to massively multiplayer games and virtual worlds [7]. They have affected the need for new analytical methods and development of data processing algorithms based on the social network analysis [11] [29]. This research among others is concentrated on the structures of social networks and their characteristics [28], their evolution in time [10] and the communities and groups [8]. Users of social networking sites are building relationships, sharing knowledge and are willing to exchange information. Social networking sites are also the environment in which there is diffusion of information and the transfer of data between users is recorded [17]. Recommendation algorithms are developed, which are an extension of earlier approaches based on collaborative filtering [1] and are oriented to the use of data from the social systems [18]. The purpose of this article is to present the concept of using graph cellular automata as a key part of recommending systems oriented at social networks. The solution was verified in a real environment and the results confirm the ability of using graph cellular automata in this field.

2 Related Work

In decision-making processes, various available sources of information are used in order to increase the rationality of decisions and reduce risk. An important role is

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 21–29, 2013.

played by information and recommendations obtained from the immediate environment, including friends and the people who can be considered a reliable source of information in a particular area. The mechanisms of this type are also characteristic for the electronic environment, which make communication easy and the possibility to get recommendations from a variety of sources. Recommending systems are applied in different areas of ecommerce platforms [20], multimedia and entertainment systems [2] and information recommendation systems based on text documents [6]. Internet platforms are increasingly adopting algorithms focused on obtaining data on user preferences and recommending them to specific products or services. Algorithms are developed based on collaborative filtering mechanisms [13] and data mining [19]. An important direction in research is increasing the computational efficiency of recommending algorithms [9], accuracy [24], modelling interfaces and testing their effects on users [16], as well as obtaining new sources of information [1]. The development of social platforms in recent years is affecting the interest in recommending algorithms that use data from social networks and provide the ability to generate recommendations based on detected relationships [12]. New mechanisms of recommendations based on the analysis of network structures and relationships between users are introduced [25]. Social influence results in information cascades because users adopt new actions after being influenced by others [22]. Social network analysis requires both static analysis [28] as well as the network evolution modelling algorithms [3]. Other areas of research are related to dynamics of ratings with the ability of predicting future links, ratings or community structures [15]. Earlier analyses of social networks were focused mainly on the relationship, but more and more research is focused on the value of attributes. This approach opens up new possibilities for recommender systems development [23]. The recommendations used in semantics and relationships connected with user ratings [12]. Recommendations can also be generated based on the analysis of multilayer networks [18]. In previous recommending systems, little attention has been paid to the use of cellular automata. Available solutions in the form of hybrid web page recommender [27] and methods based on connecting patterns of behaviour [27] and recommendation services [5] were not oriented social networks. In this paper, the use of graph cellular automata and the transfer function parameterized by the states of neighbouring nodes are proposed. The proposed solution is to generate recommendations based on the analysis in the social network environment and activation and deactivation of nodes based on the transfer function variables.

3 Conceptual Framework

Cellular automata allow to model and simulate the phenomena in the world around them, including those related to social relationships. Basic cellular automaton consists of three elements: *D* - dimensional grid of cells, the set of states of a single cell containing k elements, rule F defining state of cell in the time t+1 depending on the state of the cell and cells surrounding it at the time t [21][31]. The evolution of cellular automata occurs in discrete moments of time on the d-dimensional discrete space consisting of identical cells, where each cell can take one of the si states belonging to the set of states. Number of states is finite and arbitrarily large. Thus, the transition rule can be written as [21][31[[30]:

$$s_i(t+1)=F(s_j(t)),\ j\in O(i) \tag{1}$$

Where O (i) is a neighbourhood if the i-th cell state of the cells in step t depends on the state in step t-1 (previous state) and the state of neighbouring cells. The rule change is responsible for the transition states of the cells in the next steps of the simulation. The data from social platforms can be modelled in the form of social networks and therefore it is reasonable to use cellular graph automata. In contrast to the classical cellular automata, graph automate objects are in the relational or virtual space and their physical location in terms of the transition rule does not matter. The basis of graph cellular automata is graph with variable configuration, which is a directed weighted graph expressed in the following form [14]:

$$G=(V,E,K,\alpha) \tag{2}$$

where: V - a set of nodes, E - set of edges, K - a set of weights, α - the edge weight function. In contrast to the classical cellular automata, graph cellular automata (structurally dynamic graph cellular automata) are characterized by a dynamic relational structure of cells and neighbourhood. Ability to reconfigure allows the modelling of systems with a variable number of objects in time [14]. Graph cellular automata can be defined as follows:

$$rsdgCA=(S,d,Q,G,\ F_{rsdgCA},\ R_{rc}\ ,\ F_{rcG}) \tag{3}$$

Table 1. Rules for nodes activation and deactivations

Node activation rules
• considered node has more than *pa1* active neighbours (in the "True") or the number of active neighbours is greater than the total number of • neighbours pa_2, • number of active neighbours with weight above pa_3 is greater than pa_4, • number of active neighbours with weights above pa_5, is greater than pa_6, • number of active neighbours with weights above pa_7, is greater than pa_8, • number of active neighbours with weight above pa_9 is greater than pa_{10}.
Node deactivation rules
• the number of inactive neighbours (in the "False" state) with weights larger than pd_1 is pd_2 greater than the number of nodes compared to the number of active nodes (in the "True" state) , • the number of inactive neighbours with weight greater than pd_3, is twice the ratio of the number of nodes to the number of active nodes, • the number of inactive neighbours with weights greater than pd_4 is greater than the ratio of the number of nodes to the number of active nodes, • the number of inactive neighbours with weights greater than pd_5 is greater than the ratio of the number of nodes to the number of active nodes divided by pd_6, • the number of inactive neighbours with weights greater than pd_7 is greater than the ratio of the number of nodes divided by number of active nodes divided by pd_8.

where: S - state of automata dependent on the state of individual cells, d - dimension of the grid cells, Q - the state automate activity dependent on the activity of individual cell states, G - directed weighted graph, FrsdgCA - a rule that specifies the state of the cells, Rrc - global rule setting conditions for activating and deactivating cells and for graph reconfiguration, FrcG - a rule that specifies the configuration of the structure of cells and their neighbourhood. Both states, and the position of the objects change over the time and they depend on the relevant transition rules. In determining the state of the object by the transition rule weight, there are also important links with the objects to their neighbours. For the purposes of the application in the recommending system, algorithms of the transition rules that operate on activation and deactivation of nodes were developed with parameters pa_i and dp_j respectively. Table 1 shows the generalized conditions for activation and deactivation based on the nodes in the proposed algorithm with parameters.

Significant impacts on the methods of activation and deactivation of the nodes have weights assigned to the appropriate links between the nodes and parameters of activation and deactivation. Weight calculations are at the stage of the graph creation. The next stage of study was to design and implement the discussed algorithms in the SocialGCA system. This system is designed mainly to integrate social network data and then model them as a graph used by the graph cellular automata. The developed environment is the basis for the simulation model that allows us to evaluate the effectiveness of the recommendations. The main goal in developing this system was creating simulation and modelling an environment so that it can become the basis for graph cellular automata.

4 Empirical Research

The developed algorithms were used to carry out simulation and empirical research. The study was based on testing different values of the input parameters and rules in recommending algorithm. Studies in the experimental environment used network structures from Facebook profiles and event subscription data from users declaring interest in the same event. The method of calculating the weights of the links between nodes depends on the similarity of the attribute values. Defining these values allows the appropriate setting of the graph. During the final testing, these parameters were chosen, activation rules AP={pa_1=3, pa_2=0.2, pa_3=1, pa_4=10, pa_5=3, pa_6=5, pa_7=5, pa_8=3, pa_9=10, pa_{10}=1} and deactivation parameters DP={pd_1=1, pd_2=5, pd_3=4, pa_4=7, pd_5=15, pd_6=2, pd_7=30, pd_8=3}. The main objective of this study was to measure the effectiveness and verification of the simulation algorithms. At the beginning of each simulation, a group of users taking part in the event is chosen randomly to determine the percentage of the sample data. The selected subgroup remains unchanged, while the relationship with the other members to the event is removed. The verification involves checking each step of the simulation product of the two sets of nodes with the state of "true": a set of nodes before reduction and a set at the step of the simulation. Coverage of this product with a set of initial rate is a measure of its accuracy. In the first phase, the impact of the events was tested on the simulation results obtained. Individual events may differ primarily at the number of users assigned to them or at the spatial structure of the graph. The application database when performing each simulation on the network consisted of 3884 nodes with ten steps in each simulation. For the different stages of the simulation a different accuracy was obtained. In Table 1, the example of plots of variation for event E_1 with different sample sizes is shown.

Table 2. Visualization of simulation results for the event E1 with different percentages with initially selected nodes

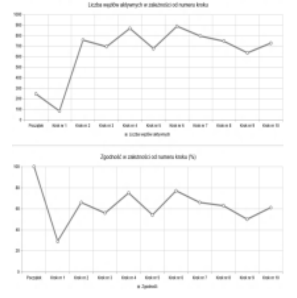
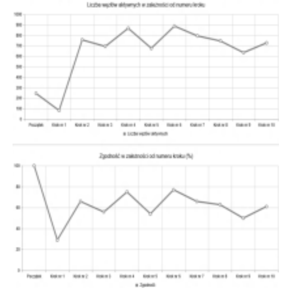
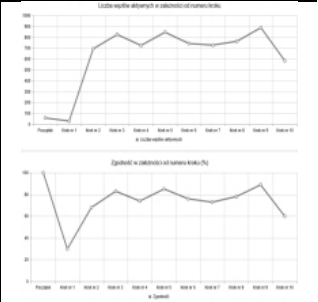
30%	*50%*	*70%*
		

Table 3 shows the aggregated results for the five analysed events. Response from the algorithm was analysed for different percentages of the initial values10, 30, 50 and 70 percent respectively.

Table 3. The final results of the simulation for the analysed events and variable percentage of initial nodes

Event	*Active nodes*	*Maximal accuracy [%] and active nodes*	**Initial nodes** **10%**	**30%**	**50%**	**70%**
E_1	198	Maximal accuracy	62	75	89	99
		Active nodes	790	910	828	838
E_2	246	Maximal accuracy	65	77	88	94
		Active nodes	894	888	958	942
E_3	84	Maximal accuracy	63	81	88	95
		Active nodes	778	766	772	882
E_4	66	Maximal accuracy	64	74	86	97
		Active nodes	882	888	738	764
E_5	60	Maximal accuracy	79	89	89	100
		Active nodes	882	888	786	904

Analysing the data for E_1, regardless of the size of the initial set of active nodes the maximum compatibility was reached with the original state at a high level. In addition, the standard compliance in any case is very similar to the compatibility of the maximum, as a result of small variations in the current compatibility regardless of the current step number. For E_2, the results are very similar to the results obtained for the preceding event but only slightly increased the average number of active nodes, and its variations, which caused the increase in volatility compliance. For E_3, the results are irregular and hard to find any specific dependencies. Fluctuations in the number of active nodes do not coincide with the fluctuations of compliance, which is a graph and in most cases is relatively smooth. The exception is the case for 30% of the initial nodes, where the number of active nodes at each step does not differ much from the average, while the compatibility graph has noticeable fluctuations in the value. The overall results are correct and turn very similar to those obtained previously for the events analysed. E_4 can be observed for the relationship between the number of active nodes, and compatibility with the original state particularly for the 10% and 30% of the initial node. The increase in the number of active nodes increases the compliance and a decrease is observed in the opposite situation. For E_5, the results obtained for this event greatly differ from those of the previous events. Already 10% of the initial node reaches the limit of 80% and the chart to 100% of the latter. It is interesting that the total number of active nodes did not increase compared to the previous simulation. The results showed that in most situations with relatively small percentage of initial nodes the effectiveness of the developed algorithms can be acceptable. A key factor is the selection of the initial nodes and parameterization algorithm.

5 Testing the Impact of the Transition Rules

The next step was carried out with simulations that were designed to investigate the effect of the modification of transition rules for the results. The transition rule used in the system consists of two rules that can be modified: rules activating nodes changing their status to "true" and submitting the link to the event and usually deactivating nodes changing their status to "false" and ultimately putting an association with the event. It should be noted that the implementation of the simulation at this stage uses an event E_1. The aim was to minimize the variables affecting the results so that the change was mainly due to the modification of the transfer function. Table 4 shows the results obtained for each transition rules.

Table 4. Final results for different transition rules

Event	*Maximal*	**Initial nodes**			
	accuracy [%] and active nodes	**10%**	**30%**	**50%**	**70%**
R_1	Maximal accuracy	71	75	88	100
	Active nodes	1074	1054	1112	1144
R_2	Maximal accuracy	35	53	78	85
	Active nodes	442	478	526	464
R_3	Maximal accuracy	34	51	70	89
	Active nodes	438	458	494	540
R_4	Maximal accuracy	31	51	75	90
	Active nodes	396	434	514	538
R_5	Maximal accuracy	34	51	70	89
	Active nodes	426	458	488	510

For rule R_1, the boundary values were decreased in the conditional statements, so that the conditions are met for a greater number of nodes. Modifications made to the rule resulted in increasing the number of active nodes at each step and a significant increase in the accuracy. The increase in compliance, however, is disproportionate to the increase in the number of active nodes, so it can be concluded that this change affected the correctness of the good results obtained. The activation rule R_2 was modified in such a way so as to meet the conditions of the nodes harder. Again, the modification was to change the values of the relevant parameters in the opposite direction. As a result, the maximum compatibility and the average number of active nodes decreased. However, this did not result in a significant reduction in accuracy of the results, which remained at a satisfactory level in many cases. The modified rule R_3 is responsible for changing the function of the node from active to inactive. This change was to increase the difficulty of the conditions in conditional statements so that in effect they reduce the number of deactivated nodes at each step of the simulation. Changes in the results are of the same order of magnitude as in the case of rule R_2 both in the number of active nodes, and the maximum compatibility. The modified R_4 for the nodes was changed in the opposite direction. To facilitate the achievement of performance conditions in the instructions and the same probability, the number of nodes with links is addressed in each step of the simulation. The deactivation rule R5 changed the nodes in such a way so as to obtain a noticeable effect of the changes in the results of the simulation, and a little more dynamic conditions contained in the function. This was achieved by implementing random elements, which has an impact on each conditional statement and is responsible for the change in knots. The introduction of random elements did not result in more irregularities in the charts, however, there was no material ratio on the total course of the simulation.

6 Summary

The main objective of this study was to develop and test the effectiveness of the recommendation system based on cellular automata with the usage of dependence graphs obtained from social networking sites. This graph is the basis for further action of graph based cellular automata whose logic has been implemented to allow complex simulations to be performed. The simulation process itself depends on a number of factors ranging from the number of nodes and the structure of the graph, the initial parameters definable simulation, up to some random elements that are specific to each particular simulation. The developed application allows experiments and simulations to be performed with the usage of real datasets. The results of the examinations and tests have shown that the designed system can be useful in predicting the behaviour of users of social networking sites and can be used as a component of recommending system. The results showed the potential of the developed simulation environment and the effects may be reflected in reality. The experiments also open up opportunities for future research. One of the possible areas is automated tuning of parameters of transition rules and adjusting them to the network characteristics.

An interesting aspect would be to enable the identification of the most influential members in the groups having the largest role in the propagation of the effectiveness of the recommendations. In the next stages of the implementation, the usage of neural networks can be assumed in order to obtain the best parameter adjustments and tuning of the simulation initial parameters as well as the transition rules.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
2. Albanese, M., d'Acierno, A., Moscato, V., Persia, F., Picariello, A.: A Multimedia Semantic Recommender System for Cultural Heritage Applications. In: Proceedings of 15th IEEE International Conference on Semantic Computing, pp. 403–410 (2011)
3. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 44–54. ACM, New York (2006)
4. Boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication 13(1), 210–230 (2007)
5. Chi-Hsuan, W., Sheng-Tzong, C.: Using cellular automata on recommendation mechanism for smart parking in vehicular environments. In: Proceedings of 2nd International Conference on Consumer Electronics, Communications and Networks, pp. 3683–3686 (2012)
6. Degemmis, M., Lops, P., Semeraro, G.: A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. User Modeling and User-Adapted Interaction 17(3), 217–255 (2007)
7. Ducheneaut, N., Yee, N., Nickel, E., Moore, R.: Alone Together - Exploring the Social Dynamics of Massively Multiplayer Online Games. In: Proceedings of ACM CHI 2006 Conference on Human Factors, Quebec, pp. 407–416 (2006)
8. Flake, G., Lawrence, S., Giles, C.L.: Efficient identification of web communities. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 150–160. ACM Press (2000)
9. Forsati, R., Meybodi, M.R.: Effective page recommendation algorithms based on distributed learning automata and weighted association rules. Expert Systems and Applications 37(2), 1316–1330 (2010)
10. Fuchs, C.: Towards a dynamic theory of virtual communities. International Journal of Knowledge and Learning 3(4/5), 372–403 (2007)
11. Hanneman, R., Riddle, M.: Introduction to social network methods. Online textbook (2005), `http://faculty.ucr.edu/~hanneman/nettext`
12. He, J., Chu, W.W.: A Social Network Based Recommender System. Annals of Information Systems. Special Issue on Data Mining for Social Network Data 12, 47–74 (2009)
13. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the ACM Conference on Computer Supported Cooperative, pp. 241–250. ACM Press, New York (2000)

14. Hołowiński, G., Małecki, K.: Grafowy automat komórkowy o zmiennych sąsiedztwach relacyjnych komórek – założenia do implementacji w FPGA. Pomiary Automatyka Kontrola 8, 861–863 (2011)
15. Jamali, M., Haffari, G., Ester, M.: Modeling the temporal dynamics of social rating networks using bidirectional effects of social relations and rating patterns. In: Proceedings of the 20th International Conference on World Wide Web, pp. 527–536. ACM, New York (2011)
16. Jankowski, J.: Modeling the structure of recommending interfaces with adjustable influence on users. In: Selamat, A., Nguyen, N.T., Haron, H. (eds.) ACIIDS 2013, Part II. LNCS, vol. 7803, pp. 429–438. Springer, Heidelberg (2013)
17. Jankowski, J., Michalski, R., Kazienko, P.: The Multidimensional Study of Viral Campaigns as Branching Processes. In: Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C. (eds.) SocInfo 2012. LNCS, vol. 7710, pp. 462–474. Springer, Heidelberg (2012)
18. Kazienko, P., Musiał, K., Kajdanowicz, T.: Multidimensional Social Network in the Social Recommender System. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 41(4), 746–759 (2011)
19. Kazienko, P.: Mining Indirect Association Rules for Web Recommendation. International Journal of Applied Mathematics and Computer Science 19(1), 165–186 (2009)
20. Kim, S., Yum, B.J., Song, J., Kim, S.M.: Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites. Expert Systems with Applications 28(2), 381–393 (2005)
21. Kułakowski, K.: Automaty Komórkowe. Kraków: Wydawnictwo Akademii Górniczo-Hutniczej (2000)
22. Leskovec, J., Singh, A., Kleinberg, J.: Patterns of Influence in a Recommendation Network. In: Ng, W.K., Kitsuregawa, M., Li, J., Chang, K. (eds.) PAKDD 2006. LNCS (LNAI), vol. 3918, pp. 380–389. Springer, Heidelberg (2006)
23. Matsuo, Y., Yamamoto, H.: Community gravity: measuring bidirectional effects by trust and rating on onlinesocial networks. In: Proceedings of the 18th International Conference on World Wide Web, pp. 751–760. ACM, New York (2009)
24. McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: CHI 2006 Extended Abstracts on Human Factors in Computing Systems, pp. 1097–1101. ACM, New York (2006)
25. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology 27, 415–444 (2001)
26. Mitrovic, M., Tadic, B.: Bloggers Behavior and Emergent Communities in Blog Space. The European Physical Journal B - Condensed Matter and Complex Systems 73(2), 293–301 (2010)
27. Talabeigi, M., Forsati, R., Meybodi, M.R.: A Hybrid Web Recommender System Based on Cellular Learning Automata. In: Proceedings of IEEE International Conference on Granular Computing, pp. 453–458 (2010)
28. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, New York (1994)
29. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press, New York (1994)
30. Wolfram, S.: A New Kind of Science. Wolfram Media, Champaign (2002)
31. Wolfram, S.: Theory and Applications of Cellular Automata. World Scientific, Singapore (1986)

The Diffusion of Viral Content in Multi-layered Social Networks

Jarosław Jankowski[1], Michał Kozielski[2], Wojciech Filipowski[2],
and Radosław Michalski[3]

[1] Faculty of Computer Science, West Pomeranian University of Technology, Szczecin, Poland
jjankowski@wi.zut.edu.pl
[2] Faculty of Automatic Control, Electronics & Computer Science,
Silesian University of Technology, Gliwice, Poland
{michal.kozielski,wojciech.filipowski}@polsl.pl
[3] Institute of Informatics, Wrocław University of Technology, Poland
radoslaw.michalski@pwr.wroc.pl

Abstract. Modelling the diffusion of information is one of the key areas related to activity within social networks. In this field, there is recent research associated with the use of community detection algorithms and the analysis of how the structure of communities is affecting the spread of information. The purpose of this article is to examine the mechanisms of diffusion of viral content with particular emphasis on cross community diffusion.

Keywords: diffusion of information, multi-layered social networks, clustering algorithms, multiplex networks, social network analysis.

1 Introduction

Social platforms have become one of the leading trends in the development of the World Wide Web. Due to their specifics and worldwide reach with huge number of users, they create new challenges in the area of analytical methods for processing large data sets. They are used for social analysis and decision support processes as well as marketing activities. The methods in the area of social network analysis are focused on different fields ranging from static analysis towards network dynamics [15]. One of the growing areas is the analysis of information diffusion with focus on diffusion models [11], analysis of the factors affecting the dynamics of the spread of information [13] and optimization of these processes as well as maximization of the reach [9]. These methods are used in marketing analysis [1] as well as for monitoring of users' activity and the flow of digital content [3]. In recent years more attention is paid to the usage of community detection algorithms oriented to the diffusion of information [2] and methods oriented to maximization of the information flow within and between communities [4].

This article attempts to clarify the mechanisms of diffusion of virtual goods in the multiplayer platform functioning as a virtual world in which users can exchange graphics, avatars or different content in viral messages. Users of such an environment

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 30–39, 2013.

create a multi-layered network, where each layer can be defined with different user activities. Analysis of such a network is an interesting but challenging problem [6]. In general, each layer of multi-layer social network can consist of different set of actors connected by different relations. Therefore, the following issues have to be considered when such complex structure is analysed: the communities identified in different layers do not have to fit to each other, some of them can be present in several layers, others can be formed only in one layer. Therefore, there are several issues concerning, e.g., common or separate layers analysis, layers compatibility or significance. These issues influencing the community identification have also impact on the analysis of information diffusion when multi-layered network structure is considered.

The goal of this work is to examine the mechanisms of diffusion of virtual goods with particular emphasis on cross community diffusion, where the communities were identified in multi-layered social network. Another goal is to verify whether the communities identified in different network layers can affect the information diffusion process. Finally, the work aims to give additional insight into the domain of information diffusion in social networks by the analysis of three real life datasets.

2 Related Work

The spread of information has become one of the main areas of the study of the phenomena associated with social media platforms. Research has been carried out on the flow of information in the systems, modelling social influence [14], studying the dynamics of viral campaigns and searching methods of optimization. Research on the theoretical and generalized mechanisms of diffusion processes including its application areas as well as the search for factors affecting the range and other parameters to evaluate their effectiveness [1] is performed.

Recently, more and more analysis related to the characteristics of the diffusion mechanisms is based on algorithms from the field of social network analysis. The static network parameters [1], the variability of the network at the time [15] and the attributes of the participants in the processes of diffusion and structures of interactions [10] are taken into account. The study of diffusion mechanisms on the basis of available influence models is focused on maximizing the range [11], selection of seeding nodes (e.g. [12]) and the analysis of the factors and parameters of social networks which affect the range [1]. The study also draws attention to the dependence of the flow of information related to the structure of the communities identified within the network. There are community detection algorithms in use which are associated with the identification of structures based on various definitions of groups [8], which can be used in networks analysis and for single as well as multi-layer networks [6]. What is recently undertaken by researchers is to try to connect the two areas and communities overlapping detection based on data on the diffusion with the assumption that the information propagation and relationships within communities are closely related [2]. Earlier experiences showed that clusters limit cascades of information and analysis of cluster structures can be a base for adjusting strategies to network structures and detected communities [7]. Other related research emphasized the ability to use interactions and the content of the transmitted messages in finding overlapping communities

[10]. Detecting communities can be used for influence maximization problems and selection of influential nodes [14]. Apart from community detection, research was performed on the flow of information inside and between communities. Integration of community detection algorithms by modelling the spread of information makes it possible to detect influence not only between individuals but between groups which can be useful for large systems where more importantly it is the global view than analysis at the micro scale [3]. Other attempts in this field are oriented towards maximizing diffusion by targeting online communities and increasing cross community communication [4].

While social platforms gather much more information than graph structures, some of the possible directions is using more information sources. An extended analysis based on communication topics, graphs topology and level of participation in communities measured by a number of interactions was performed recently [10]. The proposed method of community discovery using fuzzy modularity integrates several sources of information.

The literature review shows that cross community diffusion is one of the latest topics in this research field and still has open research challenges. The diffusion of information in the social networks is analysed in the context of communities and is usually based on information flow. In this work authors targeted the research to viral content diffusion in online community functioning in a form of virtual world in multi-layered network with the aim to determine what is the mechanism of the in different communities identified within different layers.

3 Theoretical Background and Dataset Description

The idea of multi-layered social network, sometimes called a multiplex network, assumes that there exists more than one communication type between users, e.g. between two particular nodes in the directed network there may be formed at least two edges of different type from one node to another [6]. By separating these types of links into different layers, a multi-layered social network is built. An example of multi-layered social network may be telecom network, where one layer is the layer of phone calls, the second one represents text messages and the third one – multimedia messages. What is natural, these layers may consist of different nodes and edges. This sort of networks allows to analyse layers separately or combined, depending on the desired goal. An example of multi-layered network is presented in Figure 1.

Another concept used in the latter part of this paper are clusters in social networks. Typically, a cluster is considered as a structure, in which network nodes are joined together in tightly knit groups, between which there are only looser connections [8].

The Louvain method [5] was applied to community identification in the work presented. It is based on the concept of modularity maximization and consists of two phases. The algorithm is reminiscent of the self-similar nature of complex networks and naturally incorporates a notion of hierarchy, as communities of communities are built during the process [5]. The algorithm has the computational complexity of O(m), where m is the number of nodes. The output of the method are non-overlapping groups, what enables the analysis of intra- and inter-network communication.

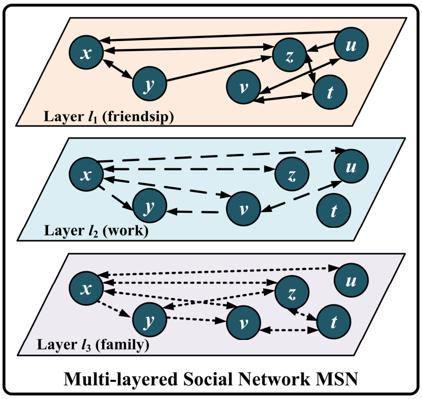


Fig. 1. An example of multi-layered social network [6]

Previous studies indicate a relationship between the processes of diffusion of information and the structures of identified communities within social networks. This area draws attention in recent publications and not all mechanisms are explained yet. In the present study, the aim was to analyse the mechanisms of diffusion in groups for different actions based on viral marketing.

This research is based on the real-life datasets presented in the following paragraphs. The analysis required the extraction of data from viral campaigns carried out in a single environment with comparable results with the ability of monitoring the media diffusion. For the purpose of verification mechanisms in the study, data from the viral campaigns (actions) denoted as V_1, V_2 and V_3 were used. The campaigns were conducted within the virtual world system in which users communicate in a synchronous chat system and their representation in the form of visual avatars. These three campaigns are described later in this section. Within the system, users can transmit messages to the other users and tracing of diffusion processes is possible.

In addition to the social network structure, a number of messages was used to estimate the relationship between a sender and a receiver. Weights W_{AB} and W_{BA} between user A and user B was represented by the number of private messages sent between them from the beginning of the relationship in both directions.

The first viral action denoted as V_1 was completed in one day. Five randomly selected users within seeding received new avatars that can be sent to other users. The mechanism of transmission of viral content did not require connection within friends list between sender and receiver.

The second action V_2 was associated with the users organizing a protest against the sharpening of legal regulations in the field of electronic media. The campaign members have the ability to upload graphical components associated with the action. The transmission mechanism of social relationships required communication between a sender and a receiver in the form of relation on the list of friends. A detailed monitoring of the media diffusion was possible and the action lasted for five days.

Action V_3 was associated with competitions in which winners received gifts and were able to give them to friends. Mechanics of diffusion required the presence of the receiver on the sender's friends list. The values characterising each campaign are presented in Table 1.

Table 1. Viral campaign characteristics

	Unique senders (A)	Unique receivers (B)	Infections sent per A	W_{AB}	W_{BA}
V_1	82	474	5.78	4.15	5.45
V_2	86	224	2.6	13.93	13.35
V_3	218	627	2.88	12.91	15.34

Preliminary analysis of the data indicates the presence of a stronger relationship between the sender and the recipient in action V_2 and V_3 than in the V_1 in action. The campaign V_1 achieved a higher rate of infection transmitted by a single user, however, it was relatively low conversion rate from receiver to sender at 0.17 while for V_2 and V_3 this factor was at 0.38 and 0.35 respectively. In all the analysed actions, senders had a higher login rate than receivers. For the action V_1 login ration of sender in relation to receiver was 1.59, in actions V_2 and V_3 was 1.8 and 1.9 respectively. Such differences indicate potential exposure of the sender as a user with a particular reputation for customer with less experience in the system usage. As shown in the preliminary analysis conducted, the actions have different characteristics, but the most noticeable differences are between the action V_1 and actions V_2 and V_3. In this context, analyses of the diffusion in communities and determining how the message was distributed within the communities can bring extensions to current research in this field.

The social network characterised above can be presented as consisting of two layers. Social layer is based on the relations derived from a list of friends of each user. Communication layer is based on the relations derived from a communication activity of each user. A message sent by a user during a viral action is represented in this case as a relation between the users. The experiments presented in the next paragraph enable the analysis the diffusion mechanisms in groups identified in these two layers.

4 Experimental Results

The main part of the study included the identification and analysis of the mechanisms of spread of information within and between communities. The results obtained show the different characteristics of the various communities within the social (Table 2) and communication (Table 3) network. The tables 2 and 3 present characteristics of the main communities identified. These characteristics cover a number of infections between the groups $C_{i,j}$, percentage of within group infections and outgoing infections and the respective percentage values.

In the social network (Table 2) for V_1 the largest community $C_{1,1}$ consists of 1,640 members and a lowest number of members at level of 277 was in community $C_{1,11}$. Most logins per user occurred in the community $C_{1,\ 19}$ which indicates the most advanced users in this group. Community $C_{1,5}$ gathered users with low activity. The largest community for V_2 was identified as $C_{2,2}$ and it consists of 1,477 members. The highest average number of logins was in group $C_{2,0}$ at the level of 835 logins per user. The three main communities associated with V_3 have the highest number of members. In community $C_{3,4}$ 3486 users were identified. The highest number of logins at 583 was detected in community $C_{3,5}$.

Analysis of the communication network (Table 3) shows the highest number of 260 users for V_1 campaign in the community $C_{1,15}$. For campaign V2 and V3 communities $C_{2,2}$ and $C_{3,3}$ got 198 and 945 users respectively. Tables 2 and 3 show the number of internal and external transmission between groups in two networks – social and communication ones.

Table 2. Intergroup and intragroup infections – Social Network

V_i		$C_{1,2}$	$C_{1,9}$	$C_{1,11}$	$C_{1,19}$	G_{IN}	G_{OUT}	SUM_{IN}	SUM_{OUT}
V_1	$C_{1,2}$	28	9	0	17	51,85%	48,15%	28	26
	$C_{1,9}$	16	28	2	27	21,92%	78,08%	28	45
	$C_{1,11}$	0	1	0	1	0,00%	100,00%	0	2
	$C_{1,19}$	101	60	13	85	39,00%	61,00%	85	174
	Total:	145	98	15	130	36,34%	63,66%	141	247
		$C_{2,0}$	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	G_{IN}	G_{OUT}	SUM_{IN}	SUM_{OUT}
V_2	$C_{2,0}$	31	5	23	15	41,89%	58,11%	31	43
	$C_{2,1}$	1	1	3	0	20,00%	80,00%	1	4
	$C_{2,2}$	25	6	39	13	30,12%	69,88%	39	44
	$C_{2,3}$	6	1	14	12	18,18%	81,82%	12	21
	Total:	63	13	79	40	42,56%	57,44%	83	112
		$C_{3,1}$	$C_{3,4}$	$C_{3,5}$		G_{IN}	G_{OUT}	SUM_{IN}	SUM_{OUT}
V_3	$C_{3,1}$	32	46	15		34,41%	65,59%	32	61
	$C_{3,4}$	56	139	58		22,13%	77,87%	139	114
	$C_{3,5}$	17	43	59		14,29%	85,71%	59	60
	Total:	105	228	132		49,46%	50,54%	230	235

The results for the infection between the groups indicated that not all campaigns behaved similarly. The campaign V_1 had more infections between groups, probably because of the impact of the mechanics of the campaign which did not require the existence of receivers on the friends list. While the number of external infections was high, at the same time infections did not reach the group $C_{1,5}$ which could be due to the isolation of the group and a major share of the beginners with low engagement in this type of actions. In campaigns V_2 and V_3 more inter-group infections were observed. A large number of infections between the groups in the observed campaigns was clearly associated with a high density network. Fig. 2 shows infections within the groups and between the groups and their variability over time. The x axis indicated the next steps determined by the individual campaigns infections. Dotted line shows an increase in external infections and continuous - internal infections within groups.

Table 3. Intergroup and intragroup infections – Communication Network

V_i		$C_{1,3}$	$C_{1,5}$	$C_{1,6}$	$C_{1,15}$	G_{IN}	G_{OUT}	SUM_{IN}	SUM_{OUT}
V_1	$C_{1,3}$	1	0	0	2	33.33%	66.67%	1	2
	$C_{1,5}$	0	3	6	2	27.27%	72.73%	3	8
	$C_{1,6}$	1	1	3	3	37.50%	62.50%	3	5
	$C_{1,15}$	25	25	41	111	54.95%	45.05%	111	91
	Total:	27	29	50	118	52.68%	47.32%	118	106
		$C_{2,2}$	$C_{2,3}$	$C_{2,4}$	$C_{2,5}$	G_{IN}	G_{OUT}	SUM_{IN}	SUM_{OUT}
V_2	$C_{2,2}$	5	0	4	0	55.56%	44.44%	5	4
	$C_{2,3}$	12	80	6	9	74.77%	25.23%	80	27
	$C_{2,4}$	1	3	16	1	76.19%	23.81%	16	5
	$C_{2,5}$	3	4	1	12	60.00%	40.00%	12	8
	Total:	21	87	27	22	71.97%	28.03%	113	44
		$C_{3,2}$	$c_{3,3}$	$c_{3,5}$	$C_{3,12}$	G_{IN}	G_{OUT}	SUM_{IN}	SUM_{OUT}
V_3	$C_{3,2}$	9	0	3	3	60.00%	40.00%	9	6
	$C_{3,3}$	2	19	8	2	61.29%	38.71%	19	12
	$C_{3,5}$	16	12	161	34	72.20%	27.80%	161	62
	$C_{3,12}$	15	21	44	205	71.93%	28.07%	205	80
	Total:	42	52	216	244	71.12%	28.88%	394	160

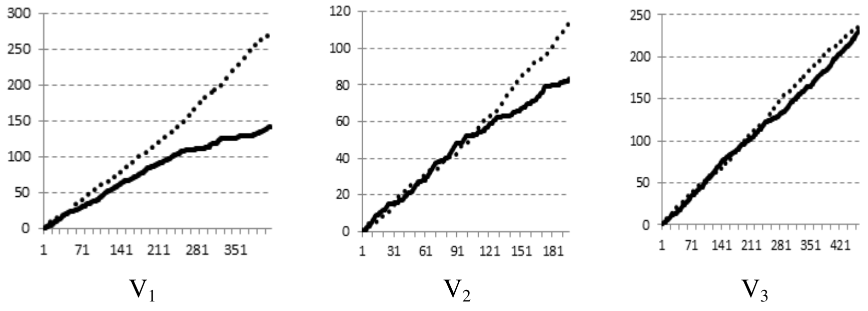
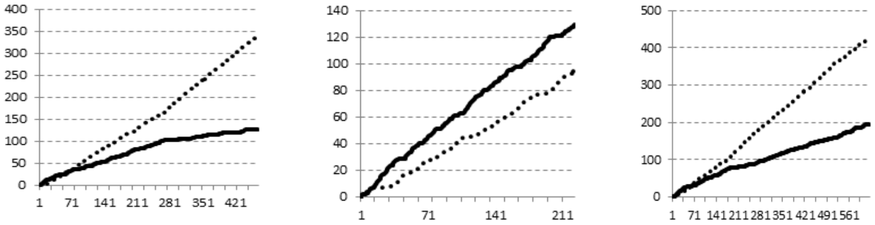


Fig. 2. Inter- and intra-group infections (dotted and solid line respectively) for campaigns V_1, V_2 and V_3 for social communities

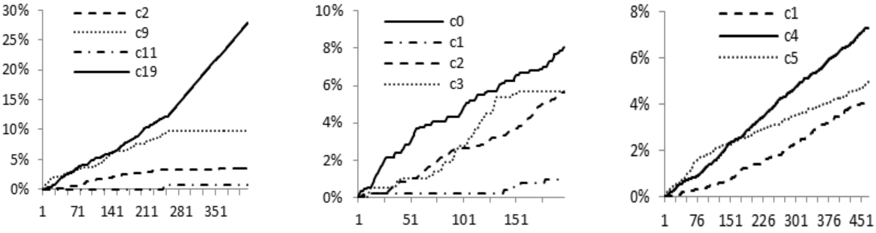
The results indicate that the infections between groups had the largest share in the campaign V_1. The V_3 campaign was characterized by a highest proportion of intergroup infections. Other relation between inter and intra group infections can be observed for groups detected in the communication layer what is presented in the Fig. 3.



V_1 V_2 V_3

Fig. 3. Inter- and intra-group infections (dotted and solid line respectively) for campaigns V1, V2 and V3 for communication communities

The next step examines growth in the number of infections in the individual groups and communities based on saturation measures. In Fig. 4 the growth rate of infection in each group and saturation is showed.



V_1 V_2 V_3

Fig. 4. Saturation in communities in social layer for viral content in V_1, V_2 and V_3 campaigns

The results indicate that for certain groups a relatively quick stabilization of infections was observed. For example, for V_1 the greatest saturation was achieved for community $C_{1,19}$ of up to 30% and dynamic of infections was highest in that community. The community $C_{1,9}$ despite the initial growth comparable to $C_{1,19}$ has stabilized at 10%. Dynamics of saturation in the campaign V_2 indicate the possibility of achieving a critical mass and a sharp increase at some point, which could be observed for the community $C_{2.3}$. Changes of dynamics in infections were observed within campaign V_3, where $C_{3,5}$ initially ran in the highest rates but after infection, 151 higher infection rates was obtained for community $C_{3,4}$. In Fig. 5 the growth rate of infection in each group for communication layer and saturation is showed.

For the communication network layer in the V_1 action the highest level of saturation was obtained for community $C_{1,3}$ and it was at two times higher level than in the social network layer. Similar pattern and relation was observed among well infected communities and those with lower saturation. For the second campaign in the communication network a single main community was identified while in the social network the communities were saturated at the similar levels. Within third campaign saturation in both network layers had similar patterns for two main communities.

Results showed differences between network layers and higher saturation in temporal communication layer.

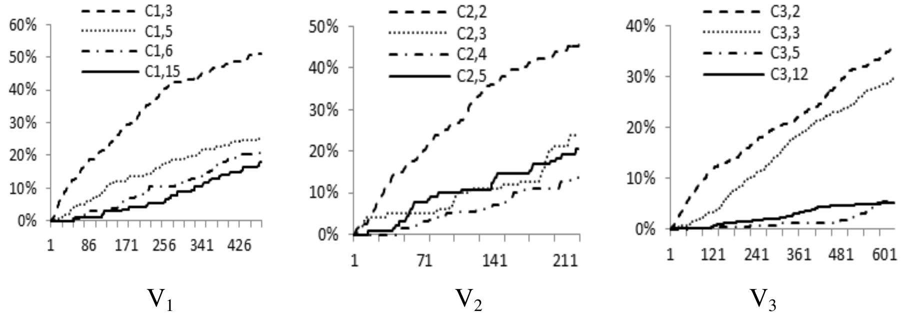


Fig. 5. Saturation in communities in communication layer for viral content in V_1, V_2 and V_3 campaigns

5 Conclusions and Future Work

The study focused on the existence of different mechanisms of diffusion based on two factors - network parameters and characteristics of users. In the presence of a weak relationship between the sender and the recipient, diffusion occurs in a wider range and increased intergroup infections can be observed. However, the spread between groups does not necessarily determine the success of the action (see V_1 action and $C_{1,5}$ community).

The results showed that communities identified in different layers did not fit to each other and different diffusion mechanisms were observed in each later on. Social network treated as more stable over the time showed higher intra group infections than within communities identified in the communication network. A relatively high density of analysed networks showed that in such a case intra community diffusion can be observed at acceptable level. Only at a very early stage, infections were propagated within the community. Next, inter community infections were observed. The performed analysis did not show compatibility of communities between layers and the community overlapping rate at both layers is low. However, for campaign V_1 the same proportion of inter and intra community infections was observed in both layers.

The results show that during viral actions targeted to communities deciding on which layer to use is more important. In the presented examples, if the campaign is addressed in a short term results would be better to target communities in the communication layer while for longer term results with higher impact on the whole community targeting communities in social layer would be more adequate. The analysis of cross community diffusion showed that temporal communication network and its specification was a better environment for inter group infections than stable network.

As a part of future works, the diffusion prediction in different network segments and the selection of seeding strategies oriented to communities can be considered. Additionally, other community identification methods are planned to be verified.

Acknowledgments. The research was partially supported by the European Commission under the 7th Framework Programme, Coordination and Support Action, Grant Agreement Number 316097, ENGINE - European research centre of Network intelliGence for INnovation Enhancement (http://engine.pwr.wroc.pl/) and the fellowship co-financed by European Union within European Social Fund.

References

1. Bampo, M., et al.: The Effects of the Social Structure of Digital Networks on Viral Marketing Performance. Information Systems Research 19(3), 273–290 (2008)
2. Barbieri, N., Bonchi, F., Manco, G.: Cascade-based Community Detection. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM 2013). ACM, Rome (2013)
3. Belák, V., Lam, S., Hayes, C.: Cross-Community Influence in Discussion Fora. In: ICWSM (2012)
4. Belák, V., Lam, S., Hayes, C.: Towards Maximising Cross-Community Information Diffusion. In: Proceedings of ASONAM 2012, pp. 171–178 (2012)
5. Blondel, V.D., et al.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 10, P10008 (2008)
6. Bródka, P., Kazienko, P., Musiał, K., Skibicki, K.: Analysis of Neighbourhoods in Multi-layered Dynamic Social Networks. International Journal of Computational Intelligence Systems 5(3), 582–596 (2012)
7. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press (2010)
8. Fortunato, S.: Community detection in graphs. Phys. Rep. 486(3-5), 75–174 (2010)
9. Goyal, A., et al.: On Minimizing Budget and Time in Influence Propagation over Social Networks. In: Social Network Analysis and Mining. Springer (2012)
10. Jankowski, J., Michalski, R., Kazienko, P.: The Multidimensional Study of Viral Campaigns as Branching Processes. In: Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C. (eds.) SocInfo 2012. LNCS, vol. 7710, pp. 462–474. Springer, Heidelberg (2012)
11. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD, pp. 137–146 (2003)
12. Ma, H., Yang, H., Lyu, M.R., King, I.: Mining social networks using heat diffusion processes for marketing candidates selection. In: Proceedings of the 17th ACM Conf. on Information and Knowledge Management, pp. 233–242 (2008)
13. Najar, A., Denoyer, L., Gallinari, P.: Predicting information diffusion on social networks with partial knowledge. In: WWW, pp. 1197–1204 (2012)
14. Wang, Y., et al.: Community-based greedy algorithm for mining top-K influential nodes in mobile social networks. In: ACM SIGKDD 2010, pp. 1039–1048. ACM, New York (2010)
15. Watts, D., Strogatz, S.: Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998)

Modelling and Analysis of Social Contagion Processes with Dynamic Networks

Alexei Sharpanskykh[1] and Jan Treur[2]

[1] Delft University of Technology, Faculty of Aerospace Engineering
Kluyverweg 1, 2629 HS Delft, The Netherlands
o.a.sharpanskykh@tudelft.nl
[2] VU University Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
treur@few.vu.nl

Abstract. In this paper an agent-based social contagion model with an underlying dynamic network is proposed and analysed. In contrast to the existing social contagion models, the strength of links between agents changes gradually rather than abruptly based on a threshold mechanism. An essential feature of the model – the ability to form clusters – is extensively investigated in the paper analytically and by simulation. Specifically, the distribution of clusters in random and scale-free networks is investigated, the dynamics of links within and between clusters are determined, the minimal distance between two clusters is identified.

Keywords: social contagion models, dynamic networks, agent-based simulation, social decision making.

1 Introduction

Social contagion models have been extensively applied to represent and analyse social decision making, opinion formation, spread of diseases and innovation [1, 2, 4, 5, 7, 8, 12, 14]. Such models describe an evolution of states of individual agents under influence of their neighbouring agents by mutual contagion of these states. In many models [4, 9, 16, 17] the links between agents and their neighbors are constant. In some other models [1, 2, 4, 7] such links may disappear abruptly when states of interacting agents are considered to be too different from each other compared to some threshold. In this paper a social contagion model for social decision making with an underlying network of agents with variable link strengths is proposed and analysed. The strength of the links in the model reflects the degree of influence of one agent to another. The higher the influence to an agent, the higher the extent to which information provided by that agent is used in the decision making; sometimes this also is related to the notion of trust (e.g., [6]). In contrast to the existing models [2, 4, 7, 8], the strength of the links changes gradually in a continuous manner, rather than in a discontinuous manner based on a threshold mechanism. Such a mechanism is supported by sociological literature; e.g., [10], in which much evidence exist that relations between individuals develop continuously.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 40–50, 2013.

Many experimental evidences exist that influence correlates positively with similarity of agents; e.g., [3, 11], either in a static sense or in a dynamic sense. This has led to the principle that the closer the opinions of the interacting agents, the higher the mutual degrees of influence of the agents is (static perspective) or will become (dynamic perspective). Such an assumption underlies most of the existing models of social influence [2, 4, 7, 8, 12, 14].

Inspired by these findings, the dynamics of the links in the proposed model is defined based on the dynamic variant of this principle: closeness of opinions leads to a positive change of connection strength over time. An important feature of the proposed model is that for certain ranges of parameter values clusters of agents emerge that are isolated from each other. A *cluster* is a set of connected agents (i.e., a connected graph) with the same states (e.g., opinions).

The dynamics of social decision making based on the model was analysed by simulation and by mathematical analysis. In particular, the formation and dynamics of clusters of agents was investigated. Both simulation and analytical findings show that the links between the agents within a cluster become stronger over time, and the corresponding degrees of influence tend to 1 (i.e., the highest strength value). At the same time, the strength of the links between the agents in different clusters degrade, and the corresponding degrees of influence tend to 0 (the lowest strength value equivalent to the absence of a connection between the agents). Furthermore, it turns out that different emerging clusters have a certain minimal distance, which was determined analytically. Cluster size distributions in random and scale-free networks were investigated by simulation. The rate of convergence of agent states to equilibrium were investigated both by simulation and analytically and are discussed in the paper. Furthermore, the cluster formation and convergence properties of the proposed model are compared with the corresponding properties of the well-cited threshold-based model developed by Hegselmann and Krause [8].

The paper is organized as follows. In Section 2 a social contagion model for social decision making with an underlying dynamic network is proposed. Results of the model analysis analytically and by simulation are presented in Section 3. Section 4 concludes the paper.

2 The Dynamical Model

The model describes dynamics of decision making by agents in a group as a process of social contagion. The opinion $q_{s,i}$ of an agent *i* for a decision option *s* is expressed by a real number in the range [*0, 1*], reflecting the degree of the agent's support for the option. For each option each agent communicates its opinion to other agents. Agents communicate only with those agents to which they are connected in a social network. In this study two network topologies are considered:

- a *scale-free network topology*: a connected graph with the property that the number of links originating from a given node representing an agent has a power law distribution. In such networks the majority of the agents have one or two links, but a few agents have a large number of links;
- a *random network topology*: a graph, in which links between nodes occur at random. Only connected graphs are considered in this study.

To compare the dynamics in both types of networks, the networks used in this study were generated with 5000 agents and the same average node degree equal to 4.5. This value is close to the average node degree of real social networks.

It is assumed that the agents are able to both communicate and receive opinions to/from the agents, to which they are connected (i.e., the links are bidirectional). Furthermore, a weight $\gamma_{i,j} \in [0,1]$, indicating the degree of influence of agent *i* on agent *j,* is associated with each link for each direction of interaction. This weight determines to which extent the opinion of agent *i* is taken into account in the update of the opinion of agent *j* for each option. These weights may or may not be symmetric.

It is assumed that the agents interact with each other synchronously, i.e., at the same time (parallel interaction mode). For a quantitative comparison of the dynamics of social contagion models with the parallel interaction mode with models with the sequential interaction mode, please refer to [16, 17].

In the parallel mode, the opinion states of the agents are updated at the same time point *t* as follows:

$$q_{s,i}(t+\Delta t) = q_{s,i}(t)+ \eta_i\ \delta_{s,i}(t)\Delta t \qquad (1)$$

Here η_i is an agent-dependent parameter within the range $[0,1]$, which determines how fast the agent adjusts to the opinion of other agents, and

$$\delta_{s,i}(t) = \sum_{j\in AG} \gamma_{j,i}(t)(q_{s,j}(t)- q_{s,i}(t))/\sum_{j\in AG} \gamma_{j,i}(t)$$

is the amount of change of the agent *i*'s opinion; *AG* is the set of all agent names.

The normalization by $\sum_{j\in AG} \gamma_{j,i}(t)$ has the effect that the agent balances by a relative comparison its own self-influence $\gamma_{i,i}(t)$ (i.e., self-assurance that its own opinion is correct) with the influences of other agents.

The degrees of influence $\gamma_{i,j}$ also change over time based on the principle: the closer the opinions of the interacting agents, the higher the mutual degrees of influence of the agents will become. This dynamic principle may be formalised by different functions as follows:

$$\gamma_{i,j}(t+\Delta t)= \gamma_{i,j}(t) + f_{i,j}(\gamma_{i,j}(t), q_{s,i}(t),q_{s,j}(t))\ \Delta t \qquad (2)$$

where for function $f_{i,j}(X, Y, Z)$ the main example used is:

$$f_{i,j}(X, Y, Z) = \text{Pos}(\alpha_{ij}\ (\beta_{ij} - (Y-Z)^2))(1-X)\ - \text{Pos}(-\alpha_{ij}\ (\beta_{ij} - (Y-Z)^2))X \qquad (3)$$

with $\text{Pos}(x) = (|x|+x)/2$, α_{ij} is a speed parameter and β_{ij} is a threshold or tolerance parameter.

Other alternatives for $f_{i,j}(X, Y, Z)$ are:

$$X\ \alpha_{ij}\ (\beta_{ij} - (Y-Z)^2)(1-X) \qquad\qquad X\ \lambda_{ij}(1-|Y-Z|)(1-X) - \zeta_{ij}\ |Y-Z|$$

Here λ_{ij} is an amplification parameter and ζ_{ij} is an inhibition parameter. Note that (1) and (2) are expressed in difference equation format. In Section 3 they are also considered in differential equation format.

A threshold-based model with abruptly changing links and threshold τ as described in [8] can be obtained by defining $f_{i,j}(X, Y, Z)$ as follows:

$$f_{i,j}(X, Y, Z) = 1-X, \quad \text{when } |Y-Z| \leq \tau \qquad (4)$$

$$f_{i,j}(X, Y, Z) = -X, \quad \text{when } |Y-Z| > \tau \qquad (5)$$

3 Model Analysis

In this section first formal analytical results for the model are presented. After that the model is analysed by simulation.

3.1 Mathematical Analysis

For a mathematical analysis, as a point of departure the following differential equations were derived from (1), (2) and (3) in Section 2. For $\gamma_{i,j}(t)$:

$$\begin{aligned} \boldsymbol{d}\gamma_{i,j}(t)/\boldsymbol{dt} = {} & \mathrm{Pos}(\alpha_{ij}\,(\beta_{ij} - (q_{s,i}(t)\text{-}\ q_{s,j}(t))^2))(1\text{-}\gamma_{i,j}(t)) \\ & - \mathrm{Pos}(-\alpha_{ij}\,(\beta_{ij} - (q_{s,i}(t)\text{-}\ q_{s,j}(t))^2))\ \gamma_{i,j}(t) \end{aligned}$$

The differential equations for the $q_{s,i}(t)$ are:

$$\boldsymbol{d}q_{s,i}(t)/\boldsymbol{dt} = \eta_i \sum\nolimits_{j\in AG} \gamma_{j,i}(t)(q_{s,j}(t)\text{-}\ q_{s,i}(t))/\sum\nolimits_{j\in AG} \gamma_{j,i}(t)$$

Equilibrium values for connection strengths $\boldsymbol{\gamma_{i,j}(t)}$.

First, the equilibrium values $\underline{\boldsymbol{\gamma}}_{i,j}$ for $\gamma_{i,j}(t)$ are addressed. The standard approach is to derive an equilibrium equation from the differential equation by putting $\boldsymbol{d}\gamma_{i,j}(t)/\boldsymbol{dt} = 0$. For the specific case for the function $f_{i,j}(X, Y, Z)$ this is

$$\begin{aligned} & \mathrm{Pos}(\alpha_{ij}\,(\beta_{ij} - (q_{s,i}(t)\text{-}\ q_{s,j}(t))^2))(1\text{-}\gamma_{i,j}(t)) \\ & \qquad - \mathrm{Pos}(-\alpha_{ij}\,(\beta_{ij} - (q_{s,i}(t)\text{-}\ q_{s,j}(t))^2))\ \gamma_{i,j}(t) = 0 \end{aligned}$$

The following lemma is used:

Lemma 1.

For any numbers α and β the following are equivalent:

(i) $\alpha\mathrm{Pos}(x) + \beta\mathrm{Pos}(\text{-}x) = 0$

(ii) $\alpha\mathrm{Pos}(x) = 0$ and $\beta\mathrm{Pos}(\text{-}x) = 0$

(iii) $x = 0$ or $x>0$ and $\alpha = 0$ or $x<0$ and $\beta = 0$. ■

Using Lemma 1 it is found that the above equilibrium equation has three solutions

$|\underline{\mathbf{q}}_{s,i} - \underline{\mathbf{q}}_{s,j}| = \sqrt{\beta_{ij}}$

$|\underline{\mathbf{q}}_{s,i} - \underline{\mathbf{q}}_{s,j}| > \sqrt{\beta_{ij}}$ and $\underline{\boldsymbol{\gamma}}_{i,j} = 0$

$|\underline{\mathbf{q}}_{s,i} - \underline{\mathbf{q}}_{s,j}| < \sqrt{\beta_{ij}}$ and $\underline{\boldsymbol{\gamma}}_{i,j} = 1$

More can be found about the circumstances under which such equilibria can occur, and for a wider class of functions $f_{i,j}(X, Y, Z)$. The following symmetry properties are relevant.

Definition.

The network is called *weakly symmetric* if for all nodes i and j at all time points it holds $\gamma_{i,j} = 0 \Leftrightarrow \gamma_{j,i} = 0$ or, equivalently: $\gamma_{i,j} > 0 \Leftrightarrow \gamma_{j,i} > 0$. The network is called *fully symmetric* if $\gamma_{i,j} = \gamma_{j,i}$ for all nodes i and j at all time points.

Note that the network is fully symmetric if the initial values for $\gamma_{i,j}$ and $\gamma_{j,i}$ are equal and $f_{i,j}(X, Y, Z)=f_{j,i}(X, Z, Y)$ for all X, Y, Z; the latter condition is fulfilled for the specific case if $\alpha_{i,j}=\alpha_{j,i}$ and $\beta_{i,j}=\beta_{j,i}$. The following lemma is used to obtain Theorem 1.

Lemma 2.
a) If for some node *i* at time *t* for all nodes *j* with $q_{s,j}(t) > q_{s,i}(t)$ it holds $\gamma_{j,i}(t) = 0,$ then $q_{s,i}(t)$ is decreasing at *t*: $\boldsymbol{dq_{s,i}(t)/dt} \leq 0.$

b) If, moreover, a node *k* exists with $q_{s,k}(t) < q_{s,i}(t)$ and $\gamma_{k,i}(t) > 0$ then $q_{s,i}(t)$ is strictly decreasing at *t*: $\boldsymbol{dq_{s,i}(t)/dt} < 0.$

Proof: a) From the expressions for $\delta_{s,i}(t)$ it follows that $\delta_{s,i}(t) \leq 0,$ and therefore $\boldsymbol{dq_{s,i}(t)/dt} \leq 0$, so $q_{s,i}(t)$ is decreasing at *t*.
b) In this case it follows that $\delta_{s,i}(t) < 0$ and therefore $\boldsymbol{dq_{s,i}(t)/dt} < 0$, so $q_{s,i}(t)$ is strictly decreasing. ■

Theorem 1 (Equilibrium values $\underline{\gamma}_{i,j}$).
Suppose the network is weakly symmetric, and $f_{i,j}(X,Y,Y) > 0$ for all *X*, *Y* with $0 < X < 1$. Then in an equilibrium state for any two nodes *i* and *j* it holds $\underline{\gamma}_{i,j} = 0$ or $\underline{\gamma}_{i,j} = 1.$ More specifically, the following hold:

a) In an equilibrium state with $\underline{q}_{s,i} \neq \underline{q}_{s,j}$ it holds $\underline{\gamma}_{i,j} = 0.$

b) In an equilibrium state with $\underline{q}_{s,i} = \underline{q}_{s,j}$ it holds $\underline{\gamma}_{i,j} = 0$ or $\underline{\gamma}_{i,j} = 1.$ If $q_{s,i}(t) = q_{s,j}(t)$ and $0 < \gamma_{i,j}(t) < 1$, then $\gamma_{i,j}(t)$ is strictly increasing at time *t*: $\boldsymbol{d}\gamma_{i,j}(t)/\boldsymbol{dt} > 0.$
Proof: Provided in the online appendix at http://iccci13.9k.com/app.pdf

Note that the criterion on the function $f_{i,j}(X, Y, Z)$ in Theorem 1 is satisfied for the specific function $f_{i,j}(X, Y, Z) = \text{Pos}(\alpha_{ij} (\beta_{ij} - (Y-Z)^2))(1-X) - \text{Pos}(-\alpha_{ij} (\beta_{ij} - (Y-Z)^2))X$ if and only if $\alpha_{ij}, \beta_{ij} > 0$, which is the case.

Equilibrium values for $q_{s,i}(t)$

In an equilibrium of the network not only the $\gamma_{i,j}$ are in an equilibrium $\underline{\gamma}_{i,j}$ but also the $q_{s,i}$. From the differential equations for the $q_{s,i}$ it follows that the equilibrium values $\underline{q}_{s,i}$ for $q_{s,i}\,(t)$ have to satisfy $\sum_{j\in AG} \underline{\gamma}_{j,i}\,(\underline{q}_{s,j} - \underline{q}_{s,i}\,) = 0.$

When $\underline{\gamma}_{j,i} = 0$ for all *j*, then from the differential equation it follows that $q_{s,i}$ is in equilibrium irrespective of what value it has. Suppose at least one node *j* exists with $\underline{\gamma}_{j,i} \neq 0.$ Then the equilibrium equations can be rewritten as

$$\underline{q}_{s,i} = \sum_{j\in AG} (\underline{\gamma}_{j,i} / \sum_{k\in AG} \underline{\gamma}_{k,i})\, \underline{q}_{s,j}$$

This provides a system of linear equations for the $\underline{q}_{s,i}$ that could be solved, unless they are trivial or dependent. To analyse this, suppose S_i is the cluster (of size s_i) of nodes with same equilibrium value as $\underline{q}_{s,i}$:

$$S_i = \{\, j \mid \underline{q}_{s,j} = \underline{q}_{s,i} \,\} \qquad\qquad s_i = \#\,(S_i)$$

In Theorem 1a) above it has been found that $\underline{\gamma}_{j,i} = 0$ if $j \notin S_i$. Therefore

$$\sum_{j\in AG} \underline{\gamma}_{j,i}\, \underline{q}_{s,j} = \sum_{j\in S_i} \underline{\gamma}_{j,i}\, \underline{q}_{s,j} = \sum_{j\in S_i} \underline{\gamma}_{j,i}\, \underline{q}_{s,i}$$

Then the equilibrium equation for $\underline{q}_{s,i}$ becomes:

$$\underline{q}_{s,i} = \sum_{j\in AG} \underline{\gamma}_{j,i}\, \underline{q}_{s,j} / \sum_{j\in AG} \underline{\gamma}_{j,i} = \sum_{,j\in S_i} \underline{\gamma}_{j,i}\, \underline{q}_{s,j} / \sum_{j\in S_i} \underline{\gamma}_{j,i} \;=\; \underline{q}_{s,i}$$

Thus these equations do not provide a feasible way to obtain information about the equilibrium values $\underline{q}_{s,i}$. However, by different methods at least some properties of the equilibrium values $\underline{q}_{s,i}$ can be derived, as is shown below.

The following conditions on the function $f_{i,j}(X, Y, Z)$ are assumed:

Definition.

The function $f_{i,j}(X, Y, Z)$ has a *threshold* τ for $Y - Z$ if

a) For all Y and Z *it holds*

$$f_{i,j}(0, Y, Z) \geq 0 \qquad\qquad f_{i,j}(1, Y, Z) \leq 0$$

b) For all X with $0 < X < 1$ and all Y and Z it holds

$$f_{i,j}(X, Y, Z) > 0 \quad \text{iff} \quad |Y - Z| < \tau$$
$$f_{i,j}(X, Y, Z) = 0 \quad \text{iff} \quad |Y - Z| = \tau$$
$$f_{i,j}(X, Y, Z) < 0 \quad \text{iff} \quad |Y - Z| > \tau$$

Note that (given that $\alpha_{ij} > 0$ is assumed) the function

$$f_{i,j}(X, Y, Z) = \text{Pos}(\alpha_{ij}\,(\beta_{ij} - (Y\text{-}Z)^2))(1\text{-}X)$$
$$- \text{Pos}(-\alpha_{ij}\,(\beta_{ij} - (Y\text{-}Z)^2))X$$

satisfies these conditions for threshold $\surd\beta_{ij}$.

Theorem 2 (Distance between equilibrium values $\underline{\boldsymbol{q}}_{s,i}$).

Suppose the network is weakly symmetric, the function $f_{i,j}(X, Y, Z)$ has a threshold τ, and the network reaches an equilibrium state with values $\underline{\boldsymbol{q}}_{s,i}$ for the different nodes i. Then for every two nodes i and j if their equilibrium values $\underline{\boldsymbol{q}}_{s,i}$ and $\underline{\boldsymbol{q}}_{s,j}$ are distinct, and the initial values for $\gamma_{i,j}$ and $\gamma_{j,i}$ are nonzero, they have a distance of at least τ: $|\underline{\boldsymbol{q}}_{s,i} - \underline{\boldsymbol{q}}_{s,j}| \geq \tau$. In particular, when all initial values for $\gamma_{i,j}$ and $\gamma_{j,i}$ are nonzero, there are at most $1 + 1/\tau$ distinct equilibrium values $\underline{\boldsymbol{q}}_{s,i}$.

Proof: Provided in the online appendix at http://iccci13.9k.com/app.pdf

In case the network is fully symmetric (i.e., $\gamma_{j,i} = \gamma_{i,j}$ for all i and j) the equilibrium values $\underline{\boldsymbol{q}}_{s,i}$ can be related to the initial values $q_{s,i}(t)$. In this case the sum $\Sigma_i\, q_{s,i}(t)$ is preserved: $\Sigma_{i\in AG}\, q_{s,i}(t) = \Sigma_{i\in AG}\, q_{s,i}(t')$ for all t and t'. From $\gamma_{j,i} = \gamma_{i,j}$ this can be established as follows:

$$\boldsymbol{d}\, \Sigma_{i\in AG}\; q_{s,i}(t)/\boldsymbol{d}t = \Sigma_{i\in AG}\; \boldsymbol{d}\, q_{s,i}(t)/\boldsymbol{d}t \;=\; \eta_i\, \Sigma_{i\in AG}\, \Sigma_{j\in AG}\, \gamma_{j,i}(t)(q_{s,j}(t)\text{-}\, q_{s,i}(t))/\Sigma_{j\in AG}\, \gamma_{j,i}(t)$$
$$= \eta_i\, [\, \Sigma_{k\in AG}\, \Sigma_{i\in AG}\, \gamma_{i,k} q_{s,k}(t)\text{-}\, \Sigma_{k\in AG}\, \Sigma_{i\in AG}\, \gamma_{i,k}\, q_{s,k}(t)]/\Sigma_{j\in AG}\, \gamma_{j,i}(t) = 0$$

The fact that $\Sigma_{i\in AG}\; q_{s,i}(t)$ is preserved can be applied to compare the equilibrium values $\underline{\boldsymbol{q}}_{s,i}$ to the initial values $q_{s,i}(t_0)$. Let $\underline{\boldsymbol{S}}$ be the set of clusters of equilibria:

$$\underline{\boldsymbol{S}} = \{S_i \mid i \text{ any node}\}$$

For $C \in \underline{\boldsymbol{S}}$ define

$$\underline{\boldsymbol{q}}_{s,C} \;=\; \underline{\boldsymbol{q}}_{s,j} \quad \text{for any } j \in C$$
$$s_C = \#(C) = s_j \text{ for any } j \in C$$

Then from the preservation it follows

$$\Sigma_{i\in AG}\; \underline{\boldsymbol{q}}_{s,i} \;=\; \Sigma_{i\in AG}\, q_{s,i}(t_0)$$

Therefore

$$\Sigma_{C\in \underline{S}}\; \Sigma_{i\in C}\; \underline{\boldsymbol{q}}_{s,i} \;=\; \Sigma_{i\in AG}\;\; q_{s,i}(t_0)$$
$$\Sigma_{C\in \underline{S}}\;\; (s_C/n)\, \underline{\boldsymbol{q}}_{s,C} \;=\; \Sigma_{i\in AG}\;\; q_{s,i}(t_0)/n$$

with $n{=}\#(AG)$ the total number of nodes. So, the weighted average over the clusters (with as weights the fraction of the total number of nodes in the cluster) is the average of the initial values $q_{s,i}(t_0)$. These are summarised in the following theorem:

Theorem 3 (Equilibria $\underline{q}_{s,i}$ in fully symmetric case).
Suppose the network is fully symmetric. Then the sum $\Sigma_{i\in AG}\ q_{s,i}(t)$ is preserved over time. Moreover, the weighted average of the equilibrium values for the clusters, with the fraction of the total number of nodes in the cluster as weights, is the average of the initial values:

$$\Sigma_{C\in \underline{S}}\ (s_C/n)\ \underline{q}_{s,C} = \Sigma_{i\in AG}\ q_{s,i}(t_0)/n \qquad \blacksquare$$

Because of the space limitations, analysis of the model behaviour around equilibria is described in an online appendix at http://iccci13.9k.com/app.pdf

3.2 Analysis by Simulation

In this section two model variants from Section 2 are analysed by simulation: model *M1* with continuously changing links (equation (3)) and a threshold-based model *M2* with abruptly changing links (equations (4) and (5)). Both models have the same threshold $\tau=\sqrt{\beta_{ij}}$. The models were simulated in Matlab.

To compare the models, 10 different random network topologies with 5000 agents and 10 different scale-free network topologies with 5000 agents were generated. The scale-free networks were obtained using the Complex Networks Package [13] with scale-free degree distribution of α=-2.2 (as in many real social networks). The average node degree of such networks with 5000 agents equals 4.5. The random networks were generated with the same average node dergree.

The agents formed opinions on some topic *s*. The parameters of the agents and of the links were uniformly distributed as follows: $\eta_i\in[0.5, 1]$; $q_{s,i}(0)\in[0,1]$; $\gamma_{i,j}(0)\in(0, 1]$ (in model *M2* $\gamma_{i,j}(0)=1$, if there was a link between *i* and *j,* and *0* otherwise). These distributions are assumed to represent the diversity that naturally occurs in real-world agent populations.

The simulation time was 300 time points and $\Delta t=1$.

In the previous section 3.1 it was proven that when all initial values $\gamma_{i,j}(0)$ in the population of agents are nonzero, then at most $1 + 1/\sqrt{\beta_{ij}}$ clusters can be formed in the model with threshold $\tau=\sqrt{\beta_{ij}}$. The minimal distance between two clusters is $\sqrt{\beta_{ij}}$, provided that these two clusters were not disconnected initially. In this section we investigate how parameters α_{ij} and β_{ij} influence the number and size of the clusters emerging in the scale-free and random networks. Furthermore, the rate of convergence of the agent opinions is determined for different parameter settings.

In the simulation study three values for β_{ij}: 0.001, 0.0025, 0.01 and three values for α_{ij}: 1, 10, 20 were used. According to the findings from Section 3.1, at most 33 clusters could emerge for β_{ij}=0.001, 21 clusters for β_{ij}=0.0025 and 11 clusters for β_{ij}=0.01. In the networks used in the simulation less clusters were formed, as these networks were not fully connected. For $\beta_{ij} > 0.01$, in most cases only one cluster was formed containing all the agents. The minimal distances between the clusters in the simulated networks were greater than $\sqrt{\beta_{ij}}$. (Table 1).

In the tables with the results small clusters have the size up to 101 agents; medium clusters contain more than 100 but less than 1001 agents, and large clusters comprise of more than 1000 agents.

Besides β_{ij}, parameter α_{ij} also influences the number of clusters with the limit $1 + 1/\sqrt{\beta_{ij}}$ (Tables 2, 3), however in a more intricate manner. For example, in the random

Table 1. The minimal distances between the clusters in 10 random and 10 scale-free networks determined by simulation and analytically

Parameter settings	**β=0.001**			**β=0.0025**			**β=0.01**		
	α=1	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**
Scale-free	0.034	0.041	0.038	0.053	0.059	0.07	1 cluster		
Random	0.041	0.034	0.033	0.1	0.084	0.2	0.4	0.15	0.23
Analytical	0.032			0.05			0.1		

networks the largest number of clusters emerges when α_{ij} takes intermediate values (around 10), whereas in the scale-free networks many clusters tend to form with low values of $\alpha_{ij.}$

Table 2. The mean and standard deviation values (in parentheses) of the numbers of small (≤ 100 agents), medium (>100 and ≤ 1000 agents) and large (>1000 agents) size clusters emerging in 10 random networks

Parameter settings	**β=0.001**			**β=0.0025**			**β=0.01**		
	α=1	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**
Small *M1*	1.1 (1)	4 (1.2)	2.9 (1.2)	0.6 (0.9)	1.9 (0.7)	1 (0.8)	0.2 (0.4)	0.3 (0.6)	0.2 (0.4)
Small *M2*	5.1 (1.1)			2.9 (0.9)			0.6 (0.5)		
Medium *M1*	0	0	0	0	0	0	0	0	0
Medium *M2*	3.8 (0.9)			2 (0.8)			0		
Large *M1*	1	1	1.1	1	1	1	1	1	1
Large *M2*	2.3 (0.6)			1.3 (0.5)			1		

Table 3. The mean and standard deviation values (in parentheses) of the numbers of small (≤ 100 agents), medium (>100 and ≤ 1000 agents) and large (>1000 agents) size clusters emerging in 10 scale-free networks

Parameter settings	**β=0.001**			**β=0.0025**			**β=0.01**		
	α=1	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**
Small *M1*	5.9 (0.9)	2.4 (0.8)	0.6 (0.6)	1.8 (0.9)	0.8 (0.7)	0.1 (0.3)	0	0	0
Small *M2*	3.1 (1.3)			1.6 (0.8)			0.1 (0.3)		
Medium *M1*	0	2.4 (0.8)	2.8 (0.6)	0	1.6 (0.5)	1.6 (0.7)	0	0	0
Medium *M2*	6.6 (1.3)			4.1 (1)			1.6 (0.5)		
Large *M1*	1	1	1	1	1.1 (0.3)	1.2 (0.4)	1	1	1
Large *M2*	1.6 (0.5)			1.5 (0.5)			1		

In the random networks simulated with *M1* model only small and large size clusters tend to form (Table 4). Only one large size cluster emerges in the *M1*-based simulations, which contains the great majority of the agents in the population. Model

M2 produces also medium size clusters, but their amount is less than the number of the small size clusters. This can be partially explained by the absence of central agents or hubs with many connections in the random networks. Such agents would be able to attract large groups of other agents and influence their opinions so that the whole agent population may become polarized in larger opposing clusters.

Table 4. The mean and standard deviation values (in parentheses) of the sizes of small (≤ 100 agents), medium (>100 and ≤ 1000 agents) and large (>1000 agents) clusters emerging in 10 random networks

Parameter	**β=0.001**			**β=0.0025**			**β=0.01**		
settings	**α=1**	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**
Small ***M1***	1.1 (0.3)	2.2 (1.6)	2.1 (1.9)	1	1.7 (1.2)	1.6 (0.8)	1	1.3 (0.6)	1
Small ***M2***	17.5 (22.1)			19.4 (25.2)			5.7 (3.4)		
Medium ***M1***	0	0	0	0	0	0	0	0	0
Medium ***M2***	405.7 (253.8)			439.7 (271.8)			0		
Large ***M1***	4998 (1)	4996 (3)	4539 (1026)	4999 (1)	4997 (2)	4998 (1)	4999 (0.4)	4999 (1)	4998 (0.4)
Large ***M2***	1465 (427)			3126 (1280)			4997 (3)		

Table 5. The mean and standard deviation values (in parentheses) of the sizes of small (≤ 100 agents), medium (>100 and ≤ 1000 agents) and large (>1000 agents) clusters emerging in 10 scale-free networks

Parameter	**β=0.001**			**β=0.0025**			**β=0.01**		
settings	**α=1**	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**	**α=1**	**α=10**	**α=20**
Small ***M1***	1.7 (0.9)	47.5 (35.9)	1	1.7 (0.7)	53.3 (34.6)	1	0	0	0
Small ***M2***	30.4 (30)			36.7 (33)			48		
Medium ***M1***	0	222 (136)	429 (166)	0	276 (173)	450 (192)	0	0	0
Medium ***M2***	424.8 (242)			462.6 (271)			543.5 (241)		
Large ***M1***	4990 (3)	4351 (228)	3797 (246)	4997 (1.8)	4105 (1051)	3566 (1200)	5000	5000	5000
Large ***M2***	1314 (165)			2030 (727)			4126 (131)		

More medium size clusters tend to emerge in the scale-free networks simulated with both *M1* and *M2* models (Tables 3 and 5). The scale-free network topology contains hub agents. When such agents have opposing opinions, they form the basis for future clusters in which the whole population is divided. Because of such agents, who group others around themselves, the small and medium size clusters in the scale-free networks are on average larger than the ones emerging in the random networks (Tables 4 and 5).

According to the analytical results from Section 3.1, the links between the agents within a cluster in model *M1* become stronger and the corresponding degrees of influence $\gamma_{i,j}(t)$ gradually tend to *1*, whereas the links between the agents from different clusters gradually disappear. In contrast to *M1*, links in *M2* appear and disappear instantaneously, depending on the states of the agents, which leads to faster and more abrupt cluster formation. Intuitively, this may be compared to instantaneous decision making (e.g., under high stress) without more detailed consideration and discussion. Because of this, in *M2* agents break links more easily than in *M1*, and thus, more clusters emerge (Tables 2, 3).

4 Conclusions and Discussion

In this paper an agent-based social decision making model based on social contagion with a dynamic network is proposed. In contrast to the existing models [2, 4, 7, 8], similarity in agent states or opinions has a dynamic effect on the strengths of the links between agents: they change gradually over time, rather than that a more static effect is used based on a threshold mechanism. The model was analysed analytically and by simulation. Cluster formation has been extensively investigated in the paper for the models with gradually and abruptly changing links: the distribution of clusters in random and scale-free networks was investigated, the dynamics of links within and between clusters were determined, the minimal distance between two clusters was identified.

In the paper simulation results were presented for the networks with the average node degree 4.5. However, also experiments for more dense networks were performed. In general, the higher the density of a network, the higher its convergence speed to an equilibrium state and the less number of clusters are formed.

Besides the parallel mode of interaction we also performed simulation for the sequential interaction mode. In the latter mode two randomly chosen agents interacted in each iteration. Small, medium and large clusters emerged in this case as well, however their numbers were lower than in the parallel case, and the convergence speed was higher.

Simulations were also performed for model variants, in which agents exchanged opinions on multiple topics at the same time. For these models two alternatives exist: 1) to introduce a separate degree of influence for each topic, or 2) to use one degree of influence for all topics. Also clusters can be considered for each topic separately or by introducing a similarity measure of the agents combining all the topics (e.g., based on the Euclidean distance). In all these cases small, medium and large clusters emerged. However, the convergence speed was significantly slower than in the case with one topic.

Acknowledgement. One of the authors was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Program of the Ministry of Economic Affairs.

References

[1] Axelrod, R.: The Dissemination of Culture: A Model with Local Convergence and Global Polarization. Journal of Conflict Resolution 41, 2023–2226 (1997)

[2] Blondel, V.D., Hendrickx, J.M., Tsitsiklis, J.N.: On Krause's consensus formation model with state-dependent connectivity. IEEE Trans. on Automatic Control 54(11), 2506–2517 (2009)

[3] Byrne, D.: The attraction hypothesis: Do similar attitudes affect anything? Journal of Personality and Social Psychology 51(6), 1167–1170 (1986)

[4] Deffuant, G., Neau, D., Amblard, F., Weisbuch, G.: Mixing beliefs among interacting agents. Adv. Complex Syst. 3, 87–98 (2000)

[5] French, J.R.: A Formal Theory of Social Power. Irvington Publishers (1993)

[6] Golbeck, J.: Computing and Applying Trust in Web-based Social Networks. PhD Thesis, University of Maryland (2005)

[7] Granovetter, M.: Threshold Models of Collective Behavior. American Journal of Sociology 83(6), 1420–1443 (1978)

[8] Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence: models, analysis and simulation. J. of Artificial Societies and Social Simulation 5, 3 (2002)

[9] Hoogendoorn, M., Treur, J., van der Wal, C.N., van Wissen, A.: An Agent-Based Model for the Interplay of Information and Emotion in Social Diffusion. In: Proceedings of IAT 2010, pp. 439–444. IEEE Computer Society Press (2010)

[10] Lewin, K.: Group Decision and Social Change. Holt and Winston, New York (1958)

[11] McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 27, 415–444 (2001)

[12] Macy, M., Kitts, J.A., Flache, A.: Polarization in Dynamic Networks: A Hopfield Model of Emergent Structure. In: Dynamic Social Network Modeling and Analysis, pp. 162–173. National Academies Press, Washington, DC (2003)

[13] Muchnik, L., Itzhack, R., Solomon, S., Louzoun, Y.: Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. Phys. Rev. E 76, 016106 (2007)

[14] Parunak, H.V.D., Belding, T.C., Hilscher, R., Brueckner, S.: Modeling and Managing Collective Cognitive Convergence. In: Proceedings of AAMAS 2008, pp. 1505–1508. ACM Press (2008)

[15] Reingold, O.: Undirected connectivity in log-space. Journal of the ACM 55(4), 17–24 (2008)

[16] Sharpanskykh, A.: Managing the Complexity of Large-Scale Agent-Based Social Diffusion Models with Different Network Topologies. In: Anthony, P., Ishizuka, M., Lukose, D. (eds.) PRICAI 2012. LNCS, vol. 7458, pp. 540–551. Springer, Heidelberg (2012)

[17] Sharpanskykh, A., Treur, J.: Group Abstraction for Large-Scale Agent-Based Social Diffusion Models. In: Zhan, J., et al. (eds.) Proceedings of the Third International Conference on Social Computing, SocialCom 2011, pp. 830–837. IEEE Computer Society Press (2011)

Abstraction of Social Contagion Models with Dynamic Networks

Alexei Sharpanskykh

Delft University of Technology, Faculty of Aerospace Engineering Kluyverweg 1,
2629 HS Delft, The Netherlands
o.a.sharpanskykh@tudelft.nl

Abstract. Social contagion models describe an evolution of states of individual agents under influence of their neighbouring agents by mutual contagion of these states. Although the behaviour of individual agents in such models is often simple, global dynamics that emerge from interaction of a large number of agents is non-trivial. In this paper abstraction techniques are proposed that allow approximating with a high accuracy global patterns of behaviour emerging in social contagion models with underlying dynamic networks. Furthermore, these techniques improve substantially the computational efficiency of social contagion models. In particular, they allow a 6 times speed up of the simulation of the model described in the paper.

Keywords: model abstraction, social contagion models, dynamic networks, social decision making.

1 Introduction

Social contagion models are often used to study social processes such as opinion formation, spread of diseases and innovation [2, 3, 5-7, 9, 10]. These models describe an evolution of states of individual agents under influence of their neighbouring agents by mutual contagion of these states. The agents in such models are connected in networks. In some models the links in networks are constant [5, 10], whereas in others the links have varying strength [3, 6, 11]. In [8, 11] it is argued that for realistic modelling of social processes the strength of the links should change gradually, rather than in an abrupt step-like manner. This conjecture is realized in a social contagion model for social decision making with an underlying network of agents with variable link strengths, proposed in [11]. In this paper we take this model as a starting point for further analysis of global properties emerging in the model. The strength of the links in the model reflects the degree of influence of one agent to another. The dynamics of the links is defined in the model based on the following principle: the closer the opinions of the interacting agents, the higher the mutual degrees of influence of the agents. This assumption underlies most of the existing models of social influence [5, 6, 7, 9] and is supported by [4].

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 51–61, 2013.

An important feature of the model is that for certain ranges of parameter values clusters of agents emerge that are isolated from each other. A *cluster* is a set of connected agents (i.e., a connected graph) with the same states (e.g., opinions). As proven in [11], the states of the agents (e.g., opinions) in each cluster converge over time to an equilibrium, which becomes the state of the cluster. The formation and dynamics of clusters in the model with different network types are extensively investigated by simulation and analytically in the companion paper [11]. In this paper we focus on model abstraction.

Model abstraction serves two purposes. On the one hand, it allows approximating emerging patterns of behaviour at the level of the whole population of agents. In particular, in the paper it is demonstrated how aggregated opinion states of clusters of agents can be determined by model abstraction. On the other hand, by model abstraction the computational efficiency of the model can be substantially improved, which is essential for running large-scale agent-based simulations. Specifically, by using model abstraction an approximate form of simulation is obtained, in which clusters of agents are considered as single entities replacing a large number of interacting agents. The proposed model abstraction procedure allows a 6 times speed up of the simulation of the model described in the paper.

The paper is organized as follows. In Section 2 a social contagion model for social decision making with an underlying dynamic network is described. A model abstraction procedure is proposed and illustrated for this model in Section 3. Section 4 concludes the paper.

2 The Model

The model describes decision making by agents in a group as a process of social contagion. The opinion $q_{s,i}$ of an agent *i* for a decision option *s* is expressed by a real number in the range [*0, 1*], reflecting the degree of the agent's support for the option. For each option each agent communicates its opinion to other agents. Agents communicate only with those agents to which they are connected in a social network. In this study two network topologies are considered:

- a *scale-free network topology*: a connected graph with the property that the number of links originating from a given node representing an agent has a power law distribution. In such networks the majority of the agents have one or two links, but a few agents have a large number of links;
- a *random network topology*: a graph, in which links between nodes occur at random. Only connected graphs are considered in this study.

To compare the dynamics in both types of networks, the networks used in this study were generated with 100 and 1000 agents and the same average node degree equal to 4.5. This value is close to the average node degree of real social networks.

It is assumed that the agents are able to both communicate and receive opinions to/from the agents, to which they are connected (i.e., the links are bidirectional). Furthermore, a weight $\gamma_{i,j} \in [0,1]$, indicating the degree of influence of agent *i* on agent *j,* is associated with each link for each direction of interaction. This weight determines to which extent the opinion of agent *i* is taken into account in the update of the opinion of agent *j* for each option. These weights may or may not be symmetric.

It is assumed that the agents interact with each other synchronously, i.e., at the same time (parallel interaction mode). For a quantitative comparison of the dynamics of social contagion models with the parallel interaction mode with models with the sequential interaction mode, please refer to [10].

In the parallel mode, the opinion states of the agents are updated at the same time point *t* as follows:

$$q_{s,i}(t+\Delta t) = q_{s,i}(t)+ \eta_i\ \delta_{s,i}(t)\Delta t \tag{1}$$

Here η_i is an agent-dependent parameter within the range $[0,1]$, which determines how fast the agent adjusts to the opinion of other agents, and

$$\delta_{s,i}(t) = \sum_{j\in AG} \gamma_{j,i}(t)(q_{s,j}(t)- q_{s,i}(t))/\sum_{j\in AG} \gamma_{j,i}(t)$$

is the amount of change of the agent *i*'s opinion; *AG* is the set of all agent names.

The normalization by $\sum_{j\in AG} \gamma_{j,i}(t)$ has the effect that the agent balances by a relative comparison its own self-influence $\gamma_{i,i}(t)$ (i.e., self-assurance that its own opinion is correct) with the influences of other agents.

The degrees of influence $\gamma_{i,j}$ also change over time based on the principle: the closer the opinions of the interacting agents, the higher the mutual degrees of influence of the agents will become. This dynamic principle may be formalised by different functions as follows:

$$\gamma_{i,j}(t+\Delta t)= \gamma_{i,j}(t) + f_{i,j}(\gamma_{i,j}(t),\ q_{s,i}(t),q_{s,j}(t))\ \Delta t \tag{2}$$

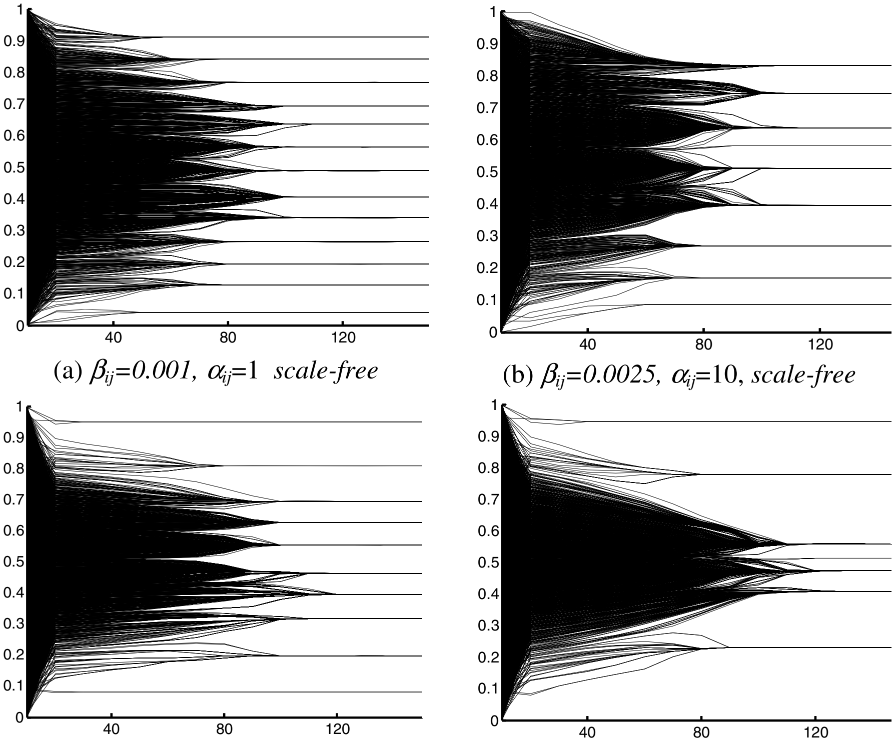
where for function $f_{i,j}(X, Y, Z)$ the main example used is:

$$f_{i,j}(X, Y, Z) = \mathrm{Pos}(\alpha_{ij}\ (\beta_{ij} - (Y-Z)^2))(1-X)\ - \mathrm{Pos}(-\alpha_{ij}\ (\beta_{ij} - (Y-Z)^2))X \tag{3}$$

with $\mathrm{Pos}(x) = (|x|+x)/2$, α_{ij} is a speed parameter and β_{ij} is a threshold parameter.

For more details about the model and its properties we refer to [11]. In this paper we provide only some examples of the simulation results (see Fig. 1). In [11] it is proven that at most $1 + 1/\sqrt{\beta_{ij}}$ clusters can be formed. Furthermore, the higher the value of α_{ij}, the more clusters emerge, with the limit $1 + 1/\sqrt{\beta_{ij}}$.

As proven in [11], the states of the agents (e.g., opinions) in each cluster converge over time to an equilibrium, which becomes the state value of the cluster. The experiments showed that the random networks converge to equilibrium states slower than the scale-free networks. Partially this can be explained by the difference in node degree distributions in these networks: There is a little variation in the degree of the nodes in the random networks. In contrast to the random networks, hubs with a high node degree in the scale-free networks facilitate intensive local interaction in each group of agents gathered around these hubs, and thus speed up the convergence of the whole network. In all the simulation experiments it was observed that the links with $\gamma_{i,j}(0) > 0$ between the agents within a cluster became stronger over time, and the corresponding degrees of influence $\gamma_{i,j}(t)$ tended to *1.* At the same time, the strength of the links with $\gamma_{i,j}(0) > 0$ between the agents in different clusters degraded, and the corresponding degrees of influence $\gamma_{i,j}(t)$ tended to *0.* Furthermore, the degrees of influence, which tend to *0,* stabilize faster than the degrees of influence tending to *1.* Moreover, whereas degrees of influence tending to *0* are in most cases monotonically nonincreasing, degrees of influence tending to *1* often show a time-varying behaviour: first decreasing and then rapidly growing.



(a) $\beta_{ij}=0.001$, $\alpha_{ij}=1$ *scale-free*

(b) $\beta_{ij}=0.0025$, $\alpha_{ij}=10$, *scale-free*

(c) $\beta_{ij}=0.001$, $\alpha_{ij}=1$ *random*

(d) $\beta_{ij}=0.0025$, $\alpha_{ij}=10$, *random*

Fig. 1. The change of the opinions $q_{s,i}(t)$ in the model with 100 agents with scale free and random networks. The horizontal axis is time (s) and the vertical axis is the degree of support for option *s*

These findings were used for defining mechanisms of model abstraction proposed in the next Section 3.

3 Model Abstraction

In the model introduced in Section 2, after some time each group of agents in an emerging cluster can be considered as a single entity. This forms a basis for abstraction of the agent-based model into a population-based model. Such an abstraction process can be performed in two steps:

Step 1: Identification of clusters of agents.

Step 2: Approximation of the equilibrium opinion states of the identified clusters.

First, in section 3.1 step 1 is considered. Then, in section 3.2 step 2 is described.

3.1 Identification of Clusters

To identify clusters of agents, it needs to be determined, which degrees of influence of the agents tend to *1* and which ones tend to *0*. After that, a standard algorithm for identification of connected components in a graph (e.g., based on breadth-first or depth-first search) can be applied to isolate the clusters.

The prediction of the values of the degrees of influence is based on observations made in Section 2 concerning $\gamma_{i,j}(t)$ with $\gamma_{i,j}(0) > 0$:

- if agents *i* and *j* belong to the same cluster, then $lim_{t\rightarrow\infty}\gamma_{i,j}(t)=1$ and $lim_{t\rightarrow\infty}\gamma_{j,i}(t)=1$
- if agents *i* and *j* belong to different clusters, then $lim_{t\rightarrow\infty}\gamma_{i,j}(t)=0$ and $lim_{t\rightarrow\infty}\gamma_{j,i}(t)=0$

Furthermore, it was observed that $\gamma_{i,j}(t)$ tending to *0* are often monotonically nonincreasing, whereas $\gamma_{i,j}(t)$ tending to *1* often first decrease and then grow rapidly. As the dynamics of decision making produced by the model is highly non-linear and erratic, it is not feasible to make early analytical predictions for the values of the degrees of influence. A heuristic approach is used instead.

The approach is based on identifying in each simulation the earliest time point after which it could be predicted with a high confidence whether each $\gamma_{i,j}(t)$ with $\gamma_{i,j}(0) > 0$ tends to *1* or to *0*. This time point is estimated by performing numerous simulations and determining the earliest time points $tp_{i,j}$ for every $\gamma_{i,j}(t)$ after which the degrees of influence tending to *0* are nonincreasing and the degrees of influence tending to *1* are nondecreasing, formally:

$$\text{For } \gamma_{i,j}(t)\rightarrow 0\text{:}\ \{tp_{i,j} \mid \exists t1,t2\ t2 > t1 \,\&\, t1 \geq tp_{i,j} \,\&\, \gamma_{i,j}(t2) \leq \gamma_{i,j}(t1) \,\&\, (tp_{i,j}=0 \mid\mid tp_{i,j}>0 \,\&\, \gamma_{i,j}(tp_{i,j}) > \gamma_{i,j}(tp_{i,j}-\Delta t))\}$$

$$\text{For } \gamma_{i,j}(t)\rightarrow 1\text{:}\ \{tp_{i,j} \mid \exists t1,t2\ t2 > t1 \,\&\, t1 \geq tp_{i,j} \,\&\, \gamma_{i,j}(t2) \geq \gamma_{i,j}(t1) \,\&\, (tp_{i,j} = 0 \mid\mid tp_{i,j}>0 \,\&\, \gamma_{i,j}(tp_{i,j}) < \gamma_{i,j}(tp_{i,j}-\Delta t))\}$$

In Table 1 the distribution (in %) of time points $tp_{i,j}$ in time intervals is provided, which is obtained by performing 1000 simulation for each setting of the model from Section 2. In each simulation a new network was generated with the uniformly distributed parameters: $\beta_{ij} \in [0.001, 0.09]$ and $\alpha_{ij} \in [1, 20]$. For the justification of the choice for the parameter ranges we refer to [11].

The values to the left and right from symbol ‘|’ in each data cell in the table indicate correspondingly the percentage of time points $tp_{i,j}$ for the degrees of influence tending to *1* and the percentage of time points $tp_{i,j}$ for the degrees of influence tending to *0*.

As can be seen from the table, starting already from time point 5 more than *95%* of the degrees of influence of the agents were either nonincreasing (in case of $\gamma_{i,j}(t)\rightarrow 0$) or nondecreasing (in case of $\gamma_{i,j}(t)\rightarrow 1$). The number of agents does not influence the distribution of $tp_{i,j}$ substantially. According to the obtained results, the degrees of influence tending to *0* become nonincreasing on average earlier (at time point 35 at latest) than the degrees of influence tending to *1* become nondecreasing (this may even occur after time point 65). This is also in accordance with the simulation results discussed in Section 2: the degrees of influence tending to *0* grow rarely after the initial stabilization period (up to time point 15). Based on the obtained simulation results the following heuristic rules for determining the limit values of the degrees of influence ($\gamma_{i,j}$*) are specified:

Rule 1: If at control point *t*

$(\gamma_{i,j}(t) - \gamma_{i,j}(t-\Delta t))/\Delta t \geq 0$ and $\gamma_{i,j}(t) > 0.5$, Then $\gamma_{i,j}* = 1$

Rule 2: If at control point *t*

$(\gamma_{i,j}(t) - \gamma_{i,j}(t-\Delta t))/\Delta t < 0$ and $\gamma_{i,j}(t) < 0.1$, Then $\gamma_{i,j}* = 0$

Table 1. The distribution (in %) in time intervals of the earliest time points after which the degrees of influence tending to 0 are nonincreasing and the degrees of influence tending to 1 are nondecreasing; the results are based on 1000 simulation trials

Setting **Time interval**	**100 agents, scale free**	**1000 agents, scale free**	**100 agents, random**	**1000 agents, random**
[0, 5]	96.7 | 100	96.7 | 98.5	95.7 | 100	95.4 | 100
(5, 15]	2.71 | 0	2.6 | 1.12	3.2 | 0	3.43 | 0
(15, 25]	0.29 | 0	0.33 | 0	0.5 | 0	0.5 | 0
(25, 35]	0.13 | 0	0.13 | 0.13	0.15 | 0	0.19 | 0
(35, 45]	0.07 | 0	0.07 | 0	0.09 | 0	0.1 | 0
(45, 55]	0.05 | 0	0.04 | 0	0.07 | 0	0.07 | 0
(55, 65]	0.03 | 0	0.004 | 0	0.05 | 0	0.04 | 0
> 65	0.07 | 0	0.11 | 0	0.24 | 0	0.27 | 0

The first control time point is set at 5, as according to Table 1 for more than 95% of the degrees of influence the correct prediction for the limit values could have been made at this point. The last control point is set at 65, and the distance between two control points is 10. The constraints $\gamma_{i,j}(t) > 0.5$ and $\gamma_{i,j}(t) < 0.1$ were identified statistically using the obtained simulation results by determining minimum (maximum) $\gamma_{i,j}(t)$ values for the degrees of influence tending to *1* (to *0*) at the control points.

Based on the identified heuristic rules the algorithm below is defined. At each control point the number of limit values $\gamma_{i,j}$* is counted (variable *count*) that can be initialized according to the rules above (lines 4-16). If this number exceeds *90%* of all $\gamma_{i,j}$ with$\gamma_{i,j}(0) >0$, the algorithm terminates and the identification of clusters (i.e., connected components in a graph) starts (lines 17,18). Otherwise, the algorithm proceeds with checking for the following control time point. Thus, the time point when clustering starts varies between simulations; it depends on the parameter and initial value settings of the model. Note that for those $\gamma_{i,j}$ for which the heuristic rules do not hold, weaker versions of the rules are applied without constraints $\gamma_{i,j}(t) > 0.5$ and $\gamma_{i,j}(t) < 0.1$. The obtained $\gamma_{i,j}$* are not taken into account in variable *count*, as they are considered to be unreliable. After the algorithm terminates, automated clustering is performed by breadth-first search.

The algorithms for determining limit values and clustering were implemented in Matlab. Again, 1000 simulations were performed for each setting from Table 1 and the number of agents was determined that were assigned to a wrong cluster; the errors (in % of agents) averaged over all simulation trials for each simulation setting are provided in Fig. 2. The clustering error of the scale-free networks is in average lower than the error of the random networks (Fig. 2). This can partially be explained by a higher convergence rate of the scale-free networks observed in the study. Furthermore, in contrast to the random networks, the clustering error of the scale-free networks does not depend significantly on the number of agents.

Algorithm 1. Determining limit values for the degrees of influence

```
m_count ← ( the number of $\gamma_{i,j}(0) > 0$)
for t=5 to 65 step 10 do
    count ← 0
    for all agents i, j do
        if $\gamma_{i,j}(0)>0$
            if $(\gamma_{i,j}(t) - \gamma_{i,j}(t-\Delta t))/\Delta t \geq 0$
                $\gamma_{i,j}^* \leftarrow 1$
                if $\gamma_{i,j}(t) > 0.5$
                    count ← count+1 endif
            else if $(\gamma_{i,j}(t) - \gamma_{i,j}(t-\Delta t))/\Delta t < 0$
                $\gamma_{i,j}^* \leftarrow 0$
                if $\gamma_{i,j}(t) < 0.1$
                    count ← count+1 endif
            endif
        endif
    end for
    if count > 0.9* m_count
        start clustering; exit endif
end for
start clustering
```

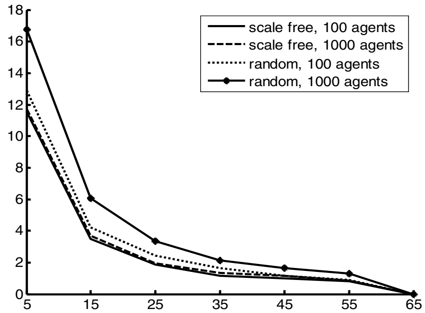


Fig. 2. The percentage of agents assigned to wrong clusters in all simulation settings. The horizontal axis is time and the vertical axis is the % of agents from the whole population.

After the agents have been divided into clusters, the values of the opinion states of the clusters are determined, which is step 2 of the abstraction procedure discussed in the following section 3.2.

3.2 Approximation of States of the Clusters

For the approximation of the states of each cluster the abstraction method based on invariant calculation from [10] is used. This method demonstrated the best precision and computational efficiency in comparison with other methods. In our abstraction procedure this method is applied for each cluster separately at the time point, when clustering of agents is performed. A concise summary of the invariant-based abstraction method is provided below. For more details please refer to [10].

For given initial values $q_{s,i}(0)$ for $i = 1, ..,$ only one of the possible equilibria with equal values will be actually reached. How this equilibrium value depends on the initial values can be described by an *invariant*: an expression in terms of the $q_{s,i}(t)$ for $i = 1, ..,$ that does not change over time. In this case an invariant inv_s as a weighted sum $inv_s = inv_s(t) = \sum_i \lambda_{s,i}\, q_{s,i}(t)$ can be obtained where the weights $\lambda_{s,i}$ depend on the coefficients η_i and $\gamma_{j,i}$ (and not on initial values). These weights can be taken (normalised) with $\sum_i \lambda_{s,i} = 1$, so that when all $q_{s,i}(t) = 1$ for all i, also the invariant is 1. Below it will be discussed how an invariant can be found.

An invariant can be considered as a kind of preservation law for the (collective) support for option s in the considered group. By internal (intragroup) interactions this collective support for s can be redistributed over persons in the group, but this does not change the collective amount. During time intervals where no external interaction is coming in, the group's collective support for s will not change. This provides an interesting use of the invariant: as a means of abstraction from the internal processes by using a descriptor at the group level.

The weights $\lambda_{s,i}$ for the invariant inv_s can be determined from the difference equations:

$$\sum_m \sum_{i\neq m} \lambda_{s,i}\, \eta_i\, \gamma_{m,i}\, q_{s,m}(t) = \sum_m \lambda_{s,m} \eta_m\, q_{s,m}(t)$$

One way to satisfy this is by taking the coefficients of $q_{s,m}(t)$ in the above expression on both sides equal; this provides the following set of linear equations for the $\lambda_{s,i}$ for all m:

$$\sum_{i\neq m} (\,\eta_i\, \gamma_{m,i}/\eta_m\ ) + 1\ )\ \lambda_{s,i} = 1$$

Thus a system of linear equations $\sum_{i\neq m} \mu_{m.i}\, \lambda_{s,i} = 1$ is found with coefficients $\mu_{m.i} = (\eta_i\, \gamma_{m,i}/\eta_m) + 1 \geq 1.$

This system can be described in matrix form as $\mathbf{A}\boldsymbol{\lambda}_s = \mathbf{1},$ where $\mathbf{1}$ is the vector with all components 1, $\boldsymbol{\lambda}_s = (\lambda_{s,1},...)$, and $\mathbf{A}$ is a square matrix with only zeros at the diagonal and all other entries ≥ 1 (expressed in η_i and $\gamma_{m,i}$). When it is assumed that the determinant $\mathbf{det(A)} \neq 0$, then this system has a unique solution. Indeed, for the general case this condition is fulfilled, and the weights $\lambda_{s,i}$ of the invariant can be obtained as a solution.

To illustrate the abstraction procedure, 1000 simulations were performed. In each simulation a new network was generated. In each simulation clusters of agents were identified using the approach from Section 3.1. Then, the invariant-based method was applied to approximate the limit values of the opinion states of the clusters. The averaged approximation error per agent for each simulation was defined as:

$$err = \sum_{i\in AGENT} |q^*_{s,i} - q_{s,i}(end\_time)|/\ |AGENT|,$$

where $AGENT$ is the set of all agents, end_time is the simulation time, $q^*_{s,i}$ is the predicted opinion state value for option s of the cluster, to which agent i belongs; it is determined by the invariant-based abstraction.

Furthermore, the time was measured for each simulation and the simulation efficiency gain was determined as the ratio of the simulation time of the model with abstraction to the simulation time of the original model without abstraction. In the model with abstraction the clusters of agents were replaced by single super-agents with the cluster states determined by the proposed abstraction approach.

The obtained results averaged over 1000 simulations are provided in Table 2. As can be seen from the table, the number of agents influences more significantly the approximation error in the random networks than in the scale-free networks. As was shown in [11], agent states propagate slowly through a random network with a low average node degree. The simulation showed that with the increase of the number of agents, the state propagation speed and the convergence speed of such networks decreases. This causes an increase of the error of the equilibrium-based abstraction employed in the paper.

Table 2. The approximation error and the efficiency gain of the abstracted model evaluated for the simulation settings. The values in brackets for the approximation error are variances.

Setting **Result**	***100 agents, scale free***	***1000 agents, scale free***	***100 agents, random***	***1000 agents, random***
Approx. error (err)	$13*10^{-4}$ $(2*10^{-4})$	$14*10^{-4}$ $(3*10^{-4})$	$27*10^{-4}$ $(3*10^{-4})$	$159*10^{-4}$ $(5*10^{-4})$
Simulation time, s	1.42	5.12	1.45	5.19
Efficiency gain	5.7	6	6.2	6

Furthermore, the approximation error is lower for the scale-free networks than for the random networks. This could also be partially explained by the difference in the network convergence speeds: the scale-free networks reach an equilibrium state faster than the random networks, and thus the equilibrium-based abstraction is more precise for the scale-free networks. The simulation efficiency gain is approximately the same (around 6) for both network types and for different numbers of agents.

Note that the values of the states for the clusters were determined at the same time point in interval [5, 65], when the clustering was performed. At this time point some degrees of influence may not have reached equilibrium yet. Thus, weak connections between clusters might still have existed, contributing to the approximation error. This issue may partially be addressed by considering clusters as aggregated agents interacting with each other, as was done in [10]. However, for some networks such interactions may be numerous, and thus would have a large negative effect on the computational efficiency. A precise performance analysis for such an aggregated form of modelling is not considered in this paper and will be addressed in the future.

4 Conclusions

In the paper an approach for abstraction of social contagion models with underlying dynamic networks is proposed. On the one hand, it enables analysis of emergent properties of a model by approximating with a high accuracy its global dynamics at the population level. On the other hand, the approach increases the computational efficiency of agent-based simulations of social contagion models (6 times for the model considered in the paper).

Several techniques for abstraction of models based on hybrid automata and differential equations [1] currently exist. However, such approaches can be efficiently applied for systems described by sparse matrixes. Social contagion models represent tightly connected systems, which do not allow a significant reduction of the state space using such techniques. In particular, a previous study showed that common model reduction techniques such as balanced truncation [1] do not allow decreasing the rank of the matrix describing the model from Section 2. In [10] abstraction of social contagion models with static networks is considered. However, the emergent behaviour of social contagion models with static networks is very different from the behaviour of models with dynamic networks (e.g., no emergence of clusters in static networks). Thus, the abstraction approach proposed in this paper differs substantially from the approach taken in [10].

In the paper simulation results were presented for the networks with the average node degree 4.5. However, also experiments for more dense networks were performed. In general, the higher the density of a network, the higher its convergence speed to an equilibrium state, and thus, the less the approximation error by the equilibrium-based abstraction. Furthermore, in average the higher the density of the network, the less number of clusters are formed.

Acknowledgement. The author was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Program of the Ministry of Economic Affairs.

References

1. Antoulas, A.C., Sorensen, D.C.: Approximation of large-scale dynamical systems: An overview. Int. J. Appl. Math. Comp. Sci. 11, 1093–1121 (2001)
2. Axelrod, R.: The Dissemination of Culture: A Model with Local Convergence and Global Polarization. Journal of Conflict Resolution 41, 203–226
3. Blondel, V.D., Hendrickx, J.M., Tsitsiklis, J.N.: On Krause's consensus formation model with state-dependent connectivity. IEEE Trans. on Automatic Control 54(11), 2506–2517 (2009)
4. Byrne, D.: The attraction hypothesis: Do similar attitudes affect anything? Journal of Personality and Social Psychology 51(6), 1167–1170 (1986)
5. Deffuant, G., Neau, D., Amblard, F., Weisbuch, G.: Mixing beliefs among interacting agents. Adv. Complex Syst. 3, 87–98 (2000)
6. Granovetter, M.: Threshold Models of Collective Behavior. American Journal of Sociology 83(6), 1420–1443 (1978)
7. Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence: models, analysis and simulation. J. of Artificial Societies and Social Simulation 5(3) (2002)
8. Lewin, K.: Group Decision and Social Change. Holt, Rinehart and Winston, New York (1958)
9. Parunak, H.V.D., Belding, T.C., Hilscher, R., Brueckner, S.: Modeling and Managing Collective Cognitive Convergence. In: Proceedings of AAMAS 2008, pp. 1505–1508. ACM Press (2008)

10. Sharpanskykh, A., Treur, J.: Group Abstraction for Large-Scale Agent-Based Social Diffusion Models. In: Zhan, J., et al. (eds.) Proceedings of the Third International Conference on Social Computing, SocialCom 2011, pp. 830–837. IEEE Computer Society Press (2011)
11. Sharpanskykh, A., Treur, J.: Modelling and Analysis of Social Contagion Processes with Dynamic Networks. In: Bădică, C., Nguyen, N.T., Brezovan, M. (eds.) ICCCI 2013. LNCS (LNAI), vol. 8083, pp. 40–50. Springer, Heidelberg (2013)

Classsourcing: Crowd-Based Validation of Question-Answer Learning Objects

Jakub Šimko, Marián Šimko, Mária Bieliková, Jakub Ševcech, and Roman Burger

Institute of Informatics and Software Engineering,
Slovak University of Technology in Bratislava,
Ilkovičova, 842 16 Bratislava, Slovakia
{name.surname}@fiit.stuba.sk

Abstract. The Web 2.0 principles reflect into learning domain and provide means for interactivity and collaboration. Student activities during learning in this environment can be utilized to gather data usable for learning corpora enrichment. It is now a research issue to examine, to what extent the student crowd is reliable in delivering useful artifacts and to bring in suitable tools to enable this. In this paper we present a method for crowd-based validation of question-answer learning objects involving interactive exercise for learners. The method utilizes students' correctness estimations of answers provided by other students during learning. We show that aggregate student crowd estimations are to big extent comparable to teacher's evaluations of provided answers.

Keywords: crowdsourcing, education, question, answer evaluation, technology enhanced learning.

1 Introduction

In this paper, we deal with a phenomenon of *crowdsourcing within learning environments*. Web 2.0-induced paradigm shift, reflected in learning, triggered collaboration and availability of learning content to the masses. Besides taking advantage of an educational system, by consuming provided content, learners produce a significant number of user-generated data that constitute a tremendous potential for further improvement of learning process: not only they leave footprints in system usage logs, they also actively contribute by adding, annotating or modifying learning materials. Some user activities in educational systems may therefore be managed to produce useful artifacts (as by-products of learning), which can be utilized to supplement learning content (e.g., student-created explanatory notes attached to original learning material). Such artifacts may either be *learning objects*[1] themselves or can be useful otherwise, e.g. as metadata describing the learning objects. Research possibilities are open here, calling for devising learning activities which (besides educational effects) lead to creation of useful artifacts. One of the key aspects such approaches must

[1] We adopt a broader definition by IEEE, which defines a learning object as "any entity, digital or non-digital, that may be used for learning, education or training" [5].

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 62–71, 2013.

consider is ensuring the quality of created artifacts, even if we involve relatively inexperienced students in the process.

In this context, a particular issue we focus on is the *acquisition of correctness information on student-created answers to exercise- and exam-like questions*. During learning exercises, students often interact with questions and answers (related to course domain). If a student answers a question during a learning course, he expects feedback on his answer correctness. Usually, this feedback is provided by the teacher (e.g. in-class learning sessions). However, the teacher's responses are not always available during *online* sessions. Instead, the online learning application comes to place, giving an automated feedback. This, however, can be done only for certain types of exercises (questions), such as multiple choice answers. The free text answers (which are sometimes the only suitable option from didactical perspective) cannot be evaluated sufficiently by today's automated means. The evaluation of free text answers can only be done by a human (and this burdens the teacher).

We present a method, which comprises learning from existing questions and answers (either correct or wrong) and at the same time, involves students into evaluation of correctness of answers for their peers during learning sessions. It is based on an interactive exercise, in which student is confronted with a question and existing student-created answer (further referred to as *question-answer learning object* or QALO). For example, in a software engineering course the question “What is the purpose of the feasibility study” may be answered with “To determine, if a problem is worth solving”. The task for the student is to evaluate correctness of this answer. After the student does so, he receives feedback based on the aggregate crowd correctness evaluation based on previous evaluations provided by other students. By deploying this method into online learning environment, following effects are achieved:

1. The students are able to exercise their course domain knowledge autonomously.
2. The correctness estimations of individual students (to the same QALO) constitute the *crowd correctness estimation*, which, as we show in our experiments, approximates the true answer correctness (according to teachers).

The outcoming crowd correctness estimation is used in the exercise itself, but may also be used for other purposes, for example to give feedback to the original author of the answer (lifting some burden from the teacher). This enables the use of *question answering* exercise, where the question answerers are fed back by their colleagues over time. However, as the input for our method, any QALOs may be used, for example results of exams or homeworks, which may be reused this way.

We devised our method and deployed it within our Adaptive Learning Framework (ALEF) [12] in the Principles of Software Engineering course. Through it, we collected usage data for the exercise, computed crowd answer correctness estimations and evaluated their accuracy against grand truth provided by teachers. We show that student-generated data obtained collectively with no prior pedagogical knowledge can to big extent substitute evaluations provided by a teacher. As a result, students can support each other during their learning sessions and teacher's efforts in learning corpora creation and feedback provision is reduced.

2 Related Work

We utilize principles of crowdsourcing, a paradigm which uses human computation for substituting machines in performing tasks hard or impossible to automate. Crowdsourcing often involves lay users, which raises the question of quality of the solutions they provide. This issue is in general solved by redundant task solving, collaborative filtering, consensus, peer-reviews etc. [10]. Our method bears a similarity with principles of community question answering systems, such as Yahoo! Answers[2], which are a subgroup of crowdsourcing approaches. In these systems, answers to questions are acquired from some crowd members and are secondarily evaluated by other crowd members. The best answers eventually emerge. To reach quality answers, some works use automated analysis of the answer texts [2], other focus more on voting and filtering [3]. Focus is also given to predicting answerer level of expertise [1]. Community question-answering systems often collect answers as solutions of problems in specific domains. With our work, we aim to explore, whether their principles could be used in a didactical scenario, where users-reviewers do not seek answers to their problems but learn and test their knowledge instead.

The crowdsourcing principles are in general, relevant for the learning domain. With the emergence of technology enhanced learning, we witness paradigm shift in learning, especially when considering web-based learning environments. Benefiting from concepts introduced by Web 2.0 and moving towards genuine Read-Write Web [9], a student becomes more autonomous and less dependent on the teacher. She is provided with more competences since she can tag, rate, share and collaborate during learning. She becomes an active contributor rather than a passive consumer of learning content [4]. The activity of a learner is "boosted" not only in relation with an educational system, but also when considering collaboration during learning [11,13].

The distributed nature of the Web allows to connect and virtually gather various learners in a convenient way – anytime, anywhere. The students can therefore be viewed as potentially useful crowd force. To participate in the creative process, they can be motivated internally (by their own will to learn) or externally (by course points or gamification). Student activities may often result into new learning materials created intentionally [8, 14], or are utilized to promote existing educational content [6]. Crowd activities are implicitly connected with collaboration or collective intelligence – in either explicit or implicit manner [7].

3 Crowd Validation of Question-Answer Learning Objects

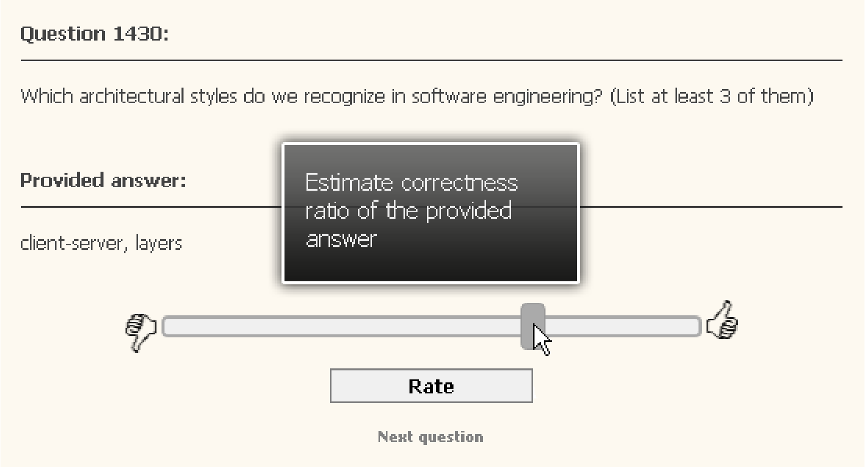
For retrieval of information on correctness of answers within *question-answer learning objects* (QALOs), we present a method consisting of interactive student exercise and a subsequent automated interpretation of student activity within this exercise. Our method retrieves the correctness information via crowdsourcing of the group of students attending a learning course. During her learning sessions, the student pulls,

[2] answers.yahoo.com

reviews and rates QALOs. By rating we mean *estimating correctness* of an answer provided in QALO. After that, she retrieves feedback in form of global QALO correctness estimation computed from estimations provided by other peers (a global "crowd truth"). The QALOs used by our method can be of any origin, provided they are relevant to the student by topic and difficulty. In a common use case, these would be questions and answers used and created during exercises in the learning course.

A typical session with QALO exercise can be described by a repetitive scenario:

1. A student makes a request for QALO (within an educational system). Usually, she does so during her "home" online learning session, either as a "starting" activity (when she wants to discover what to learn next) or "finishing" activity (when she wants to reaffirm her newly gained knowledge). It is generally expected that this scenario especially occurs prior to exams or seminaries, where students might be tested. Thus, more than one student is usually working with the system at a time.
2. The QALO selector picks a suitable QALO for the student. Its aim is to maintain an effective allocation of crowd power and to avoid the situation when many QALOs are rated by insufficient number of students. The QALO selector is described in more detail below.



Question 1430:

Which architectural styles do we recognize in software engineering? (List at least 3 of them)

Provided answer:

client-server, layers

Estimate correctness ratio of the provided answer

Fig. 1. A screenshot of the QALO interface. Using the slider, a student expresses her estimation of the correctness of the answer. After clicking the "Rate" button the estimation is stored and feedback information (the "crowd truth" correctness of the answer) is displayed.

3. The QALO is displayed to the user (see Fig. 1), consisting of a question (e.g., "Which architectural styles do we recognize in software engineering? List at least 3.") and a provided answer (e.g., "client-server, layers"). The student reviews the QALO and decides to what extent the provided answer is correct, i.e., she provides QALO correctness estimation. She expresses this by moving the marker on a slider between two extremes: incorrect (internally represented by 0) and correct (internally represented by 1). In our example, if the correct answer to the question is a list of three items and the provided answer lists only two correct options, user may

move the marker to two thirds (from left) of the slider's width, indicating the correctness estimation of the provided answer.

4. After submitting her estimation, the student is presented with the current global correctness estimation for the answer (the "crowd truth").
5. The global correctness estimation for the QALO gets updated to include student's rating. It is defined as the *average of all individual correctness estimations of this QALO* (a value between 0 and 1).
6. If the QALO received *sufficient feedback* from students, it is excluded from further processing. For the sake of experimentation (in order to acquire uniform data set) we defined this by a constant number of rating actions needed for one QALO. For the practical use though, we would rather use a dynamic metric to determine, whether the current crowd answer can be considered close to definitive. Such metric could take into account, for example, the variance of the crowd answers and exclude a QALO from the process earlier, if the variance drops under certain margin.

Computation of the global correctness for particular answer as estimated by the crowd presents core of our method. As a real values between 0 and 1, the crowd answers may be used "as is" (e.g., for feedback to students), but to give them nominal interpretation, we discretize them further into three possible values: *correct*, *incorrect* or *unknown* using two parameters: t and ε. The t (threshold) splits the correctness interval into two areas designating two possible values: *correct* and *incorrect*. The crowd answer is then determined according to which interval its real value falls. The second parameter (ε) adds a third possible value: *unknown* by inserting an "uncertainty interval" around the t value (rendering the values that fall into it as unknown), resulting in intervals $\langle 0, t-\varepsilon\rangle$, $(t-\varepsilon, t+\varepsilon)$ and $\langle t+\varepsilon, 1\rangle$ for incorrect, unknown and correct estimations, respectively. We have experimented with different values of t and ε to yield the best results.

Our method requires that a certain minimum number of students is attending the course and participates in QALO correctness estimation. In order to acquire valid global correctness estimations, the number of participating students must be equal to number needed in worst-case scenario from the *sufficient feedback per QALO* point of view (in our experiments, defined by a constant). To provide feedback to students for their estimations, the requirement is even lower: units of previous feedback actions are required – the student is always informed, how many of his peers evaluated the QALO before her and can take the feedback with adequate seriousness. Practically, in a scenario when students review potential exam questions, the motivation to interact with the QALO content is high (which was also shown in our experiments) and is therefore no problem to provide feedback in most cases.

It is important to note that our method is independent of the semantics of the QALOs and it is "portable" to any educational course where a sufficient number of QALOs is available. The questions and student answers should also be simple and test small pieces of knowledge for smoother and more controllable process.

Considering the computation of crowd answer as a valid estimation of correct answer, the process of assigning QALOs to students could be completely random if an unlimited crowd power is at hand. However, we expect that in many cases, the number of student ratings required for whole dataset exceeds the available force that the student crowd is willing to offer.

Therefore we devised the *QALO Selector* – a routine for QALO picking, executed upon each QALO request. It aims to complete the evaluation of a particular QALO in a relatively short time by assigning it to students frequently so it receives the sufficient student feedback faster. The basic heuristic to do this is to assign the QALO that has not yet been validated sufficiently but has most of the validations already done. This way, the crowd force is used effectively, leaving only a minimum number of partially validated (i.e., unusable) QALOs. Keeping this "working frame" (a set of partially validated QALOs) narrow is, however, contradictory to other requirements:

1. A QALO must be validated by different students; and
2. A single student has a need for topical diversity or adaptation to her knowledge within QALO she rates.

A student motivation to participate might drop, if she encounters the same question or even QALO. Occasional repeating of the same QALO to the same student in a short time is exploitable in many ways (speeding up the evaluation, testing the student's consistency) but due to the possible loss of motivation, we avoided it, so the student encounters each QALO only once. The same question can be encountered more times, but only after student passes a certain number of other questions.

4 Evaluation: A Real-world Experiment in Class

In order to evaluate our method, we have conducted a real-world experiment in a setting of software engineering course lectured at the Slovak University of Technology in Bratislava. Over the period of two weeks, students were free to pull, read, consider and validate QALOs, which were assembled from the questions from last term's tests and respective answers provided by last term's students. Based on obtained correctness estimations we computed average correctness estimation for each QALO. We compared the results with a gold standard – correctness estimations of QALOs provided by teachers.

Hypothesis. The correctness estimations of answers in question-answer learning objects (QALOs), obtained by student crowd using our method, are the same as teacher's correctness assignments to these QALOs. This, we measure through *accuracy* (i.e., ratio of correct crowd answers to all its answers) and *dropout* (i.e., ratio of cases where crowd reached the *unknown* answer to all cases). We conducted the experiment for multiple settings of parameters t and ε. We expected the optimal t value to be 0.5.

Environment and context. The data collection spanned over two weeks around the mid-term of the software engineering course. The course consists of weekly lectures and exercises and also comprises supplementary online learning materials within the educational system ALEF [12]. The same system provided the platform for our method for this experiment. During the course, students also undertake weekly mini-exams. These exams comprise similar or identical questions as those used in this experiment and therefore we expected a natural interest from the student side to participate in the QALO validation.

Participants. Overall, 142 students (of *Principles of software engineering* course) participated in the experiment (out of 162 students who enrolled in this course). Participation was completely voluntary. We considered no prior knowledge about the domain expertise of the participating students.

Data. We have used 200 questions, each with 20 answers to construct the initial QALO set (thus comprising 4,000 QALOs). The answers in QALOs used were taken from real exams (commenced year earlier), so we could utilize the existing teacher correctness evaluations in the experiment. According to course syllabus, each QALO has been assigned with week, when its respective topics were discussed and students were asked only those already covered by lectures. Overall, 9,939 QALO correctness estimations were collected. 479 QALOs were provided with equal or more than 16 correctness estimations (our threshold for sufficient student feedback; we used threshold equal to 16 at which further estimation would change the crowd answer only marginally in worst case).

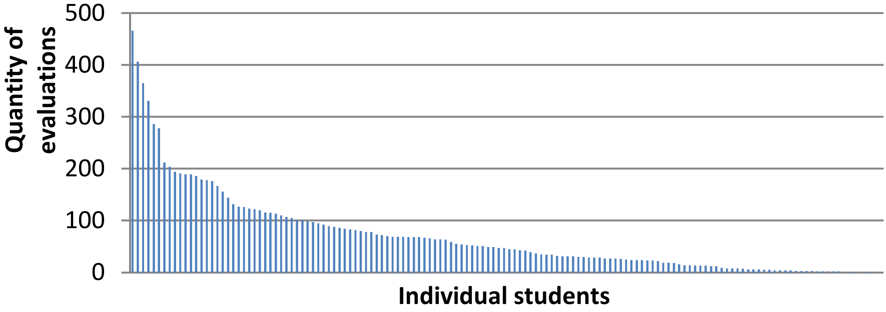


Fig. 2. Quantities of evaluations provided by individual students

In average 70 QALO evaluations were collected per student, however, students greatly differed by quantity of estimations they delivered: while few of them evaluated hundreds of QALOs (top user delivered 466), many solved only units. Such distribution follows power law (see Fig. 2). Especially, the "best performers" can be accounted to different motivations for participants: besides the educational motivation (to learn and to test one's knowledge), students were also motivated by few extra points to their course assessments and also by *gamification* mechanisms (e.g., ranking among other users) present in the used educational system.

Yet more interesting "non-uniformity" we observed in the collected data was the indication of a *tendency of the students to consider incorrect question answers as correct answers*. From all QALOs considered in the experiment, 65 % had answers marked as correct by the teacher. However, 79 % of all student evaluations (those with some tendency, i.e., those not equal to 0.5, which was also the default value) were positive about the correctness (see the histogram in the Fig. 3 which illustrates this phenomenon). This suggests, that students tend to trust the answers created by other students. This also corresponds with our pedagogical experience: when students are unsure or wrong when answering questions, they at least *try* to make their answers *look correct*. Such answers then easily confuse other students (who validate QALOs) and "trick" them into belief, that they are correct.

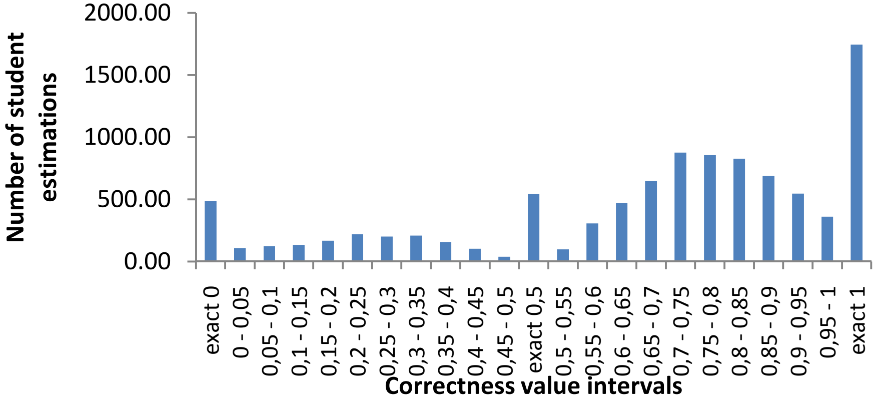


Fig. 3. Distribution of all QALO student correctness estimations according to their values. Many students state extreme estimations and also tend to consider many answers correct.

Results. The overall real correctness estimations were computed for each QALO that received exactly or more than 16 estimations. After discretization (to three possible values, according to parameters t and ε), the estimations were compared to the reference set – answer correctness information assigned by teachers in the last term: correct or incorrect. The Table 1 shows the resulting accuracy of the method (the proportion of QALOs with correctly estimated correctness – both *correct* and *incorrect* – of provided answers in all QALOs) along with the percentage of *unknown* value cases.

Table 1. Method's *accuracy* and *unknown cases percentage* (in parentheses) for different parameter setups (t – correctness threshold, ε – uncertainty factor)

t	ε = 0.0	ε = 0.05	ε = 0.10
0.55	79.60 (0.0)	83.52 (12.44)	86.88 (20.40)
0.60	82.59 (0.0)	86.44 (11.94)	88.97 (27.86)
0.65	**84.58 (0.0)**	**87.06 (15.42)**	**91.55 (29.35)**
0.70	80.10 (0.0)	88.55 (17.41)	88.89 (37.31)
0.75	79.10 (0.0)	79.62 (21.89)	86.92 (46.77)

On the contrary to the initially expected correctness threshold t = 0.5, as best parameter configuration, the correctness threshold t = 0.65 has emerged. With no uncertainty interval ($\varepsilon = 0$), the method was rendered promisingly 84.58 % accurate. With introduction of the uncertainty interval we even see an increased accuracy in crowd's decision, though dropout (unknown cases) percentages are significant too.

We consider the results very promising as reasonably high accuracy of student answer correctness can be obtained via validation performed by students themselves. The accuracy increases over 90 % if approximately 30 % of QALOs are omitted. For our purpose even a higher "loss" is affordable as our primary goal is to support learning corpora enrichment while reducing teacher's efforts and not necessarily to get correct validations for all provided answers.

One could naturally expect the threshold $t = 0.5$ to be optimal. Instead for student crowd a higher value ($t = 0.65$) was observed as better. We account this to the significantly often occurring "trusting student phenomenon" described above, where students validate incorrect answers as correct ones. The ratio of false positive and false negative crowd correctness estimations supports this assumption. With $t = 0.5$ (i.e., the expected "normal conditions"), 91 % of false crowd answers were false positives (i.e., cases when students wrongly stated that an answer is correct).

5 Discussion and Conclusions

We have presented a method for student-crowd-based acquisition of correctness information of answers to questions in the context of learning course. It benefits from collective "wisdom" of a group of lay students. Functioning also as a didactical tool, our method enables to re-use existing question-answer learning objects and to give feedback to answer creators. Since our method is not constrained in terms of the domain of the course, it is portable and applicable to any course, where question-answer learning objects are available. In our experiments, our findings were as follows:

1. In an *implicit collaboration* scenario, while undertaking their learning sessions with our method, students as a crowd are able to validate learning objects with quality comparable to their teachers.
2. An interesting effect that unfavorably skewed the student crowd answers was the "trusting student" phenomenon, where students in significant numbers evaluated incorrect question answers as correct.
3. The interactive character of the QALO correctness evaluation exercise, combined with gamification incentives successfully motivated students to participate – one student evaluated averagely 70 QALOs.

There are also several possible improvements of the base method (focusing solely on estimating correct answer without learning goals that in real case scenario are always present) that are a subject of our future work. First, it is an introduction of smarter identification of sufficient student feedback on the QALO – if the correctness estimates show only a little variance from the start, the QALO might be excluded from the process earlier and spare some student actions for other QALOs. Secondly, to further increase the speed and accuracy of our method for validating answers by student crowd we want to introduce a model of individual influence of students in average correctness estimation. The influence would source from student's level of knowledge in the course domain and would be acquired by means common in academic courses such as previous exam results.

Acknowledgements. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grants No. VG1/0675/11 and No. VG1/0971/11, the Slovak Research and Development Agency under the contract No. APVV-0208-10 and is the partial result of the Research & Development Operational Programme for the project SMART II, ITMS 26240120029, co-funded by the ERDF.

References

1. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge sharing and yahoo answers: everyone knows something. In: Proc. of the 17th Int. Conf. on World Wide Web (WWW 2008), pp. 665–674. ACM, New York (2008)
2. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proc. of the Int. Conf. on Web Search and Web Data Mining, WSDM 2008, pp. 183–194. ACM, New York (2008)
3. Chen, B.C., Dasgupta, A., Wang, X., Yang, J.: Vote calibration in community question-answering systems. In: Proc. of the 35th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2012), pp. 781–790. ACM, New York (2012)
4. Downes, S.: E-learning 2.0. eLearn magazine 2005, vol. 10 (1). ACM, New York (2005)
5. IEEE LTS: Draft Standard for Learning Object Metadata. IEEE Standard 1484.12.1. IEEE (2002) (retrieved March 2013)
6. Ghauth, K.I., Abdullah, N.A.: The Effect of Incorporating Good Learners' Ratings in e-Learning Content-based Recommender System. Educational Technology & Society 14(2), 248–257 (2011)
7. Golovchinsky, G., Qvarfordt, P., Pickens, J.: Collaborative information seeking. Information Seeking Support Systems (2008)
8. Kidd, J., O'Shea, P., Baker, P., Kaufman, J., Allen, D.: Student-authored Wikibooks: Textbooks of the Future? In: McFerrin, K., et al. (eds.) Proc. of Society for Information Technology & Teacher Education Int. Conf. 2008, pp. 2644–2647. AACE, Chesapeake (2008)
9. Lawson, M.: Berners-Lee on the read/write web. BBC, Technology (2005), `http://news.bbc.co.uk/1/hi/technology/4132752.stm` (accessed March 31, 2013)
10. Quinn, A.J., Bederson, B.B.: Human computation: a survey and taxonomy of a growing field. In: Proc. of the 2011 Annual Conf. on Human Factors in Computing Systems (CHI 2011), pp. 1403–1412. ACM, New York (2011)
11. Stahl, G., Koschmann, T., Suthers, D.: Computer-supported collaborative learning: An historical perspective. In: Sawyer, R.K. (ed.) Cambridge Handbook of the Learning Sciences, pp. 409–426. Cambridge University Press, Cambridge (2006)
12. Šimko, M., Barla, M., Bieliková, M.: ALEF: A Framework for Adaptive Web-based Learning 2.0. In: Reynolds, N., Turcsányi-Szabó, M. (eds.) KCKS 2010. IFIP AICT, vol. 324, pp. 367–378. Springer, Heidelberg (2010)
13. Šimko, M., Barla, M., Mihál, V., Unčík, M., Bieliková, M.: Supporting Collaborative Web-Based Education via Annotations. In: Proc. of W. Conf. on Educational Multimedia, Hypermedia & Telecommunications, ED-MEDIA 2011, pp. 2576–2585. AACE (2011)
14. Wheeler, S., Yeomans, P., Wheeler, D.: The good, the bad and the wiki: Evaluating student-generated content for collaborative learning. British Journal of Educational Technology 39(6), 987–995 (2008)

An Black-Box Testing Approach on User Modeling in Practical Movie Recommendation Systems

Xuan Hau Pham, Tu Ngoc Luong, and Jason J. Jung[⋆]

Knowledge Engineering Laboratory
Department of Computer Engineering
Yeungnam University, Korea
{pxhauqbu,luongngoctu,j2jung}@gmail.com

Abstract. Since there have been many practical recommendation services in real world, main research questions are *i*) how such services provide users with recommendations, and *ii*) how they are different from each other. The aim of this paper is to evaluate user modeling process in several practical recommendation systems. Black-box testing scheme has been applied by comparing recommendation results. User models (i.e., a set of user ratings) have been synthesized to discriminate the recommendation results. Particularly, we focus on investigating whether the services consider attribute selection.

Keywords: Recommendation systems, Personalization, User modeling, Black-box testing, Comparative study.

1 Introduction

Recommendation systems have been studied for a long time. There have been a lot of recommendation schemes to provide users with the most relevant information. In the real world, we have been able to access recommendation services in various domains. Particularly, movie is the most popular domain targeted by the recommendation services.

However, even though users are eager to get the recommendations from the practical systems, they have not been satisfied with the results. For example, users are consistently acting with the same preferences (i.e., the same input ratings), the results from such recommendation systems are completely difference from each other.

Users (at least system developers) want to know what kinds of recommendation mechanisms are behind the systems. Somehow, depending on their situation, more appropriate system can be selected. Thereby, in this paper, we focus on comparing the recommendation results provided from the practical recommendation systems, and investigating what kinds of recommendation schemes have been exploited in these systems. Especially, we want to show that the proposed

[⋆] Corresponding author.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 72–79, 2013.

black-box testing strategy can precisely reveal the recommendation schemes behind those practical recommendation systems. To do this, we have synthesized a number of user models by generating user ratings. Consequently, target systems returns a set of recommendations, and the systems will be differentiated with each other.

The outline of this paper is as follows. In the following Sect. 2, we will address backgrounds on recommendation systems in the literature. Sect. 3 gives a research method on comparing the results obtained from a set of selected recommendation systems. Sect. 3.2 will give the description about the recommendation systems that we have selected in this work. Most importantly, Sect. 4 show how to conduct the experiments for collecting the results from the recommendation systems. In Sect. 5, we will analyze the collected recommendation results, and draw a conclusion of this work.

2 User Modeling Strategies on Recommendation Systems

User modeling process in the recommendation systems is commonly based on analyzing various information collected by users in explicit or implicit ways [1]. Depending on the following two issues, we need to consider to build a taxonomy of user modeling in recommendation systems.

1. Recommendation systems usually ask users to explicitly input various information. First issue is what kinds of information is employed to build user models in the recommendation systems.
   - Demographic information
   - User ratings

   Given equivalently synthesized users (e.g., same age, same gender, and so on), we can compare the recommendation results to realize whether they are same or different.
2. Recommendation systems somehow exploit the information collected from the users to discover useful patterns about the users. Second issue is which recommendation strategies are applied to provide users with relevant information.
   - Personalization
   - Collaborative filtering

   Assuming that a user is interested in a certain value (e.g., "Steven Spielberg"), the recommendation results can be compared to reversely find out what kind of schemes are applied.

Hence, as shown in Table 1, the recommendation systems can be simply categorized into four different types. Of course, in practice, we are sure that the practical systems are employing a hybrid scheme combining several recommendation approaches (e.g., collaborative filtering and personalization) [2].

Especially, in the context of attribute selection-based user modeling [3], the users can be synthesized to differentiate the results from recommendation systems. Table 2 shows an example with three users (i.e., U_1, U_2, and U_3). With

Table 1. Taxonomy of recommendation systems w.r.t. user modeling processes

	Personalization	Collaborative filtering
Demographics	$\mathrm{Rec}_{\mathrm{P,D}}$	$\mathrm{Rec}_{\mathrm{C,D}}$
User ratings	$\mathrm{Rec}_{\mathrm{P,U}}$	$\mathrm{Rec}_{\mathrm{C,U}}$

two users (i.e., U_1 and U_2), we can discriminate $\mathrm{Rec}_{\mathrm{P,D}}$ and $\mathrm{Rec}_{\mathrm{P,U}}$ from $\mathrm{Rec}_{\mathrm{C,D}}$ and $\mathrm{Rec}_{\mathrm{C,U}}$ Also, with two users (i.e., U_1 and U_3), we can discriminate $\mathrm{Rec}_{\mathrm{P,D}}$ and $\mathrm{Rec}_{\mathrm{C,D}}$ from $\mathrm{Rec}_{\mathrm{P,U}}$ and $\mathrm{Rec}_{\mathrm{C,U}}$.

Table 2. Example with three synthesized users who have rated the same movies

		U_1	U_2	U_3
Age/Gender/Country		20/Male/Korea	20/Male/Korea	60/Female/France
Rating	Lincoln	5	1	Very good
	War Horse	5	1	Very good
	The Terminal	5	1	Very good

3 Research Method

In this paper, we focus on black-box testing scheme [4], since we have no information about internal strategies of the practical recommendation systems. As shown in Fig. 1, a set of users are synthesized to collect recommendations from the practical services.

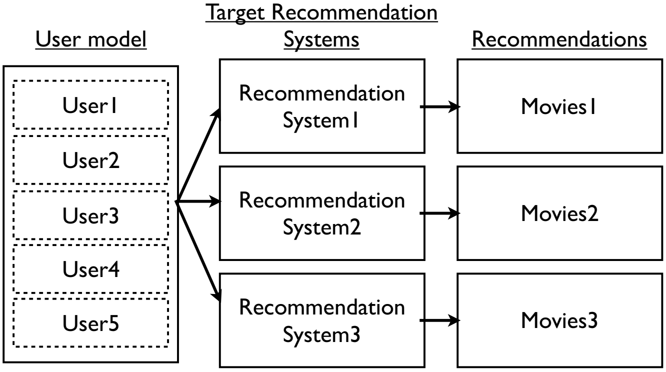


Fig. 1. Research model based on black box testing

3.1 Comparison of Recommendation Results

Once we select a set of recommendation service, recommendation results are compared with each other.

Definition 1 (Recommendation). *Given a recommendation service RS_i, the recommendation result M_i is composed of a set of movies which are regarded as the most relevant movies to user contexts. It is represented as*

$$M_i = \{m_1, m_2, \ldots, m_N\} \tag{1}$$

where N is the number of movies recommended by the system.

The recommendation results can be matched to quantify the similarity between the corresponding recommendation schemes.

Definition 2 (Similarity). *Two recommendation results M_i and M_j from recommendation services RS_i and RS_j, similarity between RS_i and RS_j can be measured by*

$$Sim(RS_i, RS_j) = \frac{M_i \cap M_j}{\max(M_i, M_j)} \tag{2}$$

where denominator can choose the maximum size of recommendation.

3.2 Selected Practical Recommendation Systems

Initially, as shown in Table 3, we have tried to select 10 movie recommendation systems [5]. Out of them, 2 recommendation systems are not available for the moment.

Table 3. Selected recommendation systems

Recommendation systems	URLs	
Jinni	`www.jinni.com`	
Taste Kid	`www.tastekid.com`	
Nanocrowd	`www.nanocrowd.com`	
Clerkdogs	`www.clerksblog.com`	Not available
Criticker	`www.criticker.com`	
IMDB	`www.imdb.com`	
Flixster	`www.flixster.com`	
Movielens	`www.movielens.org`	
Rotten Tomatoes	`www.rottentomatoes.com`	
Netflix	`www.netflix.com`	Not available in Korea

Table 4. User profiling during registration (○: required, △: optional)

Recommendation systems	Email	User name	Real name	Date of Birth	Gender	Country	City/ State	Postal Code	Marital Status
Jinni (RS_1)	○	○	△	△	△	○		△	
Taste Kid (RS_2)	○	○							
Nanocrowd (RS_3)	○	○							
Criticker (RS_4)	○	○	△	△	△	△	△		△
IMDB (RS_5)	○		○	○	○			○	
Flixster (RS_6)	○		○	○					
Movielens (RS_7)	○	○	△					△	
Rotten Tomatoes (RS_8)	○		○	○	○	○			

User Registration. Practical recommendation systems usually ask users to input various personal information during registration, depending on their recommendation schemes. Table 4 shows the list of demographic information requested by the recommendation systems.

All systems ask Email in common. It is regarded as an unique identifier for each user. Since only two kinds of personal information (e.g., Email and username) are asked in Taste Kid and Nanocrowd, these two systems are not asking any personal information. They seem to be more focused on ratings from users. In contrast, Jinni, Criticker, IMDB and Rotten Tomatoes are asking more than five types of personal information.

Table 5. Representation of user ratings

Recommendation systems	Data type	Range	Cardinality
Jinni	ordinal/ discrete	{ awful, bad, poor, disappointing, so so, ok, good, great, amazing, must see }	10
Taste Kid	ordinal/discrete	{like, dislike}	2
Nanocrowd	enumerate/discrete	{watched, not watched}	2
Criticker	numeric/integer/ discrete	{1, 2, . . . , 100 }	100
IMDB	ordinal/discrete	{1, 2, . . . , 10}	10
Flixster	ordinal/discrete	{1, 2, . . . , 5}	5
Movielens	ordinal/discrete	{1, 2, . . . , 5}	5
Rotten Tomatoes	ordinal/discrete	{1, 2, . . . , 5}	5

Representation of User Ratings. Recommendation systems ask users to rate movies. Depending on the systems, the ratings are represented in several different ways. Table 5 shows how the user ratings are represented in the recommendation systems. Three of the systems (i.e., Flixster, Movielens, and Rotten Tomatoes) are allowing users to rate the movies between 1 and 5. Jinni and IMDB are more diversified to between 1 and 10. Particularly, in Criticker, users can rate the movies from 0 to 100. On the other hand, Taste Kid and Nanocrowd are making the user ratings most simplified. Interestingly, Nanocrowd is simply asking users to record the list of movies (i.e., "watched").

Additionally, we have to consider the requirement of initial rating step, as shown in Table 6. Jinni, Criticker, and Movielens are collecting initial ratings from users, before they provide the users with recommendations.

Table 6. Initial requirement for recommendations; X indicates 'Not required'

Recommendation systems	Number of initial ratings	Rating scores
Jinni	More than 10 items	0 to 10
Taste Kid	X	0 to 1
Nanocrowd	X	0 to 1
Criticker	More than 10 items	0 to 100
IMDB	X	0 to 10
Flixster	X	1 to 5
Movielens	More than 15 items	1 to 5
Rotten Tomatoes	X	1 to 5

4 Experiments and Evaluation

In this section, we want to describe how to synthesize user models and how to collect recommendation results from the practical services.

User models have been synthesized by assuming a user is interested in a certain attribute, as follows.

- U_1: Genre "Sci-Fi"
- U_2: Director "Steven Spielberg"
- U_3: Actor "Leonardo DiCaprio"
- U_4: Actress "Angelina Jolie"

Also, the movies can be rated in two difference ways.

- Unified rating: We need to express that a user's interest is consistent in each attribute.
- Random rating: We need to express that a user's interest is not consistent in each attribute.

For example, a user is assumed to be consistently interested in a genre "Sci-Fi". As shown in Table 2, this user can be synthesized in $\{\langle m_1, 5\rangle, \langle m_2, 5\rangle, \langle m_3, 5\rangle\}$. In opposite, his random ratings can be synthesized as $\{\langle m_1, 5\rangle, \langle m_2, 1\rangle, \langle m_3, 3\rangle\}$.

Since user ratings are differently represented in recommendation services (shown in Table 5), the user ratings in different recommendation services should be normalized to be comparable.

We have collected the recommendations from the practical services. Table 7 shows evaluational results on personalization-based recommendation services. In both cases, only U$_1$ (Genre "Sci-Fi") has shown high matching ratio. The other three users (i.e., U$_2$, U$_3$, and U$_4$) are in the very low level. We found that the practical recommendation services are not considering attribute-based personalization.

Table 7. Evaluation on attribute-based personalization; (a) uniform ratings, and (b) random ratings

	RS$_1$	RS$_2$	RS$_3$	RS$_4$	RS$_5$	RS$_6$	RS$_7$	RS$_8$
U$_1$	$\frac{17}{27}$	$\frac{14}{17}$	$\frac{42}{75}$	$\frac{0}{8}$	$\frac{51}{84}$	$\frac{27}{50}$	-	$\frac{71}{135}$
U$_2$	$\frac{0}{26}$	$\frac{3}{17}$	$\frac{3}{75}$	$\frac{0}{8}$	$\frac{2}{60}$	$\frac{4}{40}$	-	$\frac{6}{191}$
U$_3$	$\frac{0}{27}$	$\frac{0}{17}$	$\frac{0}{75}$	$\frac{0}{8}$	$\frac{0}{90}$	$\frac{0}{48}$	-	$\frac{0}{171}$
U$_4$	$\frac{0}{26}$	$\frac{3}{17}$	$\frac{1}{75}$	$\frac{0}{8}$	$\frac{0}{54}$	$\frac{0}{15}$	-	$\frac{0}{103}$

	RS$_1$	RS$_2$	RS$_3$	RS$_4$	RS$_5$	RS$_6$	RS$_7$	RS$_8$
U$_1$	$\frac{11}{20}$	$\frac{13}{17}$	$\frac{42}{75}$	$\frac{2}{8}$	$\frac{35}{65}$	$\frac{24}{50}$	-	$\frac{53}{92}$
U$_2$	$\frac{2}{20}$	$\frac{2}{17}$	$\frac{0}{75}$	$\frac{0}{8}$	$\frac{1}{66}$	$\frac{6}{50}$	-	$\frac{4}{98}$
U$_3$	$\frac{0}{20}$	$\frac{0}{17}$	$\frac{0}{75}$	$\frac{0}{8}$	$\frac{0}{66}$	$\frac{0}{11}$	-	$\frac{0}{82}$
U$_4$	$\frac{0}{20}$	$\frac{0}{17}$	$\frac{0}{75}$	$\frac{0}{8}$	$\frac{0}{66}$	$\frac{0}{50}$	-	$\frac{0}{78}$

Also, Table 8 shows comparison between uniform ratings and random ratings. RS_2 and RS_8 are showing high ratio between uniform and random ratings. It means that these recommendation services are not considering user ratings as importantly as the other services are.

Table 8. Comparison of uniform and random ratings

		RS$_1$	RS$_2$	RS$_3$	RS$_4$	RS$_5$	RS$_6$	RS$_7$	RS$_8$
Common	U$_1$	18.5	35.3	-	0.0	39.7	20	-	67.4
	U$_2$	42.3	64.7	-	0.0	3.3	80	-	51.3
	U$_3$	14.8	29.4	-	0.0	14.4	6.3	-	47.9
	U$_4$	11.5	52.9	-	0.0	9.1	10	-	74.7
Similar	U$_1$	63.0	76.5	-	12.5	51.3	54	-	57.6
	U$_2$	10 .0	17.6	-	0.0	39.7	12	-	4.1
	U$_3$	0.0	0.0	-	0.0	0.0	0.0	-	0.0
	U$_4$	0.0	17.6	-	0.0	0.0	0.0	-	0.0

5 Concluding Remark and Future Work

In this work, we have investigated what kinds of recommendation schemes have been exploited in the practical recommendation systems. As a conclusion, this paper has evaluated user modeling process in several practical recommendation systems. Black-box testing scheme has been applied by comparing recommendation results. User models (i.e., a set of user ratings) have been synthesized to discriminate the recommendation results.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0017156).

References

1. Pu, P., Chen, L., Hu, R.: Evaluating recommender systems from the user's perspective: survey of the state of the art. User Modeling and User-Adapted Interaction 22(4-5), 317–355 (2012)
2. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22(1), 5–53 (2004)
3. Jung, J.J.: Attribute selection-based recommendation framework for short-head user group: An empirical study by movielens and imdb. Expert Systems with Applications 39(4), 4049–4054 (2012)
4. Beizer, B.: Black-Box Testing: Techniques for Functional Testing of Software and Systems. John Wiley & Sons (1995)
5. Reisinger, D.: Top 10 movie recommendation engines (March 2009)

Grey Incidence between Banks' Risks and Their Performance

– A Country Rating Perspective –

Ioana Bradea, Camelia Delcea, Emil Scarlat, Marcel Boloş, and Nora Chiriţă

The Bucharest Academy of Economic Studies, Department of Economic Informatics
and Cybernetics, Bucharest, Romania
{alexbradea1304,camelia.delcea}@yahoo.com

Abstract. The present paper tries to give a new perspective on the banking risks and their influence on the profitability of the banking sector as a whole, with a tremendous impact on banks survival in now-a-days economy. Due to the more and more uncertain economical environment, this impact is even harder to be measured and managed. That is why here it can be identified a highly stringent need for some methods which are best working in uncertain conditions and what could work better than grey systems theory methods? Although it is quite new, grey system theory was able to impose quickly. In practice it has been successfully applied in the analysis, modeling, prediction, control and decision making in all areas. Its advantage is that it manages to achieve a good performance in tests conducted on a small number of data. The analysis is conducted through a country rating perspective.

Keywords: grey incidence analysis, banking risks, country rating, modeling.

1 Introduction

Country risk is the risk arising from unanticipated changes in the economic or political environment prevailing in a particular country. Country risk is a reality increasingly complex, affecting the international economic transactions. Complexity is due to the fact that risks are becoming increasingly interconnected. The country risk is sensitive to political, social and economic factors and envisages losses that may occur in a business with a foreign partner.

Related to the concept of country risk, we can meet the following terms: systemic risk, sovereign risk and disasters. Systemic risk is determined by the global recession, thus affecting all countries. If the state does not interfere with measures to mitigate this risk, the risk can be considered a country risk. Sovereign risk treats the loans granted by banks to states, and natural disasters are classified as country risk if they have a regular character.

According to a report of Swissre.com," economic losses from natural catastrophes and man-made disasters reached 186 billion USD in 2012, the insured losses were 77

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 80–89, 2013.

billion USD, making 2012 the third most costly year on record and the weather events in the US dominated insured losses".

The resistance to country risk can be increased by early identification, risk prioritization and a high level of coordination between Government and policy. Monitoring and measuring risk, using key risk indicators, represent a priority for the country risk officer (CRO).

For this reason, the present paper tries to give a new perspective on the banking risks and their influences on the profitability of the banking sector as a whole, with a tremendous impact on banks survival in now-a-days economy. Due to the more and more uncertain economical environment, this impact is even harder to be measured and managed. That is why here it can be identified a highly stringent need for some methods which are best working in uncertain conditions and what could work better than grey systems theory methods?

Although it is quite new, grey system theory was able to impose quickly. In practice has been successfully applied in the analysis, modelling, prediction, control and decision making in all areas. The advantage of grey systems theory is that it manages to achieve a good performance in tests conducted on a small number of data and a large number of variables.

2 Risks in Banks and the Economic Crisis

The risk can be defined as the "uncertainty about a future outcome" [1], or „the combination of the probability of an event and its consequences". [2] Every bank faces risks, which risks that are more and more difficult to manage. [1] Next, some of the main risks that are affecting the status of a credit institution are listed.

2.1 Assets Quality Risk

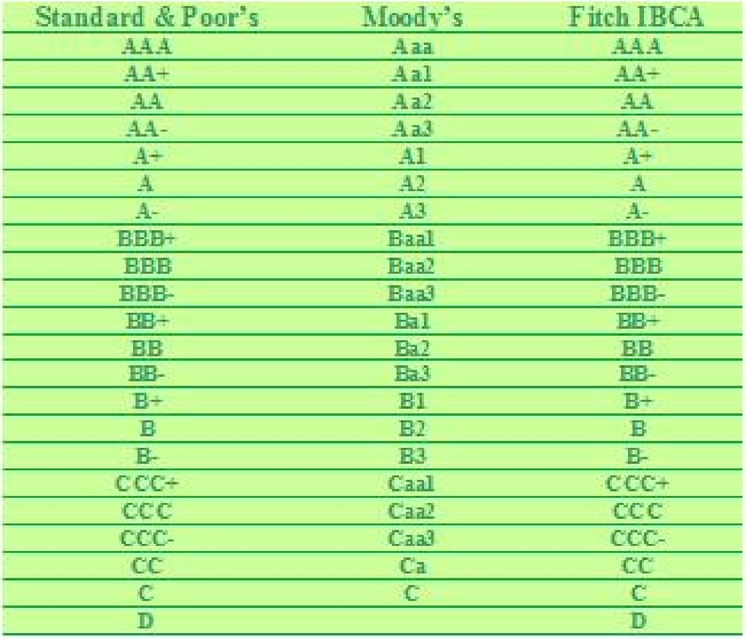
Total assets of credit institutions are mainly the sum of: cash, deposits at central banks; claims on credit institutions; claims on customers; tangible and intangible assets. The claims on credit institutions and on customers are related to granted loans. The quality of assets consists especially on the exposure at credit risk. The bad loans are the main source of banking imbalance.

Credit risk presents the loss due to a credit event and is inseparable of any payment obligation. Associated to credit risk are: the changes that appear in credit quality, credits spreads and the default. The financial derivatives have an important role in mitigation and managing the credit risk, which have as goal to transfer the risk between parts.

The concentration risk represents an unequal distribution of loans to single borrowers, or industrial or regional sectors and the dependencies between borrowers. This risk has a major impact on credit risk, because it may determines a default contagion. [3]

The credit ratings measure the propensity to default, but not illustrate the quality of assets. The principal asset of a bank consists of loans, so that bad loans may bankrupt the bank. The deterioration in quality of the loan portfolio generates important losses on short term.

An important role in analyzing the state of a bank have the credit rating agencies, which evaluate the risks and provide the results in the form of credit ratings. These ratings help the investors to decide if they will invest or not. There are well-defined rules that forbid investing in debts that are uncertain. The main problem faced by a rating agency is that it is paid by the bank to perform the classification. (Fig. 1)



Standard & Poor's	Moody's	Fitch IBCA
AAA	Aaa	AAA
AA+	Aa1	AA+
AA	Aa2	AA
AA-	Aa3	AA-
A+	A1	A+
A	A2	A
A-	A3	A-
BBB+	Baa1	BBB+
BBB	Baa2	BBB
BBB-	Baa3	BBB-
BB+	Ba1	BB+
BB	Ba2	BB
BB-	Ba3	BB-
B+	B1	B+
B	B2	B
B-	B3	B-
CCC+	Caa1	CCC+
CCC	Caa2	CCC
CCC-	Caa3	CCC-
CC	Ca	CC
C	C	C
D		D

Fig. 1. A comparison of credit ratings from the most important agencies [4]

2.2 Capital Risk

The capital risk illustrates the risk of loss an investor faces when he invested an amount of money. If the investor is an enterprise or a bank, the risk is upon the value of the invested capital. An important role is the investor's adversity at risk, depending on which he have to decide in which assets to invest the available capital.

There are three types of risk capital, which are influenced by the market, project and bank state. Capital risk occurs when there is an insufficient equity, so the bank cannot continue its activity.

2.3 Operational Risk

The Basel II accord defined the operational risk as: „the risk of loss resulting from inadequate or failed processes, people and systems or from external events". Operational risk is determined by internal or external factors. National Bank imposed since 2004 the following procedures for managing operational risk: assessment, monitoring, risk reduction.[5-6] Each credit institution has to establish its own operational risk profile for internal needs and to establish specific policies.

2.4 Liquidity Risk

The liquidity risk represents the risk that results from the inability of counterparty to settle a transaction because of the lack of liquidity. Liquidity Risk is the likelihood that the bank will not be able to make payments to customers due to the gap between long term loans and short term loans, and the lack of correlation of assets to liabilities of the bank.

There are two types of liquidity risk: immediate liquidity risk and conversion risk. The immediate liquidity risk appears when the bank cannot satisfy the massive withdrawals and have to seek overnight credits. On the other hand, the conversion risk occurs when the bank used the short term resources to make long-term placements.

3 Grey Incidence Analysis

The Grey Systems Theory appeared in 1982 after the publication of Professor Deng Julong's paper "The Control Problems of Grey Systems". [7] The name of Grey Systems comes from the nature of the subject under investigation. In control theory, the colors indicate the clarity of information. Towards information we use black to express a totally unknown information and white for full information known. Grey systems are considered to be a mix of information, partly known and partly unknown, a combination of black and white, or grey.

According to Xie and Liu, the grey incidence analysis (GIA) represents a central piece of grey system theory and is the foundation for grey modeling, decision making and control. [8] Professor Deng created this method and gave a model to calculate the degree of grey incidence. Over time, acute interest for this method led researchers from different parts of the world to study and make improvements. Among the variation of grey incidence degree should be mentioned: the degree of general incidence, the degree of grey pointed incidence, the degree of grey generalized incidence, the degree of B-mode incidence, the degree of T-mode incidence, the degree of slope incidence, the degree of grey-entropy incidence, the degree of improved T-mode incidence. Thus far, there are three important types of grey incidence: absolute, relative and synthetic. [9-12]

Grey incidence analysis is often used for solving different problems which are considered very complex due to the character of the relationship between variables. GIA gives the possibility to make predictions, clustering and decision making under conditions of great uncertainty. [13] The grey incidence analysis treats the proximity and similarity of the size between the curves, in order to establish the grade of correlation between estimated factors.

Let’s see how the three most known degrees of grey incidence are computed.

3.1 The Absolute Degree of Grey Incidence

Assume that X$_0$ and X$_j$, j=1...n, are two sequences of data with non-zero initial values and with the same length, with *t* = time period and *n* = variables: [12]

$$X_0 = (x_{1,0}, x_{2,0}, x_{3,0}, x_{4,0}, \ldots, x_{t,0}), \tag{1}$$

$$X_j = (x_{1,j}, x_{2,j}, x_{3,j}, x_{4,j}, \ldots, x_{t,j}), \tag{2}$$

The images of zero-start points are:

$$X_j^0 = (x_{1,j} - x_{1,j}, x_{2,j} - x_{1,j}, \ldots, x_{t,j} - x_{1,j}) = (x_{1,j}^0, x_{2,j}^0, \ldots, x_{t,j}^0) \tag{3}$$

The absolute degree of grey incidence is given by:

$$\varepsilon_{0j} = \frac{1 + |s_0| + |s_j|}{1 + |s_0| + |s_j| + |s_0 - s_j|} \tag{4}$$

where $|s_0|$ and $|s_j|$ are computed as follows:

$$|s_0| = \left| \sum\nolimits_{k=2}^{t-1} x_{k,0}^0 + \frac{1}{2} x_{t,0}^0 \right| \tag{5}$$

$$|s_j| = \left| \sum\nolimits_{k=2}^{t-1} x_{k,j}^0 + \frac{1}{2} x_{t,j}^0 \right| \tag{6}$$

3.2 The Relative Degree of Grey Incidence

Assume that X$_0$ and X$_j$, j=1...n, are two sequences of data with non-zero initial values and with the same length, with t = time period and n = variables: [12]

$$X_0 = (x_{1,0}, x_{2,0}, x_{3,0}, x_{4,0}, \ldots, x_{t,0}), \tag{7}$$

$$X_j = (x_{1,j}, x_{2,j}, x_{3,j}, x_{4,j}, \ldots, x_{t,j}), \tag{8}$$

The initial values images of X$_0$ and X$_j$ are:

$$X_0^{'} = (x_{1,0}^{'}, x_{2,0}^{'}, \ldots, x_{t,0}^{'}) = (\frac{x_{1,0}}{x_{1,0}}, \frac{x_{2,0}}{x_{1,0}}, \ldots, \frac{x_{t,0}}{x_{1,0}}) \tag{9}$$

$$X_j^{'} = (x_{1,j}^{'}, x_{2,j}^{'}, \ldots, x_{t,j}^{'}) = (\frac{x_{1,j}}{x_{1,j}}, \frac{x_{2,j}}{x_{1,j}}, \ldots, \frac{x_{t,j}}{x_{1,j}}) \tag{10}$$

The images of zero-start points calculated based on (9) and (10) for X_0 and X_j are:

$$X_0^{0'} = (x^{'}_{1,0} - x^{'}_{1,0}\ , x^{'}_{2,0} - x^{'}_{1,0}, \ldots, x^{'}_{t,0} - x^{'}_{1,0}) = (x_{1,0}^{'0}, x_{2,0}^{'0}, \ldots, x_{t,0}^{'0}) \tag{11}$$

$$X_j^{0'} = (x^{'}_{1,j} - x^{'}_{1,j}\ , x^{'}_{2,j} - x^{'}_{1,j}, \ldots, x^{'}_{t,j} - x^{'}_{1,j}) = (x_{1,j}^{'0}, x_{2,j}^{'0}, \ldots, x_{t,j}^{'0}) \tag{12}$$

The relative degree of grey incidence is given by:

$$r_{0j} = \frac{1 + \left|s^{'}_0\right| + \left|s^{'}_j\right|}{1 + \left|s^{'}_0\right| + \left|s^{'}_j\right| + \left|s^{'}_0 - s^{'}_j\right|} \tag{13}$$

where $\left|s^{'}_0\right|$ and $\left|s^{'}_j\right|$ are computed as follows:

$$\left|s^{'}_0\right| = \left|\sum\nolimits_{k=2}^{t-1} x_{k,0}^{'0} + \frac{1}{2} x_{t,0}^{'0}\right| \tag{14}$$

$$\left|s^{'}_j\right| = \left|\sum\nolimits_{k=2}^{t-1} x_{k,j}^{'0} + \frac{1}{2} x_{t,j}^{'0}\right| \tag{15}$$

3.3 The Synthetic Degree of Grey Incidence

The synthetic degree of grey incidence is based on the absolute and relative degrees of grey incidence obtained earlier: [12]

$$\rho_{0j} = \theta\varepsilon_{0j} + (1-\theta) r_{0j}, \tag{16}$$

with j = 2, ..., n , $\theta \in [0,1]$ and $0 < \rho_{0j} \leq 1$.

Having all these preliminaries, we can move now to the next section, where the grey incidence will be applied to the data gathered from the banking sector.

4 Application

As it has been stipulated, one of the research questions refers to determine whether there is a connection between the types of risks that arise at bank's level and its performance and how strong this connection can be. Mainly, how much of the variance of a bank profitability can be explained by analysing the levels of its risks (especially the ones identified and presented in the second subchapter of this paper, namely the Assets Quality Risk (ASR), Capital Risk (CR), Operational Risk (OR) and Liquidity Risk (LR)) and how much is due to the changes the economic environment is facing, such as firms' bankruptcy, turbulences, instability, movement in the exchange rate of the country's currency, etc. (Fig. 2)

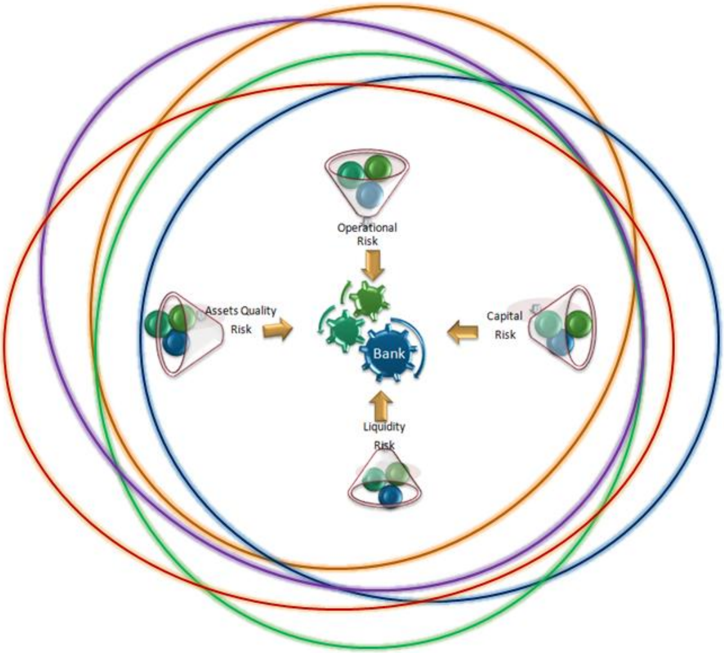


Fig. 2. Bank's Risks and the Environmental Turbulences

Moreover, another desiderate of this research is to see if there are any particularities between banks that are conducting their activity in countries that have received in 2012 a lower or a greater country rating.

For this, a number of 25 banks have been considered, divided in groups of 5 banks, each of this group representing the bank sector from a country with a specific overall country rating.

Taking into consideration the data extracted from Bankscope, the five groups were formed by countries that have received the following country rating: AA, A, BB, BBB and CCC.

For the first group of banks that are activating in countries that have received an AA overall county rating in 2012, we have been considered the most important 5 banks from Norway and Sweden.

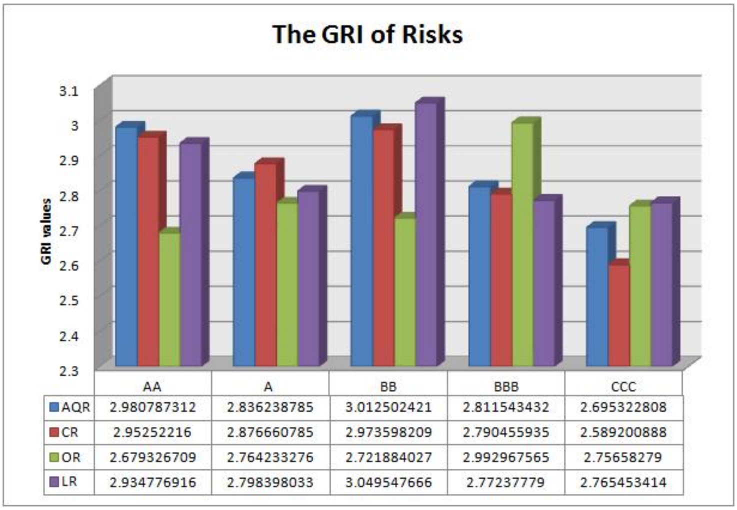
The second group (banks from country with rating A) is formed by banks from Germany, Denmark, Finland and France.

For the third group (BB), banks from Spain, Ireland and Portugal have been considered.

The fourth group (BBB) is compounded from Great Britain and Italy banks, while the last one (CCC) is formed solely by banks from Greece.

The values representing the four bank risks have been collected from Bankscope for a period of time equal to three years, from 2009 to 2011, along with the values of the banks' performance registered for the same period of time.

Grey incidence analysis has been applied to the selected data for each of the 25 banks and the results are summarised in Fig 3 and Fig 4.



	AA	A	BB	BBB	CCC
🟦 AQR	2.980787312	2.836238785	3.012502421	2.811543432	2.695322808
🟥 CR	2.95252216	2.876660785	2.973598209	2.790455935	2.589200888
🟩 OR	2.679326709	2.764233276	2.721884027	2.992967565	2.75658279
🟪 LR	2.934776916	2.798398033	3.049547666	2.77237779	2.765453414

Fig. 3. The Cumulated Grey Relational Incidence of each Risk

It can be observed that for the banks belonging to countries with an accentuated to moderate economic stability, the risk indicators considered can influence the future state of being for each bank. The only risk that fails to capture this incidence is the operational risk (OR), which is registering relative lower values in comparison with the other three risk categories. This result was quite expecting to happen, mostly because in these countries the economic environment is a stable one which grants a general stability level to each economic sector, being appropriate and facilitating a good economic development and prosperity.

For the last two categories of banks, things are happening a little bit different. For the ones that are activating in countries with BBB rating, the most incidence is reflected by the operational risk, while the three other categories of risks stand at the same level.

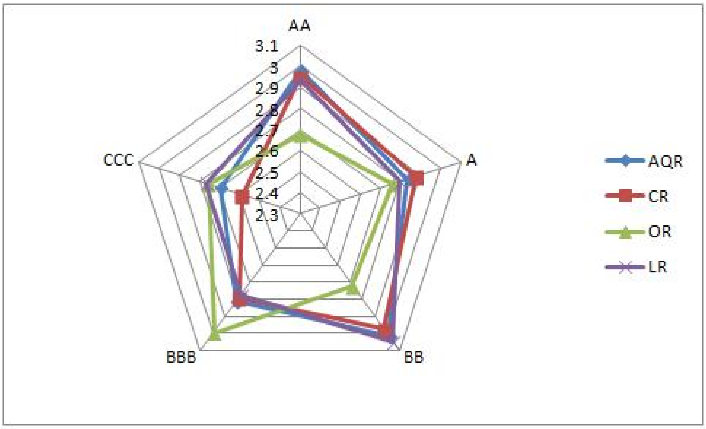


Fig. 4. Banks Risks vs. Overall Country Rating

The most relevant situation is that registered in the case of the banks from the last category considered (CCC), where the hierarchy of risks differs totally from the other four cases. Here the capital risk seems to have the lowest influence. This is not such a surprise. The most interesting achievement is that all the four types of bank risks are not succeeding in tremendously influence the actual state of the banks from this group – as it can be seen the calculated values of all four risk are having really low amounts. This can be explained through the huge flow of influences that are coming from the economic environment, which are destabilizing the banks activity. In this case, the risk indices are not sufficient anymore for explaining what is happening at a bank's level, being necessary a more detailed analysis in which the external factors should be taken into account.

5 Concluding Remarks

Based on the analysis conducted on the 25 banks selected from representative countries (5 banks for each group formed by countries with an AA, A, BB, BBB and CCC overall country rating), there have been reached the following conclusions: due to the relatively stable economical environment, for the banks acting in countries that have received a good overall country indicator, the analysis of each bank's risk can be concluding for forecasting in the future their survival or performance. The incidence of these risks is relevant with the purpose of the research.

This also happens in the forth case (for the banks acting in BBB countries), with an accent on the operational risk which succeeds, in comparison with the other three categories of risks identified, to offer a more accurate picture of each bank's evolution.

On the other hand, in the case of the banks from CCC countries, these indicators are not representing anymore the reality in banks and they can be only used partially, combined with other qualitative indicators related to the economic environment. Finding and evaluating this kind of indicators will be a goal of our following researches.

By knowing all this, an analysis conducting at a bank's level through the perspective of its main risk can be concluding only from a certain point of view, namely for the banks situated in economic stable countries. As in many other situations, the elements taken from grey systems theory have demonstrated, once more, their power of gathering the needed information.

Note: the order of the authors on this paper is random, their contribution to the achieved results being equal.

References

1. Doerig H.U.: Operational Risks in Financial Services: An Old Challenge in a New Environment, Working Paper, Credit Suisse Group (2003)
2. Spedding, L., Rose, A.: Business Risk Management Handbook – A sustainable approach. CIMA Publishing, USA (2008)
3. Lütkebohmert, E.: Concentration Riskin Credit Portfolios, p. 5. Springer-Verlag Publisher, Berlin (2009) e-ISBN 978-3-540-70870-4

4. Bank for International Settlements: Long-Term Rating Scales Comparison (2005)
5. Global Association of Risk Professionals: Operational Risk Management, ch. 12, `http://www.garp.org/media/673303/operational%20risk%20slides.pdf`
6. Moosa, I.: Quantification of Operational Risk under Basel II. Palgrave Macmillan Publisher, UK (2008)
7. Jun-Fu, C., Ren-Jie, Z., Rui, L.: Grey Incidence Analysis Based on Coefficient of Determination and Its Economic Application with the Data of Central Henan Urban Agglomeration. In: Proceedings of 2009 IEE International Conference on Grey Systems and Intelligent Services, China, pp. 32–36 (2009)
8. Xie, N.-M., Liu, S.-F.: The Parallel and Uniform Properties of Several Relational Models. Systems Engineering 25, 98–103 (2007)
9. Cotfas, L.A.: A finite-dimensional quantum model for the stock market. Physica a-Statistical Mechanics and Its Applications 392(2), 371–380 (2013)
10. Cotfas, L.A.: A quantum mechanical model for the relationship between stock price and stock ownership. In: Application of Mathematics in Engineering and Economics, vol. 1497, pp. 37–44 (2012)
11. Cotfas, N., Cotfas, L.A.: Hypergeometric type operators and their supersymmetric partners. Journal of Mathematical Physiscs 52(5) (2011)
12. Liu, S.F., Lin, Y.: Grey Systems Theory and Applications. Understanding Complex Systems. Springer, Heidelberg (2010)
13. Delcea, C., Scarlat, E.: Finding Companies' Bankruptcy Causes Using a Hybrid Grey-Fuzzy Model. Economic Computation and Economic Cybernetics Studies and Research 44(2), 77–94 (2010)

Guided Crowdsourcing for Collective Work Coordination in Corporate Environments

Ioanna Lykourentzou[1], Dimitrios J. Vergados[2], Katerina Papadaki[3], and Yannick Naudet[1]

[1] Centre de Recherche Public Henri Tudor, Luxembourg
{ioanna.lykourentzou,yannick.naudet}@tudor.lu
[2] Norwegian University of Science and Technology, Department of Telematics, Norway
dimitrios.vergados@item.ntnu.no
[3] Bank of Greece, Operational Risk Management, Greece
kpapadaki@bankofgreece.gr

Abstract. Crowdsourcing is increasingly gaining attention as one of the most promising forms of large-scale dynamic collective work. However current crowdsourcing approaches do not offer guarantees often demanded by consumers, for example regarding minimum quality, maximum cost or job accomplishment time. The problem appears to have a greater impact in corporate environments because in this case the above-mentioned performance guarantees directly affect its viability against competition. Guided crowdsourcing can be an alternative to overcome these issues. Guided crowdsourcing refers to the use of Artificial Intelligence methods to coordinate workers in crowdsourcing settings, in order to ensure collective performance goals such as quality, cost or time. In this paper, we investigate its potential and examine it on an evaluation setting tailored for intra and inter-corporate environments.

Keywords: crowdsourcing, crowd coordination, resource allocation.

1 Introduction

Crowdsourcing is a new form of user involvement on the Web. It has recently emerged as a new paradigm of collective work and as a natural result of the Web's evolution course, from a purely non-participatory system, with users in the place of content consumers, to a virtual space of full user involvement, since the Web 2.0 era and beyond.

Crowdsourcing refers to the splitting of a large, human-intelligence job into smaller micro-tasks and dynamically "outsourcing" these, not to specific individuals, but to an unknown crowd of web workers. Examples of jobs often accomplished through crowdsourcing include the translation of large corpuses of small sentences from one language to the other, the recognition of captchas, the transcription of audio files to text, but also the collective creation of articles in Wikipedia and the development of open source software artifacts by several distributed programmers [4].

The crowdsourcing technology increases rapidly. Having started only a few years ago, it is already being used at large-scale by commercial players, academics and individuals, who benefit from its ability to involve millions of users worldwide and to

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 90–99, 2013.

provide access to a scalable and on-demand workforce. Indicative of its prospective, crowdsourcing was recently included to the cycle of emerging technologies with significant foreseen potential, as predicted by professional technology watch firms like Gartner[1].

Despite its success, crowdsourcing has often been criticized for not providing guarantees critical for the requesters, such as minimum job quality, maximum cost and timeliness [6]. This is because the participating workers select the micro-tasks that they will work on with an aim to maximize individual and not system-level targets. For example workers in paid crowdsourcing seek to increase their individual profit by focusing on quantity rather than quality (i.e. submitting more in number rather than high-quality tasks). This inability to guarantee performance, and to do so simultaneously for multiple performance objectives, hinders the reliability of crowdsourcing and limits its potential. Especially for corporate environments, the above limitations make the corporate management even more skeptical in incorporating crowdsourcing approaches in vital organizational processes. Thus, recent research has started to identify the need of standardizing crowdsourcing [7] and improving in terms of system-level performance, using artificial intelligence (AI).

In this paper we present this new area of guided crowdsourcing, which can be defined as using AI methods to coordinate a user crowd towards achieving specific collective performance goals in a crowdsourcing setting. In section 2 we formulate the engineering of guided crowdsourcing solution as a 5-step process. In section 3 we present the main research streams in the area. In section 4 we showcase its capabilities for a specific application case, i.e. corporate environments. Finally, in section 5 we discuss the open research topics related to engineering efficient guided crowdsourcing solutions and conclude the paper.

2 Guided Crowdsourcing: A New Research Area for Crowdsourcing Optimization through AI-Based Coordination

2.1 Definition and Differences with Standard Crowdsourcing

Given the problems of current crowdsourcing, research has slowly started to consider what we may overall refer to as "Guided Crowdsourcing". Guided crowdsourcing can be briefly defined as *"the use of AI methods to coordinate and guide users participating in a crowdsourcing system towards achieving a collective result that meets specific performance standards, such as quality, timeliness or cost"*. The purpose of guided crowdsourcing is therefore to optimize the performance of the crowdsourcing system and provide quality, cost and time guarantees to the consumers.

Its difference with current, unguided crowdsourcing is that the latter is totally self-coordinated, with workers self-appointing themselves to the tasks that they wish to undertake, thus often resulting to poor performance results. In contrast, the coordination algorithms used in guided crowdsourcing are designed to affect the behavior of the

[1] Gartner's 2012 Hype Cycle for Emerging Technologies. Press Release. `http://www.gartner.com/it/page.jsp?id=2124315`

workers towards a specific direction, in order to achieve a specific crowdsourcing performance result.

Affecting user behavior can be implicit (increasing the price of certain tasks to make users prefer them over others) or explicit (recommending tasks to specific users).

2.2 Guided Crowdsourcing as a 5-Step Process

The basic elements that need to be defined to engineer a crowdsourcing solution out of a standard, unguided crowdsourcing system, are the following (2.2):

Goals. As goals, we define the performance aspects of the standard crowdsourcing system, which the guided approach targets at improving. They can include, as a non-exhaustive list, maximizing the quality of the accomplished jobs, minimizing the cost for each job, and meeting the deadlines of each job.

Jobs. For each crowdsourced job, we need to define one or more performance characteristics, based on the system's performance goals. These can include the jobs current quality level, current cost, deadline, as well as other more specific traits. The measurement of each characteristic of the job is either global, like its deadline, or an aggregation of the characteristics of the jobs micro-tasks. For example, if we assume a job of translating a corpus of sentences, the job current quality is the sum of the quality of the translation of each of the sentences (tasks) that have already been translated.

Workers. The workers participating in the standard crowdsourcing system are in fact its resources. For each worker, specific skills should be defined in relation to the system's goals. Such skills can for instance include the worker's expertise in relation to a knowledge-intensive task, task fulfillment speed, accuracy, judgment ability, as well as the minimum wage he would require to accomplish a given micro-task. The estimation of worker skills can be addressed through learning mechanisms, for example neural networks as used in [11].

Constraints. The constraints are the inherent characteristics of the crowdsourcing system that the guided crowdsourcing solution needs to respect. They can include: the number of micro-tasks that each worker is allowed to undertake in a given amount of time, the ability of the system to bargain or not with the workers for the price that each micro-task pays, the ability to interrupt a worker with a new task in case this suits better the objectives of the system, as well as many others.

Coordination Algorithm. As also mentioned in the definition of guided crowdsourcing, the coordination algorithm is what distinguishes the guided from a simple, standard crowdsourcing system. Overall, and following the problem formulation set above, the target of the coordination algorithm is to fulfill the objectives of the crowdsourcing system, for the amount of jobs requested, with the available workers, while respecting the crowdsourcing systems constraints. The coordination algorithm therefore is in fact an optimization technique, over the global performance of the crowdsourcing system. Depending on the exact parameterization made on each of the previous steps, different methods can be used for the design of the coordination algorithm, including queue theory, mechanism design or resource allocation, as described in the related literature section that follows.

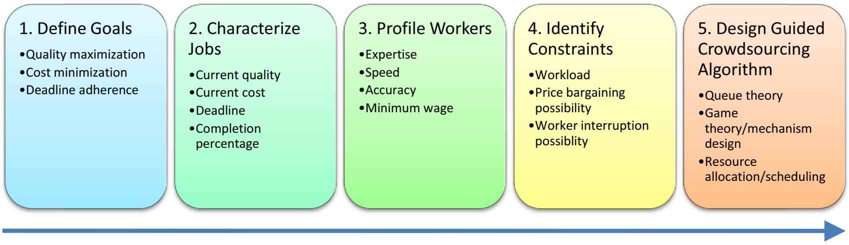


Fig. 1. The process of engineering a guided crowdsourcing solution comprises 5 main steps

3 Related Literature: Current Trends in Guided Crowdsourcing Algorithms

A queue theory-based analytical method is proposed in [3], for optimizing crowdsourcing in terms of cost. This work focuses on open crowdsourcing, with very high worker and job arrival rates, in which cost is measured in terms of task loss, i.e. those that upon arrival find no available workers to undertake them. The objective of the algorithm in this case is to calculate the optimal cost tradeoff between artificial worker retainment (paying workers to remain in the system until a task arrives) and task loss.

Game theoretic approaches are also used for the design of mechanisms that will optimize the functionality of markets with strategic resources, like the ones that emerge in crowdsourcing applications, in terms of cost. Indicative of this research stream, the works of Ghosh et al. [5] and Archak et al. [1] examine the conditions under which the implementation of contest-based mechanisms among users can reduce cost in crowdsourcing environments

Finally, resource scheduling and allocation approaches have also started to be used for the collective performance improvement of crowdsourcing systems. Psaier et al. [13] examine the improvement of task assignment in crowdsourcing environments by combining a hard/soft resource scheduling algorithm with a mediator responsible of monitoring user skills, organizing activities, settling agreements and scheduling tasks. The algorithm assumes a push crowdsourcing model, i.e. it actively sends requests to workers for the crowdsourcing tasks that need to be completed. Results obtained through simulation for various scenarios show that the algorithm produces better quality in comparison to plain random scheduling, while keeping overall task load within the set limits. In a similar spirit, Khazankin et al. [8] work with scheduled crowdsourcing, in QoS (Quality of Service)-sensitive processes. In their approach an algorithm receives tasks from ordering customers, negotiates with them for quality and temporal job requirements and once an agreement is reached, it distributes the job tasks to appropriate members from the crowd pool. Results of examining the prediction algorithm in a simulated crowdsourcing environment showed that it can efficiently predict the quality capabilities of the crowdsourced workers and therefore provide ordering customers with satisfactory quality guarantees.

4 An Application Study for Corporate Environments

4.1 Corporate Crowdsourcing: A Special Case with High Value Potential

Corporate crowdsourcing occurs when crowdsourcing is applied, instead of web workers, to the human network of a company. The main advantage of intra-corporate crowdsourcing is that it permeates the traditional departmental corporate structure, which often hinders the efficient use of human resources. In addition, it is dynamic and it can be used for on-demand tasks that the company might not want to invest with full-time dedicated resources, because of their short term nature. The notion of corporate crowdsourcing can also be extended from intra- to inter-corporate environments, i.e. the borrowing of specialized employees among companies for limited period of time.

The main differences between corporate and open crowdsourcing are three. First, corporate crowdsourcing focuses on knowledge-intensive rather than simple tasks. This is because instead of automatizing simple tasks (such as image recognition), what companies need more from crowdsourcing is to tap the innovation and knowledge creation potential of their human resource employee network [9] (example case: idea gathering for new product development). Indeed crowds can provide a much larger diversity of ideas, compared to individual experts usually hired by companies for knowledge creation, because individuals make their suggestions independently and based on more diverse knowledge backgrounds [12]. Secondly, corporate crowdsourcing allows for lower cost, because the company does not need to compete globally with others for the same worker, as it would be the case of hiring external freelancers through open crowdsourcing. Finally, the case of corporate crowdsourcing allows defining a simpler problem setting compared to open crowdsourcing, since here we can assume the presence of a fixed, easier-to-profile pool of workers and ensured worker acceptance on system recommendations. In other words, when contributing people belong to an organization, their profile (including competencies and certain motivation factors) and schedule can be known. Also, constraints such as mandatory time dedicated to contributions to a crowd-sourced problem solving can be imposed by the organization.

All of the above make corporate crowdsourcing platforms more suitable for the guided crowdsourcing approach, since a lot of information can be exploited to actually drive it. In open environments, not only is the information about workers less, but also, they are always free to refuse task assignments. Open crowdsourcing platforms can obviously be extended to gather information from people (e.g. profile, competencies, schedule, etc.), which allows guiding. However, the motivation aspect, as a means to ensure participation, needs to be taken much more into consideration in the open than in the corporate crowdsourcing case.

4.2 Problem Instantiation and Modeling

The following environment instantiates the generic guided crowdsourcing process to the specificities of a corporate crowdsourcing problem. We model the following elements:

Jobs. $J = \{j_1, \dots, j_{|J|}\}$ is the set of jobs to be crowdsourced. Each job j_i comprises a:

- Set of n micro-tasks. Each micro-task has a quality q_j, measured in the [0,1] scale, with 1 meaning perfect quality and 0 meaning no quality at all.

- Quality Q_j, the quality of the job calculated as the average quality of its micro-tasks: $Q_i = E[q_j]$, $j \in [1, |J|]$.
- Maximum cost limit C_j, which the enterprise is willing to pay for the job accomplishment. The cost of each job is initialized randomly at the beginning of the simulation.

Finally, each job belongs to one of $D_j \in D = \{D_1, D_2, \ldots, D_{|D|}\}$ "expertise domains", with each domain indicating a specific category of corporate knowledge.

Workers. We assume a population of $K = \{k_1, k_2, \ldots, k_{|K|}\}$ workers, who model the employees of the corporation participating in the crowdsourcing system. In contrast to open crowdsourcing we do not assume an infinite crowd but a large but finite pool of people. Each user k_i has:

- Expertise e_i, in each of the simulated expertise domains, measured in the [0,1] range. The quality that an employee contributes to a task is equal to his expertise on the task's domain.
- Speed s_i, i.e. time needed to accomplish a task per domain.
- Minimum "wage" w_i, below which the worker does not accept a micro-task. Since we assume a corporate environment, this wage is not necessarily monetary, but can also be "points" translatable into performance bonuses, days off or other "gamification related" rewards like charity from the part of the company.

Goals. The objectives that the guided crowdsourcing system needs to fulfill, for the specific problem setting, are to:

- Maximize the average quality of the accomplished jobs: $O_1 = max_{j \in |J|} E[Q_j]$
- Minimize the average paid cost: $O_2 = min_{j \in |J|} E[C_j]$

Constraints/Organizational Policies. The constraints depend on the organizational policies that the involved corporation(s) need to pose. For the modeled problem setting we define the constraints of:

- Maximum price. The total paid for each job cannot surpass the maximum cost set for the job: $\Sigma_{i=1}^{m} w_i \leq C_j$, $\forall j \in |J|$, $m \leq n$, where m is the total number of workers that have been given one of the n micro-tasks of the job,
- Non-preemptiveness, i.e. once an employee is occupied with a task they do not enter the system, i.e. they cannot be interrupted to undertake a new task.

Implemented Guided Crowdsourcing Algorithm. We propose a guided crowdsourcing algorithm, which uses resource scheduling to dynamically assign micro-tasks to employees according to their individual expertise and inter/intra wage. Given the objectives and constraints of the specific scenario setting (cost minimization and quality maximization), the algorithm suggests to each worker the tasks that pay less from the expertise domain that the worker is mostly expert at. Partially complete jobs (those with at least one task completed) are also preferred, starting from those with the least completion percentage, to boost job completion rate. The above problem modeling is summarized in Table 1.

Table 1. The corporate scenario problem, modeled as an instantiation of the generic 5-step guided crowdsourcing process

Problem element	Value
Goal	Maximize average job quality Minimize total cost
Workers	Expertise Speed Minimum accepted wage
Jobs	Number of micro-tasks Quality Cost
Constraints	Maximum job cost limit Non-preemptiveness

Guided crowdsourcing algorithm	For every worker that arrives:
	{
	1. Rank domains of users expertise in descending order
	2. Select first domain on list
	3. Select partially completed jobs from that domain
	4. Rank the tasks of the selected jobs in ascending cost order
	5. Allocate first task on the list
	}

4.3 Evaluation

First we parameterize the variables of the above problem modeling (Table 2). Corporate crowdsourcing is expected to work mostly with knowledge-intensive rather than simple tasks, as mentioned above. Therefore, for the selection of the number of users, worker and job arrival rates and total simulation time, we data-mine a real-world system focusing on the crowdsourcing of knowledge-intensive tasks, namely the Data Hub[2]. The extracted dataset covers a timespan of 67 months and features the contributions of 1600 users, and therefore we set the simulation time equal to 67 simulation units and the simulated population to 1600 workers. Worker expertise is initialized using a beta distribution function, calibrated so that, for each domain, few people are experts and there is long tail of semi or non-experts, as it is the typical case of expertise distribution in enterprise corpora [2]. Worker speed per domain is initialized randomly through a uniform distribution. Worker wage is linearly analogous to expertise (i.e. the more expert a worker, the higher wage they require to fulfill a task). Wage also depends on whether the employee will work for his company (intra-crowdsourcing) or as external worker for another company (inter-crowdsourcing). For intra-crowdsourced work it is set equal to ones expertise, while it doubles if the worker is externally hired.

The interaction of workers with jobs is performed as follows: Workers enter the crowdsourcing platform with an arrival rate λ and jobs are generated with a generation rate μ. Both rates increase exponentially with time. As soon as a worker enters the platform they select a micro-task to fulfil. This selection depends on whether the simulated system works under a guided or an unguided crowdsourcing manner. In the unguided

[2] `http://datahub.io/`

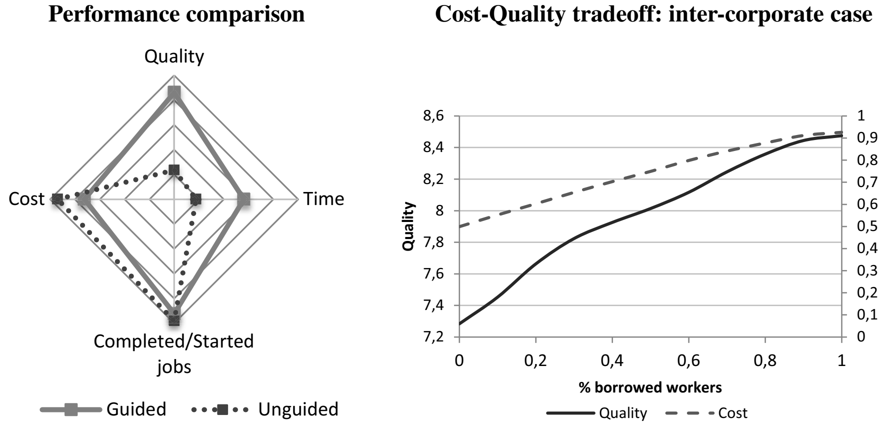
Table 2. Parameters used for the evaluation

Parameter	Value
Simulation time	67 simulation units
Workers	1600
Domains	20
Tasks per job	3
Job arrival rate μ	$\alpha \cdot e^{\beta t}$, with $\alpha = 130$ and $\beta = 0.05$
Maximum job cost	[0, 2] according to uniform distribution
Worker arrival rate λ	$\gamma \cdot e^{\delta t}$, with $\gamma = 30$ and $\delta = 0.05$
Worker wage	ρ·expertise, with $\rho = 1$
Worker speed	[0, 1], according to uniform distribution
Worker expertise	[0,1] according to beta distribution
Worker wage	α·expertise, with $\begin{cases} \alpha = 1, \text{ if internally hired} \\ \alpha = 2, \text{ if externally hired} \end{cases}$
Companies	50 (for the inter-corporate scenario)

version, which serves as our benchmark, people seek to maximize their individual profit and therefore they select the task that pays the most, from the ones that surpass their minimum requested wage. In the guided version of the system they can select only among tasks that are recommended to them by the guided crowdsourcing algorithm.

Scenario 1. Intra-corporate Crowdsourcing. In the first scenario all employees belong to the same organization. We examine the performance of the guided crowdsourcing algorithm according to four criteria: average quality, cost, completed versus started tasks, and time until completion (Fig. 2a, with all results normalized to the [0,1] scale). As it can be observed, it performs better than the unguided system in terms of quality (0.86 instead of 0.23) and cost (0.72 instead of 0.93), which are the two criteria that the algorithm is designed to optimize. The guided crowdsourcing algorithm also achieves to keep the completion rate of finished versus started jobs at comparable levels (0.92) with that of the unguided system (0.98). However, the above come at the cost of timeliness (0.56 instead of 0.18 average time units), since the algorithm gives each worker the task that he is most expert at, therefore "spreading" user contributions across jobs, in comparison to the unguided system where users all target the same, high-paying jobs.

Scenario 2. Inter-corporate Crowdsourcing. In the second scenario, workers belong to multiple companies, which can lend them to one another. In this case, for each worker we assume two wages: an intra-corporate one, equal to the workers expertise, and an inter-corporate one, which is double the intra-corporate one. Accordingly, each company has an upper limit for the percentage of employees it can borrow externally. We simulate 50 companies and examine the average quality gained and the extra cost paid, for different upper employee borrowing limits. As it may be observed, and intuitively expected, the more employees a company borrows the more qualitative tasks it achieves, but at a higher cost (Fig. 2b). Therefore, although guided crowdsourcing can be used to augment job quality, often significantly, it remains at the disposal of each organization, to determine the best tradeoff suitable for its crowdsourcing needs, according to its needs, expertise availability and cost constraints.


Performance comparison

Cost-Quality tradeoff: inter-corporate case

Fig. 2. a) Performance comparison between the examined guided crowdsourcing algorithm vs. unguided crowdsourcing. All axes are presented in % ratios in respect to their maximum value. b) The tradeoff between cost and quality for inter-corporate crowdsourcing.

5 Conclusion and Perspectives

Guided crowdsourcing is a new, emerging domain with high potential. It refers to the optimization of the performance of crowdsourcing systems, in terms of quality, cost and timeliness, by using AI-based methods to coordinate the involved user crowd. In this paper we present the notion of guided crowdsourcing, formulate the process of engineering a guided crowdsourcing solution as a 5-step process, present the main research streams in the area and examine its potential on the application case of inter and intra corporate crowdsourcing. Results are promising, indicating that guided crowdsourcing can help achieve better performance in comparison to typical unguided crowdsourcing.

A number of topics need to be further investigated. Firstly, in the problem instantiation treated in this paper a fixed pool of workers and jobs is assumed. This may hold true for certain cases, like the corporate one (where the worker/job pool is either fixed or predictable with high accuracy), but in other environments, like open crowdsourcing, there is a need to model the uncertainty in the size and availability of the worker pool, as well as on the load of job demands. In this case, a dynamic scheduling problem formulation and algorithm may be more appropriate. Also, in contrast to corporate environments (where worker expertise can be considered known or easy to obtain, e.g. using data from the employees' job description or previously undertaken tasks within the organization), an extension of the proposed approach to open crowdsourcing environments would necessitate the incorporation of an adaptive skill learning mechanism, such as the one proposed in [10]. Other optimization goals may also be considered, regarding corporate crowdsourcing. For example, instead of maximizing average quality while remaining under a certain cost limit, a company might prefer to minimize the total project cost and keep a minimum quality baseline or optimize a weighted combination of both, which would pose the need to define and solve the problem as multi-objective resource scheduling. The proposed approach can also be compared to more benchmarks, extending the comparison with the fully unguided benchmark used in this paper. These benchmarks may include the partial filtering of workers based on the quality of their previous contributions (e.g. for certain types of jobs in Amazon Mechanical Turk, the requesters can allow the participation of only certain ”qualified” workers). It would be also interesting to compare the proposed approach with the algorithms proposed by other studies of the related literature, and in particular those that use resource allocation as their main algorithmic technique. Finally, further research needs to consider the broader context of integrating guided crowdsourcing in the enterprise, investigating issues related to internal regulation changes that are necessary to accomodate in-house crowdsourcing, ethical issues, as well as the topic of incentive engineering.

Summarizing, guided crowdsourcing is a technology and research area with significant potential, but also with much room for improvement, both in terms of algorithmic efficacy, as well as in terms of harmonization with the human factor that it entails.

References

1. Archak, N.: Optimal design of crowdsourcing contests. In: ICIS 2009 Proceedings, vol. 200(512), p. 16 (2009)
2. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006, pp. 43–50. ACM, New York (2006)
3. Bernstein, M.S., Karger, D.R., Miller, R.C., Brandt, J.: Analytic methods for optimizing realtime crowdsourcing. CoRR, abs/1204.2995 (2012)
4. Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the world-wide web. Commun. ACM 54(4), 86–96 (2011)
5. Ghosh, A., Hummel, P.: Implementing optimal outcomes in social computing: a game-theoretic approach. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012, pp. 539–548. ACM, New York (2012)
6. Ipeirotis, P.G.: Analyzing the amazon mechanical turk marketplace. XRDS 17(2), 16–21 (2010)
7. Ipeirotis, P.G., Horton, J.J.: The need for standardization in crowdsourcing. In: CHI 2011 Crowdsource Workshop, pp. 1–4 (2011)
8. Khazankin, R., Schall, D., Dustdar, S.: Predicting QoS in scheduled crowdsourcing. In: Ralyté, J., Franch, X., Brinkkemper, S., Wrycza, S. (eds.) CAiSE 2012. LNCS, vol. 7328, pp. 460–472. Springer, Heidelberg (2012)
9. Kittur, A.: Crowdsourcing, collaboration and creativity. XRDS 17(2), 22–26 (2010)
10. Liu, X., Lu, M., Ooi, B.C., Shen, Y., Wu, S., Zhang, M.: Cdas: a crowdsourcing data analytics system. Proc. VLDB Endow. 5(10), 1040–1051 (2012)
11. Lykourentzou, I., Papadaki, K., Vergados, D.J., Polemi, D., Loumos, V.: Corpwiki: A self-regulating wiki to promote corporate collective intelligence through expert peer matching. Inf. Sci. 180(1), 18–38 (2010)
12. Poetz, M.K., Schreier, M.: The value of crowdsourcing: Can users really compete with professionals in generating new product ideas? Journal of Product Innovation Management 29(2), 245–256 (2012)
13. Psaier, H., Skopik, F., Schall, D., Dustdar, S.: Resource and agreement management in dynamic crowdcomputing environments. In: 2011 15th IEEE International Enterprise Distributed Object Computing Conference (EDOC), pp. 193–202 (2011)

Google Trends for Data Mining. Study of Czech Towns

Jiří Horák, Igor Ivan, Pavel Kukuliač, Tomáš Inspektor,
Branislav Devečka, and Markéta Návratová

Institute of Geoinformatics, Faculty of Mining and Geology, VSB – Technical University
of Ostrava, 17. listopadu 15, 708 33, Ostrava-Poruba, Czech Republic
{jiri.horak,igor.ivan,pavel.kukuliac,tomas.inspektor,
branislav.devecka,marketa.navratova}@vsb.cz

Abstract. Selected web search engines provide statistics of user activities according to the topics, time and locations. The utilization requires well prepared phrases and searching range. The system of etalons for calibration searching frequencies provided by Google Trends is proposed. It was applied for evaluation of searching names of Czech towns. The regression analysis proved high correlation with population. Highlighted anomalies were explored. K-means cluster analysis enabled a categorization of selected towns. The geographical network analysis of relationships among towns suffers from low quality of locations provided by Google. The discussion includes an overview of main pros and cons of Google Trends and provides recommendations.

Keywords: Google, web search engine, data mining, Czech towns.

1 Introduction

Nowadays, the Internet is the primary information source, although many authors criticise the reliability of acquired information. The number of Internet users grows rapidly. The share of Internet users in the total population is about 35% [1], five times more than 10 years ago. Over the last five years, developing countries have increased their share of the world's total number of Internet users from 44% in 2006 to 62% in 2011 [1]. Nevertheless, large differences still endure between countries. Internet users represent ¾ of population 16+ in developed countries, while only ¼ in developing countries [2]. In 2011, the Czech Republic (hereinafter CR) reached the European average in the percentage of Internet users in adult population (73% for population aged 16-74, which represents 5.8 millions of users). In 2005, the share was only 35% of the population over the age of 15[3].

Along with the growing use of the Internet, the use of search engines for obtaining information is also increasing. To locate online information or services, more than 80% of users who were searching for information on websites [4] used a search engine; now it is even higher.

The Internet in the CR is utilized mainly for communication and searching information [2]. The topics most frequently searched for are goods and services (4.5 millions of Internet users in 2011) and travelling and accommodation (3 millions of Internet users in 2011).

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 100–109, 2013.

The search engine providers keep records of the frequencies of particular keyword phrases and related information such as time, associated topics or geographical locations where the search was requested. In 2008, Google made the search statistics available using the web application Google Insights for Search (hereinafter GI). Although, GI was merged with Google Trends (GT) in October 2012, all the features of searching statistics have remained the same. The application allows researchers to assess the frequency of a particular search query across time, geographical units and to obtain associated phrases. Such statistics provide new possibilities for investigating behaviour and interests of users, and to a certain extent, give evidence about the use of the Internet, the required information and perception of the target objects.

This data source is utilized in the health sector where, for example, Breyer et al. [5] demonstrated the presence of a correlation between the geographic and temporal distribution of searches for information about kidney stones, and the factual known occurrences of kidney stones in the population. Furthermore, Brownstein et al. [6] proved a strong correlation between the frequency of searches of the bacteria salmonella and of the current number of Sallmonella Typhimurium infections in the U.S. Ibuka et al. [7] used GI as one of the data sources to analyse the dynamics of risk perception due to the imminent pandemic H1N1 virus in 2009 and demonstrated the rapid growth of search keywords across the U.S. within 14 days and the subsequent rapid return to the original low values. Similarly Ginsberg et al. [8] tried to estimate the weekly activities of the influenza virus. The data of GI is commonly used in the field of marketing and can be one of the important sources for a variety of business strategies (i.e. [9]). Webb [10] proved the correlation between a search term „domestic execution“ and the actual number of home executions in the U.S. GT (former GI) is currently used in many other disciplines as well. However, it is necessary to take into account the specific aspects of the data sources that are discussed further in the paper. The aim of the paper is to introduce Google Trends as a data source for data mining and demonstrate the utilization in the case study of Czech towns.

2 Web Search Engines and Google Trends

The most popular web search engines are Google (80.32 % of the global market), Bing (9.35%) and Yahoo! Search (7.07%) (Sep 2012 - Feb 2013; [11] StatOwl, 2013). Locally other search engines may play an important role (i.e. „Seznam“ (www.seznam.cz) in CR). The share of Google in CR is about 51% (Jan 2011) and it continuously grows, while the rest of the market is almost fully covered by Seznam.

Searching algorithms of Google enable the usage of two basic types of quantitative cyberspace analysis of selected objects [12]. The Web Content Analysis records the absolute numbers of hits or hyperlinks for selected object and ranks it accordingly. The Web Activity Analysis (WAA), on the other hand, illustrates the users' interest in information on these topics across time. The search engine supports such analysis, providing statistics of the users' activity (GI, later GT). Also other web search engines (i.e. Seznam, Yahoo! Search) provide statistics of user activities. I.e. Seznam offers outputs of statistics (for exact matching or extended searching) in absolute values portrayed in a graph, but this data is limited to the period of two last months and no data export is available. Currently, Google Trends seems to be the most advanced tool. The WAA requires conducting following steps.

First it is necessary to select appropriate terms for identifying searched objects. If the searched term consists of more words, the statistics provides the frequency of user's queries containing all words in an arbitrary order (Boolean AND between words). Upper-case and lower-case letters are not distinguished. Non-Latin characters are not matched with equivalents in Latin alphabet (if matching exists), so the searched term with diacritical characters is different from the one with equivalent non-diacritical characters (resulting in different statistics).

Searching can be filtered by temporal range (by months), by type of source (web, images, news, products), by geographical units of entering queries (countries and regions, usually NUTS2), and by thematic categories (25 categories, but not available for all countries including CR).

Google outputs of WAA provide only normalised and scaled data. The normalisation means recalculating data according to the common variable in the searched regions to eliminate different bases in regions [13]. After that, the data is rescaled to the range 0-100. Each value is divided by the maximal value within the range of searching and then multiplied by 100. Finally, the mean frequency of searching (MFS) represents an average of all values in the given range of searching.

Outputs can be obtained in the form of time series (fig. 1), geographical statistics (countries, regions) or thematic statistics (the most frequent phrases). Customised data processing and evaluation is enabled using data exported into CSV format.

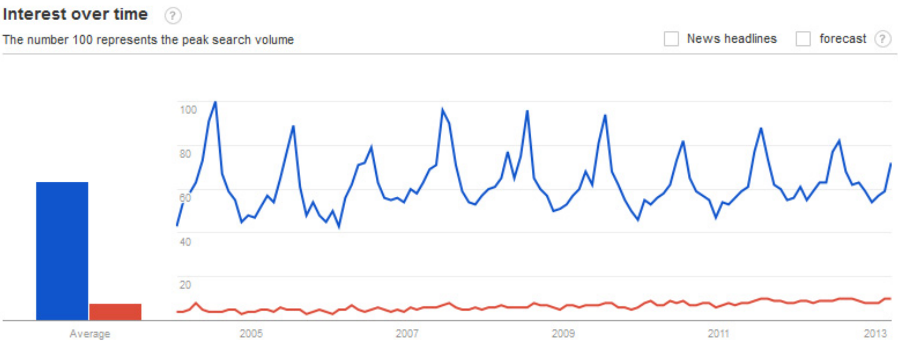


Fig. 1. Frequency of searching terms for two Czech spas (the upper curve for *karlovy vary*, the lower curve for *mariánské lázně*) by Google Trends

Google Trends only provides statistics for searched terms above a certain (but not known) threshold of frequency. The system also eliminates repeated queries from a single user (a single IP) over a short period of time. The credibility of Google Trends was tested using automated generating requests for two queries („Cimrman" to test possibility to increase existing results, and „8gh5a7sw1r" to establish a new statistics for a meaningless query) using WAPT software (SoftLogica LLC). WAPT is a load and stress testing tool for testing web sites and intranet applications with web interfaces. Google have recognised this artificial attempt and rejected 5/6 of requests. Remaining answered queries (12331, resp. 47992) had no impact on final statistics.

During one session it is possible to search for up to five terms (or groups of terms). If more terms are required to evaluate, it seems necessary to use a system of established **etalons** and to calibrate obtained results to a common scale. Good candidates for the etalon's role only contain Latin characters and provide stable results in time. The most frequent searched terms should be used as first etalon, while the last etalon should be a term with MFS close to minimal value (to exclude extrapolations). Other etalons should be distributed equally – we recommend that MFSs of following etalons should be close to ratio 1:10. To create the common scale, GT is used to obtain the value of MFS for each etalon (MFS_{etalon}) in the comparison of searching frequency with the superior etalon. Next, calibrated values CFS are enumerated using an arbitrary large value as a maximum of the scale (table 1). Calibration coefficients CC_{etalon} are calculated by dividing CFS_{etalon} with MFS_{etalon}. After the establishing system of etalons, the required terms are searched together with etalons and the obtained results for terms (MFS_{term}) are calibrated according to (1).

$$CFS_{term} = CC_{etalon} * MFS_{term} \qquad (1)$$

where CFS_{term} is Calibrated Frequency of the Searched term, CC_{etalon} is a calibration coefficient for the applied etalon and MFS_{term} is the Mean Frequency of the Searched term (Google output for the term from synchronous searching it and the etalon).

Table 1. System of etalons and calculation of CFSs

1st etalon	2nd etalon	$MFS_{1st.etalon}$	$MFS_{2nd.etalon}$	$CFS_{1st.etalon}$	CC_{etalon}
praha				100 000	
pardubice	praha	8	79	10 127	1 265.875
poruba	pardubice	8	80	1 013	126.625
mohelnice	poruba	37	76	493	13.324
radonice	mohelnice	6	58	51	8.5

3 Case Study: Evaluation of Interest in Czech Towns

The case study is focused on the evaluation of interest in Czech towns (with population above 10 thousand) due to the possibility of comparing Google statistics with a theoretical model and also exploiting Google identification of the entering queries' location. The object of the study is Czech towns. The geographical filter for searching was set to the CR to exclude queries from abroad. The temporal range represents a period from November 2004 to October 2010. The searched term (or the group of terms) for a town was always coupled with an etalon used for calibration.

The system of etalons was based on geographical names. We expect improved consistency of results, using minimal and maximal values in data series and better understanding of readers. The temporal stability of selected etalons was tested. During one month the requests for frequency of searching etalons in pairs (*praha* and *brno*, *praha* and *pardubice*, *pardubice* and *poruba*, *poruba* and *mohelnice*, *mohelnice* and *radonice*) were repeated every day. The normalised and scaled data was calibrated to obtain CFS for etalons. Differences among CFS for etalons enumerated for every day of searching were evaluated using the coefficients of variation. These values are quite low (0.6-7.6%), so proposed etalons are considered as stable in time.

The statistics provided by GI for names of Czech towns in Google for six-year long period were explored and analysed according to the frequency and associated topics. Data was calibrated using the system of six etalons.

To obtain relevant statistics it is needed to prepare appropriate searched terms. The terms should enable selecting and combining keywords, best emulating users' behaviour in entering queries for searched towns. The occurrence of diacritical characters in towns' names was solved by creating terms which consist of diacritical and non-diacritical variants of names. The similar composition was used for names including a hyphen (original name + names of parts). In several cases, the composition of the most popular keywords for the town was used (i.e. name of the town + popular objects like cinema, shopping centre). The selection of appropriate meaning from the set of synonyms was based on subtraction (the ambiguous name minus keywords identifying other inappropriate meanings). Such excluding keywords were selected from the list of the most searched topics related to the ambiguous name. Long names were shortened to the most frequent lexical form (i.e. Dvůr Králové nad Labem → <dvůr králové+dvur kralove>). Any substitution of original names obviously increases uncertainty of the results, which is necessary to take into account during interpretation.

The Google statistics for the searched terms were exported, stored to database and then processed. MFS for searched term was calibrated according to (1). Enumerated CFSs were used for the **quantitative evaluation**.

The values of CFS for Czech towns depicted in the map (fig. 2) indicate an existing correlation with the local population. The suitability of the population model was proved by Spearman r = 0.81 (<0.72,0.87> with p=0.05) and by logical reasoning - a higher population generates a higher number of queries, part of them aimed at objects inside the home city, plus more populated places are equipped with more facilities attracting searching.

The regression analysis confirmed this relationship (fig. 3). The highest coefficient of determination is for cubic relationship (R^2=0.966), but the shape is spurious. Both quadratic and linear relationship provide satisfactory results ($R^2_{\text{quadratic}}$=0.952; R^2_{linear}=0.936). Towns were classified using 95% and 90% confidence intervals of the linear regression and also according to the distance from the regression line. Small towns were evaluated using differences in the ordered lists of population and CFS. Results of analysis help to distinguish towns deviating from the common population model and focus on explanations of results in these cases.

The overall quantitative evaluation was supplemented by a **thematic evaluation**. Google does not provide a categorization of searched terms in CR. The thematic classification was conducted using the text analysis of the most frequent phrases of users' queries (provided by Google Trends), related to searched terms.

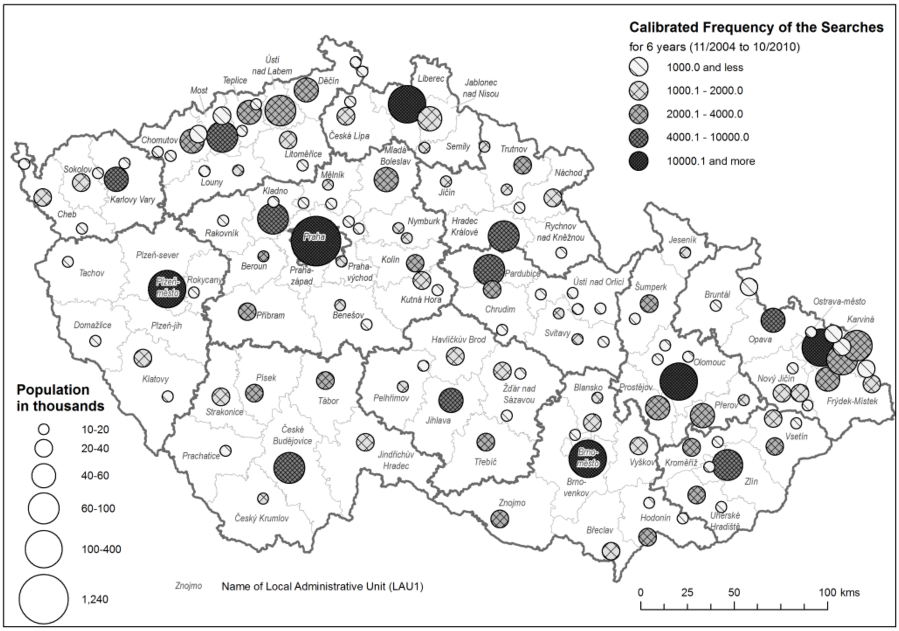


Fig. 2. The Calibrated Frequency of the Searched names of Czech towns by Google

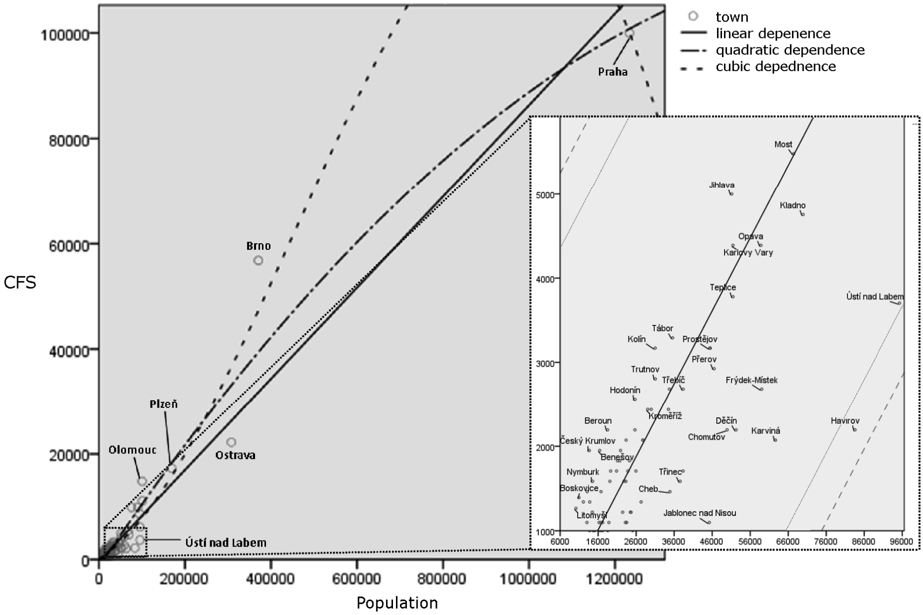


Fig. 3. The dependence of CFS and the population for Czech towns, the right detail with a linear regression and the confidence intervals 90 and 95 % identifying the most deviating towns

K-means clustering method was used to explore the internal groups of regional centres with similar searching topics. The first cluster contains only one town, Ústí nad Labem, with a high share (42.7 %) of searched terms unrelated to any of the four main categories. Prague is the only member of the second cluster, specific for the reason of searching for topics generally entitled as common services (i.e. offices, public transport) and extra services (mainly related to travelling like hotels, restaurants, maps, airport etc.). Both clusters have a similarly large portion of searching common services. The last two clusters contain eleven towns with similar number of searches related to carrier (i.e. job, universities) and common services. Additionally, the third cluster is specific for the higher rate of searches to extra services (similarly to Prague). This cluster is not homogenous; Karlovy Vary is the most deviating town inside this cluster with the highest level of searches from the extra services category. This high level of searches in towns classified to this group is also related to leisure activities (i.e. cinema, theatre, shops) but it is not as high as in the case of six towns belonging to the fourth cluster. This cluster is typical for very high rate of searches related to leisure activities (almost one half of all searches).

Table 2. Share of searched topics of clusters of Czech regional centres (values in %)

Cluster	1	2	3	4
Members	Ústí nad Labem	Praha	Brno, České Budějovice, Hradec Králové, Zlín, Karlovy Vary	Olomouc, Ostrava, Pardubice, Plzeň, Jihlava, Liberec
Carrier	12.821	8.185	21.902	23.214
Common.Services	31.624	58.363	16.992	15.142
Extra.Services	0.000	14.235	15.816	8.414
Leisure	12.821	7.829	35.231	48.311
Others	42.735	11.388	10.059	4.919

The **geographic evaluation** utilizes the statistics for locations where the users enter the query. Standardised data of MFS enables to create flowline maps (fig. 4) and to analyse relationships among places, especially a regional share of searches or influences of large cities (Prague, Brno etc.). Unfortunately, the domestic location of query identified by Google is not satisfactory yet (see discussion later). To eliminate issues, we recommend to evaluate aggregated results. One option is to assess an internal factor (the perception of Czech towns by residents) and an external factor (the interest of non-resident Google users) according to associated topics and the share of queries located by Google in the target town to all queries.

4 Discussion

The main advantages of Google Trends utilization are the objectiveness (most of the users do not know about monitoring and their behaviours respect a natural setting), availability (public data, fully accessible, free of charge), continual monitoring, large volume of respondents, timeliness (immediate publication of statistics), readability of searched results, internationality (territorial comparisons, searched terms in different languages, different targets). Scheitle [14] emphasizes that, unlike traditional surveys, search engine data are non-invasive and do not force users into particular response categories. It also eliminates Hawthorne effect.

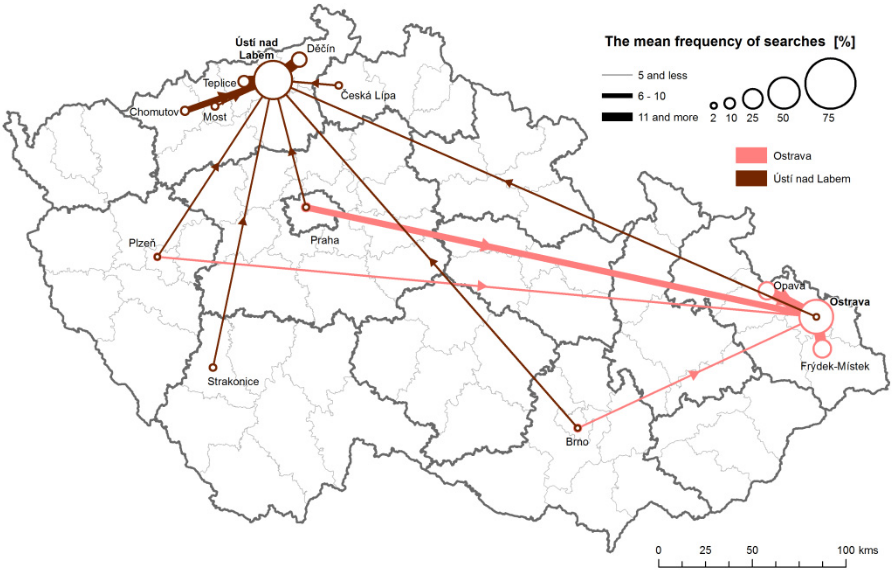


Fig. 4. The locations from where users search terms *Usti nad Labem* and *Ostrava*

Nevertheless, there are also many serious drawbacks which need to be taken into account during the research. The statistics do not inform about users of other web search engines and do not include other ways to reach the required information like usage of direct web links, call numbers for information etc. Of course, it is difficult to approximate results for non-internet users or the whole population. Systematic biases issued mainly from differences in the internet population according to geographical factors (i.e. rural versus urban areas), wealth, age and education. The basic recommendation is to prefer relative comparisons (relative thematic, territorial or temporal differences). The geographical biases in our case study do not seem to be significant, due to small differences between towns in two investigated population categories (66.8 % for cities with population above 50000, 62.6 % for towns populated between 10 000 and 49 999 (CSU, 2010)).

Outputs are normalized and scaled, intended for relative comparisons inside the territorial, temporal and thematic range. The maximum of five searched terms (groups) in one session limits research capabilities. We have proposed to use a system of etalons for calibration.

Outputs strongly depend on the way of searching. It is necessary to effectively elaborate the strategy of asking GT, especially to test various searched terms and analyse provided results with the aim to develop an appropriate system of terms and parameterization of GT. International comparisons are complicated due to the existing language dependency (see differences between English and Arabic in Al-Eroud et al. [15]). Currently, English is preferred in querying, but in the close future, Google envisages to separate the language and the content to reach a language independent evaluation of users' interest.

The detail geographical evaluation is limited due to the low accuracy of the user location provided by Google. In the case of cable connections, the location depends on the precision and the accuracy of network configuration description. The errors and lack of details lead to raise anomalies especially for the municipal level. The wireless connection to Internet significantly improves location of users. Mobile devices can be located using a GPS module, the strength of signal from available Basic Transfer Stations, or using WiFi module and identification of SSID and MAC addresses around [16]. In July 2011, Google announced the plans to improve the geographical location of all data since the beginning of 2011. We recommend to test and verify provided locations before using local statistics.

5 Conclusion

Web Activity Analysis of web search engines represents a new source of information valuable to the social research. The utilization of Google Trends requires an elaborated strategy based on refined searched phrases and well established temporal, geographical and thematic ranges. Extended research may use the proposed system of etalons, suitable for calibrating frequencies of searched terms. The credibility of Google Trends outputs was verified by testing.

The case study of Google searches analysis of names of Czech towns demonstrates the process of research preparation, processing of primary data and final evaluation. The quantitative part of evaluation includes regression analysis of the size of the population, identification and interpretation of deviations from the basic theoretical model. The processing of large volume of textual terms requires multidimensional data mining tools. K-means cluster analysis enabled a categorization of selected towns into main classes, one of them with a large share of extra services linked especially to travelling and the second one with dominating searching of leisure activities.

Finally, the geographical analysis tried to understand network relationships among towns. It was discovered that the quality of location places where user entered his/her search into Google is not satisfactory yet, especially for small places. The integrated evaluation helped to discover noticeable differences in Google searching frequency of Czech towns.

Main advantages of utilization Google Trends are objectiveness, availability, continual monitoring, large volume of respondents, timeliness, readability of searching results and internationality. But this tool also encounters many drawbacks. The biases issued from geographical and temporal differences in Google users' volume have to be taken into account. Suitable relative comparisons have to be preferred. Results depend on language of searching. The identification of places is less reliable.

Acknowledgements. The research is supported by the Specific University Research project SP 2012/168 and by ESF CZ.1.07/2.2.00/15.0276 (Geocomputation).

References

1. ITU: The World in 2011: ICT Facts and Figures (2011), `http://www.itu.int/ITU-D/ict/facts/material/ICTFactsFigures2011.pdf`
2. Mana, M.: Internetová populace. Statistika & My 3/2012, pp. 30–33 (2012), `http://www.czso.cz/csu/2012edicniplan.nsf/t/BB0040343D/$File/1804120330_33.pdf`
3. CSU: Využívání informačních a komunikačních technologií v domácnostech a mezi jednotlivci (2010), `http://www.czso.cz/csu/2010edicniplan.nsf/t/E4003156C1/$File/970110.pdf`
4. Nielsen Media: Search engines most popular method of surfing the web. Commerce Net (1997), `http://www.commerce.net/news/press/0416.html`
5. Breyer, B.N., Sen, S., Aaronson, D.S., Stoller, M.L., Erickson, B.A., Eisenberg, M.L.: Use of Google Insights for Search to Track Seasonal and Geographic Kidney Stone Incidence, the United States. Urology 8(78(2)), 267–271 (2011)
6. Brownstein, J.S., Clark, C.F., Madoff, L.C.: Digital Disease Detection—Harnessing the Web for Public Health Surveillance. N. England. J. of Medicine 360, 2153–2157 (2009)
7. Ibuka, Y., Chapman, G.B., Meyers, L.A., Li, M., Galvani, A.P.: The dynamics of risk perceptions and precautionary behavior in response to 2009 (H1N1) pandemic influenza. BMC Infectious Diseases 10(1) (2010)
8. Ginsberg, J., Mohebbi, M., Patel, R., Brammer, M., Brilliant, L., Letter, L.: Detecting influenza epidemics using search query data. Nature (457), 1012–1014 (2009)
9. Clipp, C.: An Exploration of Multimedia Multitasking: How Television Advertising Impacts Google Search, 39 p. Duke University, North Carolina (2011)
10. Webb, G.K.: Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures. Issues in Information Systems X(2), 82–87 (2009)
11. StatOwl (2013),
`http://www.statowl.com/search_engine_market_share.php`
12. Boulton, A., Devriendt, L., Brunn, S.D., Derudder, B., Witlox, F.: City Networks in Cyberspace and Time: Using Google Hyperlinks to Measure Global Economic and Environmental Crises. In: Firmino, R.J., Duarte, F., Ultramari, C. (eds.) ICTs for Mobile and Ubiquitous Urban Infrastructures: Surveillance, Locative Media and Global Networks, pp. 67–87. IGI Global, Hershey (2011)
13. Google Trends (2013), `http://support.google.com/trends/topic/13975?hl=en&ref_topic=13762&uls=en`
14. Scheitle, C.P.: Google's Insights for Search: A Note Evaluating the Use of Search Engine Data in Social Research. Social Science Quarterly (92), 285–295 (2011)
15. Al-Eroud, A.F., Al-Ramahi, M.A., Al-Kabi, M.N., Alsmadi, I.M., Al-Shawakfa, E.M.: Evaluating Google queries based on language preferences. Journal of Information Science 37(3), 282–292 (2011)
16. Krejcar, O., Janckulik, D., Motalova, L.: Dataflow Optimization Using of WiFi, GSM, UMTS, BT and GPS positioning in Mobile Information Systems on Mobile Devices. In: 2nd International Conference on Computer Engineering and Applicatons (ICCEA 2010), vol. 2, pp. 127–131. IEEE Comp Soc., Bali Island (2010)

Automatic Web-Based User Interface Delivery for SOA-Based Systems

Marek Kopel, Janusz Sobecki, and Adam Wasilewski

Faculty of Computer Science and Management, Wroclaw University of Technology
Wyb.Wyspianskiego 27, 50-370 Wrocław, Poland
{Marek.Kopel,Janusz.Sobecki,Adam.Wasilewskii}@pwr.wroc.pl

Abstract. In the paper a method for ontology enhanced web-based user interface generation for SOA systems is presented. SOA paradigm defines a set of methodologies for building a software out of interoperable services. These services usually have an interactive character so in consequence they need a corresponding user interface to be delivered. Today these interfaces are implemented by the software designers and programmers manually or semi-automatically. We present in this paper the corresponding functionalities of the up-to-date sophisticated systems for service oriented systems design and implementation. Today, however there is a great need for more flexible user interface authoring and automatic generation, which is based on the services input and output parameters and their ontology-based description.

Keywords: User Interface, Ontology, SOA.

1 Introduction

SOA (Service Oriented Architecture) paradigm delivers general methodology to create computerized information systems in form of interoperable services. Systems that are designed accordant to this paradigm are composed of separate services, which are executed on user's demand. Each service is an implementation of a business functionality, which can be used in different business processes [9]. The system that has been built according to the SOA paradigm can combine composite services out of atomic ones according to the specified rules of composition.

The composite service is built from at least two different reusable services. The consequence of this feature is that the particular user interface differs depending on the composition (sequence of services) and the context of use of the particular service.

The problem of design and implementation of the personalized and automatically generated user interface (UI) for services in SOA-oriented systems is quite new. Quite similar solutions in the area of user interface adaptation have been addressed in many papers before, for example [13] and [15]. The user interface is usually adapted to the personal user's needs and/or environment settings [6].

Usually in the up-to-date implementations of the SOA-based system [16], the user interface for each composite service and for each user role is designed and then implemented manually or semi-automatically using for example CASE-tools, however

C. B?dic?, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 110–119, 2013.

then in this case the precise information flow has to be specified. A consequence of this is considerable time and money consumption on user interface implementation, which is rising with the number of different user roles and the system environments to be programmed. These problems may be overcome by the application of the automatic user interface authoring procedures, which may be additionally enhanced by the application of the ontologies.

The idea of distributed architecture poses new challenges for user interface. Although the functionality of a SOA system is implemented in a scattered manner, user who operates the system need a consistent look and intuitive interaction with the interface. One way to deal with this problem is considering any set of services that deliver a certain functionality a single, composed service. This way each user interface for a certain task is communicating with a single service endpoint, which simplifies the problem from the designer's and developer's point of view.

The motivation of this work lies in implementing GUI for PlaTel platform, which is an application of SOA paradigm in service composition and execution [16]. We assume that - while using PlaTel - system designer defines a business process. Then the system automatically suggests corresponding services according to their functional and nonfunctional parameters. These services form the composite service. The problem to which the solution is presented in this paper is how to implement user interface to the composite service.

2 IT Tools for BPM and User Interface Prototyping in IBM BPM

Gartner Group defines Business Process Management (BPM) as: "a management discipline that treats business processes as assets that directly contribute to enterprise performance by driving operational excellence and business agility" [1]. According to [3] key benefits of BPM include the following:

- formalization of existing processes and spotting needed improvements,
- facilitation of automated, efficient process flows,
- increasing of productivity and decreasing number of employees,
- allowing people to solve the hard problems.

The potential of BPM approach shows Forrester Group. According to their report [12] 81% of financial/insurance companies, 78% of construction companies, 73% of manufacturing companies and 65% of public sector and healthcare companies already implemented BPM IT tools or are interested in such projects.

BPM IT tools allow their users to:

- model business processes, implement and execute those models and refine the models based on as-executed data,
- provide transparency into business processes, the centralization of corporate business process models and KPI (Key Performance Indicator) execution metrics,
- develop interfaces (human-computer and computer-computer) which help with the implementation of Service Oriented Architecture (SOA) paradigm.

Some BPM IT tools that support user interface implementation are listed in Tab. 1.

Table 1. BPM IT tools supporting implementation of user interfaces

Vendor	Tool
BOC Information Technologies Consulting	*Adonis Business Edition* (extra functionality)
Software AG (IDS Scheer)	*webMethods Business Process Management Suite* *webMethods Composite Applications Framework*
BTC Business Technology Consulting	*Bonapart® Proffesional*
SourceCode Technology	*K2 Blackpearl*
Metastorm	*Metastorm BPM*
Rodan Systems	*OfficeObjects@e-Forms*
Altar	*Piramid WorkFlow*
IBM Corporation	*IBM BPM* (*Process Designer*, former *WebSphere Lombardi Edition*)

According to Forrester Research report [12] one of the market leaders is IBM that has two products (*Process Designer* – former Lombardi's *Teamworks* and *Integration Designer*) integrated into IBM BPM platform, based on Websphere Application Service (WAS). For this reason IBM BPM Process Designer was selected as an example to describe user interface prototyping possibilities in BPM IT tools.

User interface prototyping in IBM BPM products is called *Human Task* that includes one or more *coaches*. Coaches are created within Human Services editor in the Process Designer tool and include forms and associations with data structures.

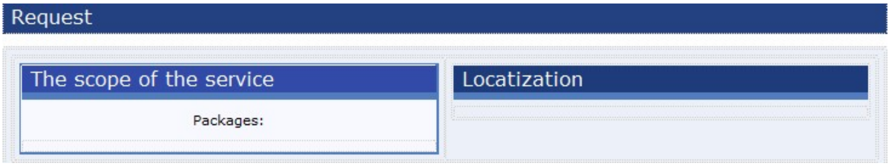
In the newest release of IBM BPM (version 8.0, May 2012) two different ways to coaches development are available: the old one, called *Heritage coach*, well known from the previous releases of IBM and Lombardi BPM products and the new one, default in the actual release.

Heritage coach, built in Coach Designer, is a XML document that contains a description of the user interface design. Then this XML document is passed through XSLT transformation and converted to a HTML document which is sent to the browser and visualized to the end user [7].

Heritage coaches allow to divide the web page that is presented to the user into several sections (containers in which the control objects can be placed). Default styles of the sections are following:

- One-Column with title
- One-Column
- Two-Column
- Three-Column.

Sections have some attributes that can be changed if necessary, e.g.: tile, numbers of columns and heading. Fig. 1 presents an example of the coach named *Request* that includes three sections: a section without visible title (Two-Column), *The scope of the service* (One-Column with title) and *Localization* (One-Column with title). The attributes of the section and the scope of the service are presented in the Fig. 2.



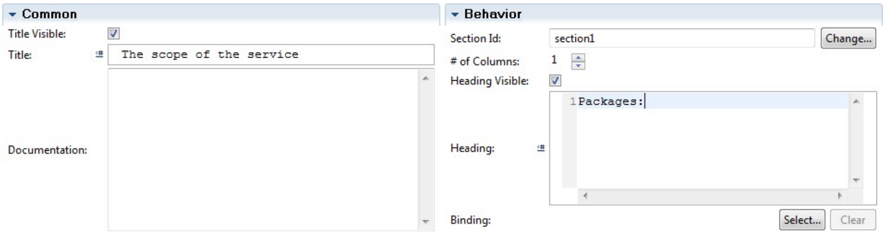
Request

The scope of the service

Packages:

Locatization

Fig. 1. Sections in the heritage coach *Request*



Common

Title Visible: ☑

Title: The scope of the service

Documentation:

Behavior

Section Id: section1

# of Columns: 1

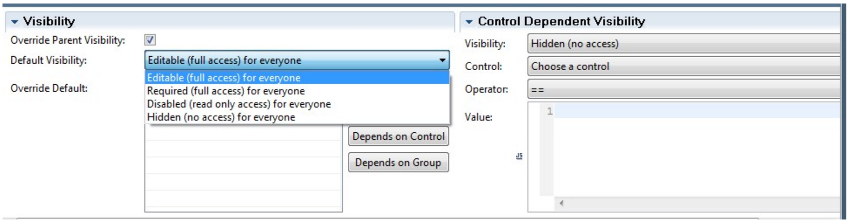
Heading Visible: ☑

Heading: Packages:

Binding:

Fig. 2. Attributes of the one section in the heritage coach *Request*

Heritage coaches allow to use several pre-defined control objects. Such blocks are used to construct the visual representation of the data structures that are necessary to perform interaction with the user. Each control object has the set of attributes that let to customize its appearance and activity. Common attributes of control objects include: label (visible description of the object), documentation (description for developers), control id (unique), control type (type of the object), binding (association with a variable). All the objects have corresponding attributes that allow to control their visibility (Fig. 3).



Visibility

Override Parent Visibility: ☑

Default Visibility: Editable (full access) for everyone

Editable (full access) for everyone
Required (full access) for everyone
Disabled (read only access) for everyone
Hidden (no access) for everyone

Override Default:

Control Dependent Visibility

Visibility: Hidden (no access)

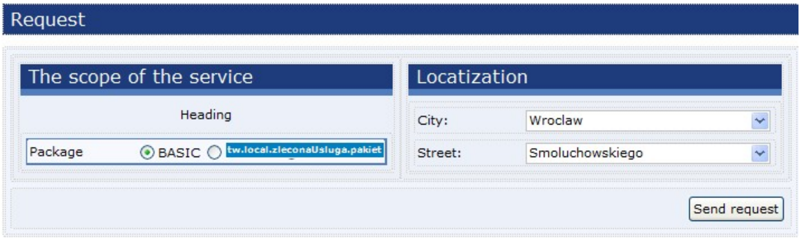
Control: Choose a control

Operator: ==

Value:

Fig. 3. Visibility controlling in Heritage coaches

Default visibility can be set to the one of the following options: Editable, Required, Disabled, Hidden; however with the possibility of dynamic change depended on a control or on the group. Moreover object's visibility can be control by custom Java Script code. Most of the objects can be associated with the events that trigger selected action. Supported events include: *onClick*, *onFocus*, *onBlur*, *onChange*, *onDblClick*, *onMouseDown*, *onMouseUp*, *onMouseOver*, *onMouseOut*, *onKeyPress*, *onKeyDown*, *onKeyUp*. Specific attributes depend on the type of the control object. Usually they are connected with the object representation on the screen. For example, for a single-select object can be set: widget type (radio button, drop-down list, multi-line list), orientation, data values and display text for each defined value and optional dynamic


Request

The scope of the service

Heading

Package ● BASIC ○ tw.local.zleconaUsluga.pakiet

Locatization

City: Wroclaw

Street: Smoluchowskiego

Fig. 4. Heritage coach *Request* with sections and control objects

data association (data values and description taken from selected variables). An example of heritage coach with sections and control objects is shown in the Fig. 4.

Heritage coaches may be further customized by Java Scripts but with some restrictions. Moreover there is no possibility to develop own control objects because the concept of *reusability* is not supported by heritage coach technology. Those problems are solved by new coach technology that is available from the release 8.0.

3 User Interfaces Enhanced by Ontology

Ontologies have been used in information systems design and development on at least several different levels, such as databases integration, business logic or Graphical User Interfaces (GUI) [14]. Application of the ontologies in the process of user interface delivery is not a new idea, they have been applied to the problem of unification of user interface parameters in the user interface recommendation method using the ant colony metaphor [15]. The work [13] presents an approach for mapping formal ontologies to GUI, which supports device independent GUI construction and semi-automatic GUI modeling. In the following work [5], however ontology-based approach of user interface development was designed for eliminating the demerits of the model-based approach, while preserving its merits, by exchange models of different interface components.

By application of some formal ontologies we may specify the structure of interaction. The applied ontology delivers the vocabulary of any given entity concept and corresponding sub-concepts. It may also specify the domain of the entity values. By defining automatic mapping functions from ontology to GUI we can deliver a tool for user interface generation.

Automatic user interface delivery for software services in the SOA systems is possible according to the formal service description and the user model. The formal description given in the SSDL (Smart Service Definition Language) [16] defines functional and non-functional description, service input, and service output. The most important element that is necessary for UI construction is the service input attribute names and the type of values as well as output attribute names and values. However, to present the input and output properly, we should possess the proper user model that decides, for example, what language is appropriate for the given user and how the specified input values should be entered (i.e. directly or using a select box) and how the output values should be presented (i.e. using a table of values or charts). In some cases, we should also apply ontologies to resolve some complex concepts into simple ones (i.e. address could be divided into *street*, *city*, *state* and *zip code* – in the case of an address in the U.S.). The general framework for UI generation for the service is given in Fig. 5. The user interface first enables the user to enter the input data, the input data labels, as well as their form and is worked out according to the user model; the ontology as well as the output data is presented according to the same elements.

An ontology-enhanced user interface is defined as a user interface, which visualization capabilities, interaction possibilities, or development processes are enabled or (at least) improved by the employment of one or more ontologies [10]. In the user interface enhancement, we may use different types of ontologies. We can distinguish the following ontologies that concern: the real world, the IT system, users, and roles [10]. The real world ontology characterizes its part, especially the area of the system application (i.e., e-commerce, travel, etc.), in order to identify the central concepts and relations among them. The IT system itself delivers formalized ontology by containing categories such as Software Module, Web Service, etc. We may further divide the IT ontology into hardware and software ontologies. Finally, users and roles ontology characterizes the users, their preferences, and their roles, which have an influence on the rights and possibilities they have in using a system.

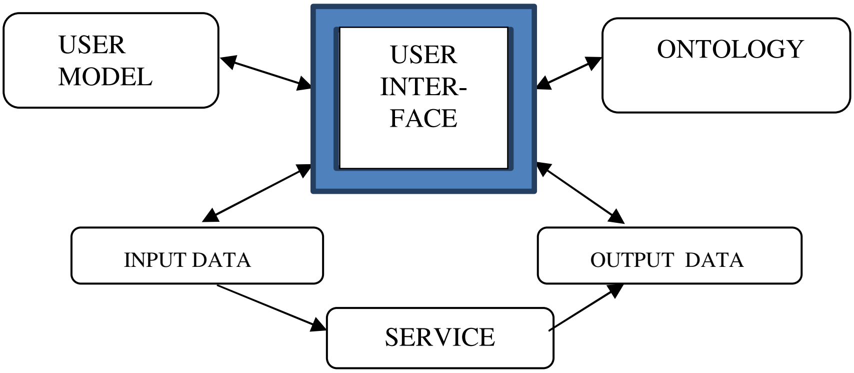


Fig. 5. General architecture for UI delivery for SOA based system

Nowadays in order to formally define the UI, specialized language is used, and the most popular languages are based on XML; some of the most popular are XUL and XIML, but other UI description languages are also used such as: UsiXML, UIML, Maria, LZX, WAI ARIA, and XForms. It has been discussed, however, that the application of ontology will not replace the application of user interface description languages, but rather deliver its valuable enhancement [11].

The application of ontologies in the delivery of UI for composite services is based on several assumptions [16]. First we assume that the composition is made out of some atomic services and the new service satisfy all the functional and non-functional requirements. These requirements are specified by the semantic query that describes the composite service and is often referred to as an Service Level Agreement (SLA).

Description of semantic query requires also application of some domain ontologies representing the domain knowledge, and the services that are described in terms of associated inputs, outputs and functionalities with the concepts from the ontology. Having composite service describe in terms of input and output data described with the domain knowledge represented by ontologies, as well as user goals and roles we may built an automatic mechanism for the UI delivery.

4 Generating User Interfaces in SOA Systems Based on WSDL Extension

The general idea of automatic generating user interfaces for Web Services is based on the assumption that user interaction can be serialized or simplified to some kind of an HTML form. So the starting point for building a UI generator is the ability to produce a Web form. Such a Web form when submitted would deliver all the necessary inputs from user and make the interaction with a Web service possible [8]. Generating User Interfaces in SOA Systems based on WSDL Extension

In order to help or even make possible for the user to fill the form some metadata are needed. In case of the numerical data some metadata may deliver parameters for range and granularity of the data, i.e. ‘min’, ‘max’ and ‘step’. Given the restrictions it is possible to validate the data on the client side, e.g. using HTML5 Forms and setting attributes for input tags. Another application of the metadata, which also embeds validation, would be the possibility of visualizing that numerical input with a specialized widget, e.g. a slider as shown in Fig. 6.

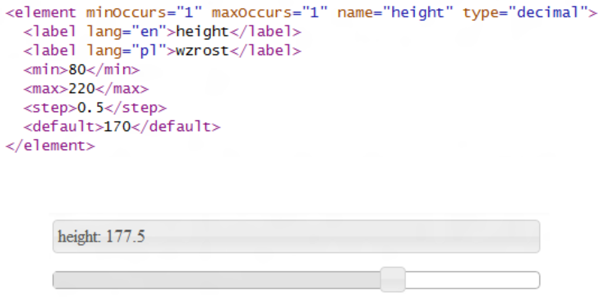
When building a user interface in SOA architecture, either in automated or semi-automated way, the metadata needed for that purpose must be stored in some common format. When a UI is supposed to support interaction with a single Web Service and each composition of services may be represented as a single, composite service, then extending the service’s WSDL file seems a natural decision.

The extension design is based on populating WSDL nodes, which describe Web Service method parameters, with child nodes describing the metadata. An example of extending an input parameter description in WSDL is presented in Fig. 7. The input named voivodeship is a string type parameter. But defining it as a select with prevalidated domain values (lines 3-7, 9-13) allow visualizing it in a user interface as a widget like a singleselect version (maxOccurs=”1”) of those from Fig. 6 or Fig. 7.

Of course, in accordance with SOA paradigm, domain values need not be enlisted in the WSDL file. The <domainvalues> node can have an additional attribute specifying a URL of a corresponding set of values. This way the WSDL extension is enabled to use external, decentralized ontologies in order to get up-to-date domain value sets.

The metadata are also localized for 2 languages: English (lines: 2-7) and Polish (lines: 8-13). More complex conditions for domain values can be defined in analogous way to the multilingual conditionals: for each condition a new set of values or a corresponding URL can be entered.

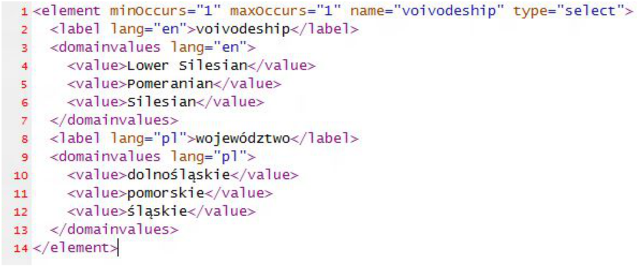
As mentioned earlier user interaction with an interface is a flow. In may be linear, so called wizard interface, in which users interact a step at a time. This type of the user interface is usually used for configuration interfaces, where next step depends on previous ones. It is also ideal for SOA based systems, because each step may be treated as a simple WS call, where both inputs and outputs are defined and not changed during the call. However there is often a need for a non-linear interaction and building an interface to fulfil that need in the SOA based systems is a real challenge.



```
<element minOccurs="1" maxOccurs="1" name="height" type="decimal">
  <label lang="en">height</label>
  <label lang="pl">wzrost</label>
  <min>80</min>
  <max>220</max>
  <step>0.5</step>
  <default>170</default>
</element>
```

height: 177.5

Fig. 6. XML formatted input validation metadata and an example visualization of the input as a slider widget.[8]



```
<element minOccurs="1" maxOccurs="1" name="voivodeship" type="select">
  <label lang="en">voivodeship</label>
  <domainvalues lang="en">
    <value>Lower Silesian</value>
    <value>Pomeranian</value>
    <value>Silesian</value>
  </domainvalues>
  <label lang="pl">województwo</label>
  <domainvalues lang="pl">
    <value>dolnośląskie</value>
    <value>pomorskie</value>
    <value>śląskie</value>
  </domainvalues>
</element>
```

Fig. 7. An extended WSDL described Web Service input parameter with label and domain values metadata.[8]

The nonlinear interaction is analogous to multithreaded computing or windowed interface. We propose using a design pattern of a Web dashboard interface where interaction can take place simultaneously in each widget window.

In Fig. 8 an example of a dashboard interface is presented. The interface intermediates user with a WS of a PlaFIT application. PlatFIT application is an implementation of the PlaTel platform and smart services described in the work [16]. The presented view in Fig. 8, allows monitoring the training process of players by their coach.

Assuming the whole dashboard is the UI for the one complex WS call, each widget is a simple WS called via the main WS, but operated independently. Still, since each widget is a part of one composite WS flow, they share input and output events. This way user interaction can be synchronized throughout the dashboard and in each the widget.

In Fig. 9 different charts show data from different sources i.e. sensors for: heart rate, EMG, geolocation, etc. When playing back the training each chart shows the

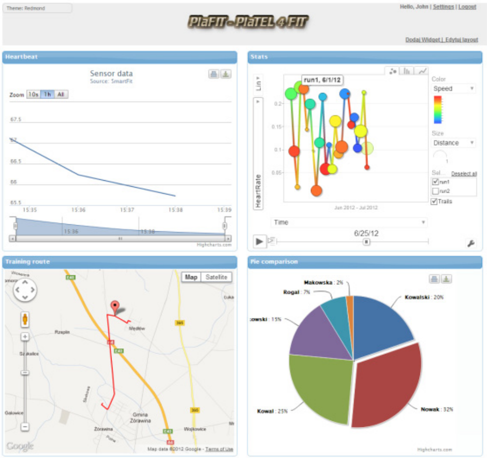


Fig. 8. Example of a dashboard with synchronized chart widgets. While interacting with one chart - other may also respond to the changes.

read from a sensor at the same moment. When moving a slider along the time axis all chart should synchronize and show the data according to the chosen point in time: end of a heart rate line should point to the "current" heart rate, the map marker should move to the position, corresponding to the player's current position on the route, etc.

5 Summary

In this paper we presented ontology enhanced user interface delivery for composite services in SOA systems. First we mentioned some the most popular Business Process Modeling tools. Then we presented how the user interface is prototyped using IBM BPM. With this tool we may prototype the user interface along with the business process modeling, however more complicated data inputs and its validation should be coded manually in some scripting language.

In the following sections we presented general information about the application of ontologies in user interfaces automatic delivery. We discussed mappings between levels of user experience and types of ontologies. Then we resented the general model of user interface delivery for services in the SOA based systems. We distinguished the following elements: user model, ontology, input data and output data. Then we discussed the types of ontologies that are applied in the whole process of user interface delivery enhanced by ontologies.

The last sections present some detailed solutions of the GUI automatic generation. We discussed specified GUI elements such as, sliders, comboboxes, input dependencies and multiple language selection. We propose to use WSDL for describing the user interface definition that is then rendered in a web browser to deliver working user interface. We also presented more sophisticated user interfaces that are built of several widgets that are dependent on one another. In the SOA based systems services are usually run one after another, however in many of today's interactive systems several services should work in parallel and exchange data between themselves. This necessitates the application of some special control mechanisms, which are less common for SOA based systems. In the future some work on solving these problems within SOA paradigm should be done.

Acknowledgements. The research was partially supported by Grant no. N N519 444939 funded by Polish Ministry of Science and Higher Education (2010-2013) and the European Commission under the 7th Framework Programme, Coordination and Support Action, Grant Agreement Number 316097, ENGINE - European research centre of Network intelliGence for INnovation Enhancement (http://engine.pwr.wroc.pl/).

References

[1] Business Process Management. Downloaded (November 2012), http://www.gartner.com/it-glossary/business-process-management-bpm/

[2] Garret, J.J.: Elements of user experience, Downloaded from http://www.jjg.net/elements/pdf/elements.pdf (April 2012)

[3] Havey, M.: Essential Business Process Modeling. O'Reilly (2005) ISBN: 0-596-00843-0

[4] Hickson, I.: HTML - Living Standard: Web Forms 2.0: DataList, http://www.whatwg.org/specs/web-forms/current-work/#datalist (last Updated April 9, 2012)

[5] Kleshchev, A., Gribowa, V.: From an Ontology-Oriented Approach Conception to User Interface Development. International Journal "Information Theories & Applications" 10, 87–93 (2004)

[6] Kobsa, A.: Personalized Hypermedia and International Privacy. Communications of the ACM 45(5), 64–67 (2002)

[7] Kolban, N.: IBM BPM Book on DeveloperWorks, http://www.neilkolban.com/IBM/

[8] Kopel, M., Sobecki, J.: Web-Based User Interface for SOA Systems Enhanced by Ontology. In: Zgrzywa, A., Choroś, K., Siemiński, A. (eds.) Multimedia and Internet Systems: Theory and Practice. AISC, vol. 183, pp. 239–247. Springer, Heidelberg (2013)

[9] Newcomer, E., Lomow, G.: Understanding SOA with Web Services. Addison Wesley Professional (2004)

[10] Paulheim, H., Probst, F.: Ontology-Enhanced User Interfaces: A Survey. International Journal on Semantic Web and Information Systems (IJSWIS) 6(2), 36–59 (2010)

[11] Paulheim, H., Probst, F.: A Formal Ontology on User Interfaces – Yet Another User Interface Description Language. In: SEMAIS, Palo Alto, CA, USA (February 13, 2011)

[12] Richardson, C.: The Forrester Wave: Business Process Management Suites, Q3, Forrester Research (August 26, 2010)

[13] Shahzad, S.K.: Ontology-based User Interface Development: User Experience Elements Patterns. Journal of Universal Computer Science 17(7), 1078–1088 (2011)

[14] Sobecki, J.: Hybrid Adaptation of Web-Based Systems User Interfaces. In: Bubak, M., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2004. LNCS, vol. 3038, pp. 505–512. Springer, Heidelberg (2004)

[15] Sobecki, J.: Ant colony metaphor applied in user interface recommendation. New Generation Computing 26(3), 277–293 (2008)

[16] Stelmach, P., Juszczyszyn, K., Prusiewicz, A., Świątek, P.: Service Composition in Knowledge-based SOA Systems. New Generation Computing 30(2&3), 165–188 (2012)

Developing a Multilingual Application Using Linked Data: A Case Study

Shingo Kinomura[1,*] and Kazuhiro Kuwabara[2]

[1] Graduate School of Science and Engineering, Ritsumeikan University
[2] College of Information Science and Engineering, Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577 Japan

Abstract. This paper describes development of a multilingual application using the concept of linked data. Linked data is a basis of Web of Data, which has the potential to overcome the language barrier since it focuses on data, not documents described in a natural language. To investigate how the linked data can be applied in a multilingual application, we have developed a prototype application where multilingual knowledge sources are extracted from existing parallel translations represented in the spreadsheet format. This application provides information in four languages in the domain of apartments rented to international students living in Japan. We discuss how the linked data can be made use of in the development of a multilingual application.

Keywords: multilingual application, linked data, RDF, design pattern.

1 Introduction

There is a lot of information available on the Internet. Web documents are typical examples. They are intended to be consumed by human users and are written in a natural language. Machine translation services are becoming easily available on the Web and can be used to access web contents in a different language. There is also a framework to use various web-based machine translation services [7].

To overcome the language barrier, a cross-lingual information query has been widely researched [5]. Multilingual web is also becoming a reality. At W3C, the MultilingualWeb-LT Working Group[1] has been formed and is addressing the standards for multilingual web.

At the same time, linked data has attracted a lot of attention [1], which is to achieve a so-called Web of Data. Since linked data's main focus is on data, not on web documents as in conventional web, it has the potential to be a basis of a multilingual application [4].

In this paper, we describe a case study on developing a multilingual application using linked data. This application is intended to provide useful information

[*] Currently with Central Japan Railway Company.

[1] `http://www.w3.org/International/multilingualweb/lt/`

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 120–129, 2013.

on renting and living in an apartment in Japan to international students. It provides frequently asked questions and their answers in four languages, namely English, Chinese, and Korean as well as Japanese.

In this application, the original source of information was prepared as a list of frequently asked questions in Japanese, and translated by human experts into the other three languages. In this sense, the main research question is how to make use of existing multilingual data to develop a multilingual application using linked data.

This paper is structured as follows. The next section briefly describes related research on the linked data applied to the multilingual area. The following sections describe an application built as a case study, and discuss the benefits of using the linked data framework in building a multilingual application. The final section concludes the paper.

2 Related Works

The linked data is based on the Resource Description Framework (RDF) [10]. In RDF, each *resource* is assigned a URI. The properties of a resource are represented as a triple of subject, predicate, and object. By allowing a URI of the resource to be dereferenced, Web of Data can be achieved. As a query language of RDF, SPARQL is defined. SPARQL basically executes pattern matching of triples to find relevant information from the RDF dataset.

Since the linked data focuses on data, it is inherently language-independent, and can be a basis of a multilingual application. Gracia et al. discuss issues involving multilingual Web of Data [4]. In the Web of Data, ontologies play a major role. An ontology often contains a lexical part, and is usually language-dependent. To develop a multilingual application, mappings between ontologies described in different languages need to be defined. One of the proposed approaches is CLOVA (Cross-Lingual Ontology Visualization Architecture) [8]. By separating the language-independent layer and the language-dependent layer, developing a cross-lingual application becomes easier. The localization of an application is done using the lexicon ontology such as LexInfo [2], which is based on the lemon model [9].

In this paper, in contrast to the ontology-based approach, we consider a lightweight approach that makes use of tag words in the existing multilingual datasets comprising parallel translations.

3 Construction of RDF Datasets

The multilingual application we consider here contains frequently asked questions (FAQ) in the domain of rental applications in four languages. The original FAQ data was compiled by an apartment management company and contain about 300 questions and 700 sentences for their answers. In addition to this, a glossary of technical terms in the domain of rental apartments was also created. This glossary includes about 2,000 words and expressions. The FAQ and

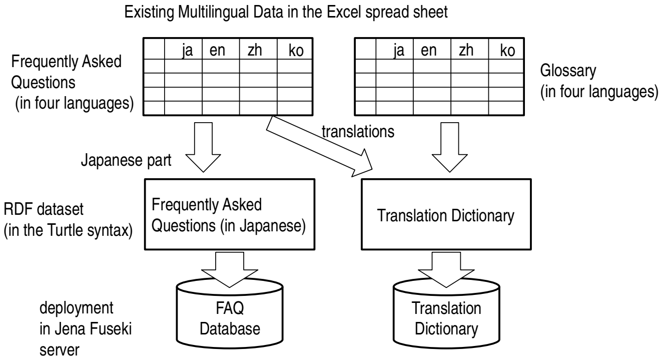

Existing Multilingual Data in the Excel spread sheet
Frequently Asked Questions (in four languages)
ja
en
zh
ko
Glossary (in four languages)
translations
Japanese part
RDF dataset (in the Turtle syntax)
Frequently Asked Questions (in Japanese)
Translation Dictionary
deployment in Jena Fuseki server
FAQ Database
Translation Dictionary

Fig. 1. Construction of RDF datasets

glossary were compiled in Japanese by a human expert and then translated into English, Chinese, and Korean by human translators. They are represented in the Excel spreadsheet format. Each row in the spreadsheet corresponds to an entry in the FAQ or glossary, and each column corresponds to its representation in each language.

In the example application, we separated the FAQ data described in Japanese from corresponding translations [6]. The translations extracted from the FAQ data were merged with the glossary and converted into the translation dictionary in the RDF format. In this way, adding an entry to the FAQ data can be made easier since the translations need not be accompanied. Compared to the approach that the FAQ contents in multiple languages are stored in a single RDF dataset, it would also be easier to expand the FAQ to another language since only an additional translation dictionary in a different language is needed, and FAQ data itself need not be changed.

As shown in Fig. 1, the original spreadsheet data was first converted into the RDF format in the Turtle syntax. The resultant files are to be loaded into an RDF server that implements SPARQL endpoints. In the prototype implementation, we used an Apache Jena Fuseki server[2].

3.1 Frequently Asked Questions in Rental Apartments

The RDF structure of the FAQ database is as shown in Fig. 2. In this figure, an example of "an air conditioner does not work" is shown. The question and answer pair is given a URI (in this case `qa:298`), and several properties to denote a question and answer are defined. To handle the case where the question or answer consists of multiple sentences, we attached the `qid` property for a question and `aid` property for an answer sentence where an RDF node is assigned to each sentence. The question and answer sentences are in Japanese since this is the language originally used to describe the FAQ.

[2] https://jena.apache.org/documentation/serving_data/index.html

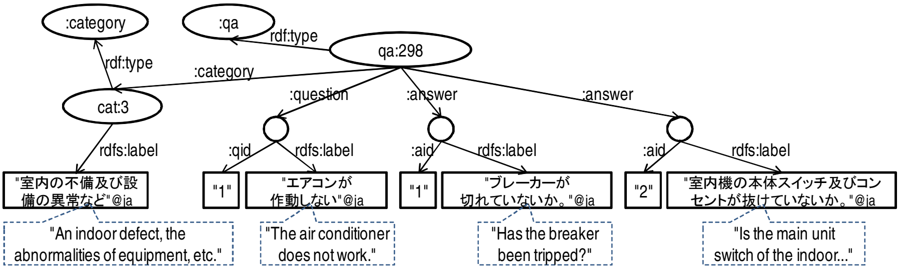

:category
:qa
rdf:type
qa:298
rdf:type
:category
cat:3
:question
:answer
:answer
rdfs:label
:qid
rdfs:label
:aid
rdfs:label
:aid
rdfs:label
"室内の不備及び設備の異常など"@ja
"1"
"エアコンが作動しない"@ja
"1"
"ブレーカーが切れていないか。"@ja
"2"
"室内機の本体スイッチ及びコンセントが抜けていないか。"@ja
"An indoor defect, the abnormalities of equipment, etc."
"The air conditioner does not work."
"Has the breaker been tripped?"
"Is the main unit switch of the indoor..."

Fig. 2. RDF representation of frequently asked questions and their answers (partial)

In addition, each question and answer pair has a category property. There are eight categories defined in the original FAQ data such as *room search*, *indoor defects*, *manners*, and *contract*. A category is attached to each question and answer pair, and is used to search for the relevant questions.

3.2 Translation Dictionary

The translation dictionary contains technical words and expressions regarding apartments in four languages from the original glossary and translations extracted from the FAQ data. The RDF representation of this translation dictionary is shown in Fig. 3. In this figure, the term (floor plan) is depicted. We defined a property to denote the transformation (in the figure it is denoted by `mlt` prefix).

In addition to a simple translation, several properties are defined to describe phonetic transcription ("yomi") in Japanese, and fluctuations in the representation such as "floor plan" and "floor-plan."

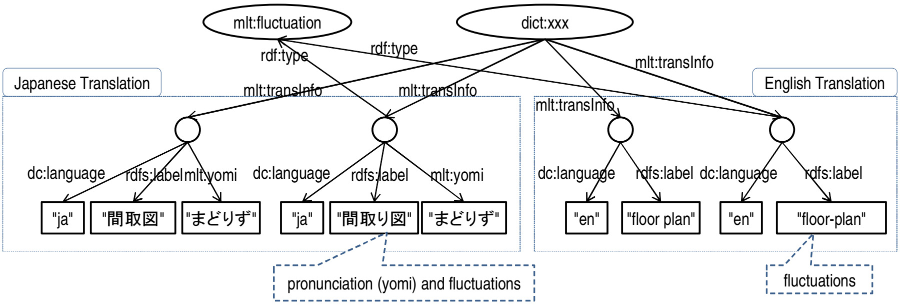


Fig. 3. RDF representation of translations (partial)

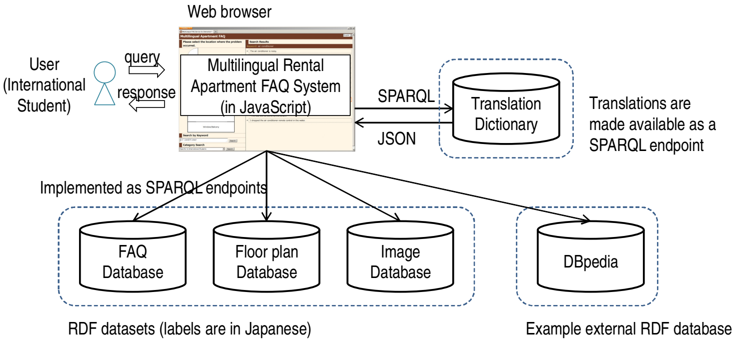

Web browser
User (International Student)
query
response
Multilingual Rental Apartment FAQ System (in JavaScript)
SPARQL
JSON
Translation Dictionary
Translations are made available as a SPARQL endpoint
Implemented as SPARQL endpoints
FAQ Database
Floor plan Database
Image Database
DBpedia
RDF datasets (labels are in Japanese)
Example external RDF database

Fig. 4. Overview of prototype application

4 Prototype Application

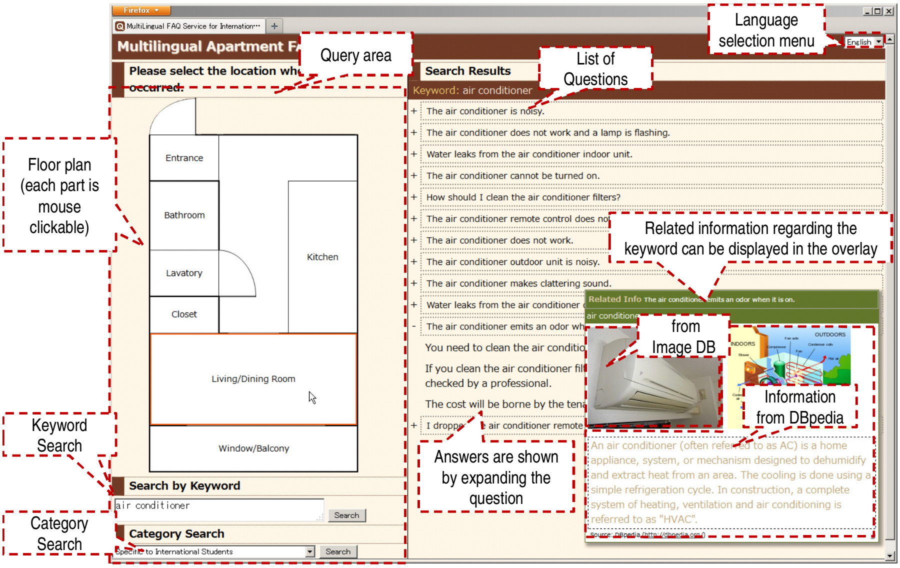
The prototype application has a browser-based user interface to search for the relevant questions and answers from the FAQ data. In addition, related information with the questions is also shown. To demonstrate retrieval of the related information from multiple SPARQL endpoints, a sample image database was also implemented. In addition, we used a DBpedia as another source of the related information. Figure 4 shows an overview of the prototype application.

In this prototype, a JavaScript program accesses the RDF datasets that are published as SPARQL endpoints. To simulate multiple SPARQL endpoints, we assigned a different URL to each RDF dataset. As mentioned in the previous section, the datasets are prepared as Turtle files, which are loaded into the Apache Jena Fuseki server when the server starts up.

Since most modern browsers implement the functionality of Cross-Origin Resource Sharing [11], a JavaScript program can make a query to SPARQL endpoints directly. Thus, most of the functions of the prototype application were implemented in JavaScript.

There are three search methods for FAQ data. One method is to specify a related area of a typical apartment floor plan. Another is to enter a keyword in the input form, and the last one is to select a category of a question from the menu. Figure 5 shows a screenshot of the prototype application. The left part of the screen is used as an area to input search conditions. When the relevant questions are found, they are displayed in the right part of the screen. By clicking the question, corresponding answers are shown. In addition, by right-clicking on the question, related information is shown in a pop-up window.

The prototype can be used in four languages (English, Chinese and Korean as well as Japanese). In the upper-right corner of the browser window, there is a menu to select the language to use. In the implementation, all of the HTML elements that need to be translated into different languages are attached with a custom CSS (Cascading Style Sheet) class (called `multilingual`). Another custom HTML property of `data-label-ja` is used to store a label in Japanese. When the



Multilingual Apartment F

Please select the location wh occurred.

Search Results

Keyword: air conditioner

+ The air conditioner is noisy.
+ The air conditioner does not work and a lamp is flashing.
+ Water leaks from the air conditioner indoor unit.
+ The air conditioner cannot be turned on.
+ How should I clean the air conditioner filters?
+ The air conditioner remote control does not
+ The air conditioner does not work.
+ The air conditioner outdoor unit is noisy.
+ The air conditioner makes clattering sound.
+ Water leaks from the air conditioner
- The air conditioner emits an odor wh

You need to clean the air conditio

If you clean the air conditioner filt checked by a professional.

The cost will be borne by the tena

+ I droppe e air conditioner remote

Related Info The air conditione emits an odor when it is on.

air conditione

An air conditioner (often referred to as AC) is a home appliance, system, or mechanism designed to dehumidify and extract heat from an area. The cooling is done using a simple refrigeration cycle. In construction, a complete system of heating, ventilation and air conditioning is referred to as "HVAC".

Source: DBpedia (http://dbpedia.org/)

Search by Keyword

air conditioner

Category Search

Specific to International Students

Fig. 5. Screenshot of prototype application

user selects a different language, all the HTML elements with the `multilingual` CSS class will be updated by translating the value of the `data-label-ja` property into the specified language. To translate the words, the RDF database of the translation dictionary is used. To minimize the access to the RDF dataset, a simple cache mechanism is also incorporated.

5 Example Usage

Let us here explain a possible usage of the prototype application. When a user encounters a problem while living in an apartment, the user searches for the relevant question from the FAQ data. Let us suppose that the air conditioner placed in the living room has a problem, and that the user selects English to access the system.

5.1 Search by Floor Plan

An image of a typical floor plan is used to specify the place where the trouble has occurred. In this example scenario, the living room is selected by clicking a corresponding part in the floor-plan image. To search the FAQ data for the relevant question from the apartment floor plan, we defined a simple RDF dataset to represent the vocabulary of the floor plan. From the typical floor-plan layout of the apartment, parts of the apartment such as “living/dining”, “bathroom”, “window/balcony”, “entrance”, “kitchen”, and “closet” were extracted manually. In addition, 65 keywords were extracted from the FAQ data and attached to a

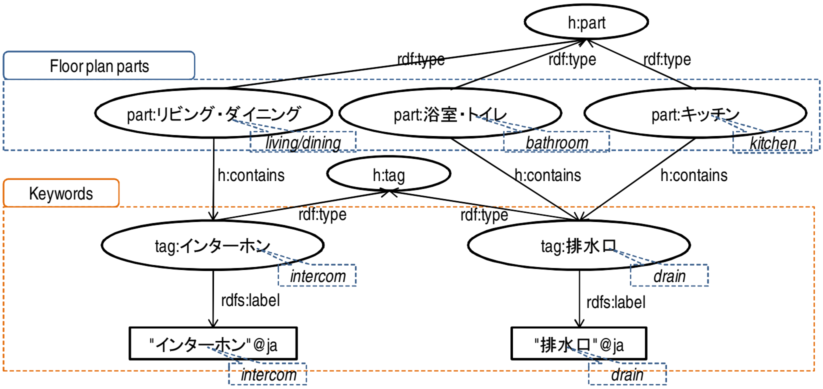


Fig. 6. Floor-plan categories (partial)

corresponding floor-plan part also by hand. Figure 6 shows a part of the floor-plan RDF dataset. In this figure, three floor-plan parts are depicted. The keyword ("intercom") is attached to the living/dining room, and the keyword ("drain") is attached to the part types of "bathroom" and "kitchen."

Going back to the example scenario, questions related to the living room are searched for with the keywords associated with the specified floor-plan part. Since the resultant questions are written in Japanese and the language that the user selected is English, the labels in the questions are translated into English by accessing the translation dictionary RDF dataset. Then, a list of questions is presented to the user.

5.2 Keyword Search

The prototype application also offers a function to search by specifying a keyword. There is an input box into which a keyword can be entered. In the example scenario, "air conditioner" is entered since an air conditioner is the source of the trouble. The prototype application first searches for a corresponding keyword in Japanese to "air conditioner", which the user entered. Based on the corresponding keyword in Japanese, the relevant question is obtained.

5.3 Selection of the Question

The user selects a question most relevant to the problem from a list shown. In the example scenario, let us suppose that a question of "The air conditioner emits an odor when it is on" is selected. When the user clicks on the question sentence displayed, the corresponding answers are retrieved by issuing a SPARQL query to the FAQ dataset and translated into English. Then, they are displayed below the question.

5.4 Related Information

In addition to the answers to the frequently asked question, the related information is retrieved from other RDF databases. Since each question is associated with the floor-plan part, a SPARQL query is made to the FAQ dataset to obtain the URI of the associated floor-plan part of the question sentence. In the example scenario, for the question ("The air conditioner emits an odor when it is on") the result is part: リビング・ダイニング where part: denotes the prefix defined for this application. Here, Internationalized Resource Identifier (IRI) is used to represent a name of the floor-plan part as discussed in the multilingual linked open data patterns [3]. More specifically, Japanese is used in this IRI since the FAQ itself is written in Japanese (this IRI means "living/dining room"). By accessing the floor-plan RDF dataset with this IRI, keywords associated with the part are retrieved.

By using the SERVICE clause defined in SPARQL's federated query[13], this series of queries can be made into a single query. After the keywords are obtained, the application program compares the keywords and the question sentence. Only the keywords contained in the question sentences are used to retrieve the related information. In the example, "air conditioner" is used as a keyword to retrieve the related information from DBpedia and a simple image database we created.

Use of DBpedia. In the prototype application, a DBpedia resource URI of the corresponding keyword was obtained before-hand using DBpedia's lookup service,[3] and stored in the RDF database. Using this URI, we retrieve the dbo:thumbnail property for an image and an explanation from the dbo:abstract property. Since DBpedia contains abstract information in different languages, we look at the language tag. In the prototype application, the four languages are checked using the SPARQL OPTIONAL construct since the information may not be available in all the languages specified. The result of this search is cached by the application program for later use.

Use of Image Database. As an example of utilizing a different RDF database, we have constructed a simple image database that contains images that are related to the housing fixtures. Figure 7 shows part of the image database, which is represented as an RDF dataset. This data set basically includes tag words along with the file name that corresponds to the image file name. The tag words are written in Japanese as the language tag of @ja is attached.

In the prototype application, we assigned an IRI to each tag word. Using the Simple Knowledge Organization System (SKOS) [12] vocabulary, skos:closeMatch relationships are also defined to connect the tag word and URI of the image in the image database so that the image can easily be retrieved from the image database.

[3] http://lookup.dbpedia.org/api/search.asmx/

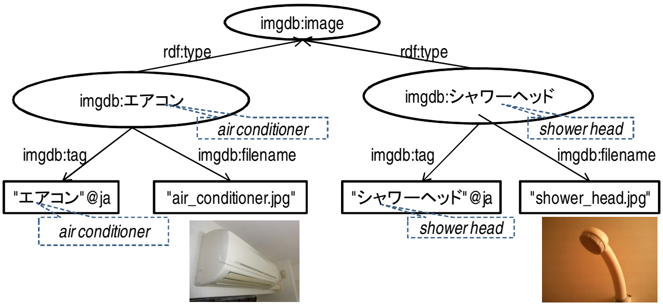


Fig. 7. Sample Image Database (partial)

6 Discussion and Future Work

There are several ways to store texts in multiple languages. In the prototype, basically we use the `rdfs:label` and the text is accompanied by a language tag such as `@en`. However, the translation dictionary adopted a complex structure to handle fluctuations in literal expression and Japanese phonetic transcription. In the representation of the translation dictionary (Fig. 3), a blank node is created to hold various entries in the translation dictionary. Currently, we need to know how the multiple languages are represented when we compose a query to a SPARQL endpoint. It would be desirable to be able to unify the representation of multiple languages.

The main FAQ data is currently described in Japanese. When accessed in a different language, the translation dictionary is utilized to translate the sentences in Japanese. In this sense, the current application deals with multiple languages at the lexical level. The differences in the cultural background are not well addressed. The construction of ontologies in multiple languages with their mappings among them will be needed for more intelligent search.

The proposed approach is also applicable to other applications such as multilingual information services. In such a case, since the translation resources are separated from the information contents, it would be easier to share the translation resources by different applications.

7 Conclusion

This paper described an implementation of a multilingual application using the linked data concept. The goal of this application is to provide information related to rental apartments in Japan to international students. The application makes use of the existing multilingual data provided in the Excel spreadsheet format. They were converted into RDF datasets that represent the contents and translations separately. Using the linked data concept, the multilingual application was easier to develop.

The current application only provides a search function for a relevant question and answer from the existing FAQ data. We plan to add a user interface to enable updating of the FAQ data easily by a user, and evaluate the current approach of implementing RDF datasets in a multilingual application.

Acknowledgments. The authors wish to thank Toshihiko Nakai for the multilingual FAQ and glossary data for rental apartments used in the example application.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
2. Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M.: LexInfo: A declarative model for the lexicon-ontology interface. Web Semantics: Science, Services and Agents on the World Wide Web 9(1), 29–51 (2011)
3. Gayo, J.E.L., Kontokostas, D., Auer, S.: Multilingual linked open data patterns, `http://www.semantic-web-journal.net/system/files/swj406.pdf`
4. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web 11, 63–71 (2012)
5. Kishida, K.: Technical issues of cross-language information retrieval: a review. Information Processing & Management 41(3), 433–455 (2005)
6. Kuwabara, K., Kinomura, S.: Mediating accesses to multiple information sources in a multi-lingual application. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part I. LNCS, vol. 7653, pp. 326–334. Springer, Heidelberg (2012)
7. Ishida, T.: Language grid: An Infrastructure for Intercultural Collaboration. In: IEEE/IPSJ Symposium on Applications and the Internet (SAINT 2006), pp. 96–100 (2006)
8. McCrae, J., Campana, J.R., Cimiano, P.: CLOVA: An architecture for cross-language semantic data querying. In: Buitelaar, P., Cimiano, P., Montiel-Ponsoda, E. (eds.) 1st International Workshop on the Multilingual Semantic Web (MSM 2010), pp. 5–12 (2010)
9. McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging lexical resources on the semantic web. Language Resources and Evaluation 46, 701–719 (2012)
10. Resource Description Framework (RDF), `http://www.w3.org/RDF/`
11. W3C Candidate Recommendation, Cross-Origin Resource Sharing, `http://www.w3.org/TR/cors/`
12. W3C Recommendation, SKOS Simple Knowledge Organization System Reference, `http://www.w3.org/TR/skos-reference/`
13. W3C Recommendation, SPARQL 1.1 Federated query, `http://www.w3.org/TR/sparql11-federated-query/`

Integrating Smartphone's Intelligent Techniques on Authentication in Mobile Exam Login Process

Zhaozong Meng, Joan Lu, and Ahlam Sawsaa

School of Computing and Engineering, University of Huddersfield, UK
{z.meng,j.lu,a.sawsaa}@hud.ac.uk

Abstract. The emerging build-in sensing techniques create opportunities for Human-Computer Interaction capability of the mobile devices. This investigation explores novel build-in sensing techniques and relevant computation intensive algorithms to enhance the operational efficiency and usability of mobile applications. A case study on mobile application authentication involving touch screen manual input, camera barcode scanning, NFC recognition is implemented. Qualitative and quantitative evaluations on usability, efficiency, stability, and accuracy of user operation are examined on mainstream mobile platforms. Result illustrates the advantage of the proposed scheme and verifies the feasibility to enhance user-mobile interaction with mobile sensing techniques.

Keywords: Mobile Computing, Mobile Application, Human-Computer Interaction (HCI), Barcode scanning, Near Field Communication (NFC).

1 Introduction

The Smartphone today can be used for some computing tasks intelligently assisting people's activities with build-in sensors in many different scenarios in data tailoring, navigation, recommendation, etc. [1]. However, the mobile devices are limited in user interaction, that is information representation and data input [2]. Usability, simplicity, and efficiency become crucial concerns of user experience of mobile applications.

However, mobile devices create new ways to enhance the human phone interaction of mobile applications with the novel sensing techniques [3]. Smartphone oriented applications can be reached by: (1) Employing novel build-in hardware and sensing techniques to replace the traditional complex operations, and (2) Utilizing the computation intensive algorithms to reduce the time consuming and error prone tasks. The intelligent sensing technique is critical support for the Smartphones to be able to attract over one billion users worldwide [4].

There are investigations attempting to use the speech and visual information on mobile devices for user interaction, and some motion sensors and recognition modules such as accelerometer and NFC are also drawing attentions [5]. Feng et al. reviewed mobile search and presented speech-based method and multimodal interaction to optimize the efficiency of mobile search [6]. Hannuksela et al. designed an interaction for camera enabled devices to write just by moving the device, using discrete

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 130–142, 2013.

cosine transform and k-Nearest Neighbour rule to discriminate feature and classification, and created a recognition rates ranged from 92% to 98% [7]. Broll et al. presented a Java ME prototype of Physical Mobile Interaction (PMI) paradigm supporting and facilitating mobile interaction with services through the interaction with physical objects, such as through NFC, pointing through visual marker recognition, and direct input through standard input widgets [8]. These techniques convert human interaction to a form that can be easily captured by mobile devices. Thus, it improves the interaction and makes it easy to operate for some use cases. However, we can hardly find studies evaluating the promising novel user interaction techniques, especially for the heterogeneous mobile platforms which are limited in user interaction.

This investigation aims to verify the feasibility of utilizing the build-in sensing techniques to create innovative user interaction approaches for the mobile applications, and determine the affecting factors and underlying limitations of these techniques on different mobile platforms. This work employs the mobile application user authentication as a use case and implements innovative sensing techniques to enhance the usability and efficiency of operation compared with traditional touch screen input.

2 A Case Study - User Authentication with Smartphone Sensing Technologies

2.1 Background

Unlike traditional examination entries, this project implements Smartphone sensing technology into mobile examination system (MES) to authenticate users entering the examination system. As each student has a unique ID card with a barcode on the card and a NFC chip, the camera and NFC technique can be used for user identification and authentication to simplify the operation and improve the speed and accuracy. The accelerometer can be used for user interaction with specific actions such as shaking (see Fig. 1).

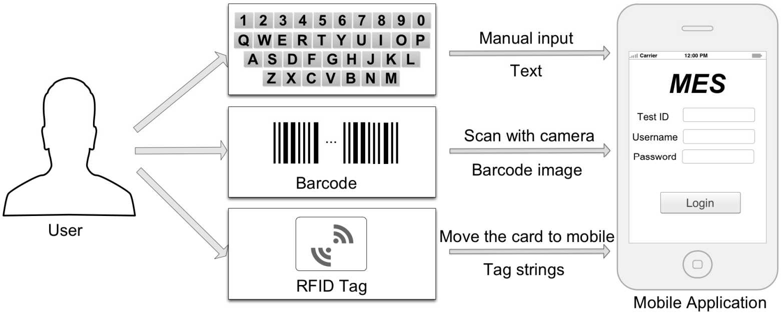


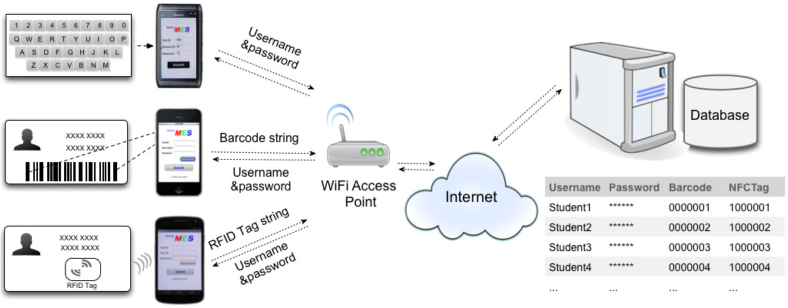
Fig. 1. Technique methods employed in the case study

In order to verify the methods to improve the efficiency and usability of mobile based applications, three user interaction techniques such as touch screen input, student ID card barcode scanning, and student ID barcode NFC recognition are designed and implemented on the latest mobile platforms as shown in Fig.1.

The attributes that can be used to assess the efficiency and usability of the sensing technologies include easy to use, operation speed, error rate, vulnerable to inference, computation intensity, arbitrary degree, and use cases [9]. The evaluation of the user interaction in the case study is focused on the comparison of performance of these approaches and the determination of how the performance of these techniques varies on the heterogeneous mobile platforms.

2.2 System Design

2.2.1 System Architecture of MES User Authentication Module



Username	Password	Barcode	NFCTag
Student1	******	0000001	1000001
Student2	******	0000002	1000002
Student3	******	0000003	1000003
Student4	******	0000004	1000004
...	...	...	...

Fig. 2. System architecture of user authentication

The MES is a cross-platform system for the mainstream mobile operating systems, such as Apple iOS, Google Android, Windows Phone 7, and Symbian^3. The system architecture is described in Fig. 2. The accounts of the authorized users are stored in the database of the backend server. When the user login with barcode or NFC magnetic card, the mobile device first obtains data from the barcode or NFC and then accesses the bounded username and password from remote server for login.

2.2.2 Barcode Scanning – A Camera Based Computation Intensive Approach

1. Student ID card Barcode Format

The formats for barcodes vary in different areas, the student ID barcode for most universities in the UK is a proprietary format widely used in libraries, which is named Telepen. The Telepen symbology was devised by George Sim, Chairman of SB Electronic System Limited in early 1972. The specification of this symbology is presented in [10]. Compactness and security are advantages of the Telepen symbology.

2. Encoding Format of Telepen Barcode

The information of the barcode formats is encoded and represented by the wide and narrow bars and spaces. The barcode of Telepen format consists of only four patterns of bar and space pairs (see Fig. 3). Let the barcode shown in Fig. 4 be a sample barcode, the encoding method is introduced as follows.

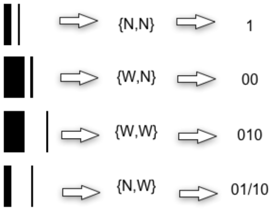


Fig. 3. Patterns of Telepen format

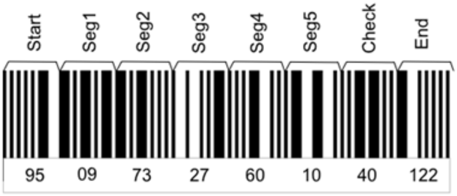


Fig. 4. A sample Telepen barcode

In each Telepen barcode there is a specific beginning and an ending segment. As shown in Fig. 4, the first and the last segments are the start and end of the whole barcode respectively. Besides, seg1 - seg5 are data segments, and segment 6 is the check segment. Each segment is encoded with the patterns pairs in Fig. 3.

Take Seg1 in Fig. 4 for example, the corresponding patterns in the segment are:

Patterns [Seg1] = {{W,N}, {N,N}, {W,N}, {N,N}, {W,N}}

Then, comparing with the encoding digit pairs of each pattern in Fig. 3, the binary digit of the segment Binary[Seg1] can be obtained: 00100100. The binary digit is 8-bit even-parity, the initial value can be obtained with equation (1):

$$\text{value\_init= decoded\_byte \& 0x7F} \qquad (1)$$

Then, the decimal data of segment 1 Dec[Seg1] is reversed digit of the initial value of Binary [Seg1]: $1\times2^5 + 1\times2^2 = 36$. Using the same method, the decimal data of the data segment Dec[barcode] can be obtained: {95,36,100,54,87,37,67,122}. For verification, the data should meet equation (2):

$$(127 - \sum \text{Dec[Seg i]})\ \%127 = \text{Dec[Check]} \qquad (2)$$

For sample barcode, the verification value is: 127 - (36+100+54+87+37)%127 = 67, which equals to the value of the check segment. Then, the data segments subtract 27, the data of the barcode Dec_barcode[dataSeg] can be obtained: {09,73,27,60,10}. Then, the string '0973276010' is the encoded data of the sample barcode in Fig. 4.

3. Barcode Scanning based on Image Processing.

Most of the barcode readers are laser scanners which work with PC applications. On mobile devices, the barcode can be gathered with the build-in camera, and the decoding can be done with image processing algorithm with the workflow in Fig. 5.

The decoding algorithm extracts the encoded data from the camera gathered images. It comprises the format recognition, pattern matching, 8-bit parity checking, result verification, and length checking. In addition, the system needs some preparations such as camera initialization, camera lens focusing, image acquisition, etc. shown in Fig. 5. The data normalization and pattern matching algorithm are the major concerns. The barcode scanning module is embedded into mobile client applications of MES project on the four mobile platforms with some resource from [11].

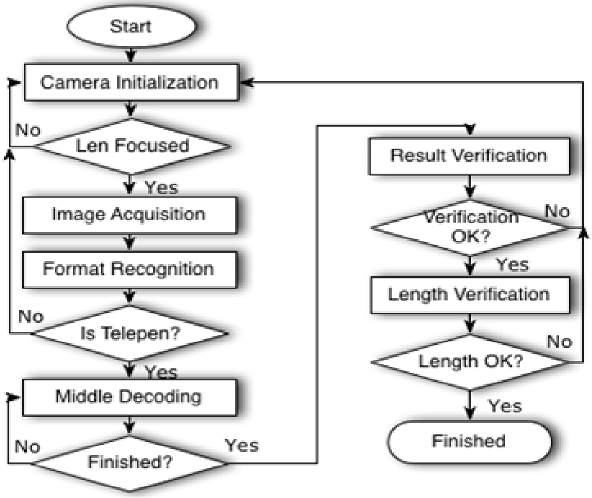


Fig. 5. Flow chart of barcode scanning module

2.2.3 NFC Recognition – A Radio Frequency Identification Technique

The NFC technique is a subset of RFID and limits the range of communication within 4 inches which is strong in security for some use cases. The strength of NFC technology is its ability to precisely identify objects and at low cost [12]. Thus, it becomes a fast growing area of research and application in the recent years.

The application of NFC recognition on mobile devices is a new technology for the Smartphone and Tablets. In this investigation, a NFC enabled mobile phone Samsung Google Nexus is selected as a test platform. The main purpose is to compare the different methods on the same device. When the NFC module is enabled, students just move their student ID cards close the mobile device to login the system immediately.

3 Experiments and Evaluation

Barcode scanning and NFC recognition is implemented for the authentication in MES application. The barcode scanning is implemented on four mobile platforms iOS, Android, Windows Phone 7, and Symbian^3. The NFC magnetic card login is implemented on Android Phone Galaxy Nexus. Experiments are carried out to test and compare the operation, speed and error rate of the three interaction approaches. Performance evaluation of the different methods on different platforms is conducted.

3.1 Simplified Operations Using the Interaction Techniques

3.1.1 Barcode Scanning Approach

The login operation can be done with such steps: firstly, launch the MES application on mobile devices and input the Test ID. Then, follow the steps1, 2, and 3 in Fig. 6. (1) Shake the phone or press "Barcode scan" button, then the camera is opened and scanning is initiated; (2) Target the red line in the camera vision over the barcode till the scanning finishes; (3) Confirm matched ID to login.

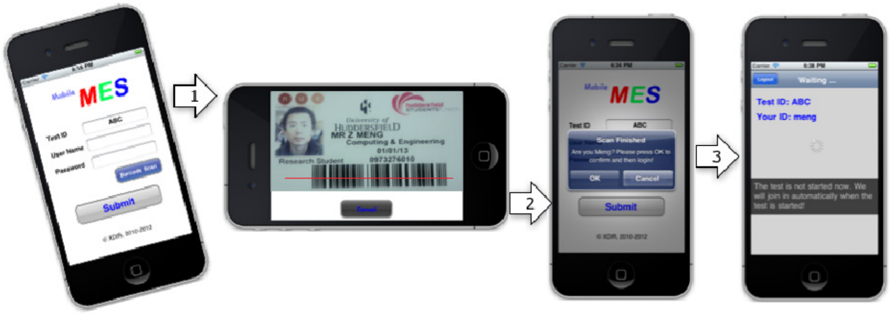


Fig. 6. Work flow of the barcode scanning module on iOS platform

With barcode scanning, the user just needs to shake the phone and target the barcode in the camera vision. It is much easier compared with operations on soft keypad.

3.1.2 NFC Recognition Approach

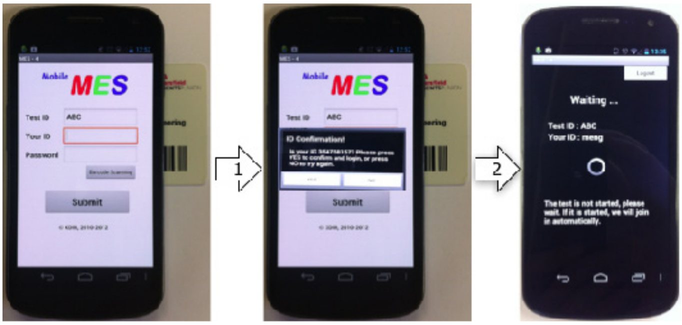


Fig. 7. Work flow of NFC recognition approach

The NFC recognition is implemented on an Android device Nexus S, which also largely simplifies the operations of login process. The operation steps of the login process with a student ID card are: Launch the MES application on mobile devices and input the Test ID. Then, follow steps 1 and 2 in Fig. 7: (1) Move the student ID card close to the phone, till the recognition is finished; (2) Confirm the ID to login.

The NFC recognition approach converts the operation on the soft keyboard to moving the student ID card close to the mobile, which is much easier and convenient.

3.2 Testing of the Interaction Techniques

The testing aims to provide the answers to the following questions:

- How the technologies enhance the usability and efficiency of mobile applications?
- What are the strengths and weakness of these technologies in mobile applications?
- What are the influencing factors of these technologies?
- How they differ among the different mobile hardware and software platforms?

With these questions, the approaches implemented in the system including the touch screen manual input, barcode scanning, and NFC recognition are tested separately. The parameters to be measured are: speed (time spent) and error rate. The mean value and standard deviation (STDEV) of time spent can be calculated accordingly for further analysis. Mean value can be used to estimate the real value of some parameters, and standard deviation is a parameter that can be used to evaluate dispersion degree. Thus, they are used to evaluate efficiency and stability of the approaches.

The strategies for the testing and evaluation are descried as below:

- Test the techniques on the same platform to evaluate the interaction techniques
- Test the same technique on the different platforms to examine its performance
- Discuss the underlying influencing factors based on the testing results

As some methods such as the barcode scanning is easy to be affected by the ambient environment such as brightness of illumination and light reflection, the testing of different devices are expected to be done under the identical condition. The account to be used is randomly generated by specially designed algorithm, and each of the method on different device is repeatedly tested for 20 times.

Table 1. Devices used in the testing

Platforms	Devices	OS Version	Camera (Pixels)	Screen(Inches)
iOS	Apple iPhone 4s	iOS 4.3.5	5M	3.5
Android	Samsung Galaxy Nexus	Android 4.1.1	5M	4.65
WP7	HTC HD2	WP 7.1	5M	4.3
Symbian	Nokia N8	Symbian^3	12M	3.5

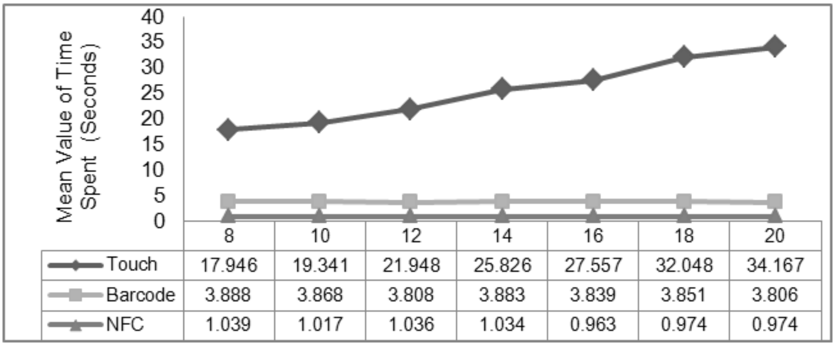
The devices to be used in the test are required to cover the mainstream mobile operating systems. The devices used in the testing are listed in Table 1.

The speed of the three user interaction techniques is tested. The lengths of username and password are set from 8 to 20 in even values. The touch screen, barcode scanning, and NFC recognition are tested separately with the same user. The time is gathered using the timer function on the mobile platform, and the accuracy is 1 millisecond. Timer starts and ends in the tests are:

- Touch screen: From launching the application to pressing the submit button
- Barcode scanning: From initiation of scanning till account is shown and confirmation is required
- NFC recognition: From launching the application till account is shown and confirmation is required.

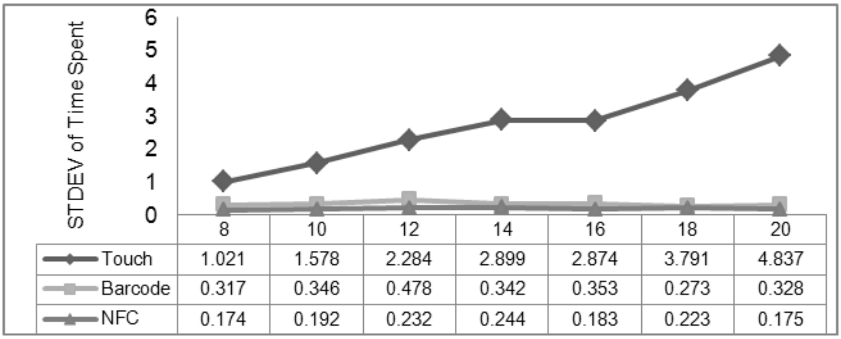
3.2.1 Performance Analysis of the Three Interaction Techniques

The testing results of the mean values and standard deviations of the time spent on Android device Galaxy Nexus are given in Fig. 8 and Fig. 9.



	8	10	12	14	16	18	20
Touch	17.946	19.341	21.948	25.826	27.557	32.048	34.167
Barcode	3.888	3.868	3.808	3.883	3.839	3.851	3.806
NFC	1.039	1.017	1.036	1.034	0.963	0.974	0.974

Fig. 8. Mean value of time spent (Seconds) of three techniques on Galaxy Nexus



	8	10	12	14	16	18	20
Touch	1.021	1.578	2.284	2.899	2.874	3.791	4.837
Barcode	0.317	0.346	0.478	0.342	0.353	0.273	0.328
NFC	0.174	0.192	0.232	0.244	0.183	0.223	0.175

Fig. 9. Standard deviation of time spent of three techniques on Galaxy Nexus

From Fig. 8, it is found that NFC approach is the fastest (Average[MEAN, STDEV]=[1.005, 0.203]), barcode scanning is medium (Average[MEAN, STDEV]= [3.849, 0.348]), and the touch screen manual input is the slowest (Average[MEAN, STDEV]=[25.548, 2.755]). Thus, the barcode and NFC technique is more efficient than touch screen for the interaction. From the curves in Fig. 9, it is clear that the time spent in the interaction regularly increases with the length of information input using the touch screen manual input approach. While it is consistent for the barcode scanning and NFC recognition, no matter how long the information to input is.

From Fig. 9, it is found that the standard deviation of the gathered data varies. The standard deviation of touch screen manual input increases with the length of the input data. That means the longer the input data is, the more uncertainty of the time spent of the input will be. While the standard deviation of the other two approaches remains small and varies smoothly. Namely, no matter how long the data to input, the time spent is constant. The barcode scanning and NFC is more stable in the interaction.

3.2.2 Performance Analysis of Interaction Techniques on Different Platforms

The touch screen manual input and barcode scanning on different mobile platforms are also tested with user name and password length of 8 digits that are randomly generated, and the results are given in Fig. 10 and 11. The condition of the testing environment is considered to guarantee that the external affect can be ignored and will not become a critical problem influencing the performance of the system.

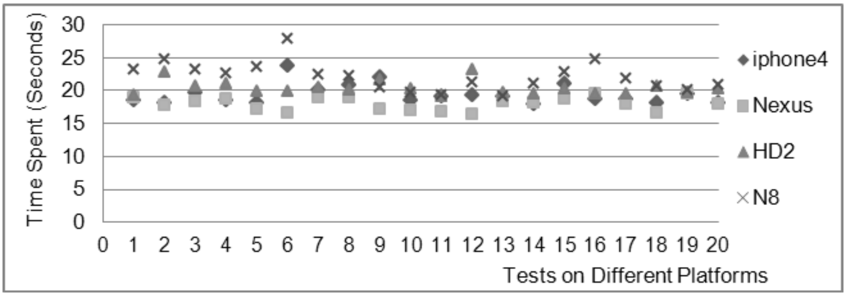


Fig. 10. Time spent (Seconds) by touch screen manual input

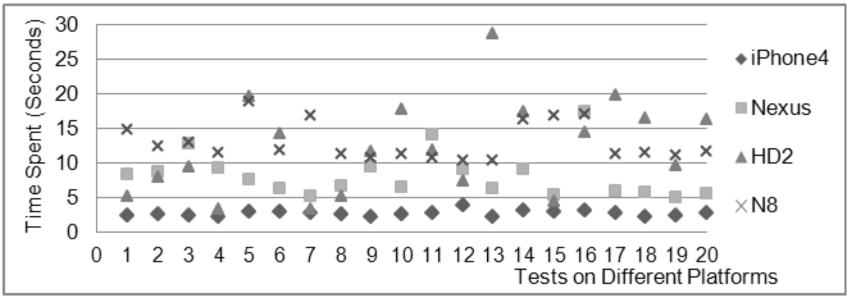
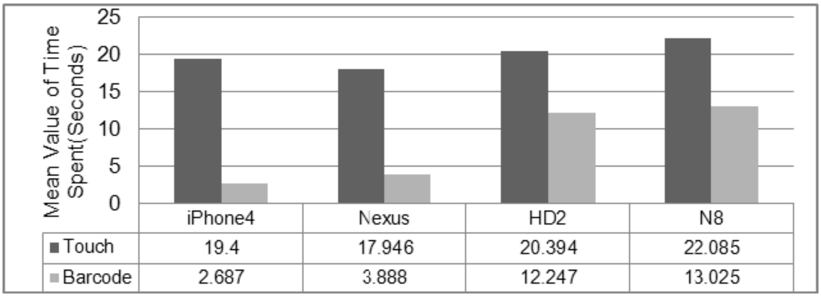


Fig. 11. Time spent (Seconds) by barcode scanning

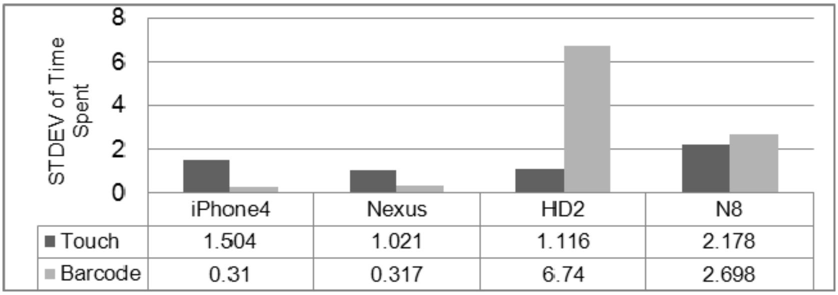
With the raw data, the mean value of parameters and standard deviations are calculated to evaluate the efficiency and stability. The mean value, standard deviation, and error rate of the two technologies barcode scanning and NFC recognition on different platforms are presented in Fig. 12-14.



	iPhone4	Nexus	HD2	N8
■ Touch	19.4	17.946	20.394	22.085
■ Barcode	2.687	3.888	12.247	13.025

Fig. 12. Time spent (seconds) of touch screen and barcode scanning

Fig. 12 shows the mean time spent in the user authentication in touch screen manual input and barcode scanning. It reveals that the barcode scanning approach is faster than the touch screen, especially on iPhone4 (MEAN=2.687S) and Galaxy Nexus (MEAN=3.888S). Moreover, the operation speed of an approach varies on the different mobile platforms. For manual input method, the operation on Galaxy Nexus (MEAN=17.946S) is the fastest because of the biggest screen size (4.65 Inches). Conversely, the operation on Nokia N8 (MEAN=22.085S) is the slowest due to the smallest screen size (3.5 Inches) and the interface is not convenient as that of iPhone4 (MEAN=19.400S) with the same size. For barcode scanning, the operation speeds vary remarkably across mobile platforms. The reason may lie in the control efficiency of the camera and the quality of camera lens. The HTC HD2 (MEAN=12.247S) and Nokia N8 (MEAN=13.025S) are much slower, because the control of the camera is not efficient and it adjusts again and again in the scanning.



	iPhone4	Nexus	HD2	N8
■ Touch	1.504	1.021	1.116	2.178
■ Barcode	0.31	0.317	6.74	2.698

Fig. 13. Standard deviation and of touch screen and barcode scanning

Fig. 13 shows the standard deviation of time spent in user authentication of these two technologies. The results illustrate that the standard deviation of the manual input on different devices is basically consistent. That is because manual input is not easily affected by some objective conditions. Conversely, the barcode approach is high in standard deviation except the iPhone4 (STDEV=0.310) and Galaxy Nexus (STDEV=0.317) as the camera on iPhone 4 and Galaxy Nexus is of high quality and focuses fast. The Nokia N8 (STDEV=2.698) is high in standard deviation because the control of the camera is not very fluent, and the HTC HD2 (STDEV=6.74) is high in speed standard deviation because the control of the camera is very slow. Besides the focus speed of the camera, the quality of image is also an influencing factor.

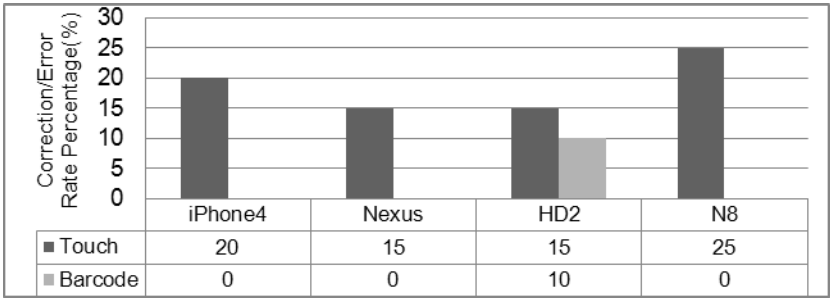


Fig. 14. Correction rate (touch screen) and error rate (barcode scanning)

Finally, Fig. 14 depicts the correction rate in touch screen manual input and error rate in barcode scanning operations. The correction rate is calculated as the percentage of the corrected logins, which is determined by the screen size and the usability of the soft keypad. The Galaxy Nexus and HTC HD2 are lower in correction because the screen sizes are bigger, and Nokia N8 is the highest because its keys on soft keypad are too small. On the other hand, the error rate of HTC HD2 is high possibly because the camera is susceptible to light intensity and the image quality is usually not good.

3.3 Summary

Table 2 summarizes the three methods and their performance in user interaction. All the listed attributes may affect the efficiency and usability of the interaction technologies applied in mobile applications. The touch screen is slow and complex in operation, high in error rate of operation, vulnerable to interference such as user behavior, its strength is low in computation intensity, input arbitrary information, and can be used in any environment and any kind of applications. The camera barcode scanning is fast in operation speed and low in operation complexity and error rate, but it is vulnerable to interference, high in computation intensity, input only fixed information, and can be used only in appropriate environment and with mobile native applications.

Table 2. Mobile sensing technologies and the attributes

Attributes	Touch Screen	Camera(Barcode)	NFC
Operation speed	Slow	Fast	Fast
Operation complexity	High	Low	Low
Error rate	High	Low	Low
Vulnerable to interference	True	True	False
Computation intensity	Low	High	Low
Input data arbitrary	High	Low	Low
Use cases (environment)	Arbitrary	Limited	Arbitrary
Generality &Compatibility	Arbitrary	Native App	Native App

The NFC recognition approach is similar to camera barcode scanning, but it is resistant to interference and low in computation intensity.

4 Conclusion

This investigation employs the novel hardware and sensing technologies to simplify the user interaction and promote the efficiency and usability of mobile applications. The barcode scanning and NFC recognition methods are successfully implemented on the mainstream mobile operating systems. Result verifies the feasibility and advantages they are used to improve the user interaction of mobile applications.

From the evaluation, it is evident that the application of the technologies can effectively help relieve the inherent problems in operational efficiency of touch screen manual input because of the limited screen size. These technologies convert the manual keyboard input operation into simple and convenient actions. It firstly simplifies the operations and accelerates the speed of input, and then reduces error rate in the operations. The efficiency and usability of the interaction is therefore raised.

The performance of interaction techniques differs in speed and accuracy among the mobile platforms, depending on the capability of the hardware module and software platform. Due to the heterogeneity in mobile devices, how much the sensing technology can enhance user interaction dependents on the coordination of hardware and related processing algorithms. In addition, from the testing we also find that sensing technologies may be affected by ambient environments. Therefore, the technologies can make sense only when the interference in the ambient environment is acceptable.

Acknowledgements. This research project has been funded with support from the European Commission, Project Number: 511485-LLP-1-2010-1-NO-KA3-KA3MP. The views expressed in this publication are those of the authors and do not necessarily represent the views of the Commission. The Commission cannot be held responsible for any use which may be made of the information contained therein.

References

1. Kanjo, E., Bacon, J., Landshoff, P., Roberts, D.: MobiSens: Making Smart Phones Smarter. IEEE Pervasive Computing 8(4), 50–57 (2009)
2. Zhou, D., Chander, A., Inamura, H.: Optimizing User Interaction for Web-based Mobile Tasks. In: Proc. of 2010 10th Annual International Symposium on Applications and the Internet, pp. 68–76 (2010)
3. Khan, W., Xiang, Y., Aalsalem, M., Arshad, Q.: Mobile Phone Sensing System: A Survey. IEEE Communications Survey & Tutorials 99, 1–26 (2012)
4. Alexander, A.: Smartphone Usage Statistics 2013 (2012), http://ansonalex.com/infographics/smartphone-usage-statistics-2012-infographic/
5. Hinckley, K., Pierce, J., Sinclair, M., Horvitz, E.: Sensing Techniques for Mobile Interaction. In: Proc. of the 13th Annual Symposium on User Interaction Software and Technology, pp. 91–100 (2000)

6. Feng, J., Johnston, M., Bangalore, S.: Speech and Multimodal Interaction in Mobile Search. IEEE Signal Processing Magazine 28(4), 40–49 (2011)
7. Hannuksela, J., Sangi, P., Heikkila, J.: Motion-based Handwriting Recognition for Mobile Interaction. In: Proc. of 18th International Conference on Pattern Recognition, vol. 4, pp. 397–400 (2006)
8. Broll, G., Siorpaes, S., Rukzio, E., Paolucci, M., Hamard, J., Wagner, M., Schmidt, A.: Supporting Mobile Service Usage through Physical Mobile Interaction. In: Proc. of Fifth Annual IEEE International Conference on Pervasive Computing and Communications, pp. 262–271 (2007)
9. Hornbæk, K.: Current Practice in Measuring Usability: Challenges to Usability Studies and Research. International Journal of Human-Computer Studies 64, 79–102 (2006)
10. SB Electronics: Barcode symbology – Information and History. SB Electronics Systems Ltd. (2012), `http://www.telepen.co.uk/telepen_symbology.htm`
11. ZXing Group: ZXing – Multi-format 1D/2D Barcode Image Processing Library with Clients for Android, Java (2012), `http://code.google.com/p/zxing`
12. Sheng, Q.Z., Li, X., Zeadally, S.: Enabling Next-generation RFID Applications: Solutions and Challenges. IEEE Computer 41(9), 21–28 (2008)

Horn-TeamLog: A Horn Fragment of TeamLog with PTime Data Complexity

Barbara Dunin-Kęplicz[1], Linh Anh Nguyen[1,2], and Andrzej Szałas[1,3]

[1] Institute of Informatics, University of Warsaw
Banacha 2, 02-097 Warsaw, Poland
{keplicz,nguyen,andsz}@mimuw.edu.pl
[2] Faculty of Information Technology, VNU University of Engineering and Technology
144 Xuan Thuy, Hanoi, Vietnam
[3] Dept. of Computer and Information Science, Linköping University
SE-581 83 Linköping, Sweden

Abstract. The logic TeamLog proposed by Dunin-Kęplicz and Verbrugge is used to express properties of agents' cooperation in terms of individual, bilateral and collective informational and motivational attitudes like beliefs, goals and intentions. In this paper we isolate a Horn fragment of TeamLog, called Horn-TeamLog, and we show that it has PTime data complexity.

1 Introduction

Horn fragments of applied logics have attracted a lot of attention. They allow to express rules in the form of implication $\varphi_1 \land \ldots \land \varphi_k \to \psi$ which are widely used in practice of logic programming, deductive databases as well as knowledge representation and reasoning. Moreover, Horn fragments typically enjoy PTime combined or data complexity, in contrast to (at least) NP-hardness of full logics.

Horn fragments of some basic modal logics were studied in [1, 8, 10]. In [12] Nguyen studied a deterministic Horn fragment of test-free PDL (propositional dynamic logic [9]). Following his approach, in [3] we formulated a Horn fragment of serial PDL with test and proved that it has PTime data complexity. In [13] Nguyen studied a Horn fragment of serial regular grammar logics without converse. Later, he extended the method to formulate some Horn fragments of regular description logics with PTime data complexity [14]. Recently, Nguyen and Szałas [16] proved that Horn fragments of serial regular grammar logics with converse enjoy PTime data complexity.

TeamLog [5–7] is a multimodal logic for specifying teamwork in terms of agents' individual, bilateral and group beliefs, goals and intentions. Individual beliefs, goals and intentions are expressed by using modal operators of $KD45$, K and KD, respectively. That is, beliefs satisfy positive and negative introspection, beliefs and intentions are consistent, while goals may be inconsistent. Interaction between is specified by TeamLog axioms stating positive and negative introspection for goals and intentions as well as by an axiom ensuring that intentions are included in goals. Moreover, TeamLog has modal operators for expressing common beliefs and mutual intentions of groups of agents.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 143–153, 2013.

TeamLog, as studied in this paper, is extended with state names and ABoxes. A knowledge base in TeamLog is a pair $\langle \Gamma, \mathcal{A} \rangle$, where Γ is a finite set of global assumptions and $\mathcal{A}$ is a finite set of assertions about named states (called an ABox). Thus, in our setting, a knowledge base contains information about a number of possible worlds. This follows the approach of description logics and contrasts with the traditional setting of modal logics, where a knowledge base usually contains only information about the actual world. Data complexity is measured with respect to the size of the ABox $\mathcal{A}$, while assuming that the assertions in $\mathcal{A}$ are in the "reduced form" (i.e., atomic formulas) and the set Γ of global assumptions is fixed. Note that the complexity of full TeamLog is ExpTime while its known restrictions are, at best, NP-hard.

In this paper we isolate a Horn fragment of TeamLog, called Horn-TeamLog, with PTime data complexity. Our Horn-TeamLog is a general Horn fragment of TeamLog, where we only impose restrictions necessary to eliminate nondeterminism. We develop an algorithm with PTime data complexity for checking satisfiability of a Horn-TeamLog knowledge base. It adopts Nguyen's techniques of using minimal states [10, 13], global caching [12–14] and dealing with non-seriality [11, 12, 14], applied for Horn fragments of modal and description logics. Other key techniques for our algorithm are the ones proposed recently by Nguyen and Szałas [16] for dealing with converse modalities and by Nguyen et al. [15] for dealing with non-seriality in the presence of converse.

2 Syntax and Semantics of TeamLog

We consider a version of TeamLog [2, 5–7] extended with state names and ABoxes. It uses a set Σ_P of *propositions* (called *atomic formulas*), a finite set Σ_A of *agents* and a set Σ_S of *state names*. We use letters like p, q to denote propositions, letters like σ, ϱ to denote agents, and letters like a, b to denote state names. *Formulas* of TeamLog are defined using the following BNF grammar, with $p \in \Sigma_P$, $\sigma \in \Sigma_A$ and $\emptyset \subset g \subseteq \Sigma_A$:

$$\begin{aligned}\varphi ::= \top \mid \bot \mid p \mid \neg\varphi \mid \varphi \to \varphi \mid \varphi \land \varphi \mid \varphi \lor \varphi \mid \textsc{Bel}_\sigma(\varphi) \mid \text{E-}\textsc{Bel}_g(\varphi) \mid \\ \mid \text{C-}\textsc{Bel}_g(\varphi) \mid \textsc{Goal}_\sigma(\varphi) \mid \textsc{Int}_\sigma(\varphi) \mid \text{E-}\textsc{Int}_g(\varphi) \mid \text{M-}\textsc{Int}_g(\varphi)\end{aligned}$$

The intuitive meaning of modalities is the following:

- $\textsc{Bel}_\sigma(\varphi)$ – agent σ has the belief that φ,
- $\text{E-}\textsc{Bel}_g(\varphi)$ – every agent in group g believes that φ,
- $\text{C-}\textsc{Bel}_g(\varphi)$ – group g has the common belief that φ,
- $\textsc{Goal}_\sigma(\varphi)$ – agent σ has the goal to achieve φ,
- $\textsc{Int}_\sigma(\varphi)$ – agent σ has the intention to achieve φ,
- $\text{E-}\textsc{Int}_g(\varphi)$ – every agent in group g has the individual intention to achieve φ,
- $\text{M-}\textsc{Int}_g(\varphi)$ – group g has the mutual intention to achieve φ.

We define $\text{C-}\textsc{Int}_g(\varphi) = \text{M-}\textsc{Int}_g(\varphi) \land \text{C-}\textsc{Bel}_g(\text{M-}\textsc{Int}_g(\varphi))$. It states that group g collectively intends to achieve φ. For $\sigma \in \Sigma_A$, let B_σ, G_σ, I_σ be binary predicates corresponding to accessibility relations between states of agent σ w.r.t. beliefs, goals, intentions, respectively. Let $\Sigma_{m_+} = \{B_\sigma, G_\sigma, I_\sigma \mid \sigma \in \Sigma_A\}$.

An *interpretation* is a pair $\mathcal{I} = \langle \Delta^\mathcal{I}, \cdot^\mathcal{I} \rangle$, where $\Delta^\mathcal{I}$ is a set of *states* (or *possible worlds*), and $\cdot^\mathcal{I}$ is an interpretation function that maps each state name $a \in \Sigma_S$ to an element $a^\mathcal{I} \in \Delta^\mathcal{I}$, each proposition $p \in \Sigma_P$ to a subset $p^\mathcal{I}$ of $\Delta^\mathcal{I}$, and each predicate $R \in \Sigma_{m_+}$ to a binary relation $R^\mathcal{I}$ on $\Delta^\mathcal{I}$.

Let $g \subseteq \Sigma_A$, $\mathcal{I}$ be an interpretation, and x be a state of $\mathcal{I}$. A state y of $\mathcal{I}$ is called g_B*-reachable* (respectively, g_I*-reachable*) from x if $\langle x, y \rangle \in (\bigcup_{\sigma \in g} B_\sigma^\mathcal{I})^+$ (respectively, $\langle x, y \rangle \in (\bigcup_{\sigma \in g} I_\sigma^\mathcal{I})^+$). The *satisfaction relation* $\mathcal{I}, x \models \varphi$ is defined as usual for classical connectives and as follows for modalities:

- $\mathcal{I}, x \models \textsc{Bel}_\sigma(\varphi)$ iff $\forall y\ [B_\sigma^\mathcal{I}(x,y) \to \mathcal{I}, y \models \varphi]$
- $\mathcal{I}, x \models \textsc{Goal}_\sigma(\varphi)$ iff $\forall y\ [G_\sigma^\mathcal{I}(x,y) \to \mathcal{I}, y \models \varphi]$
- $\mathcal{I}, x \models \textsc{Int}_\sigma(\varphi)$ iff $\forall y\ [I_\sigma^\mathcal{I}(x,y) \to \mathcal{I}, y \models \varphi]$
- $\mathcal{I}, x \models \text{E-}\textsc{Bel}_g(\varphi)$ iff $\forall \sigma \in g\ [\mathcal{I}, x \models \textsc{Bel}_\sigma(\varphi)]$
- $\mathcal{I}, x \models \text{C-}\textsc{Bel}_g(\varphi)$ iff $\mathcal{I}, y \models \varphi$ for all y that are g_B-reachable from x
- $\mathcal{I}, x \models \text{E-}\textsc{Int}_g(\varphi)$ iff $\forall \sigma \in g\ [\mathcal{I}, x \models \textsc{Int}_\sigma(\varphi)]$
- $\mathcal{I}, x \models \text{M-}\textsc{Int}_g(\varphi)$ iff $\mathcal{I}, y \models \varphi$ for all y that are g_I-reachable from x.

We define $\varphi^\mathcal{I} = \{x \in \Delta^\mathcal{I} \mid \mathcal{I}, x \models \varphi\}$ and, for a set X of formulas, $X^\mathcal{I} = \bigcap\{\varphi^\mathcal{I} \mid \varphi \in X\}$.

An interpretation $\mathcal{I}$ is a TEAMLOG *interpretation* if the following conditions hold for every $\sigma \in \Sigma_A$:

$$\forall x \exists y\ B_\sigma^\mathcal{I}(x,y) \tag{1}$$

$$\forall x \exists y\ I_\sigma^\mathcal{I}(x,y) \tag{2}$$

$$\forall x \forall y \forall z\ [(B_\sigma^\mathcal{I}(x,y) \land B_\sigma^\mathcal{I}(y,z)) \to B_\sigma^\mathcal{I}(x,z)] \tag{3}$$

$$\forall x \forall y, z\ [(B_\sigma^\mathcal{I}(x,y) \land B_\sigma^\mathcal{I}(x,z)) \to B_\sigma^\mathcal{I}(y,z)] \tag{4}$$

$$\forall x \forall y \forall z\ [(B_\sigma^\mathcal{I}(x,y) \land G_\sigma^\mathcal{I}(y,z)) \to G_\sigma^\mathcal{I}(x,z)] \tag{5}$$

$$\forall x \forall y \forall z\ [(B_\sigma^\mathcal{I}(x,y) \land G_\sigma^\mathcal{I}(x,z)) \to G_\sigma^\mathcal{I}(y,z)] \tag{6}$$

$$\forall x \forall y \forall z\ [(B_\sigma^\mathcal{I}(x,y) \land I_\sigma^\mathcal{I}(y,z)) \to I_\sigma^\mathcal{I}(x,z)] \tag{7}$$

$$\forall x \forall y \forall z\ [(B_\sigma^\mathcal{I}(x,y) \land I_\sigma^\mathcal{I}(x,z)) \to I_\sigma^\mathcal{I}(y,z)] \tag{8}$$

$$\forall x \forall y [G_\sigma^\mathcal{I}(x,y) \to I_\sigma^\mathcal{I}(x,y)]. \tag{9}$$

Conditions (1)-(9) correspond respectively to the following properties, which are valid in every TEAMLOG interpretation:

- belief consistency: $\neg\textsc{Bel}_\sigma(\bot)$
- intention consistency: $\neg\textsc{Int}_\sigma(\bot)$
- positive introspection for beliefs: $\textsc{Bel}_\sigma(\varphi) \to \textsc{Bel}_\sigma(\textsc{Bel}_\sigma(\varphi))$
- negative introspection for beliefs: $\neg\textsc{Bel}_\sigma(\varphi) \to \textsc{Bel}_\sigma(\neg\textsc{Bel}_\sigma(\varphi))$
- positive introspection for goals: $\textsc{Goal}_\sigma(\varphi) \to \textsc{Bel}_\sigma(\textsc{Goal}_\sigma(\varphi))$
- negative introspection for goals: $\neg\textsc{Goal}_\sigma(\varphi) \to \textsc{Bel}_\sigma(\neg\textsc{Goal}_\sigma(\varphi))$
- positive introspection for intentions: $\textsc{Int}_\sigma(\varphi) \to \textsc{Bel}_\sigma(\textsc{Int}_\sigma(\varphi))$

- negative introspection for intentions: $\neg\textsc{Int}_\sigma(\varphi) \to \textsc{Bel}_\sigma(\neg\textsc{Int}_\sigma(\varphi))$
- intention implies goal: $\textsc{Int}_\sigma(\varphi) \to \textsc{Goal}_\sigma(\varphi)$.

An *ABox* is a finite set of *assertions* of the form $p(a)$, $\neg p(a)$ or $R(a,b)$, where $p \in \Sigma_P$, $a, b \in \Sigma_S$ and $R \in \Sigma_{m+}$. A *knowledge base* is a pair $\langle \Gamma, \mathcal{A} \rangle$, where Γ is a finite set of formulas and $\mathcal{A}$ is an ABox. The formulas of Γ are called *global assumptions* of the knowledge base.

An interpretation $\mathcal{I}$ *validates* a global assumption φ if $\varphi^\mathcal{I} = \Delta^\mathcal{I}$. It *validates* an assertion $p(a)$ (respectively, $\neg p(a)$, $R(a,b)$) if $a^\mathcal{I} \in p^\mathcal{I}$ (respectively, $a^\mathcal{I} \notin p^\mathcal{I}$, $\langle a^\mathcal{I}, b^\mathcal{I} \rangle \in R^\mathcal{I}$). A TeamLog *model* of a knowledge base $\langle \Gamma, \mathcal{A} \rangle$ is a TeamLog interpretation validating all global assumptions of Γ and all assertions of $\mathcal{A}$. A knowledge base is TeamLog *satisfiable* if it has a TeamLog model.

We say that a state name *a satisfies* the property φ w.r.t. a knowledge base KB, denoted by $KB \models \varphi(a)$, if, for every TeamLog model $\mathcal{I}$ of KB, we have that $a^\mathcal{I} \in \varphi^\mathcal{I}$.

3 Horn-TeamLog

We will denote modal operators $\textsc{Bel}_\sigma$, E-$\textsc{Bel}_g$, C-$\textsc{Bel}_g$, $\textsc{Goal}_\sigma$, $\textsc{Int}_\sigma$, E-$\textsc{Int}_g$, M-$\textsc{Int}_g$ using the forms $[B_\sigma]$, $[EB_g]$, $[CB_g]$, $[G_\sigma]$, $[I_\sigma]$, $[EI_g]$, $[MI_g]$, respectively. These are *universal* modal operators. We use also dual *existential* modal operators $\langle B_\sigma \rangle$, $\langle EB_g \rangle$, $\langle CB_g \rangle$, $\langle G_\sigma \rangle$, $\langle I_\sigma \rangle$, $\langle EI_g \rangle$, $\langle MI_g \rangle$, with semantics defined as usual, e.g., $\langle B_\sigma \rangle \varphi$ is treated as $\neg[B_\sigma]\neg\varphi$.

As TeamLog does not require "goal consistency", for the Horn-TeamLog defined here, we use also the modal operator $[G_\sigma]_\diamond$ defined by:

$$[G_\sigma]_\diamond \varphi \equiv [G_\sigma]\varphi \wedge \langle G_\sigma \rangle \varphi \ \ (\equiv [G_\sigma]\varphi \wedge \neg[G_\sigma]\neg\varphi).$$

Let $\Sigma_{M+} = \{B_\sigma, G_\sigma, I_\sigma, EB_g, CB_g, EI_g, MI_g \mid \sigma \in \Sigma_A \text{ and } \emptyset \subset g \subseteq \Sigma_A\}$.

A formula is called a *positive* $\Box_\diamond$*-formula* if it is constructed from $\top$ and propositions of Σ_P using $\wedge$, $\vee$ and the modal operators $[R]$, $\langle R \rangle$, $[G_\sigma]_\diamond$, $\langle G_\sigma \rangle$, where $R \in \Sigma_{M+} \setminus \{G_\varrho \mid \varrho \in \Sigma_A\}$ and $\sigma \in \Sigma_A$.

A *negative* $\Box_\diamond$*-formula* is a formula $\neg\varphi$ with φ being a positive $\Box_\diamond$-formula. A Horn-TeamLog *formula* is a formula of one of the following forms:

- $\top$, a proposition of Σ_P, a negative $\Box_\diamond$-formula,
- $\varphi \wedge \psi$, where φ and ψ are Horn-TeamLog formulas,
- $\varphi \vee \psi$, where φ is a negative $\Box_\diamond$-formula and ψ is a Horn-TeamLog formula,
- $\langle R \rangle \varphi$, where φ is a Horn-TeamLog formula and $R \in \Sigma_{m+}$,
- $[R]\varphi$, where φ is a Horn-TeamLog formula and $R \in \Sigma_{M+}$.

Note that, if φ is a Horn-TeamLog formula, then so is C-$\textsc{Int}_g(\varphi)$.

A knowledge base $\langle \Gamma, \mathcal{A} \rangle$ is called a Horn-TeamLog *knowledge base* if Γ consists of Horn-TeamLog *formulas*. The following proposition is trivial.

Proposition 3.1. *Given a* Horn-TeamLog *knowledge base* $\langle \Gamma, \mathcal{A} \rangle$*, a positive* $\Box_\diamond$*-formula* φ *and* $\sigma \in \Sigma_S$*,* $\langle \Gamma, \mathcal{A} \rangle \models \varphi(a)$ *iff the* Horn-TeamLog *knowledge base* $\langle \Gamma \cup \{\neg\varphi \vee p\}, \mathcal{A} \cup \{\neg p(a)\} \rangle$ *is* TeamLog *unsatisfiable, where p is a fresh proposition.*

A *modal context* is a (possibly empty) sequence of modal operators of the form $[R]$ with $R \in \Sigma_{M_+}$. A Horn-TeamLog *clause* is a formula of the form

$$\boxdot(A_1 \land \ldots \land A_k \to B) \;\text{ or }\; \boxdot\,(A_1 \land \ldots \land A_k \to \bot), \;\text{ where:}$$

- $\boxdot$ is a modal context and $k \geq 0$,
- $A_1, \ldots, A_k$ are formulas of the form p, $[R]p$, $\langle R\rangle p$, $[G_\sigma]_\diamond p$ or $\langle G_\sigma\rangle p$, where $p \in \Sigma_P$, $R \in \Sigma_{M_+} \setminus \{G_\varrho \mid \varrho \in \Sigma_A\}$ and $\sigma \in \Sigma_A$,
- B is a formula of the form p, $[R]p$ or $\langle S\rangle p$, where $p \in \Sigma_P$, $R \in \Sigma_{M_+}$ and $S \in \Sigma_{m_+}$.

Notice that in Horn-TeamLog clauses:

- at the left hand side of $\to$, the operator $[G_\sigma]_\diamond$ is used instead of $[G_\sigma]$,
- at the right hand side of $\to$, the operators $\langle EB_g\rangle$, $\langle CB_g\rangle$, $\langle EI_g\rangle$ and $\langle MI_g\rangle$ are disallowed.

Every Horn-TeamLog clause can be represented as a Horn-TeamLog formula. It can be shown that the language of Horn-TeamLog clauses is as expressive as the language of Horn-TeamLog formulas [4].

By the *size* of an ABox we mean its cardinality. The *data complexity* class of Horn-TeamLog is defined to be the complexity class of the problem of checking satisfiability of a Horn-TeamLog knowledge base $\langle \Gamma, \mathcal{A}\rangle$, measured in the size of $\mathcal{A}$ when assuming that Γ is a fixed set consisting of Horn-TeamLog clauses.

4 Automaton-Modal Operators

For $R \in \Sigma_{M_+}$, let $\overline{R}$ be a new predicate standing for the *converse* of R. Let $\Sigma_{M_-} = \{\overline{R} \mid R \in \Sigma_{M_+}\}$, $\Sigma_M = \Sigma_{M_+} \cup \Sigma_{M_-}$, $\Sigma_{m_-} = \{\overline{R} \mid R \in \Sigma_{m_+}\}$ and $\Sigma_m = \Sigma_{m_+} \cup \Sigma_{m_-}$. For $R = \overline{S} \in \Sigma_{M_-}$, let $\overline{R}$ stand for S. For a word $w = R_1 \ldots R_k$ over alphabet Σ_m, the converse of w is defined to be $\overline{w} = \overline{R_k} \ldots \overline{R_1}$.

Recall that a *finite automaton* A over alphabet Σ_m is a tuple $\langle \Sigma_m, Q, q_0, \delta, F\rangle$, where Q is a finite set of states, $q_0 \in Q$ is the initial state, $\delta \subseteq Q \times \Sigma_m \times Q$ is the transition relation, and $F \subseteq Q$ is the set of accepting states. A *run* of A on a word $R_1 \ldots R_k$ over alphabet Σ_m is a finite sequence of states $q_0, q_1, \ldots, q_k$ such that $\delta(q_{i-1}, R_i, q_i)$ holds for every $1 \leq i \leq k$. It is an *accepting run* if $q_k \in F$. We say that A *accepts* a word w if there exists an accepting run of A on w.

Given an interpretation $\mathcal{I}$ and $R \in \Sigma_{m_+}$, define

$$\overline{R}^{\mathcal{I}} = (R^{\mathcal{I}})^{-1} = \{\langle y, x\rangle \mid \langle x, y\rangle \in R^{\mathcal{I}}\}.$$

For a finite automaton A over alphabet Σ_m, define

$$\begin{aligned}\mathsf{A}^{\mathcal{I}} = \{\langle x, y\rangle \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \mid\ & \text{there exist a word } R_1 \ldots R_k \text{ accepted by } \mathsf{A} \\ & \text{and elements } x_0 = x,\ x_1,\ \ldots,\ x_k = y \text{ of } \Delta^{\mathcal{I}} \\ & \text{such that } \langle x_{i-1}, x_i\rangle \in R_i^{\mathcal{I}} \text{ for all } 1 \leq i \leq k\}.\end{aligned}$$

We will use auxiliary modal operators $[\mathsf{A}]$ and $\langle \mathsf{A}\rangle$, where A is a finite automaton over alphabet Σ_m. We call $[\mathsf{A}]$ (respectively, $\langle \mathsf{A}\rangle$) a *universal* (respectively, *existential*) *automaton-modal operator*. In the *extended language*, if φ is a formula then $[\mathsf{A}]\varphi$ and $\langle \mathsf{A}\rangle\varphi$ are also formulas.

The semantics of $[\mathsf{A}]\varphi$ and $\langle\mathsf{A}\rangle\varphi$ are defined as follows:

$$([\mathsf{A}]\varphi)^{\mathcal{I}} = \big\{x \in \Delta^{\mathcal{I}} \mid \forall y\big(\langle x, y\rangle \in \mathsf{A}^{\mathcal{I}} \text{ implies } y \in \varphi^{\mathcal{I}}\big)\big\}$$
$$(\langle\mathsf{A}\rangle\varphi)^{\mathcal{I}} = \big\{x \in \Delta^{\mathcal{I}} \mid \exists y\big(\langle x, y\rangle \in \mathsf{A}^{\mathcal{I}} \text{ and } y \in \varphi^{\mathcal{I}}\big)\big\}.$$

For a finite automaton A over Σ_m, assume that $\mathsf{A} = \langle \Sigma_m, Q_{\mathsf{A}}, q_{\mathsf{A}}, \delta_{\mathsf{A}}, F_{\mathsf{A}}\rangle$.

If q is a state of a finite automaton A then by A_q we denote the finite automaton obtained from A by replacing the initial state by q.

For $R \in \Sigma_{M+}$, let $\mathit{RegExp}(R)$ be the regular expression over alphabet Σ_m defined as follows:

$$\begin{array}{ll}
\mathit{RegExp}(B_\sigma) & = (B_\sigma \cup \overline{B}_\sigma)^*; B_\sigma \\
\mathit{RegExp}(G_\sigma) & = (B_\sigma \cup \overline{B}_\sigma)^*; G_\sigma \\
\mathit{RegExp}(I_\sigma) & = (B_\sigma \cup \overline{B}_\sigma)^*; (I_\sigma \cup G_\sigma) \\
\mathit{RegExp}(EB_{\{\sigma_1,\ldots,\sigma_k\}}) & = \mathit{RegExp}(B_{\sigma_1}) \cup \ldots \cup \mathit{RegExp}(B_{\sigma_k}) \\
\mathit{RegExp}(CB_g) & = (\mathit{RegExp}(EB_g))^*; \mathit{RegExp}(EB_g) \\
\mathit{RegExp}(EI_{\{\sigma_1,\ldots,\sigma_k\}}) & = \mathit{RegExp}(I_{\sigma_1}) \cup \ldots \cup \mathit{RegExp}(I_{\sigma_k}) \\
\mathit{RegExp}(MI_g) & = (\mathit{RegExp}(EI_g))^*; \mathit{RegExp}(EI_g)
\end{array}$$

For $R \in \Sigma_{M+}$, let $\mathbf{A}_R$ be a fixed finite automaton recognizing the regular language over alphabet Σ_m expressed by $\mathit{RegExp}(R)$, and let $\mathbf{A}_{\overline{R}}$ be a fixed finite automaton accepting only the words $\overline{w}$ such that w is accepted by $\mathbf{A}_R$.

5 Checking Satisfiability of Horn Knowledge Bases

In this section we present an algorithm that, given a Horn-TeamLog knowledge base $\langle \Gamma, \mathcal{A}\rangle$ with Γ consisting of Horn-TeamLog clauses, checks whether $\langle \Gamma, \mathcal{A}\rangle$ is TeamLog satisfiable.

Let X be a set of formulas. The *saturation* of X (w.r.t. Γ), denoted by $\mathsf{Satr}(X)$, is defined to be the least extension of X such that:

- if $[R]\varphi \in \mathsf{Satr}(X)$ then $[\mathbf{A}_R]\varphi \in \mathsf{Satr}(X)$,
- if $[\mathsf{A}]\varphi \in \mathsf{Satr}(X)$ and $q_{\mathsf{A}} \in F_{\mathsf{A}}$ then $\varphi \in \mathsf{Satr}(X)$,
- $\{\top, \langle B_\sigma\rangle\top, \langle I_\sigma\rangle\top\} \subseteq \mathsf{Satr}(X)$ for every $\sigma \in \Sigma_A$,
- if $[G_\sigma]_\diamond$ occurs in Γ then $[\mathbf{A}_{\overline{G}_\sigma}]\langle G_\sigma\rangle\top \in \mathsf{Satr}(X)$,
- for every $R \in \Sigma_{M+}$, if $p \in \mathsf{Satr}(X)$ and $\langle R\rangle p$ occurs at the left hand side of '$\to$' in some clause of Γ then $[\mathbf{A}_{\overline{R}}]\langle \mathbf{A}_R\rangle p \in \mathsf{Satr}(X)$.

For $R \in \Sigma_{m+}$, the *transfer* of X through $\langle R\rangle$ is defined as follows:

$$\mathsf{Trans}(X, R) = \{[\mathsf{A}_q]\varphi \mid [\mathsf{A}]\varphi \in X \text{ and } \langle q_{\mathsf{A}}, R, q\rangle \in \delta_{\mathsf{A}}\}.$$

Algorithm 1 checks satisfiability of a given Horn-TeamLog knowledge base $\langle \Gamma, \mathcal{A}\rangle$, where Γ consists of Horn-TeamLog clauses. It uses the following data structures:

- Δ_0 : the set of all state names occurring in the ABox $\mathcal{A}$,

Function Find(X)

```
if there exists $z \in \mathit{\Delta} \setminus \mathit{\Delta}_0$ with $Label(z) = X$ then return $z$ else add a new element $z$ to $\mathit{\Delta}$ with $Label(z) := X$ and return $z$
```

Procedure ExtendLabel(z, X)

```
if $\mathsf{Satr}(X) \subseteq Label(z)$ then return if $z \in \mathit{\Delta}_0$ then $Label(z) := Label(z) \cup \mathsf{Satr}(X)$ else // simulate changing label
    $z_* := \mathtt{Find}(Label(z) \cup \mathsf{Satr}(X))$;
    foreach $y$, $R$, $\varphi$ such that $Next(y, \langle R\rangle\varphi) = z$ do $Next(y, \langle R\rangle\varphi) := z_*$
```

Function CheckPremise($x, A_1 \wedge \ldots \wedge A_k$)

```
foreach $1 \leq i \leq k$ do
    if $A_i = p$ and $p \notin Label(x)$ then return false else if $A_i = \langle R\rangle p$ and $\langle \mathbf{A}_R\rangle p \notin Label(x)$ then return false else if $A_i = [G_\sigma]_\diamond p$ then
        if $Next(x, \langle G_\sigma\rangle\top)$ is not defined or $p \notin Label(Next(x, \langle G_\sigma\rangle\top))$ then
            return false
    else if $A_i = [R]p$ and ($R = B_\sigma$ or $R = I_\sigma$) then
        if $Next(x, \langle R\rangle\top)$ is not defined or $p \notin Label(Next(x, \langle R\rangle\top))$ then
            return false
    else if $A_i = [EB_g]p$ then
        foreach $\sigma \in g$ do
            if $Next(x, \langle B_\sigma\rangle\top)$ is not defined or $p \notin Label(Next(x, \langle B_\sigma\rangle\top))$ then
                return false
    else if $A_i = [EI_g]p$ then
        foreach $\sigma \in g$ do
            if $Next(x, \langle I_\sigma\rangle\top)$ is not defined or $p \notin Label(Next(x, \langle I_\sigma\rangle\top))$ then
                return false
    else if $A_i = [CB_g]p$ then
        if there exist $y$ $CB_g$-reachable from $x$ and $\sigma \in g$ such that $Next(y, \langle B_\sigma\rangle\top)$ is not defined or $p \notin Label(Next(y, \langle B_\sigma\rangle\top))$ then
            return false
    else if $A_i = [MI_g]p$ then
        if there exist $y$ $MI_g$-reachable from $x$ and $\sigma \in g$ such that
        $Next(y, \langle I_\sigma\rangle\top)$ is not defined or $p \notin Label(Next(y, \langle I_\sigma\rangle\top))$ then
            return false
return true
```

Algorithm 1. checking satisfiability of a Horn-TeamLog knowledge base

```
Input: a set $\Gamma$ of Horn-TeamLog clauses and an ABox $\mathcal{A}$.
Output: $true$ if $\langle \Gamma, \mathcal{A} \rangle$ is TeamLog satisfiable, or $false$ otherwise.

let $\Delta_0$ be the set of all state names occurring in $\mathcal{A}$;
if $\Delta_0 = \emptyset$ then $\Delta_0 := \{\tau\}$;
set $\Delta := \Delta_0$ and set $Next$ to the empty mapping;
foreach $a \in \Delta_0$ do $Label(a) := \mathsf{Satr}(\{\varphi \mid \varphi(a) \in \mathcal{A}\} \cup \Gamma)$;
repeat
    foreach $R(a,b) \in \mathcal{A}$ do $Label(b) := Label(b) \cup \mathsf{Satr}(\mathsf{Trans}(Label(a), R))$;
    foreach $x$, $\langle R \rangle \varphi$, $y$ s.t. $Next(x, \langle R \rangle \varphi) = y$ and $x$ is reachable from $\Delta_0$ do
        $y_* := \mathtt{Find}(Label(y) \cup \mathsf{Satr}(\mathsf{Trans}(Label(x), R)))$;
        $Next(x, \langle R \rangle \varphi) := y_*$;
    foreach $\langle x, R, y \rangle \in Edges$ such that $x$ is reachable from $\Delta_0$ do
        $\mathtt{ExtendLabel}(x, \mathsf{Trans}(Label(y), \overline{R}))$
    foreach $x \in \Delta$ reachable from $\Delta_0$ and $\langle R \rangle \varphi \in Label(x)$ do
        if $Next(x, \langle R \rangle \varphi)$ is not defined then
            $Next(x, \langle R \rangle \varphi) := \mathtt{Find}(\mathsf{Satr}(\{\varphi\} \cup \mathsf{Trans}(Label(x), R) \cup \Gamma))$
    foreach $x \in \Delta$ reachable from $\Delta_0$ and $(\varphi \to \psi) \in Label(x)$ do
        if $\mathtt{CheckPremise}(x, \varphi)$ then $\mathtt{ExtendLabel}(x, \{\psi\})$
    if there exist $x \in \Delta$ and $\{p, \neg p\} \subseteq Label(x)$ or $\bot \in Label(x)$ then
        return $false$
until no changes occurred in the last iteration;
return $true$;
```

- Δ : a set of states containing Δ_0,
- *Label* : for every $x \in \Delta$, $Label(x)$ is a set of formulas called the *label* of x,
- *Next* : $\Delta \times \{\langle R \rangle \top, \langle R \rangle p \mid R \in \Sigma_{m_+}, p \in \Sigma_P\} \to \Delta$ is a partial function with the meaning that: $Next(x, \langle R \rangle \varphi) = y$ means $\langle R \rangle \varphi \in Label(x)$, $\varphi \in Label(y)$, and $\langle R \rangle \varphi$ is "realized" at x by going to y.

Define $Edges = \{\langle x, R, y \rangle \mid R(x,y) \in \mathcal{A}$ or $Next(x, \langle R \rangle \varphi) = y$ for some $\varphi\}$.

We say that $x \in \Delta$ is *reachable* from Δ_0 if there exist $x_0, \ldots, x_k \in \Delta$ and elements $R_1, \ldots, R_k$ of Σ_{m_+} such that $k \geq 0$, $x_0 \in \Delta_0$, $x_k = x$ and $\langle x_{i-1}, R_i, x_i \rangle \in Edges$ for all $1 \leq i \leq k$. Similarly, $x \in \Delta$ is CB_g*-reachable* (respectively, MI_g*-reachable*) from x_0 if $x = x_0$ or there exist $x_1, \ldots, x_k \in \Delta$ and elements $R_1, \ldots, R_k$ of $\{B_\sigma \mid \sigma \in g\}$ (respectively, $\{I_\sigma \mid \sigma \in g\}$) such that $x_k = x$ and $\langle x_{i-1}, R_i, x_i \rangle \in Edges$ for all $1 \leq i \leq k$.

Algorithm 1 tries to construct a TeamLog model of $\langle \Gamma, \mathcal{A} \rangle$. The intended model extends $\mathcal{A}$ with disjoint trees rooted at state names occurring in $\mathcal{A}$. The trees may be infinite. However, we represent such a semi-forest as a graph with global caching: if two unnamed states (i.e., states not occurring in $\mathcal{A}$) in a tree or in different trees have the same label, then they should be merged. In other words, for every finite set X of formulas, the graph contains at most one state (a node) $z \in \Delta \setminus \Delta_0$ such that $Label(z) = X$. The function $\mathtt{Find}(X)$ returns such a state z if it exists, or creates such a state z otherwise. A tuple $\langle x, R, y\rangle \in \mathit{Edges}$ represents an edge $\langle x, y\rangle$ with label $R \in \Sigma_{m_+}$ of the graph. The notions of *predecessor*, *successor*, *ancestor* and *descendant* are defined as usual.

For each $x \in \Delta$, $\mathit{Label}(x)$ is a set of requirements to be "realized" at x. To realize such requirements of states, sometimes we have to extend their labels. Suppose we want to extend the label of $z \in \Delta$ with a set X of formulas. Consider the following cases:

- Case $z \in \Delta_0$ (i.e., z is a named state occurring in $\mathcal{A}$): As z is "fixed" by the ABox $\mathcal{A}$, we have no choice but to extend $\mathit{Label}(z)$ directly with $\mathsf{Satr}(X)$.
- Case $z \notin \Delta_0$ and the requirements X are directly caused by z itself or its successors: If we directly extend the label of z (with $\mathsf{Satr}(X)$) then z will possibly have the same label as another state not belonging to Δ_0 and global caching is not fulfilled. Hence, we "simulate" changing the label of z by using $z_* := \mathtt{Find}(\mathit{Label}(z) \cup \mathsf{Satr}(X))$ for playing the role of z. In particular, for each y, R and φ such that $\mathit{Next}(y, \langle R\rangle\varphi) = z$, we set $\mathit{Next}(y, \langle R\rangle\varphi) := z_*$.

Extending the label of z for the above two cases is done by Procedure $\mathtt{ExtendLabel}(z, X)$. The third case in considered below.

Suppose that $\mathit{Next}(x, \langle R\rangle\varphi) = y$. Then, to realize the requirements at x, the label of y should be extended with $X = \mathsf{Satr}(\mathsf{Trans}(\mathit{Label}(x), R))$. How can we realize such an extension? Recall that we intend to construct a forest-like model for $\langle \Gamma, \mathcal{A}\rangle$, but use global caching to guarantee termination. There may exist another $\mathit{Next}(x', \langle R'\rangle\varphi') = y$ with $x' \neq x$. That is, we may use y as a successor for two different states x and x', but the intention is to put x and x' into disjoint trees. If we directly modify the label of y to realize the requirements of x, such a modification may affect x'. The solution is to delete the edge $\langle x, R, y\rangle$ and reconnect x to $y_* := \mathtt{Find}(\mathit{Label}(y) \cup X)$ by setting $\mathit{Next}(x, \langle R\rangle\varphi) := y_*$. The extension is formally realized by steps 7-9 of Algorithm 1.

Consider the other main steps of Algorithm 1:

- Step 6: If $R(a, b) \in \mathcal{A}$ then we directly extend the label of b with $\mathsf{Satr}(\mathsf{Trans}(\mathit{Label}(a), R))$.
- Steps 10-11: If $\langle x, R, y\rangle \in \mathit{Edges}$ then we extend the label of x with $\mathsf{Trans}(\mathit{Label}(y), \overline{R})$ by using the procedure $\mathtt{ExtendLabel}$ discussed earlier. Notice the converse $\overline{R}$.
- Steps 12-14: If $\langle R\rangle\varphi \in \mathit{Label}(x)$ (where $R \in \Sigma_{m_+}$) and $\mathit{Next}(x, \langle R\rangle\varphi)$ is not defined yet then to realize the requirement $\langle R\rangle\varphi$ at x we connect x via R to a node with label $X = \mathsf{Satr}(\{\varphi\} \cup \mathsf{Trans}(\mathit{Label}(x), R) \cup \Gamma)$ by setting $\mathit{Next}(x, \langle R\rangle\varphi) := \mathtt{Find}(X)$.
- Steps 15-16: If $(\varphi \to \psi) \in \mathit{Label}(x)$ and φ "holds" at x then we extend the label of x with $\{\psi\}$ by using the procedure $\mathtt{ExtendLabel}$ discussed earlier. Suppose $\varphi = A_1 \wedge \ldots \wedge A_k$. How to check whether φ "holds" at x? It "holds" at x if A_i "holds" at x for every $1 \leq i \leq k$. There are the following cases:
  - Case $A_i = p$: A_i "holds" at x if $p \in \mathit{Label}(x)$.
  - Case $A_i = \langle R\rangle p$: Whenever p appears in $\mathit{Label}(z)$ for some state z, we include $[\mathbf{A}_{\overline{R}}]\langle \mathbf{A}_R\rangle p$ in $\mathit{Label}(z)$ by saturation. Note that $p \to [\mathbf{A}_{\overline{R}}]\langle \mathbf{A}_R\rangle p$ is valid. Then, to check whether $A_i = \langle R\rangle p$ "holds" at x we just check

whether $\langle \mathbf{A}_R \rangle p \in \mathit{Label}(x)$. (Semantically, $\langle \mathbf{A}_R \rangle p$ is equivalent to $\langle R \rangle p$.) The reason for using this technique is due to the use of global caching (in order to guarantee termination). We do global caching to represent a possibly infinite semi-forest by a finite graph possibly with cycles. As a side effect, direct checking "realization" of existential automaton-modal operators is not safe. Furthermore, we cannot allow universal modal operators "run" along such cycles. "Running" universal modal operators backward along an edge is safe, but "running" universal modal operators forward along an edge is done using a special technique, which may replace the edge by another one as in the steps 7-9 of Algorithm 1.

- Case $A_i = [R]p$ with $R = B_\sigma$ or $R = I_\sigma$: To check whether A_i "holds" at x we just check whether $p \in \mathit{Label}(y)$ with $y = \mathit{Next}(x, \langle R \rangle \top)$. The intuition is that, y is the "least R-successor" of x, and if $p \in \mathit{Label}(y)$ then p occurs or will occur in all R-successors of x.
- Case $A_i = [G_\sigma]_\diamond p$: A_i "holds" at x iff both $[G_\sigma]p$ and $\langle G_\sigma \rangle \top$ "hold" at x. By including $[\mathbf{A}_{\overline{G_\sigma}}]\langle G_\sigma \rangle \top$ in saturations, if $\langle G_\sigma \rangle \top$ "holds" at x then we can expect that $\langle G_\sigma \rangle \top \in \mathit{Label}(x)$ and $\mathit{Next}(x, \langle G_\sigma \rangle \top)$ is defined. Hence, similarly to the above case, to check whether A_i "holds" at x we just check whether $p \in \mathit{Label}(y)$ with $y = \mathit{Next}(x, \langle G_\sigma \rangle \top)$.
- Case $A_i = [R]p$ with $R \in \{\mathit{EB}_g, \mathit{CB}_g, \mathit{EI}_g, \mathit{MI}_g\}$: The idea for checking whether A_i "holds" at x is to check whether p belongs to the labels of all "minimal" R-successors of x. "Minimal" successors are the ones created for realizing requirements of the form $\langle S \rangle \top$ with $S \in \Sigma_{m+}$.

Formally, checking whether φ "holds" at x is done by Function `CheckPremise`(x, φ).

Expansions by modifying the label of a state and/or setting the mapping *Next* are done only for states that are reachable from Δ_0. Note that, when a node z is simulated by z_* as in Procedure `ExtendLabel`, the node z becomes unreachable from Δ_0. We do not delete such nodes z because they may be reused later.

When some $x \in \Delta$ has $\mathit{Label}(x)$ containing $\perp$ or a pair p and $\neg p$, Algorithm 1 returns *false*, which means that the knowledge base $\langle \Gamma, \mathcal{A} \rangle$ is TeamLog unsatisfiable. When the graph cannot be expanded any more, the algorithm terminates in the normal mode with result *true*, which means $\langle \Gamma, \mathcal{A} \rangle$ is TeamLog satisfiable. See [4] for a proof of the following theorem.

Theorem 5.1. *Algorithm 1 correctly checks satisfiability of* Horn-TeamLog *knowledge bases in* TeamLog *and has* PTime *data complexity.*

Corollary 5.2. *The problem of checking satisfiability of* Horn-TeamLog *knowledge bases in* TeamLog *has* PTime *data complexity.*

6 Conclusions

We have formulated Horn-TeamLog as a general Horn fragment of TeamLog with PTime data complexity. To estimate its data complexity, we have developed an algorithm for checking TeamLog satisfiability of a Horn-TeamLog knowledge base. Our algorithm uses global caching and avoids nondeterminism.

Horn-TeamLog differs considerably from all of the Horn fragments studied in [3, 11–14] because it deals with negative introspection of beliefs, which is closely related with converse modalities. It also differs considerably from the Horn fragments of serial regular grammar logics with converse studied in [16] and the Horn fragments of basic monomodal logics [1, 8, 10] because of the nature of goals (non-seriality) and the presence of common beliefs and mutual intentions.

In a forthcoming paper some examples and a case study illustrating the use of Horn-TeamLog will be provided.

Acknowledgements. This work was supported by the Polish National Science Centre under Grants No. 2011/01/B/ST6/02769 and 2011/01/B/ST6/02759.

References

1. Chen, C.C., Lin, I.P.: The computational complexity of the satisfiability of modal Horn clauses for modal propositional logics. Theoretical Computer Science 129, 95–121 (1994)
2. Dunin-Kęplicz, B., Nguyen, L.A., Szałas, A.: A framework for graded beliefs, goals and intentions. Fundamenta Informaticae 100(1-4), 53–76 (2010)
3. Dunin-Kęplicz, B., Nguyen, L.A., Szałas, A.: Tractable approximate knowledge fusion using the Horn fragment of serial propositional dynamic logic. Int. J. Approx. Reasoning 51(3), 346–362 (2010)
4. Dunin-Kęplicz, B., Nguyen, L.A., Szałas, A.: A long version of the current paper (2013), `http://www.mimuw.edu.pl/~nguyen/HornTeamLog-long.pdf`
5. Dunin-Kęplicz, B., Verbrugge, R.: Collective intentions. Fundam. Inform. 51(3), 271–295 (2002)
6. Dunin-Keplicz, B., Verbrugge, R.: Teamwork in Multi-Agent Systems: A Formal Approach. John Wiley & Sons, Ltd. (2010)
7. Dziubiński, M., Verbrugge, R., Dunin-Kęplicz, B.: Complexity issues in multiagent logics. Fundam. Inform. 75(1-4), 239–262 (2007)
8. Fariñas del Cerro, L., Penttonen, M.: A note on the complexity of the satisfiability of modal Horn clauses. Logic Programming 4, 1–10 (1987)
9. Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. MIT Press (2000)
10. Nguyen, L.A.: Constructing the least models for positive modal logic programs. Fundamenta Informaticae 42(1), 29–60 (2000)
11. Nguyen, L.A.: A bottom-up method for the deterministic horn fragment of the description logic $\mathcal{ALC}$. In: Fisher, M., van der Hoek, W., Konev, B., Lisitsa, A. (eds.) JELIA 2006. LNCS (LNAI), vol. 4160, pp. 346–358. Springer, Heidelberg (2006)
12. Nguyen, L.A.: On the deterministic Horn fragment of test-free PDL. In: Hodkinson, I., Venema, Y. (eds.) Advances in Modal Logic, vol. 6, pp. 373–392. King's College Publications (2006)
13. Nguyen, L.A.: Constructing finite least Kripke models for positive logic programs in serial regular grammar logics. Logic Journal of the IGPL 16(2), 175–193 (2008)
14. Nguyen, L.A.: Horn knowledge bases in regular description logics with PTime data complexity. Fundamenta Informaticae 104(4), 349–384 (2010)
15. Nguyen, L.A., Nguyen, T.-B.-L., Szałas, A.: Horn-DL: An expressive Horn description logic with PTime data complexity. In: Faber, W. (ed.) RR 2013. LNCS, vol. 7994, pp. 259–264. Springer, Heidelberg (2013)
16. Nguyen, L.A., Szałas, A.: On the Horn fragments of serial regular grammar logics with converse. In: Proceedings of KES-AMSTA. Frontiers of Artificial Intelligence and Applications, vol. 252, pp. 225–234. IOS Press (2013)

A Multiagent System Generating Complex Behaviours

Florin Leon

Department of Computer Science and Engineering, Technical University of Iaşi,
Bd. Mangeron 27, 700050 Iaşi, Romania
fleon@cs.tuiasi.ro

Abstract. In this paper we describe the design of a multiagent system based on simple interaction rules that can generate different overall behaviours, from asymptotically stable to chaotic, verified by the corresponding largest Lyapunov exponent. We show that very small perturbations can have a great impact on the evolution of the system, and we investigate some methods of controlling such perturbations in order to have a desirable final state.

Keywords: multiagent system, chaotic behaviour, perturbations, chaos control.

1 Introduction

Chaos has been extensively studied in physical systems, including methods to control it for uni-, bi- and multi-dimensional systems [1]. Also, concepts such as causality and the principle of minimal change in dynamic systems have been formalized [11].

Many human-related e.g. social or economic systems are nonlinear, even when the underlying rules of individual interactions are known to be rational and deterministic. Prediction is very difficult or impossible in these situations. However, by trying to model such phenomena, we can gain some insights regarding the fundamental nature of the system. Surprising or counterintuitive behaviours observed in reality can be sometimes explained by the results of simulations.

Therefore, the emergence of chaos out of social interactions is very important for descriptive attempts in psychology and sociology [7], and multiagent systems are a natural way of modelling such social interactions. Chaotic behaviour in multiagent systems has been investigated from many perspectives: the control of chaos in biological systems with a map depending on growth rate [13], the use of a chaotic map by the agents for optimization [2] and image segmentation [10], or the study of multiagent systems stability for economic applications [3]. However, in most of these approaches, chaos is explicitly injected into the system, by using a chaotic map, e.g. the well-known logistic map, in the decision function of the agents.

The main goal of this work is the design of simple interaction rules which in turn can generate, through a cascade effect, different types of overall behaviours, from stable to chaotic. We believe that these can be considered metaphors for the different kinds of everyday social or economic interactions, whose effects are sometimes entirely predictable and can lead to an equilibrium while some other times fluctuations can widely affect the system state, and even if the system appears to be

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 154–164, 2013.

stable for long periods of time, sudden changes can occur unpredictably because of subtle changes in the internal state of the system. We also aim at investigating how very small changes can non-locally ripple throughout the system with great consequences and if it is possible to reverse these changes in a non-trivial way, i.e. by slightly adjusting the system after the initial perturbation has occurred.

The paper is organized as follows. Section 2 presents the interaction protocol of the multiagent system and its mathematical formalization. Section 3 discusses the deterministic and chaotic behaviours that emerge from the system execution. Section 4 presents an experimental study regarding the effects of small perturbations in the initial state of the system and the possibility of cancelling them through minimal external interventions. The final section contains the conclusions of this work.

2 The Design of the Multiagent System

The main goal in designing the structure and the interactions of the multiagent system was to find a simple setting that can generate complex behaviours. A delicate balance is needed in this respect. On the one hand, if the system is too simple, its behaviour will be completely deterministic and predictable. On the other hand, if the system is overly complex, it would be very difficult to assess the contribution of the individual internal elements to its observed evolution. The multiagent system presented as follows is the result of many attempts of finding this balance.

The proposed system is comprised of n agents; let A be the set of agents. Each agent has m needs and m resources, whose values lie in their predefined domains D_n, $D_r \subset \mathbb{R}^+$. This is a simplified conceptualization of any social or economic model, where the interactions of the individuals are based on some resource exchanges, of any nature, and where individuals have different valuations of the types of resources involved.

In the present model, it is assumed that the needs of an agent are fixed (although an adaptive mechanism could be easily implemented, taking into account, for example, previous results [8,9]), that its resources are variable and they change following the continuous interactions with other agents.

Also, the agents are situated in their execution environment: each agent a has a position π_a and can interact only with the other agents in its neighbourhood Λ_a. For simplicity, the environment is considered to be a bi-dimensional square lattice, but this imposes no limitation on the general interaction model – it can be applied without changes to any environment topology.

2.1 Social Model

Throughout the execution of the system, each agent, in turn, chooses another agent in its local neighbourhood to interact with. Each agent a stores the number of previous interactions with any other agent b, $i_a(b)$, and the cumulative outcome of these interactions, $o_a(b)$, which is based on the profits resulted from resource exchanges, as described in the following section.

When an agent a must choose another agent to interact with, it chooses the agent in its neighbourhood with the highest estimated outcome: $b^* = \arg\max_{b \in \Lambda_a} o_a(b)$.

The parallelism of agent execution is simulated by running them sequentially and in random order. Since one of the goals of the system is to be deterministic, we define the execution order from the start. Thus, at any time, it can be known which agent will execute and which other agent it will interact with. When perturbations are introduced into the system, the same execution order is preserved. It has been shown that the order of asynchronous processes plays a role in self-organisation within many multi-agent systems [4]. However, in our case this random order is not necessary to generate complex behaviours. Even if the agents are always executed in lexicographic order (first A1, then A2, then A3 etc.), sudden changes in utilities still occur, although the overall aspect of the system evolution is much smoother.

2.2 Bilateral Interaction Protocol

In any interaction, each agent tries to satisfy the needs of the other agent as well as possible, i.e. in decreasing order of its needs. The interaction actually represents the transfer of a resource quantum γ from an agent to the other. Ideally, each agent would satisfy the greatest need of the other.

For example, let us consider 3 needs (N) and 3 resources (R) for 2 agents a and b: $N_a = \{1, 2, 3\}$, $N_b = \{2, 3, 1\}$, $R_a = \{5, 7, 4\}$, $R_b = \{6, 6, 5\}$, and $\gamma = 1$. Since need 2 is the maximum of agent b, agent a will give b 1 unit of resource 2. Conversely, b will give a 1 unit of resource 3.

In order to add a layer of nonlinearity, we consider that an exchange is possible only if the amount of a resource exceeds a threshold level θ and if the giving agent a has a greater amount of the corresponding selected resource r_{sel} than the receiving agent b: $R_a(r_{sel}) > R_b(r_{sel})$ and $R_a(r_{sel}) > \theta$.

In the previous situation, if we impose a threshold level $\theta = 5$, agent a will still give b 1 unit of resource 2, but b will only satisfy need 1 for agent a.

Based on these exchanges, the resources are updated and the profit p_a is computed for an agent a as follows:

$$p_a = \gamma \cdot N_a(r_{sel}) \cdot \frac{R_b(r_{sel})}{R_a(r_{sel})}. \tag{1}$$

A bilateral interaction can bring an agent a profit greater or equal to 0. However, its utility should be able to both increase and decrease. For this purpose, we can compute a statistical average of the profit, p_{avg}, and increase the utility of an agent if the actual profit is above p_{avg}, and decrease the utility if the profit is below p_{avg}.

Thus, the equation for updating the utility level of an agent a is:

$$u_a \leftarrow \frac{u_a \cdot i_a^{adj} + \eta \cdot (p_a - p_{avg})}{i_a^{adj} + 1}, \tag{2}$$

where the adjusted number of interactions is: $i_a^{adj} = \min\left(\sum_{b\in A} i_a(b), i_{mem}\right)$, i_{mem} is the maximum number of overall interactions that the agent can "remember" (i.e. take into account) and η is the rate of utility change. At the beginning, the utility of the agent can fluctuate more, as the agent explores the interactions with its neighbours. Afterwards, the change in utility decreases, but never becomes too small.

For example, if $i_{mem} = 20$, $u_a = 0.1$, $p_a = 8.5$, $\eta = 1$, $p_{avg} = 7.5$ and the sum of all previous interactions is 2, the utility will change to: $u_a' = (0.1 \cdot 2 + (8.5 - 7.5) \cdot 1) / 3 = 0.4$. If the sum of all previous interactions is 100, the same utility will change only to: $u_a' = (0.1 \cdot 20 + (8.5 - 7.5) \cdot 1) / 21 = 0.14$.

Similarly, the social outcome of an agent *a* concerning agent *b* is updated as follows:

$$o_a(b) \leftarrow \frac{o_a(b) \cdot i_a(b) + \eta \cdot \left(p_a - p_{avg}\right)}{i_a(b) + 1}. \tag{3}$$

In this case, the social model concerns only 1 agent and thus the use of the actual number of interactions can help the convergence of the estimation an agent has about another.

Regarding the computation of the average profit, we used a statistical approach where we took into account 100 continuous interactions between two randomly initialized agents, which exchange resources for 100000 time steps. The average profit depends on the number of resources, their domain and the interaction threshold.

3 Types of Behaviours

A key challenge in applied dynamical systems is the development of techniques to understand the internal dynamics of a nonlinear system, given only its observed outputs [5]. As the observed output of our multiagent system, we consider only the agent utilities. We can view this output as a discrete time series, one for each agent. In the following, we analyse the evolution of these time series over time. Since there is no stopping condition for the agent interactions, we restrict our study to a predefined, finite time horizon, e.g. 1000, 2000 or 10000 time steps.

Depending on the number of agents and the initial state of the system, several types of behaviours can be observed:

- *Asymptotically stable:* When only 2 agents exist in the system, we noticed that they can perform an indefinite number of interactions. They can stabilize to a continuous exchange of resources, possibly the same resource in both cases (γ units of the same resource are passed back and forth between the 2 agents). With 2 agents, the system quickly converges to a stable state (figure 1). Depending on the initial state, a stable state can also be reached by some agents in a system with multiple agents. The typical behaviour in the latter case is a high frequency vibration around the value of convergence. However, it is also possible that multiple agents all converge to stable states and the system remains in equilibrium afterwards;

- *Quasiperiodic:* With more interacting agents in the system, usually their utilities no longer converge to a stable value. Instead, the values belong to a certain range, with few, predictable oscillations around the mean value. Figure 2 shows the evolution of the utility of 4 agents over 10000 time steps. In order to smooth out short-term fluctuations and highlight longer-term trends, a simple moving average method is used, with a window size of 10 time steps;
- *Chaotic:* With a high number of agents (e.g. over 10), the complexity of their interactions usually exceeds the deterministically predictable level. The utilities of some agents widely fluctuate, even after the initial period where a part of the system approaches a stable zone. Figure 3 displays the behaviour of 100 agents over 10000 time steps. A simple moving average is applied here again, with a window size of 100 time steps. One agent (with a utility value around -3) has unpredictable great changes, although they appear to be governed by a higher-level order of some kind. Another agent has a sudden drop in utility around time step 9000, although it has been fairly stable before.

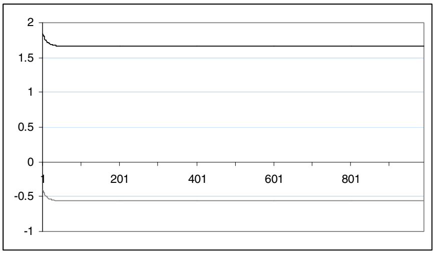


Fig. 1. Asymptotically stable behaviour - 2 agents, 1000 time steps

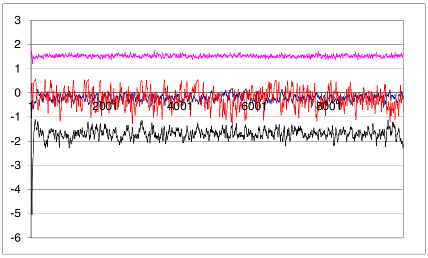


Fig. 2. Quasiperiodic behaviour - 4 agents, 10000 time steps

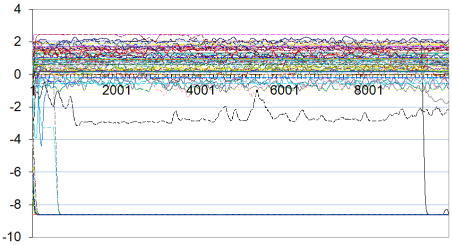


Fig. 3. Chaotic behaviour - 100 agents 10000 time steps

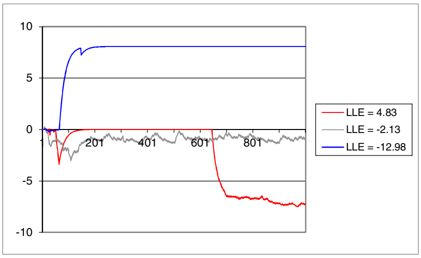


Fig. 4. Chaotic and non-chaotic variations

We consider that the third type of behaviour is chaotic, since it satisfies the typical features of chaos [6]:

- *Nonlinearity:* Given the nonlinearity caused by the minimum threshold for resource exchange, the system can be viewed as a hybrid one, with transitions between different ways of operation. Also, the maximum of the social outcome can change, thus an agent can interact with different neighbours, which results in different profits and further changes to the social outcomes;
- *Determinism:* Apart from the random initialization of the agent parameters (which nevertheless can be controlled by using the same seed for the random number generator), all the interaction rules are deterministic;
- *Sensitivity to initial conditions:* As we will show in the experimental study presented in section 4, very small changes in the initial state of the system can lead to a radically different final state;
- *Sustained irregularity and mostly impossible long-term predictions:* These are also characteristic of observed behaviours.

Regarding the effect of small perturbations, which in general can be used to control a chaotic system, out of many runs under different configurations, we noticed that a perturbation can affect the overall system behaviour in more ways:

- *No effect* within a predefined time horizon: depending on the agent positions, the system state and the place where the perturbation occurs, some changes can have no effect at all;
- *A temporary effect* which is later cancelled out within the time horizon;
- *A permanent effect* which reflects in the final state of the system, within the predefined time horizon.

We can make a parallel between these kinds of effects and the choices we make in everyday life. Out of the many alternatives that we have, a different choice can have no effect or sometimes we may not know that something is different until a later time when the different choice becomes relevant. Other times, a different choice impacts our environment immediately. Even if something changes, the overall environment can eventually reduce the perturbation, or the system can toggle to a whole different state indefinitely. All these kinds of behaviours have been observed in the designed multiagent system.

We can measure the degree of chaos introduced by a perturbation by considering the difference between the changed system and the original system as a time series, and computing the largest Lyapunov exponent (LLE) of the variation in an agent utility. Basically, LLE describes the predictability of a dynamical system. A positive value usually indicates that the system is chaotic [12]. There are methods, e.g. [14], which compute the LLE from the output of the system regarded as a time series.

Figure 4 displays three situations. The variation with a positive LLE (4.83) can be considered to be chaotic. We can notice the sudden change in utility after the half of the simulation, although the perturbation has occurred in the first time step. A small negative LLE (-2.13) indicates an almost deterministic behaviour, which can correspond to a quasiperiodic variation. Finally, a high negative LLE (-12.98) indicates a deterministic behaviour, when the time series converges to a value and remains stable there. Positive LLEs are not only found in some utility variations, but also in some of the original utility evolutions, depending on the system initial state.

4 Experimental Studies

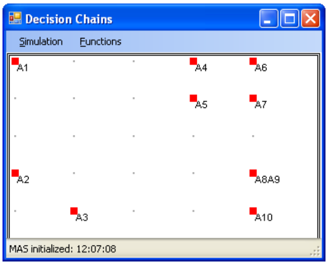
A mathematical analysis of a nonlinear hybrid system is usually very difficult. Therefore, in the following, we will present an empirical experimental study, where we will emphasise different cases or settings which reveal certain types of behaviour.

Since one of the characteristics of a chaotic system is that small changes in its initial state can greatly affect the final state through a cascade effect, we observe the influence of perturbations on the system behaviour. We also reflect on the question of when it is possible to correct some distortions with the smallest amount of external energy, such that, after a perturbation, the system should reach again a desired state within a corresponding time horizon, through small changes.

In all the case studies presented in this section, the following parameters were used: the number of agents n = 10, the number of needs and resources m = 10, their domains $D_n = D_r =$ [0, 10), the resource transfer quantum $\gamma = 1$, the resource exchange threshold $\theta = 5$, the interaction memory $i_{mem} = 20$, the utility change rate $\eta = 2$, the side length of the agent square neighbourhood Λ is 4 and the computed average profit $p_{avg} = 7.5$.

4.1 Original Behaviour

The configuration under study is composed of 3 subgraphs (figure 5): one agent, A1, is isolated and cannot interact with any other agent. Two agents, A2 and A3, form their own bilateral subsystem and seven agents can interact with one another in their corresponding neighbourhoods. A change in any of those agents can affect any other one in this subgraph, because, for example, A4 can influence A7, A7 can influence A9 and A9 can influence A10. The evolution of the agent utilities for 2000 time steps is displayed in figure 6.



MAS initialized: 12:07:08

Fig. 5. The positions of the agents

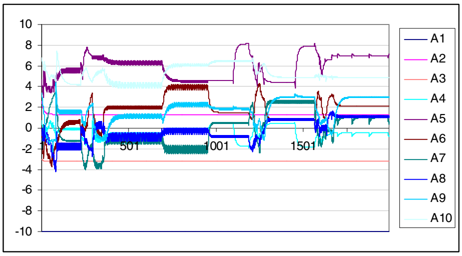


Fig. 6. The original evolution of agent utilities with no perturbation

4.2 The Effect of Small Perturbations

In this section, we observe the evolution of the utilities when a very small perturbation is added to or subtracted from a resource of an agent. Figure 7 shows the difference between the changed behaviour due to the presence of the perturbation and the original behaviour seen in figure 6, with a slightly larger perturbation of 0.1, and when the agents execute in lexicographic order. Figure 8 shows this difference for a perturbation of only 10^{-5} and when agents execute in a predefined random order. With 10 agents and 10 resources, this corresponds to a 10^{-7} change in the initial system state. We can see that, in general, the smaller a perturbation is, the longer it takes for its effect to accumulate and impact the observed behaviour of the system.

The actual number of an agent or resource is not very important, as we study the overall performance of the system. However, one can notice that the effects are non-local, and a change in one agent can affect other agents in its subgraph. Also, even if the perturbation has occurred in the first time step, big differences can appear later on, after 686 and 1873 time steps, respectively.

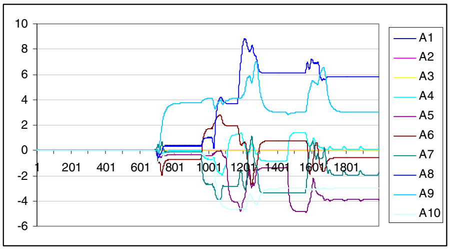


Fig. 7. The consequences of a perturbation of 0.1 in resource 5 of agent A3

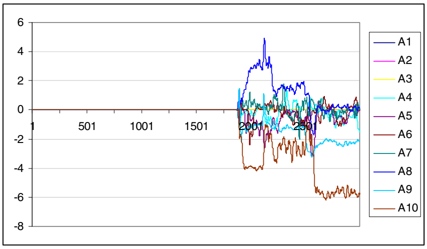


Fig. 8. The consequences of a perturbation of -10^{-5} in resource 6 of agent A6

4.3 Perturbation Correction

Given a perturbation in the initial time step with a non-null effect on the system, we are interested in finding a way to cancel or greatly reduce its impact, as observed on the time horizon and even beyond. Since this correction must be done from outside the system, and consists in changing the amount of a resource of an agent, it is also important that we find the minimum (or a small) amount of change needed to return the system to its final state as it would have been with no perturbation.

We would also like to find flexible solutions. A trivial solution would be to reverse the perturbation in the first time step. However, it is more interesting to see if there can be changes in later steps of the simulation which can tackle the effect of the initial perturbation.

Because the effects of change are non-local and can propagate throughout the subgraph of an agent's neighbours, we have applied, so far, the following search methods:

- *Exhaustive search with one correction point:* trying all the resources of all the agents in each step of the simulation, adding or subtracting a small amount (e.g. 0.1, 0.5), and observing the maximum utility variation in the final state of the system. If this maximum variation is below a desired threshold (e.g. 1), then a solution has been found;
- *Random corrections with one or multiple points:* considering 1 or more (e.g. 3) sets of quadruples (agent, resource, simulation step, correction amount) which inject changes into the simulation, and seeing if the final state of the system matches the final state in the original setting. The random search is by far faster than exhaustive search, but it cannot tell if any solution exists at all.

Besides considering only the state of the system at the time horizon (e.g. 2000 time steps), it is also important to verify if the system behaviour continues to be desirable. Figure 9 shows the effect of a 1 point correction for the situation presented in figure 7, which remains stable for a test period of 100 more time steps after the initial 2000 ones. However, if the system is chaotic, it is impossible to guarantee that this difference will remain small forever.

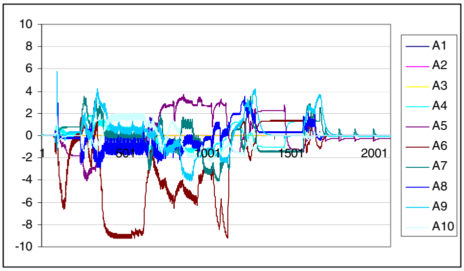


Fig. 9. A perturbation correction with an amount of -0.5 in resource 2 of agent A4 in step 70, leading to a maximum difference of 0.48 utility units from the original final state within the test period of 100 time steps

5 Conclusions

In this paper we presented the design of a multiagent system that can display different types of behaviours, from asymptotically stable to chaotic. In this case, chaos arises only from the agent interactions, and it is not artificially introduced through a chaotic map.

As future directions of research, we aim at further analysing the results of the interactions in order to see whether some probabilistic predictions can be made, taking into account the system state at a certain moment. It is important to determine when small perturbations have visible effects and when they can be controlled. Also, one must investigate whether classical chaos control techniques used for physical systems such as the OGY method, can be applied as well for this multiagent system.

Another fundamental question is whether the chaos in the system is only transient and eventually stabilises into a steady state or its behaviour remains chaotic forever.

Out of many experiments, it seems that sometimes the system converges to a stable state. In other cases, chaos doesn't seem to be only transient, e.g. with 50 agents executing in lexicographic order (which corresponds to fewer fluctuations), there are still sudden changes occurring in the utility variation even after 50000 time steps. One needs to distinguish between these cases as well.

So far, the proposed system is mainly of theoretical importance, but one can investigate if it can be used for social or economic simulations, for the modelling of biological processes or other typical applications.

Acknowledgment. This work was supported by the "Partnership in priority areas – PN-II" program, financed by ANCS, CNDI - UEFISCDI, project PN-II-PT-PCCA-2011-3.2-0732, No. 23/2012.

References

1. Boccaletti, S., Grebogi, C., Lai, Y.C., Mancini, H., Maza, D.: The Control of Chaos: Theory and Applications. Physics Reports 329, 103–197 (2000)
2. Charrier, R., Bourjot, C., Charpillet, F.: A Nonlinear Multi-agent System designed for Swarm Intelligence: The Logistic MAS. In: First IEEE International Conference on Self-Adaptive and Self-Organizing Systems, pp. 32–44 (2007)
3. Chli, M., De Wilde, P., Goossenaerts, J., Abramov, V., Szirbik, N., Correia, L., Mariano, P., Ribeiro, R.: Stability of Multi-Agent Systems. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 551–556 (2003)
4. Cornforth, D., Green, D.G., Newth, D.: Ordered asynchronous processes in multi-agent systems. Physica D (Nonlinear Phenomena) 204, 70–82 (2005)
5. Dawes, J.H.P., Freeland, M.C.: The '0–1 test for chaos' and strange nonchaotic attractors, `http://people.bath.ac.uk/jhpd20/publications/sna.pdf` (preprint)
6. Ditto, W., Munakata, T.: Principles and Applications of Chaotic Systems. Communications of the ACM 38(11), 96–102 (1995)
7. Katerelos, I.D., Koulouris, A.G.: Is Prediction Possible? Chaotic Behavior of Multiple Equilibria Regulation Model in Cellular Automata Topology, Complexiy, Wiley Periodicals 10(1) (2004)
8. Leon, F.: Multiagent Role Allocation Based on Cognitive Dissonance Theory. International Review on Computers and Software 6(5), 715–724 (2011)
9. Leon, F.: Self-Organization of Roles Based on Multilateral Negotiation for Task Allocation. In: Klügl, F., Ossowski, S. (eds.) MATES 2011. LNCS, vol. 6973, pp. 173–180. Springer, Heidelberg (2011)
10. Melkemi, K.E., Batouche, M., Foufou, S.: Chaotic MultiAgent System Approach for MRF-based Image Segmentation. In: Proceedings of the 4th International Symposium on Image and Signal Processing and Anlysis, ISPA 2005, Zagreb, Croatia, pp. 268–273 (2005)
11. Pagnucco, M., Peppas, P.: Causality and Minimal Change Demystified. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, USA, pp. 125–130 (2001)

12. Rothman, D.: Nonlinear Dynamics I: Chaos (2006), http://ocw.mit.edu/courses/earth-atmospheric-and-planetary-sciences/12-006j-nonlinear-dynamics-i-chaos-fall-2006/lecture-notes/lecnotes15.pdf
13. Solé, R.V., Gamarra, J.G.P., Ginovart, M., López, D.: Controlling Chaos in Ecology: From Deterministic to Individual-based Models. Bulletin of Mathematical Biology 61, 1187–1207 (1999)
14. Wolf, A., Swift, J.B., Swinney, H.L., Vastano, J.A.: Determining Lyapunov exponents from a time series. Physica D (Nonlinear Phenomena) 16, 285–317 (1985)

Using Norm Emergence in Addressing the Tragedy of the Commons

Sorin Dascalu, Tudor Scurtu, Andreea Urzica,
Mihai Trascau, and Adina Magda Florea

University "Politehnica" of Bucharest,
313 Splaiul Independentei, 060042, Bucharest
{sorin.dascalu,tudor.scurtu}@cti.pub.ro,
{andreeea.urzica,mihai.trascau,adina.florea}@cs.pub.ro
http://acs.pub.ro

Abstract. In this paper we propose an approach for the experimental analysis of the classical problem of the Tragedy of the Commons. This approach revolves around norms within multi-agent systems, without regimenting them. The proposed model allows norms to emerge between selfish agents that compete for the same resource. The paper shows how the emerging norms lead to a balanced distribution of resources among the individuals, diminishing the effects of the tragedy.

Keywords: multi-agent systems, tragedy of commons, norm emergence.

1 Introduction

The Tragedy of the Commons (ToC) is a well-known problem that dates from man's early days. The issue can be summarized in short: several individuals share a common *resource* which they can harvest for their own benefit. None of them is responsible for the regeneration of the resource, it is a common responsibility.

This paper shows how, by simulating the ToC within a system, and by observing the possible outcomes, that the tragedy can be diminished. In order to achieve this and so that a norm can emerge, some design considerations are yielded.

Norms emerge when the agents placed in the same environment, being able to interact with one-another, directly or indirectly, come to a common agreement of certain rules which they should follow in order to achieve global welfare.

Our system is a generalization of the herders problem and we have constructed an architecture starting from it, where agents are indirectly punished by the environment if they tend to overexploit it. This system is highly extendable and can be instantiated in several modern instances of the problem.

Some problems along with related current efforts are presented in Section 2. In Section 3 we introduce our proposal by describing a scenario and how it can be applied. Section 3 explains the architecture, while Section 4 and 5 cover design and implementation considerations. We show and interpret the results in Section 6. Lastly, we conclude and present some ways of improving our system and extending it to every-day issues.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 165–174, 2013.

2 Related Work

This section provides an insight on the *Tragedy of the Commons (ToC)* issue and on the mechanism used to obtain a resolution to this problem - Norm Emergence.

The ToC is a well-known problem that arises from an individual's goal to achieve their own, local interest, the best way they can, rather than finding an optimal solution for their system. This problem was first described by ecologist Garett Hardin [1] in 1968. Some modern instances can be: exploitation of forests, traffic on freeways, bandwidth sharing or any type of vandalism.

An interesting simulation was considered in Imitation and Inequity in Avoiding the ToC [2]. The authors consider a model where each *agent*'s actions rely on a three weight neural network that maps agent wealth and local resources to a decision of what fraction of the available resource to consume, producing several behavioural strategies. The aim is to create a quasi-balance, between maximizing the take of some and minimizing the take of others. A practical study of the ToC, found in [3], considers networking traffic and how the overall congestion of a network can lead to worse performance as opposed to the case in which every network member would not use all of its allocated bandwidth.

There are different ways that social norms emerge in multi-agent systems and topologies, as well as different natures of the problems and different attitudes of the agents, affect the outcome of these ways. The authors of [4] study the emergence of norms in a virtual environment, where having a central authority to impose a norm would be computationally more expensive than an alternative in which the agents would create themselves the norm and adhere to it.

In their papers, Norm Emergence in Spatially Constrained Interactions [5] and Norm Emergence under Constrained Interactions in Diverse Societies [6], the authors show how the constraints of a grid topology and various agent learning strategies influence norm emergence. The modelled problem is a simple one: deciding on which side of the road to drive. When two players meet and both choose the same side of the road, they receive equal payoffs. In case they choose to drive on opposite sides, a collision occurs and they both receive equal penalties.

These studies confirmed that social norms can spontaneously emerge without the presence of leaders, in different topologies and in heterogeneous societies.

3 The Simulation Environment

We consider a society that revolves around a resource desired by all agents. This resource is distributed evenly over a number of domains, hereafter called plots. Each agent has a certain initial resource quota it wants to harvest from the whole system. At each step agents will want to fulfil their quota harvesting from different plots. The resources are fully renewable and regenerate completely at the start of a new iteration step.

The individuals follow certain harvesting rules:

1. Each agent can harvest as much as it wants from a *plot*, provided the plot can offer how much the agent desires;

2. Agents choose the plots from which to harvest in a blind manner, without knowing how much is available in the current run;
3. An agent decides by itself the quota ratio it intends to take from a plot;
4. The agent will acquire utility equal to how much it has farmed from a plot. However if it wants more than the plot can offer we will consider this as a form of starvation and the agent will lose utility equal to the quantity it was not able to harvest.

The ToC implies a destruction of the common. In this case the destruction consists of the overexploitation of some plots while others are barely touched.

4 System Dynamics

An important part is the assessment of system performance. To this end, we have devised several metrics upon which we will base our conclusions:

1. *Agent Quota* – the agents dynamically update their quota based on how much utility they have managed to harvest; the system norm consists in the average overall quota;
2. *Environment Damage* – measures the efficiency with which the environment is harvested; it is relative to the quota required by agents and the available quantity on the plots;
3. *Loss per Agent* – the average difference between the agents quota and the managed harvested utility.

In the proposed set-up, the agents are able to learn step by step. We expect to see them agree on a strategy setting that allows all of them to achieve their quota or at least maximize their utility.

For the purpose of this evaluation we separated the agents in two categories:

- agents with a set of well-built strategies to choose from, that imply harvesting in a somewhat equal manner from all the plots; the goal is for each such agent to end up with its best strategy; the process of each individual choosing and continuing with such a strategy will determine an even distribution of resource gathering from the setting; we will call them *static*;
- agents that do not have fixed strategies and choose on their own how to gather resources; they start by randomly generating strategies and continue to update them as the simulation proceeds; we will call them *adaptive.*

By providing the simulation with mostly static agents and a small ratio of adaptive individuals we hope to see a trend of the adaptive agents to imitate the strategies of the static ones.

5 System Architecture

The system was designed in a way that permits a clear separation between the different components described in the Proposed Scenario chapter. In short, an

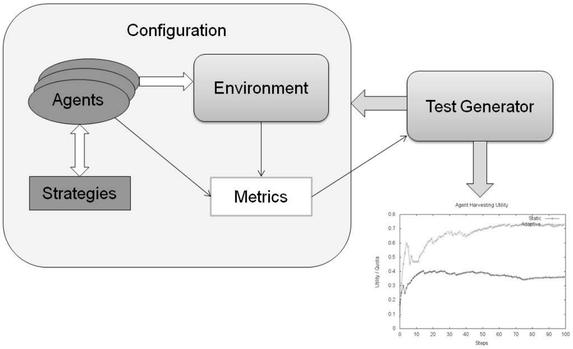


Fig. 1. Architectural overview

overview of the architecture can be seen in Figure 1. The main points of interest are the environment, the agents and the strategies they use. The Environment is composed of several plots, each having the same iteration capacity in term of resource units. These resources will regenerate at the beginning of a turn. In order to efficiently generate relevant tests, we have created a simulation suite that automatically generates scenarios, runs the program and plots graphs of metrics related to agent behaviour and environment impact.

The number of each type of agents is customizable. At each iteration step, the agents will try to harvest the land. As to maintain fair odds for all individuals, we considered a random order in each turn for who is first and who is last in the exploitation process. This simple principle is a form of avoiding the appearance of greedy, privileged individuals that can acquire their quota as they please. This simplification should not impose a limitation on the system as from a statistical point of view each agent has equal chances of harvesting the target resource.

Each actor will have a memory of its chosen actions over the last N turns. This is mostly a mechanism of observing if they can decide on a good strategy and be consistent with it. The central element in all agents is the Strategies they use to achieve high utility. Thus, they are judged based on the utility they have provided to the agent over the course of the last iterations. The Strategies component can be separated in two subcomponents:

- Strategy Generation - used to create new strategies for adaptive actors or to update generated ones based on how efficient they have proved to be; there is no strategy generation for static agents - instead they will begin the simulation with a number of predefined strategies, obtained by distributing the quota in a relative equal manner over random plots of the setting;
- Strategy Selection - how each agent reasons in choosing its play in turn; the choice is a flexible one - an actor will not pick the highest performing strategy but will have the highest probability to do so.

We have several governing parameters for each type allowing us to play a more aggressive or conservative strategy generation method and observe its influence on the overall performance of the system. At each step, after a run, an agent will decrease or increase the played strategy's utility based on how well it has performed and helped it in achieving its quota.

6 Design Considerations

6.1 Modelling the Environment as a Graph

We have chosen to model the environment as an undirected graph, in which the nodes correspond to the plots and the edges to the paths between the plots.

The practical usage is immediate: most current problems that would be modelled in a 2d space have an implicit form of cost:

- a form of energy cost: gas/energy; consider the ToC in a common fish area form, divided in fishing areas; each sailor would use their engine-powered boat to get from one area to another; to do this, the engine uses gas that is proportional to the distance travelled, strength of the current etc.
- time cost for reaching from one parcel to another: if ToC takes the classical form of a number of herders trying to graze their cattle on the same land;
- risk of data loss if we were to consider the usage of network bandwidth as a common resource by multiple clients in a Local Area Network.

Consider the graph in Figure 2. Each vertex represents a common resource; each edge is a penalty that an agent would incur should it not be able to achieve its goal on a vertex. Hardin's classical problem of herders sharing common land can be easily modelled on the example: the vertices are common pieces of fertile land; the edges are the distances between those pieces of land.

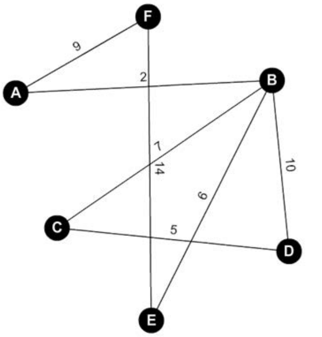


Fig. 2. One of the graphs used for testing the simulation

A harvesting scenario implies a herder takeing his herd on several pieces of land for grazing; if he takes his herd on an already grazed piece of land, then the cows would starve directly proportional to the distance they have travelled (the penalty - the weight of an edge on the graph). I.e., if farmer John takes his herd from land A to land F, but land F is not profitable enough to feed his cows, the herd will starve because they had travelled 9 miles to get from A to F, and perhaps make the herd thinner with on average of 9 pounds per cow.

6.2 Quota Variation

The system norm is interpreted as the average quota of the agents. We introduce the following quota emergence/strategy negotiation system: each agent starts out with an initial quota; each agent will attempt to satisfy this quota and create a strategy that allows them to achieve it; there will be a number of steps in which the agents will basically be allowed to negotiate and obtain a high percentage of their goal; after the mentioned number of steps, the agent compares the actual amount of resources it managed to exploit to what it should have harvested; following this comparison, several scenarios are possible:

1. the agent managed to harvest an amount of resources larger than a high percentage of its intended quota (for example, 90%); in this case it will raise its quota as this implies that the environment might be further exploited;
2. the agent harvested a very small amount of resources (for example less than 10%of the intended harvest); in this scenario, the agent tried to obtain too much utility; it makes sense that it should lower its expectations, the quota;
3. the agent hasn't been exceptional in either way; it will maintain its quota and continue on perfecting its strategies;
4. after this, an agent will reiterate the last 2 steps a fixed number of times; ideally we would repeat this process until all of the agents would find their own quota and no changes would appear; however the system is yet to be perfect, so we will only be testing it for a fixed amount of steps and observe how the quotas evolve and whether a norm emerges in a quota form or not.

Real-life parallelism would include situations in which economic agents would not invest in a sector that deals with the harvesting of a dwindling resource, new investors would come into a newly developed domain, as well as the situation in which a company is forced to make cutbacks in times of economic hardship.

6.3 Strategy Types

To differentiate real life scenarios we have created two types of strategy classes:

1. Order Significant Strategy – this strategy class assumes that the agent is a "nomad" – it does not have a home plot and will gather resources as it navigates the graph;this scenario could be appropriate if in the case of the common fishing waters problem we assume that the fishing boats are large

industrial vessels that have the autonomy to fish in the entire area before returning home; thus, the utility loss for a failed harvest attempt consists only of the edge value between the former plot and the now harvested plot;

2. Point of Origin Strategy – the second strategy class assigns each agent a residency plot; it would be the plot to which an agent has to return after each harvest attempt; this could be fitting if we consider small fishing boats that have to return to refuel after each fishing session; in the implementation the difference consists of the fact that the penalty is now the value of the shortest path from the point of value to the unsuccessfully harvested plot.

6.4 Implementation Considerations

The entire simulation for a test can be described with the following algorithm:

```
1.Randomly generate environment as a graph;
2.Generate agents (static/dynamic);
3.Initialize metrics used for simulation outcome;
4.For finite number of steps or until quota convergence
    4.1. Regenerate resources in environment;
    4.2. For each agent:
        4.2.1. Choose strategy to use for harvest;
        4.2.2. Harvest environment with that strategy;
        4.2.3. Improve strategies after relevant number of uses;
        4.2.4. Adjust quota after relevant number of steps;
    4.3. Compute metrics for simulation step;
```

7 Results

The following graphs are classified after discriminating configuration parameters. They represent the evolution of described metrics over the course of simulations.

7.1 Single-Type Agent Behaviour Validation

When only one type of agents is present in the environment, they behave just as expected before the tests were run. The Static Agent type maintains a fixed quota and reaches a utility plateau. The Adaptive Agents vary their quota depending on resource abundance, continuously improving the average harvested utility.

As expected, the static agents converge on the optimal relative strategy, while the dynamic agents succeed in continuously generating strategies that outperform or are at least as good as the existing strategies.

7.2 System Evolution with Two Types of Agents

In resource-rich environments - those in which the amount of resources surpass the total initial agent quotas, as presented in Figure 3, it can be observed that

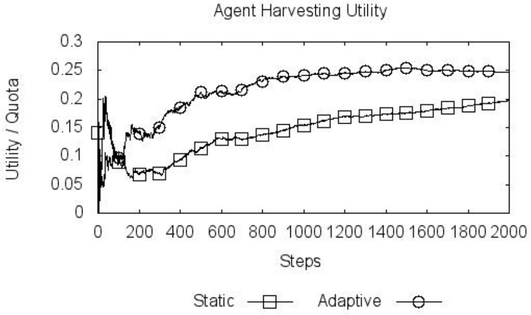


Fig. 3. Comparative agent utility evolution in a scenario with both types of agents in a relatively resource-rich environment

the adaptive agents quota increases in each iteration in order to attempt to harvest the leftover resources.

The environment damage steadily decreases, as seen in Figure 4, and the average loss increases at a lower rate than the quota norm. Despite the loss increase, the norm also increases because the marginal payoff is positive.

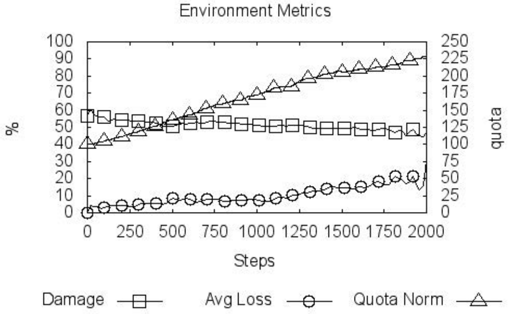


Fig. 4. Example of environment metrics evolution in the scenario from Figure 3

In resource-poor environments - those in which the total initial agent quotas surpass the amount of resources, as seen in Figure 5, the quota norm decreases in order to try to minimize the average loss. The environment damage is kept low because of uniform harvesting, and the average loss decreases at a lower rate than the quota.

The Adaptive Agents fare better than their static counterparts in dividing the resources among themselves and in harvesting them without incurring a large damage to the environment in cases in which the environment is a relatively low-resource one.

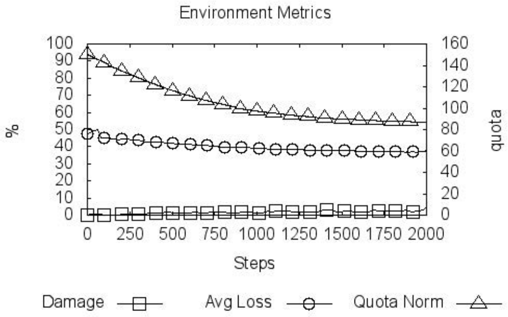


Fig. 5. Example of environment metrics evolution in scenario with both types of agents in a relatively resource-poor environment

7.3 Final Remarks

The Static Agents have the same behaviour for both types of strategies. The Adaptive Agents obtain better results with the Point of Origin Strategy, because they can find paths between their plot of origin and potential harvest plots more efficiently than Static Agents. An interesting observation is that in these iterated simulation conditions the loss is divided among a majority of the agents. Another observation is that the process generally converges within 2000 steps, but false plateaus may occur before the strategy updates can begin to influence agent behaviour.

In the current context, the emerging social norm can be interpreted as the average value of the quota set by each agent in the environment. The best scenario seems to be the one with only adaptive agents: they manage to harvest most of the environment's resources, establish a high quota and at the time make sure that free resources remain scattered equally, resulting a in a low environmental damage.

The main conclusions that can be drawn from the graphs are that Adaptive Agents succeed in environments with scarce resources. They do this by successfully dividing the plots between themselves and the other agents, regardless of type.

8 Conclusions

This paper presents an experimental approach for the analysis of a classical problem, the Tragedy of the Commons. The paper investigates how norms could emerge within the well-known scenario of resource harvesting and finds the basic design specifications that would allow agents to maintain the integrity of the environment even while being self-interest driven.

The system has an easy-to-understand architecture and the output is straightforward and significant. The presented system has been tested over multiple scenarios and results have been interpreted in accordance with the given input. The simulations reveal that agents succeed in dividing the resource plots among themselves. Although the Tragedy of the Commons is not always and fully averted, agents seem to have a better response of how they should behave in order to achieve a higher long term utility. Furthermore, adaptive agents manage to outperform their static counterparts and they even succeed in environments in which resources are scarce.

A plausible scenario to which our system may be extended is trading in the modern age. Plots could be considered cities in which agents sell their merchandise. If there is an oversaturation of a product in a city the price will significantly drop, the agents might even not be able to sell anymore. A scheduled improvement to the proposed model concerns how the penalty of an agent is computed in case of a failed harvest. This penalty should take into consideration that although the agent cannot take as much as it intended, it can still use some of the utility from that plot.

Acknowledgements. The work presented in this paper has been funded by Project 264207, ERRIC- Empowering Romanian Research on Intelligent Information Technologies / FP7- REGPOT- 2010-1.

References

1. Hardin, G.: The Tragedy of the Commons. Science 162(3859), 1243–1248 (1968)
2. Van Belle, T., Ackley, D.H.: Imitation and Inequity in Avoiding the Tragedy of the Commons. In: Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems, vol. 9, p. 274 (2004)
3. Cole, R., Dodis, Y., Roughgarden, T.: Bottleneck links, variable demand, and the tragedy of the commons. In: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm, pp. 668–677 (2006)
4. Savarimuthu, B.T.R., Purvis, M., Purvis, M., Cranefield, S.: Social norm emergence in virtual agent societies. In: Baldoni, M., Son, T.C., van Riemsdijk, M.B., Winikoff, M. (eds.) DALT 2008. LNCS (LNAI), vol. 5397, pp. 18–28. Springer, Heidelberg (2009)
5. Mukherjee, P., Sen, S., Airiau, S.: Norm emergence in spatially constrained interactions. In: Working Notes of the Adaptive and Learning Agents Workshop at AAMAS, vol. 7, pp. 2–6 (2007)
6. Mukherjee, P., Sen, S., Airiau, S.: Norm emergence under constrained interactions in diverse societies. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, vol. 2, pp. 779–786 (2008)

Assessing Agents Interaction Quality via Multi-agent Runtime Verification

Najwa Abu Bakar and Ali Selamat

Faculty of Computing
Universiti Teknologi Malaysia
81300, Skudai, Johor, Malaysia
najwa.abakar@gmail.com, aselamat@utm.my

Abstract. Interaction between agents is the main process executed within multi-agent systems. Agents autonomously and proactively interact with each other to perform tasks. However, there are possibilities that during execution interaction could fail due to external factors such as modified user inputs, inactive hosts and compromised network. Based on this observation, it is important to assess the quality of interaction process during multi-agent systems execution in order for the systems to be continuously trusted to perform critical tasks and improved in terms of its security and reliability from time to time. Thus, in this research, a solution called Multi-agent Runtime Verification framework is proposed to assess agents interaction quality. In this paper, the availability and trustability metrics are addressed based on supporting contextual information. The verification process assesses the quality of the messages transmitted during interaction by assigning scores for each of the defined metrics. The scores are used to determine whether the communication is going to fail or succeed. The framework aims to reduce agents interaction failure during runtime. Finally, an experiment is set up to evaluate the effectiveness of the proposed solution.

Keywords: Multi-agent systems, runtime verification, agents interaction, message passing, software quality assurance.

1 Introduction

Analysis towards existing multi-agent systems (MAS) verification state-of-the-art shows that MAS interaction correctness has been assessed either during design or during development [2][3][11]. Firstly, MAS interaction verification during design assesses the correctness of MAS design model using techniques such as automated theorem proving and model checking [6][9][14][17]. Secondly, MAS verification during development analyses codes for bugs using programming language debugging tools and tests system functionalities using MAS testing tools [12]. Each approach implements certain techniques and performs verification towards selected MAS interaction properties. However, since some of agents interaction failures can only be discovered when agents interaction errors occur during execution, MAS verification during runtime is needed.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 175–184, 2013.

Agents in actions are autonomous and perform tasks without being controlled and supervised [20][21]. From our observation, as the multi-agent systems executed, agents interaction quality changes due to external factors such as modified user inputs, host status and network infrastructure. Agents that are supposed to receive messages can be unavailable due to inactive hosts or compromised network. Likely, trust of agents can evolve during execution as it depends on agent authorization, authentication and location. New runtime interaction properties need to be specified and verified to address these factors. Therefore, in this paper, MAS interaction verification during runtime is proposed to monitor and determine effectiveness of MAS interaction from time to time. The changes of the agents availability and trustability status due to the external factors that could possibly affect the interaction quality between agents are considered.

In this study, based on the above mentioned problem, we address the question of how the effectiveness of MAS interaction can be improved. In this paper, we propose Multi-agent Runtime Verification (MARV) framework to assess agents interaction quality during runtime. The framework aims to improve the effectiveness of agents interaction by reducing interaction failures. In this framework, the defined agents interaction quality requirements and the developed metrics are used to perform MAS interaction verification. The metrics rules are used to assess the quality scores of MAS captured messages and the scores are used to filter out MAS interaction that will not succeed. Finally, the effectiveness of MARV framework will be evaluated.

The rest of the paper is organized as follows. Section 2 provides the related works. Next, Section 3 describes the methodology of the proposed solution. Then, in Section 4, we present the case study and finally, in Section 5, we provide the conclusion and the future work.

2 Related Works

2.1 Existing Works for MAS Interaction Runtime Verification

During runtime, to perform tasks, agents communicate to each other by sending requests and responses, transmitting data and sharing critical information to negotiate, collaborate or compete [20][21]. Currently, there are several tools that perform monitoring and analysis for verifying agents interaction correctness during runtime. These runtime monitoring or debugging tools have been performed either for single (intra) agent level or multi (inter) agent level. Single agent verification can be performed using debugging tools for certain agent development languages such as JAVA while multi agent verification includes modeling and specification of interactions during design phase. When the number of agents are low (less than 100), agent platform such as JADE (Java Agent Development) can be used to visualize the interactions between agents [5][10]. The RMA (Remote Monitoring Agent) graphical user interface (GUI) of the JADE framework comes with the sniffing feature to see interaction between agents within the same platform and host as well as between agents within multiple platforms and hosts. From the sequence diagram, the correctness of the message sequence is checked and analyzed to ensure that the messages are sent and received correctly by the sender and the receiver agents [5].On the other hand, when the number of agents is large and the MAS is getting more complex, higher level runtime verification, debugging and analysis using techniques such as data mining are performed towards the agents ACL (Agent Communication Language) messages exchanged. A tool called ACLAnalyser sniffs all messages exchanged between agents in a MAS, store them in a database, and recover them later for analysis [8].

There are also existing runtime verification works that implement agent within MAS application itself to perform verification or debugging processes. Each approach has its strengths and limitations. The debugging agent proposed by Poutakidis [16] monitors the interactions and exchanged messages between other agents and checks them against interaction protocols. Next, Lockman and Selamat [13] and Abdul Bujang and Selamat [1] implement verification agents alongside other agents within MAS applications to check certain message content properties i.e. format and size. The checked properties are implemented (hard coded) within the verification agent as part of the MAS applications. Thus, the verification is only suitable for those particular applications. Osman [15] performs verification towards MAS properties during runtime using model checking technique. Model checking is performed to check the design without considering the runtime agents status and messages format. Finally, Gomez-Sanz et al. [11] presents tool for analysis, design and development of MAS that is suitable for finding errors during both design and development levels of the multi-agent systems.

The existing works presented have significant contributions in debugging and verifying multi-agent systems. However, these works do not consider agent interaction properties that evolve during runtime due to external factors. Therefore, in this paper, we address agents availability and trustability properties. The properties depend on the status and location of the agents as well as the authorization levels of the agent roles that may change during execution.

2.2 Interaction Quality Properties

During agents interaction, there are requirements that need to be fulfilled as well as constraints that need to be addressed to ensure the success of the communication. For example, correctness criteria based on message sequence and interaction protocols are the required properties that have been verified by the existing works mentioned in the previous subsection. In this paper, we focus on the agent availability and trustability properties. Agent availability to receive sent messages is the crucial condition for a successful interaction. Trustability of the agents to send or receive messages is also important to ensure confidentiality and integrity of the communication. The requirements and constraints of agents interaction in multi-agent applications have been discussed in the literatures [18][19][20][21].

3 Methodology

3.1 Multi-agent Runtime Verification (MARV)

MARV framework adapts Total Quality Management (TQM), a quality assurance method [7]. The strong point or novelty of this approach that distinguishes this research from the existing solutions is the integration of data quality principles in the effort to help improve quality of agents interaction. Instead of using visualization or display method to monitor and analyze messages as implemented by JADE Sniffer and ACAnalyser, our proposed framework monitors and verifies agents interaction by working on the "data" e.g. captured messages, hosts and network monitoring information and MAS contextual knowledgebase. Agents interaction issues are tackled and analyzed at root problem that is at the data level instead of diagram or other graphical format. These data can later be enriched and enhanced for further analysis using techniques such as artificial intelligence and data mining.

3.2 MARV Architecture

There are four main components implemented to verify MAS interaction that are the *Definition*, *Measurement*, *Analysis* and *Improvement*. First, during *Definition*, based on the identified MAS interaction data and requirements, agents interaction quality criteria are defined. Metrics and rules are constructed for the criteria. Second, during agents interaction quality *Measurement* process in MARV, a few steps are performed that are getting the input and assessing the metric scores. Figure 1 shows the measurement component within MARV architecture and its connection with other multi-agent runtime verification components. The metrics assessment scores for the criteria are used in the *Analysis* stage to decide whether the interactions are correct. Finally, *Improvement* is suggested to correct the interactions among agents within the MAS. Besides the four stages of methodology, there are other components that support the verification process that are 1) *Capturing the Interaction Messages*, 2) *Gathering the Agents Profile* and 3) *Updating the MAS Knowledgebase*. The interaction messages are captured and each attribute in the messages are saved in the database. These attributes are assessed to evaluate whether the interactions are correct. The supporting contextual information, the agents profile and MAS knowledgebase that are gathered and updated by the other components are also saved in the database. These supporting information determine whether the interactions are going to be successful or not.

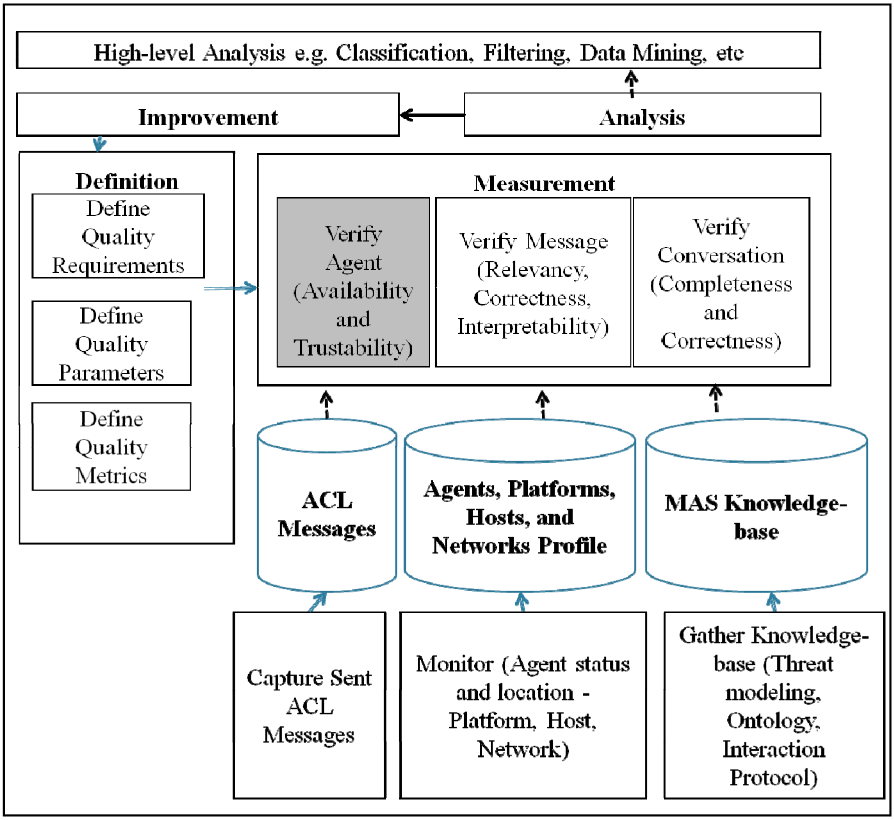

High-level Analysis e.g. Classification, Filtering, Data Mining, etc
Improvement
Analysis
Definition
Define Quality Requirements
Define Quality Parameters
Define Quality Metrics
Measurement
Verify Agent (Availability and Trustability)
Verify Message (Relevancy, Correctness, Interpretability)
Verify Conversation (Completeness and Correctness)
ACL Messages
Agents, Platforms, Hosts, and Networks Profile
MAS Knowledge-base
Capture Sent ACL Messages
Monitor (Agent status and location - Platform, Host, Network)
Gather Knowledge-base (Threat modeling, Ontology, Interaction Protocol)

Fig. 1. MARV Architecture

Figure 1 shows the MARV architecture. The process flow of the framework is enlisted below:

1) *Definition*; Agents interaction quality requirements, parameters, and metrics are defined. As shown in the architecture, there are several procedures performed that are the definition of 1) *agents interaction quality requirements* (criteria) 2) *agents interaction parameters* (data) and 3) *agents interaction quality metrics* (rules). The requirement definition is performed by identifying the factors that determine the success of agents interaction according to the agents context and MAS supporting information. Based on the adapted quality assurance approach, the selected criteria should be measurable and can be used to quantify a message as having the characteristics to be successful in agent conversation.

2) *Capturing ACL Messages*; ACL messages (the data) are captured from executed MAS to be transferred into database.

3) *Monitoring*; Agents and MAS infrastructure such as platforms, hosts, and networks are monitored and the information related to agents activeness and host aliveness are stored in a database as supporting information.

4) *Gathering Knowledgebase*; Threat model, ontologies and interaction protocols of the executed MAS in which information related to agent roles, authentication and authorization levels are extracted from the MAS and stored in knowledgebase database.

5) *Measurement*; The raw agents interaction data to be assessed are the captured messages. The message attributes include the *SenderName*, *SenderHost*, *ReceiverName*, *ReceiverHost*, *Performative* and *Content* [5][10]. The measurement component takes the captured raw messages from the database as the input. At the same time it queries agents contextual (profile) information that are the *location (host) status*, *host aliveness status*, *agent activeness status*, *authorization levels of agent roles*, and *authentication of message content*. The MAS knowledgebase that are the *threat model*, *ontologies* and *interaction protocols* as the supporting information for the assessment are also accessed. In this paper, the metrics used are the agent availability and trustability (explained further in Section 3.3). The metric rules defined in the definition component are used in this measurement component. The measurement process flow is further explained in Section 3.4.

6) *Analysis*; The analysis phase evaluates the effectiveness of the verification in reducing agents interaction failures.

7) *Improvement*; Based on the analysis, improvement can be made by redefining the interaction requirements, parameters, rules, and formulation of the runtime verification metrics.

8) *High-level Analysis*; The verification output in the form of quality scores are stored in a database and can be used to analyze the quality of agents interaction. Further analysis can also be performed during high-level analysis stage to classify, filter, or performing data mining to make advance conclusion and decision making.

3.3 Interaction Quality Metrics

The constructed metrics for the defined criteria includes rules that determine the scores for each message in which the scores are used to classify the messages into valid or invalid messages during the analysis stage. In this paper, our discussion focuses on two metrics that are the availability and trustability metrics. *Availability* metric refers to the availability of the *receiver agent* (A_{rec}) to perform tasks upon receiving the sent messages. A_{rec} score depends on the *agent activeness* (A_{ac}) status and *host aliveness* (H_{al}) status (as shown in Equation 1).

$$A_{rec} = A_{ac} \wedge H_{al} \tag{1}$$

Trustability metrics refer to the ability of the sender (T_{sen}) and receiver (T_{rec}) agents to be trusted to send and receive the message, respectively. The T_{sen} and T_{rec} scores depend on the *role* of the receiver and sender agents, R_{rol} and S_{rol}, respectively and the *location* of the receiver and sender agents, R_{loc} and S_{loc}, respectively. If the roles of the receiver and sender agents are authorized to send and receive the message content, S_{rol} and R_{rol} scores are 1, else the scores are 0. Similarly, if their locations (hosts) are within the trusted network, S_{loc} and R_{loc} scores are 1, else the scores are 0 (as shown in Equation 1 and 2).

$$T_{sen} = S_{rol} \wedge S_{loc} \tag{2}$$

$$T_{rec} = R_{rol} \wedge R_{loc} \tag{3}$$

Finally, the quality of the message, M can be assessed using Equation 4 below. Authentication of the message content, T_{mes} is also considered besides A_{rec}, T_{sen} and T_{rec}.

$$M = A_{rec} \wedge T_{sen} \wedge T_{rec} \wedge T_{mes} \tag{4}$$

As shown by the equation 1,2,3 and 4, the quality of the message depends on the availability and trustability metrics (rules). First, to ensure that a transmitted message has its receiver available to response to it, the availability of the receiver is checked (refer Equation 1). Next, in order to ensure trustability of agents interaction, both the sender agent and receiver agent must be the agents that are authorized (based on their roles) to send and receive the message. The sender agent must have the authority to send the particular message in order for the message content to be trusted. If the roles of the sender and receiver agent do not match authorization level of the receiver agent and sender agent, then the message receive low (0) score while high (1) score is given when the roles match the receiver and sender authorization levels.

Trustability and availability are the examples of the whole set of the defined metrics that also includes *correctness of timing*, *correctness of syntax*, *relevancy of message*, *interpretability of semantic*, *correctness of message sequence* and *completeness of conversation*. For this paper, the above two metrics and the rules are presented as an example. The rules are constructed by considering the MAS requirements previously discussed [18][19].

3.4 Measurement Process Flow

As shown in MARV architecture (Figure 1), this measurement stage is a part of the whole verification process. After all the inputs are gathered, the agents interaction quality metric scores are assessed for each captured transmitted message based on the agents context and systems supporting information that are agents profile and MAS

knowledgebase. The rules are the simple if-else statement (refer Section 3.3) implemented as the verification algorithm. For each metric, the rules check via its if-else statement the message attributes value against the monitored information in agent profiles and MAS knowledgebase resources, and then assign score accordingly using the rules. The assessment steps are:
1) Get message from the captured messages database
2) Get rule from metric rules database
3) Get the related message attribute values from the captured messages database
4) Get the related agent profile and MAS knowledgebase required to execute the rules
5) Execute the if-else statement of the rule
6) Assign rule score to the message
7) Repeat step 2 to 5 until all the rules are executed
8) Repeat step 1 to 6 until all the captured messages are assessed

Figure 2 shows the messages scores assessment process flow diagram. A verification algorithm is designed based on the illustrated process.

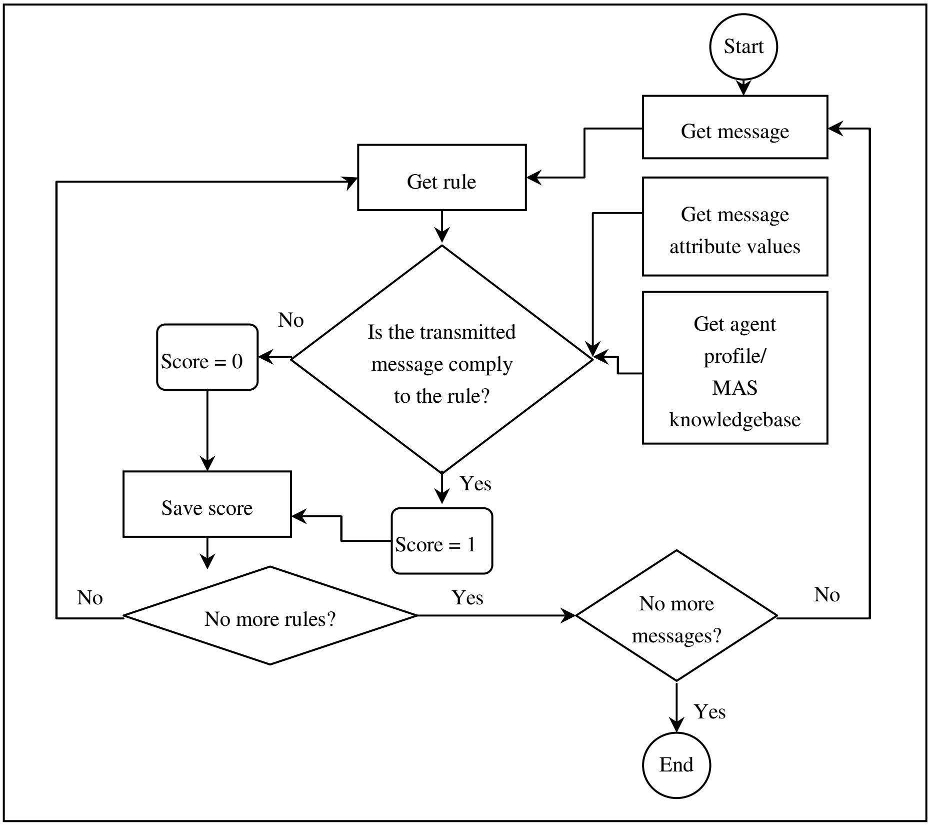

Start
Get message
Get rule
Get message attribute values
Get agent profile/ MAS knowledgebase
No
Score = 0
Is the transmitted message comply to the rule?
Yes
Save score
Score = 1
No
No more rules?
Yes
No more messages?
No
Yes
End

Fig. 2. The measurement process flow diagram

4 Case Study: Book Trading

The book trading system is a multi-agent system developed in JADE platform [5]. There are two types of agents in the system that are the seller and buyer. The buyer agents take the book title as input and try to buy the book from seller agents at a reasonable price. The seller agents keep the list of books to sell and try to sell them at the highest possible price. Both agents perform negotiation in achieving the best result. Table 1 below shows the sample of book-trading agents message attributes and the scores that could be assigned to each message using the availabilty and trustability metrics presented in this paper. The scores can be 1 or 0 depending on the runtime contextual and supporting information captured during the execution.

Table 1. Sample of book-trading agents message attributes and scores

Sender Name	Sender Host	Receiver Name	Receiver Host	Performative	Content	A_{rec}	T_{sen}	T_{rec}	T_{mes}	M
Buyer1	Host1	Seller1	Host2	CFP	bookTitle	1/0	1/0	1/0	1/0	1/0
Buyer1	Host1	Seller2	Host3	CFP	bookTitle	1/0	1/0	1/0	1/0	1/0
Seller1	Host2	Buyer1	Host1	PROPOSE	price	1/0	1/0	1/0	1/0	1/0
Seller2	Host3			REFUSE	not-available	1/0	1/0	1/0	1/0	1/0

5 Conclusion and Future Work

The measurement process of MARV framework developed to verify MAS interaction during runtime is presented. The measurement process includes the assessment of agents interaction quality scores by considering agents contextual information and MAS knowledgebase collected during agents interaction activities. In this paper, we have presented agents interaction quality metrics that are the availability of the receiver agent, the trustability of the receiver agent, and the trustability of the sender agent that consider the verification requirements at the agent level. An experiment is set up using JADE environment by developing the metric rules as algorithm, incorporating the assessment process, capturing the transmitted messages and collecting the supporting information. In the future, more metrics will be identified at the conversation level that includes the interpretability of the message content and the completeness of the conversation.

Acknowledgments. The authors thank Ministry of Higher Education Malaysia (MOHE) under Fundamental Research Grant Scheme (FRGS) Vot 4F031 and Universiti Teknologi Malaysia under Research University Funding Scheme (Q.J130000.2510.03H02) for supporting the related research.

References

1. Abdul Bujang, S., Selamat, A.: Verification of Mobile SMS Application with Model Checking Agent. In: ICIMT 2009 Proceedings of the 2009 International Conference on Information and Multimedia Technology, pp. 361–365. IEEE Computer Society, Washington, DC (2009)

2. Abu Bakar, N., Selamat, A.: Analyzing model checking approach for multi agent system verification. In: 2011 5th Malaysian Conference on Software Engineering (MySEC), pp. 95–100 (2011)
3. Abu Bakar, N., Selamat, A.: Agent-based Model Checking Verification Framework. In: 2012 IEEE Conference on Open Systems (ICOS), pp. 1–4 (2012)
4. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York (1999)
5. Bellifemine, F.L., Caire, G., Greenwood, D.: Developing multi-agent systems with JADE. John Wiley & Sons, Ltd., West Sussex (2007)
6. Berard, B., Bidoit, M., Finkel, A., Laroussinie, F., Petit, A., Petrucci, L., et al.: Systems and Software Verification: Model-Checking Techniques and Tools (1999)
7. Besterfield, D., Besterfield-Michna, C., Besterfield, G., Besterfield-Sacre, M., Urdhwareshe, H., Urdhwareshe, R.: Total Quality Management. Pearson Education (2011)
8. Botía, J.A., Gómez-Sanz, J.J., Pavón, J.: Intelligent data analysis for the verification of multi-agent systems interactions. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 1207–1214. Springer, Heidelberg (2006)
9. Clarke, E.M., Grumberg, O., Peled, D.A.: Model Checking. The MIT Press, Cambridge (1999)
10. FIPA: The Foundation for Intelligent Physical Agents (2012), `http://www.fipa.org/` (retrieved May 18, 2012)
11. Gómez-Sanz, J.J., Botía, J., Serrano, E., Pavón, J.: Testing and debugging of MAS interactions with INGENIAS. In: Luck, M., Gomez-Sanz, J.J. (eds.) AOSE 2008. LNCS, vol. 5386, pp. 199–212. Springer, Heidelberg (2009)
12. JAT Jade Agent Testing framework, `http://jattool.sourceforge.net` (retrieved March 18, 2013)
13. Selamat, A., Lockman, M.T.: Multi-agent Verification of RFID System. In: Nguyen, N.T., Katarzyniak, R.P., Janiak, A. (eds.) New Challenges in Computational Collective Intelligence. SCI, vol. 244, pp. 255–268. Springer, Heidelberg (2009)
14. Lomuscio, A., Qu, H., Raimondi, F.: MCMAS: A model checker for the verification of multi-agent systems. In: Bouajjani, A., Maler, O. (eds.) CAV 2009. LNCS, vol. 5643, pp. 682–688. Springer, Heidelberg (2009)
15. Osman, N.: Runtime Verification of Deontic and Trust Models in Multiagent Interactions. Phd Thesis (2008)
16. Poutakidis, D.: Debugging Multi-Agent Systems With Design Document. Phd Thesis (2008)
17. Raimondi, F.: Model checking multi-agent system. Phd Thesis (2006)
18. Silva, C., Pinto, R., Castro, J., Tedesco, P.: Requirements for Multi-Agent Systems. In: Workshop em Engenharia de Requisitos (WER), Piracicaba-SP, Brasil, pp. 198–212 (2003)
19. Singh, M.P., Chopra, A.K.: Correctness Properties for Multiagent Systems. In: Baldoni, M., Bentahar, J., van Riemsdijk, M.B., Lloyd, J. (eds.) DALT 2009. LNCS, vol. 5948, pp. 192–207. Springer, Heidelberg (2010)
20. Sycara, K.P.: Multiagent Systems. AI Magazine 19(2), 79–92 (1998); Association for the Advancement of Artificial Intelligence
21. Wooldridge, M.: An Introduction to MultiAgent Systems, 2nd edn. John Wiley & Sons, United Kingdom (2009)

Flexible and Emergent Workflows Using Adaptive Agents

Arcady Rantrua, Marie-Pierre Gleizes, and Chihab Hanachi

Institut de Recherche en Informatique de Toulouse
University of Toulouse
Toulouse, France
{rantrua,gleizes,hanachi}@irit.fr

Abstract. Most of existing workflow systems are rigid since they require to completely specify processes before their enactment and they also lack flexibility during their execution. This work proposes to view a workflow as a set of cooperative and adaptive agents interleaving its design and its execution leading to an emergent workflow. We use the theory of Adaptive Multi-Agent Systems (AMAS) to provide agents with adaptive capabilities and the whole multi-agent system with emergent "feature". We provide a meta-model linking workflow and AMAS concepts, and the specification of agent behavior and the resulting collaborations. A simulator has been implemented with the Make Agent Yourself platform.

Keywords: workflow, multi-agent system, flexibility, adaptation, emergence.

1 Introduction

In some dynamic and unpredictable process-oriented applications, processes can not be defined by explicitly specifying a priori all possible alternatives of execution. For example, in crisis management or innovative design, some guidelines should be followed and the goal reached, but the process definition and its execution with its possible alternatives should be subject to consultation and collective choices at run-time. In other words, the process emerges since its definition and execution interleave. A partial and high-level process is set up, detailed, maintained and adapted in a collective way according to the crisis evolution, appreciated through the feedback information from the field, the availability of resources and the tasks pre-conditions. Also, tasks pre-conditions could be relaxed to be more reactive and efficient, and resources may evolve in an unpredictable way (damages or reinforcements).

Although, these kinds of applications require adaptiveness and emergence, they also have to guarantee that we have followed a logical process (a set of coordinated tasks) since they could engage the responsibilities of actors in crisis situations and also because the process followed is a result to be reused, explained, improved (in both crisis and design situations).

Given this context, the problem addressed in this work is: *how to build a flexible workflow system supporting emergent processes i.e. that ensures the interleaving of definition, execution and adaptation of processes?*

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 185–194, 2013.

Research in the area of workflow flexibility is becoming increasingly important as a result of organizational demands to develop applications that have to face dynamic and unpredictable situations. A comprehensive survey of the area can be found in [2] and [3]. Four types of flexibility have been identified: flexibility by design, flexibility by deviation, flexibility by under-specification and flexibility by change. While most of them are still based on procedural languages that are rigid to represent domains with a lot of variability, we have to mention the case-handling approach, based on declarative rules, that supports the personalization and adaptation of pre-defined processes [4]. Also, tools (like WADE [5]) and applications [6] based on the agent paradigm have been built to support workflow flexibility, implementing them as cooperative agents. However, in the best of our knowledge, the idea to focus on emergent workflows and to support the interleaving of definition, execution and adaptation have only been investigated by a few studies (see [1]).

Our approach, described in this paper, is to agentify the components involved in a workflow according to the agent paradigm and have them cooperate and adapt to ensure the emergence of a relevant process. Our model is grounded on the AMAS theory (see Section 3).

The contribution of the paper is threefold:

- The definition of a meta-model that links the AMAS and the workflow concepts.
- A specification of an agent-based workflow system based on the AMAS theory.
- The WOrkflow Adaptive Simulator (WOAS) compliant with this specification and implemented with the MAY[1] Platform.

The paper is organized as follows. Section 2 provides a short state of the art on flexibility in workflows. We then introduce the AMAS theory and its adequacy to deal with adaptability and emergence. Section 4 presents the specification of our system: the meta-model, the agent structure and behaviour and agent collaboration protocol. Section 5 provides an overview of the WOAS simulator and experimentation. Finally, we discuss our approach and conclude the paper.

2 State of the Art

Workflows are designed through three different and interrelated perspectives giving rise to three conceptual models: process (behavior), organization and information. Also, design and execution are most of the time two separated phases.

In this state of the art, we will compare existing works about workflow flexibility according to three criteria: the design approach followed (procedural, event driven, emergent), the flexibility techniques followed and their impact: life-cycle phase (design and/or execution), workflow model (process, organization, information).

Design Approach. Traditional workflows are most of the time represented by procedural languages (Petri Net based formalism such YAWL, BPMN, BPEL, ...) relevant for routines, but too rigid to face dynamic and unpredictable applications.

[1] Make Agents Yourself (`http://irit.fr/MAY-en`).

Some other systems, called declarative, are designed through an event-based approach that specifies the workflow behavior through reactions to perform when some events occur. The Wide system of Casati [7] follows this approach by using the active rule formalism. The two previous approaches assume a centralized workflow engine controlling the execution of the workflow, and do not provide means to ease interactions between actors to negotiate and solve problems due to interdependant tasks. A third approach, called agent-based workflow, takes a more radical choice, fully rethinking the system in terms of cooperative and problem-solving agents and distributing the control over the agents. Even in this last approach, we can distinguish two classes:

1. *Design-execution separation:* the process models (control, information, organization) are pre-defined and distributed over different agents, each one being able to play a role, deciding by itself or after negotiation how to play it. The WADE system [5], one of the most well-known platforms, follows this approach.
2. *The emergent approach* involves agents having capabilities to design and execute the process. Even if some tasks, information and roles could pre-exist, agents have the capabilities to change or adapt all of them, and to elaborate the whole process in real time, interleaving design, adaptation and execution in an opportunistic way.

The ADEPT system [6], stands between these two cases. Indeed, in ADEPT, Agents' organization (abstracted as agencies), tasks, and the handled information by each task are predefined. On the contrary, the whole process is neither predefined nor designed, but the result of agents' interactions and negotiations. To the best of our knowledge, very few agent related works have fully investigated the emergent approach providing support for interleaving design and execution as required in dynamic and unpredictable situations. However, we can mention the Dynamic Engineering Design Processes (DEDP) approach used in the PSI project by [1]. Our work differs from [1] by its more autonomous and emergent approach made possible thanks to its theoretical background: the AMAS Theory in our case (see Section 3).

Flexibility Techniques. In addition to the previous approaches or paradigms, several works on workflow flexibility have been focused on adding flexibility to procedural languages. Four types of flexibility have been identified in [2]: flexibility by design, flexibility by deviation, flexibility by underspecification (late binding and late modeling) and flexibility by change.

The two most interesting techniques for our purpose are underspecification and flexibility by change, that are the most dynamic since they integrate the possibility to model or change part of a process during its execution. One of the most advanced concrete solutions, following the flexibility by change technique, is the "case handling" technique formalized by Van Der Aalst et al. [4] and implemented in the FLOWer system that allows the actors to change the process by focusing on what can be done instead of what should be done. This technique is dynamic but not emerging.

Although all these techniques improve workflow flexibility, they remain applied to the procedural approach that constraints their expressive power.

Flexibility Impacts. Most research works discussed in this section has been focused on bringing flexibility to the workflow process model. The agent-based approach is probably the most dynamic and flexible in terms of control structure and organization. The Case Handling technique is one of the very few approaches that addresses flexibility regarding the three models. In fact, most of these works separate design and execution time, and whatever the approach followed (procedural, event based or agent), main decisions about the three models are decided at design time and the user has little influence on the three models during execution.

In conclusion, even if numerous works have addressed the workflow flexibility issue, none of them deals with emergent workflow which authorizes to interleave design and execution. This is regretful as this feature is known to be useful to face dynamic and unpredictable situations. However, the agent-based approach and some ideas from the case-handling technique are useful to support emergent workflows as we will show in our proposition.

3 Background on Adaptive Multi-Agent System

MAS are composed of interacting autonomous heterogeneous software entities or components called agents [9]. This approach makes the modeling, design and simulation of complex systems easier thanks to their capabilities: local representation of the knowledge of the agents (constraints and objectives); use of interaction / negotiation between agents; physical and processing distribution; decentralized decision making; autonomous behaviors, openness (new agents can appear and disappear at any time). All these capabilities contribute to accurately represent complex systems and simulate them. This is needed when the global behaviors (such as dynamic workflows) emerge from interdependent local interactions and reactions (such as resources breakdowns or unpredictable events).

The AMAS (Adaptive Multi-Agent System) theory is based on the general property asserting that the system realizes the required function when all its components (the agents) are in cooperative situation [10]. Consequently, a self-organized process locally guided by a cooperative attitude leads to a dynamic equilibrium such that each agent reaches (or tends to) its individual goals and the system reaches an adequate collective functionality. This assumption is the foundation of the AMAS approach that considers cooperation as the criterion for self-organization [11]. Basically, every agent pursues a local objective and interacts with others while respecting cooperative rules that guide its interactions in order to avoid and possibly remove situations that are judged as non cooperative from its local point of view. These situations are called Non-Cooperative Situations (NCS for short).

The benefits could be expressed in terms of efficiency, decentralization (minimizing shared knowledge, maximizing privacy), openness (adding/removing agents or knowledge) and robustness (minimizing problem-dependant parameters). This means that the system has the capability, through the agents composing it, to adapt to global or local changes at runtime, and to provide close solutions between

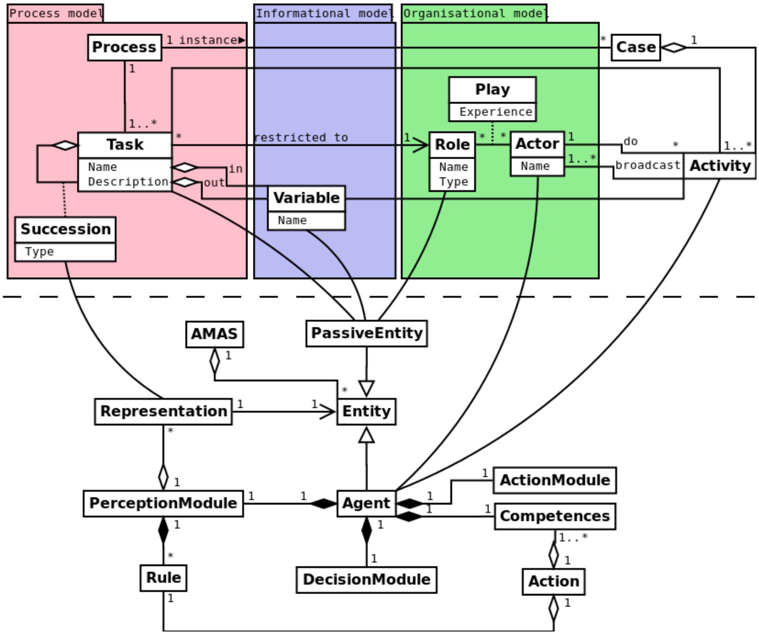


Fig. 1. Metamodels of AMAS and Workflows linked together

changes, since changes are locally processed by agents; contrary to global centralized algorithms requiring starting the resolution process again from scratch at each change. This cooperative approach has already been applied to several domains such as adaptive profiling, multi-criteria optimization, dynamic control of complex systems or manufacturing control.

4 Modeling Emergent Workflow by Means of AMAS Theory

4.1 The Principles of Our Approach

The required function of our system is to produce a workflow. We model this workflow as a set of entities where activities and actors are active entities (or agents) able to decide by themselves actions to perform and tasks, roles and variables are passive entities (or resources) handled by the agents.

More precisely, we consider two categories of agents:

- *Software agents* solve non-cooperative situations to try to stabilize their neighborhood following the AMAS theory. Their individual goal is to guarantee the local consistency of the workflow. They also automatically make emerge parts of the workflow that ease the agents' perception of the design evolution.
- *Human agents* are in charge of building the whole workflow. In our experimentation, they will be represented by artificial agents. Their goal is to reach a dynamic equilibrium corresponding to the situation where the designed workflow is consistent and accepted by every actor.

The components of the workflow cooperate to decide the workflow structure. The workflow itself may be viewed as a multi-agent system producing its own design through self-organization.

4.2 The Metamodel for Linking Adaptive Multi-Agent Systems and Workflow Concepts

Figure 1 links AMAS concept (top part) and workflow concepts (bottom part) with cross-domain association. Actors are active entity, or agent (entity that can autonomously make decisions). Actors have been agentified because they have knowledge about what can and should be done. Activities are also agentified because they actively work to solve the conflicts that can occur between the agents. The other entities, namely variables, tasks and roles, are passive. They are part of a library of resources which is used by the agents to represents the environment (the workflow).

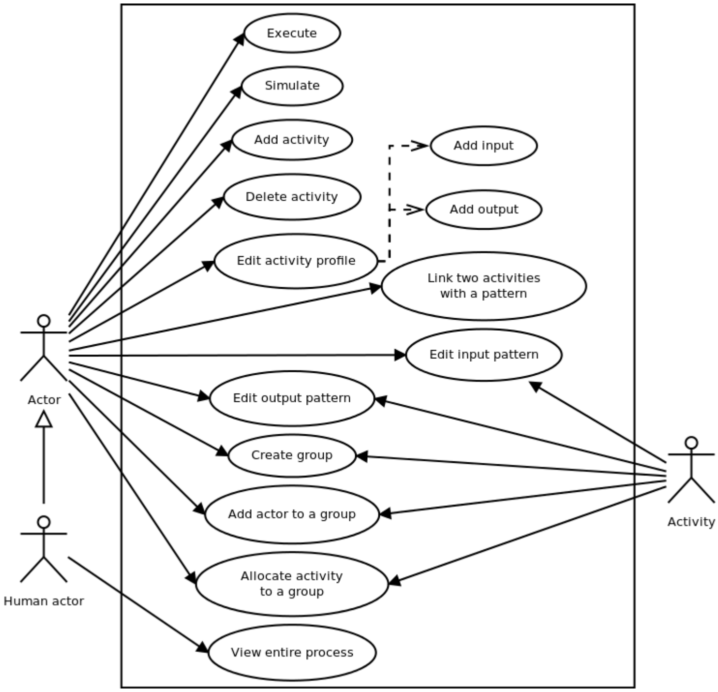


Fig. 2. Use Case Diagram of the System

4.3 Agents' Actions and Behavior

Agents' Actions. As shown in figure 2 we have three types of agents that can perform several actions impacting the three perspectives through the manipulation of data (inputs, outputs, profile), groups, activities and coordination pattern.

The "simulate" actions implements the interleaving of runtime and design time concepts by allowing a software agent to simulate the execution of an activity and see the consequences. That is how our system can check the validity of the design and try to fix it before the real execution.

Activity Structure. An activity has a name (usually a verb representing the action that it does, like “firefighting”) and a set of input/output data named profile. Those activities can be executed by an actor if he/she has the corresponding capacity. In order to represent the different ways for executing an activity our model accepts the coexistence of multiple capacities with the same name but with different profiles. In case of a conflict, actors trigger a negotiation protocol to agree on a profile.

An activity also features two coordination patterns, linking it to its previous and following activities. It may be: a sequence, an alternative (if), a parallel control structure or a synchronization. An activity is associated to a group that contains actors with common interests (it usually means that each group member can execute the activity).

4.4 Agent Collaboration and Conflict Resolution

Cooperation is triggered by NCS. The Petri net of figure 3 describes the protocol that handles an NCS from its detection to its repair. When an agent detects a NCS, it triggers a repair action to reach a cooperative situation.

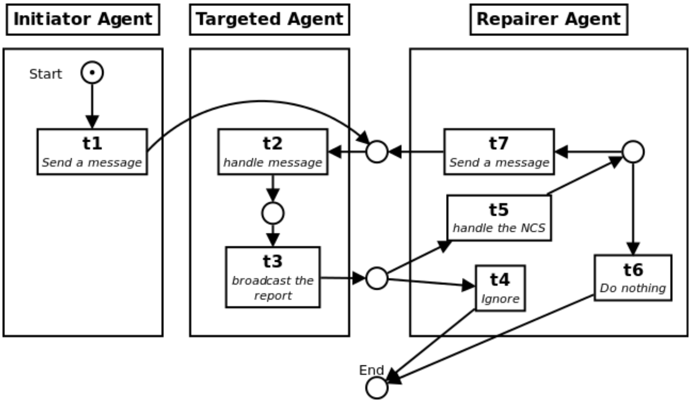


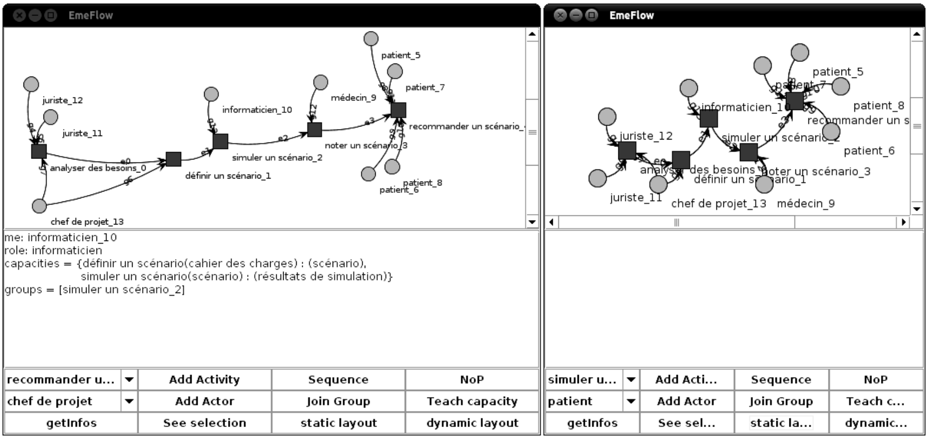
Fig. 3. Petri net of NCS handling

At the beginning an initiator agent (IA) sends a message to a targeted agent (TA) asking it for a modification (e.g. if the targeted agent is an activity it could be asked to adapt its profile). TA handles the requests (t2) and then broadcasts the report (t3) to all the agents interested by the change (e.g. any member of the activity’s group). One agent among them will play the role of repairer. The repairer agent (RA) can find the change cooperative and ignore it (t4). Or, it could discover an NCS and handle it (t5). It can choose between: a) doing nothing (t6), sometimes there is no other way and one of the agents must accept a sub-optimal situation for itself, and b) asking TA to modify itself again by sending it a message (t7). This protocol could loop until one of the following situations is met:

- All the potential repairer agents agree with the change: there are no more conflicts (they all perform t4).
- Some repairer agents don't agree but decide to do nothing (they perform t6) to help the convergence of the system.

Repairer agents detect when they are in an NCS and trigger the corresponding rule. We distinguish two categories for these rules:

- Those that keep the workflow valid. For example, finding a compatible profile when actors' capacities are in conflict. Currently the (very trivial) repair action is to have the profile of any agent be a subset of the activity's profile.
- And those that allow the workflow to be built faster. For example, if an actor A participates in an activity T and realizes that the outputs of T are compliant with the inputs of one of its capacities. Then A will instantiate this capacity as a new activity T' and connect it to T.



me: informaticien_10
role: informaticien
capacities = {définir un scénario(cahier des charges) : (scénario),
simuler un scénario(scénario) : (résultats de simulation)}
groups = [simuler un scénario_2]

Fig. 4. Two WOAS' windows representing the same workflow (left: 4a and right: 4b)

5 Validation: The WOAS Simulator

To validate our approach we have built the WOAS simulator and tested it on a case study. This case study is about the co-design of an healthcare product (improving homecare). It involves actors from several domains: medical, law, computer scientists, sociologists and patients.

During our tests some of those roles were played by human actors. Each of them could access a shared board (see the interface in figure 4a) in which he/she could co-design a workflow.

Figure 5 shows part of a design session as a sequence diagram with real human actors. The computer scientist creates a group (1) and adds himself to this group (2).

He then allocates this group to the activity “user requirement analysis” (3). As a default reaction to any allocation, the activity, which is a software agent decides to respond (4) with its information (profile, allocated actors, …). The computer scientist (a human agent) decides to add the project manager to his/her group (6) so they can work together on the “user requirement analysis”. Again, the activity responds with its information (7). By accessing to the profile of the activity, the project manager checks if he/she accepts the profile of the activity (8). This is not the case and he/she decides to modify it (9, 10).

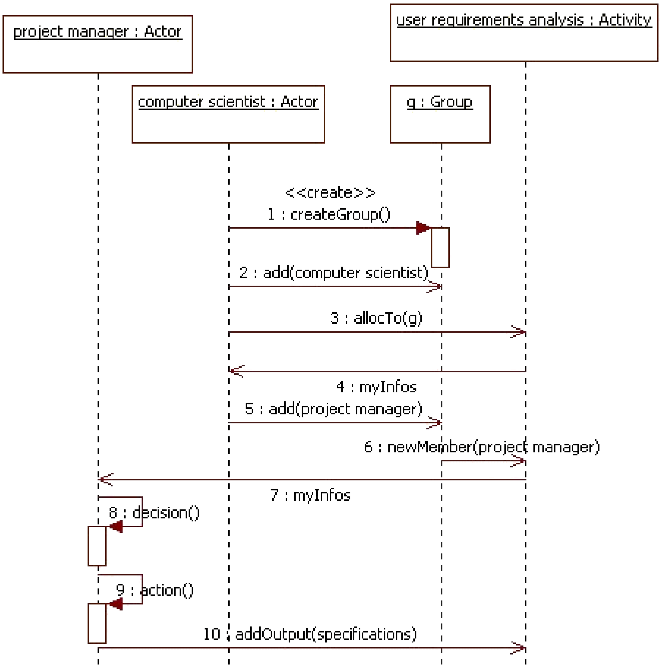

project manager : Actor
user requirements analysis : Activity
computer scientist : Actor
g : Group
<<create>>
1 : createGroup()
2 : add(computer scientist)
3 : allocTo(g)
4 : myInfos
5 : add(project manager)
6 : newMember(project manager)
7 : myInfos
8 : decision()
9 : action()
10 : addOutput(specifications)

Fig. 5. Sequence diagram of a co-design session

Our system provides the following services:

- It allows each agent to add an activity in a shared space, connect it to other activities with a coordination pattern (sequence, parallelism, synchronization, conditional structure), allocate it to a group of actors, modify the profile (input data, output data) of an activity.
- It allows the user to add actors to the workflow and inside a group.
- It allows actors to negotiate the activities’ profile and the workflow structure.
- It allows visualizing the workflow evolving in real-time.

Our system is implemented using SpeADL/MAY [8], a tool allowing to build architectures that are adequate for the development and the execution of multi-agent systems. It also comes with a set of reusable components providing features like communications or scheduling leading to a more lightweight development.

6 Discussion and Conclusion

This paper defines and specifies a new approach for supporting flexible and emergent workflow systems. This is made possible with the interleaving of design, adaptation and execution phases. Actors can decide and specify in a cooperative way, using a negotiation process, the tasks profiles and the group allocated to a task. This work is grounded on the AMAS theory that supports adaptation and conflict resolution. AMAS and workflow concepts have been linked through a unifying meta-model. A simulator called WOAS has been implemented and it demonstrates the feasibility of our solution in a distributed mode. Nevertheless several issues remain open:

- WOAS remains a simulator. It requires, for real world applications, the integration of an execution module and to take into account real world feedbacks that could probably lead to improvements of the tool.
- In our model and simulator some negotiations protocols are explicit and simple, others are complex and emergent. It would be interesting to discover and represent the emergent ones to better understand the emergence phenomenon and to make them available as resources for agents.

References

1. Ermolayev, V., Jentzsch, E., Karsayev, O., Keberle, N., Matzke, W.-E., Samoylov, V.: Modeling Dynamic Engineering Design Processes in PSI. In: Akoka, J., et al. (eds.) ER Workshops 2005. LNCS, vol. 3770, pp. 119–130. Springer, Heidelberg (2005)
2. Schoneneberg, H., Mans, R., Russell, N., Mulyar, N., van der Aalst, W.: Process flexibility: asurvey of contemporary approaches. In: International Workshop on CIAO/EOMAS, International Conference on Advanced Information Systems, pp. 16–30 (2008)
3. Nurcan, S.: A survey on the flexibility requirements related to business process and modeling artifacts. In: Hawaii International Conference on System Sciences, vol. 378 (2008)
4. van der Aalst, W., Weske, M., Grünbauer, D.: Case handling: a new paradigm for business process support. Data Knowl. Eng. 53(2), 129–162 (2005)
5. Bergenti, F., Caire, G., Gotta, D.: Interactive workflows with WADE, WETICE (2012)
6. Jennings, N.R., Norman, T.J., Faratin, P., O'Brien, P., Odgers, B.: Autonomous agents for business process management. Applied Artificial Intelligence 14(2), 145–189 (2000)
7. Casati, F., Baresi, L., Castano, S., Fugini, M.G., Mirbel, I., Pernici, B.: WIDE Workflow Development Methodology, pp. 19–28 (1999)
8. Noël, V.: Component-based software architectures and multi-agent systems: mutual and complementary contributions for supporting software development, Ph.D. Thesis (2012)
9. Weiss, G.: Multiagent systems. A modern approach to distributed artificial intelligence. The MIT Press (1999)
10. Camps, V., Gleizes, M.-P., Glize, P.: A theory of emergent computation based on cooperative self-organization for adaptive artificial systems. In: Fourth European Congress of Systems Science (1999)
11. Glize, P., Picard, G.: Self-organisation in constraint problem solving. In: Serugendo, G.D.M., Gleizes, M.-P., Karageorgos, A. (eds.) Self-organising Software from Natural to Artificial Adaptation, pp. 347–377. Springer (2011)

A Framework for Conflict Resolution in Multi-Agent Systems

Ghusoon Salim Basheer, Mohd Sharifuddin Ahmad, and Alicia Y.C. Tang

Universiti Tenaga Nasional, Jalan IKRAM-UNITEN,
43000 Kajang, Selangor, Malaysia
rawagy2013@gmail.com, {sharif,aliciat}@uniten.edu.my

Abstract. In this paper, we analyze the effects of agents' confidence in conflict resolution of a multi-agent system and propose a resolution technique to possible conflicts in the environment. It establishes a framework on agents' confidence and conflict strength and shows the possible limits of agents which provide uncertain information. This paper offers a confidence model that depends on some factors, which suggest that agents' opinions do not have equal confidence level. We exploit such circumstance in selecting appropriate strategy to resolve conflicts between agents. Consequently, we propose a multi-agent framework which examines agents' confidence using specific factors and exploiting the level of confidence for detecting the best strategy to conflict resolution.

Keywords: Multi-agent System, Conflict resolution, Confidence.

1 Introduction

In the concept of teamwork in multi-agent systems, agents must communicate with each other [1]. When an agent helps a user in some decision, it must ensure that the information is filtered to avoid irrelevant details, as in real world situations [2]. In multi-agent environments, when agents work as a team, they occasionally select one of multiple strategies to eliminate conflicts between agents. Such strategies include negotiation, arbitration, and voting. Equipping agents with the capability to choose one or more strategies gives them more flexible behavior [3]. Conflict resolution strategies in multi-agent systems need to simulate the natural resolution of conflict in real life. Results of previous research on various conflict resolution strategies do provide a foundation to solve the conflict problem, but there is limited research focusing on how agents should select the most appropriate conflict resolution strategy given the goals and current situational context. Most state-of-the-art techniques have not considered all the possible states of conflict occurrences [4], [5].

In this paper, inspired from the social theory of conflict resolution [6], we exploit the theory to build a conflict resolution algorithm to resolve conflicts that occur between agents. The algorithm considers the possibilities of conflict occurrence in the agents group and the use of a management model that demonstrates the rich variety of humans' conflict behavior".

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 195–204, 2013.

The first part of this paper contains a review of the previous papers, their approaches, and current models. The review discusses the proposed strategies of conflict resolution in multi-agent system environment followed by a deliberation on confidence and conflict concepts, their theories, both theoretically and practically. We then propose our technique, in which we suggest new criteria for resolving agents' conflicts based on their confidence. We organize the rest of the paper as follows: Section 2 reviews the concept of conflicts in humans and multi-agent systems. In Section 3, we highlight the state-of the-art of conflict resolution. Section 4 proposes the concept of confidence in multi-agent systems. Section 5 discusses our proposed framework and Section 6 concludes the paper.

2 The Concept of Conflict

2.1 Conflict in Human

In human society, if conflict appears, there are two ways of handling it: continue with the conflict or solve it [4]. Rummel [7] detects power as a source of conflict in social conflict, and he classified conflicts into three types. The First type is when two individuals interest the same thing, and to solve this conflict one of them must be excluded (conflict of appropriate interests). The second type of conflict includes i wants x that j do not wants it (inverse Interest). The third type occurs when two individuals (i and j), i wants A, and j wants B, where A and B are opposed (incompatible interests). Tessier et al. [8] identified six types of conflict resolution model in human society as follows:

1. *Flight*: represents fleeing one of two opponents.
2. *Destruction*: takeovers one of opponents.
3. *Subservience*: gives up by one of opponents.
4. *Delegation:* adds a third party to judge between opponents.
5. *Compromising*: obtains the result of negotiation.
6. *Consensus*: obtains the agreement of opponents.

The following table shows the handling methods of human conflict as mentioned by Bohinc [9].

Table 1. Handling Methods of Human Conflict [8]

Conflict	Example
Flight	Shelving
Destruction	Murder
Subservience	Talking over
Delegation	Judge
Compromise	Negotiation outcome

2.2 Conflicts in Agents

A multi-agent system can be considered as a collection of entities communicating and interacting with each other to achieve individual or collective goals [10]. Through coordinating operation, agents work jointly to guarantee coherent process. Conflict resolution is the fundamental process for coordinated agent attitude [3].

Conflict between agents arises in multi-agent environment in many cases, and it is solved depending on its type and dimension, Tessier [8] classified conflicts into two main classes: physical conflicts and knowledge conflicts. Physical conflicts are consequences of external and resource conflicts. Knowledge conflicts (or epistemic conflicts) occur when each agent has its own information that is different from other agents. In this class of conflict, agents conflict in beliefs, knowledge and opinions.

Inspired from human's conflict resolution strategies, we proposed a framework for conflict resolution as follows [8]:

- **Forcing**: corresponds to Destruction/Flight in some conflict state. We recognize that there is no chance to resolve the conflict.
- **Submitting/Ignoring**: corresponds to Subservience. In this case, there is no force, but inducement between both sides.
- **Delegation**: corresponds to Delegation when the conflict cannot be resolved, both opponents request a third party that has deep knowledge to judge.
- **Negotiation**: corresponds to Compromising through negotiation when one of the opponents is willing to yield. This state includes an agreement in a different style.
- **Agreement**: corresponds to Consensus. Each opponent must give all details about its decision to a third party. For this reason, this process comes as a result of a delegation process.

3 State-of-the-Art in Conflict Resolution

The search space of conflict resolution strategy can be reduced by understanding the nature of a conflict. An agent can focus only on the attitudes that are most suitable for the type of conflict [10]. However, there are many different approaches associated with conflict resolution strategies, but the important question is how an agent selects the most suitable strategy for its situation and aims. Few researchers discussed agents switching between multiple conflict strategies [10], [11], [12]. Liu et al. [10] opined that agents should select an appropriate strategy for conflict resolution depending on three factors: type of conflict, agent's rule, and preference solution. They classified conflicts into three classes: goal conflicts, plan conflicts, and belief conflicts. After classifying conflicts that appeared in the system, many modifications such as goal modification, plan modification, and desire modification are performed to resolve the conflicts. Adler et al. [11] allowed an agent to select a specific strategy from many other strategies such as priority agreement, negotiation, arbitration, and self-modification. They detected network performance as criteria that control selecting suitable strategy. For example, if there is heavy network traffic, an agent selects the arbitration strategy to resolve conflict, but if there is light traffic, the agent selects negotiation or other strategy. Liu et al. [12], mentioned the importance of allowing agents to select an appropriate conflict resolution strategy based on many factors such as conflict's nature (if there is conflict in goal, plan or belief), the agent's autonomy level, and the agent's solution preferences. Barber et al. [3] produced multiple strategies for conflict resolution such as negotiation, self-modification and voting. Selecting each one of these strategies depends on several characteristics like cost and required time. Jung et al. [13] tried to solve agents' conflict problem by implementing a new system called CONSA (Collaborative Negotiation System based on Argumentation), based on agent negotiation strategy. Through negotiation, agents propose arguments as justifications or elaborations to explain their decisions.

In [4], the authors defined three strategies to conflict resolution: Negotiation, Mediation and Consensus. Giret et al. [5] used alternative dispute resolution (ADR) strategy that is applied in grievance protocols [14] to solve conflict in the Open MAS. The important point in their framework is that any grievance includes negotiation. In this way, the result of a conflict resolution can be an agreement among the conflicting parties by which they voluntarily settle the conflict, or a decision from an arbitrator (a neutral third party) which is final and binding on both conflicting parties. The framework is designed in such a way that multiple agreement mechanisms may be implemented concurrently. Wagner et al. [1] focused on the distinguishing between potential conflict and real conflict in multi-agent system. They use expanded classical Petri net (CPN) as a modeling tool to describe potential conflicts in MAS. They introduce it as a new model called *potential arc*. Through a MAS design model, an agent designer creates several paths for each action depends on the system resources. Each one of these paths may include potential arc that contain potential conflicts, which are still probable until they appear at runtime.

In Fan[15] PhD thesis, he proposed a novel plan to resolve conflicts. His model examines the beliefs of agents instead of using utility function to detect the winning party, and then detects the reason of participating agents. Using this processing strategy, his model is able to resolve the conflict. Jung [16] developed new strategies that increase the speed of convergence to conflict resolution. His strategy depends on discovering the auxiliary communication between neighboring agents and selecting new choice from them to speed up conflict resolution.

4 Confidence Models

In [17], [18], [19], the authors build their models by using agent confidence and agent reputation, which depends on past experiences. In situations where no past experiences are available, their model gathers information from other agents. Detecting agent confidence depends on other agents' experience. Hermoso et al. [20] defined agent confidence based on the agent's experience through interacting with other agents. Their model stores the agents' confidence values from past connections in a table, which are exploited to detect agents' confidence values.

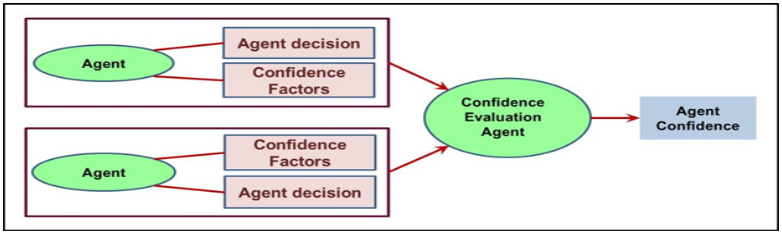


Fig. 1. Agent Confidence Model

We propose a new model to build agent confidence based on specific pre-defined factors. The agent confidence level depends on satisfying these factors. Our model provides a Confidence Evaluation Agent (CVA), which receives defined factors for each agent to calculate the agent's confidence value (see Figure 1).

5 The Proposed Conflict Resolution Framework

From the review of current research works, there are many proposed strategies for conflict resolution, but there is no one strategy that works best for all situations [10]. Current agent-based approaches that resolve conflict are offered from many different perspectives. In multi-agent environments, conflicts occur when agents are assigned to conflicting goals or when resources are scarce. Such situations create the need for some strategy to resolve the conflicts. In order to understand the issues of conflicts in multi-agent environments, we analyze the social theory of conflict and propose a conflict resolution strategy. In our approach, agents assign confidence values to their opinions from domain specific pre-defined factors. We analyze conflict resolution strategies to discover the outcome of each strategy. We then utilize the domain parameters to reliably classify the components that we consider important for conflict management. We propose two dimensions of agent's conflict resolution framework:

- **Confidence Level**. which is the level of certainty on an opinion. This varies from one agent to another depending on the pre-defined confidence factors. There are two levels:
  - High level Confidence (HLC).
  - Low Level Confidence (LLC).

- **Conflict Strength**. which defines the magnitude of conflict between two agents' opinions.

Figure 2 depicts the analytical process of classifying the two dimensions of conflict resolution framework. For instance, when the Evaluation agent (shown in Figure 1) must resolve opinions conflict over many collected opinions, it could depend on the confidence level of each agent and the strength of that conflict to gain helpful information to for selecting appropriate strategy for conflict resolution.

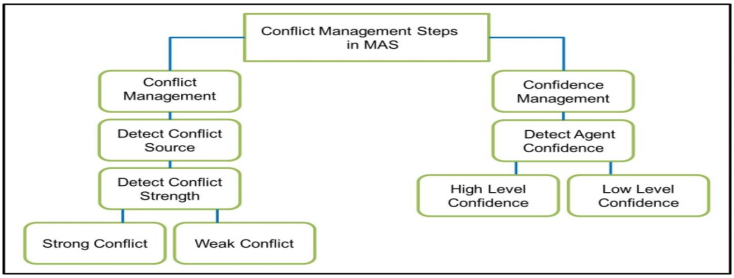


Fig. 2. A framework for Classifying Conflicts and Confidence in MAS

Figure 3 shows an overview of the proposed strategy to categorize all possible conflict resolution choices based on the social conflict theory.

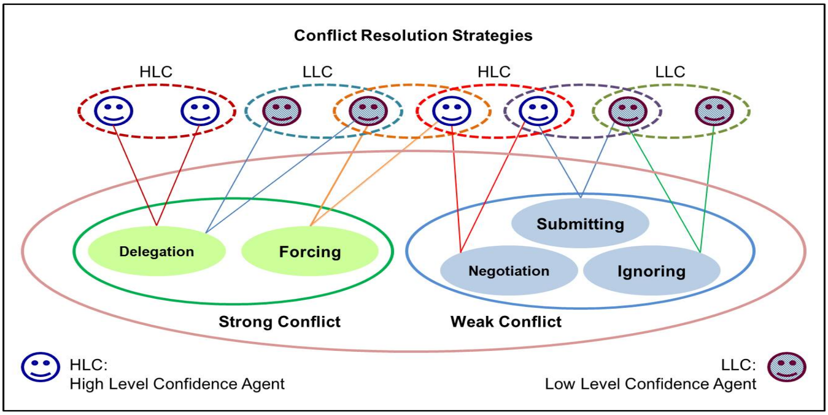


Fig. 3. A Proposed Model to Select Appropriate Conflict Resolution Strategy

We define two levels of conflict strength:

- **Strong Conflict (SC):** When two agents conflict more than 50% of their decisions or opinions. Under this type of conflict we have two situations:
  - (a) Conflicting agents with same level of confidence, i.e., a High Level Confidence agent (HLC) with another High Level Confidence agent (HLC) or LLC with LLC. The proposed strategy for such conflict is Delegation.
  - (b) Conflicting agent with different level of confidence, i.e., a High level Confidence agent (HLC) with a Low Level Confidence agent (LLC). The proposed strategy for such situation is Forcing.

- **Weak Conflict (WC):** When two agents conflict less than 50% of their decisions or opinions. Under this type of conflict we have three situations:
  - (a) Conflicting agents with High level Confidence (HLC) with High Level Confidence (HLC). The proposed strategy is Negotiation.
  - (b) Conflicting agents with High Level Confidence (HLC) with Low Level Confidence (LLC). The suitable strategy is Submitting.

(c) Conflicting agents with Low level Confidence (LLC) with Low Level Confidence (LLC). The proposed strategy to apply is Ignoring.

In [21], a negotiation will take place when neither party in a conflict is strong enough to impose its decision or to resolve the conflict unilaterally. The authors in [21] also mentioned that in situation of perceived asymmetry, the stronger party tends to act exploitatively while the weaker acts submissively in order to reach a situation for effective and satisfying negotiations. This situation corresponds to the "forcing strategy" in our framework. Hazleton [22] defines "argue" as withdrawing (avoiding) happens when a person does not pursue her/his own concerns. He mentioned that "avoiding" is suitable when the issue is trivial. We are inspired by the concept of detecting the "submitting" state for addressing the issue of not important (i.e. weak conflict) and one of the two conflicted parties is not strong enough. It was also argued in [21] that two low-power conflicted parties will cause a deadlock because they do not have the power to make the other move. Based upon these ideas we classified the state into two choices: "ignoring" in the state of weak conflict and "delegation" in the state of strong conflict.

5.1 Selecting Conflict Resolution Strategies

The conflict resolution strategies used in our confidence model are based on the following definitions:

Definition 1. Given a set of agents, $A= \{a1, a2. . . an\}$, each agent $a_i \in A$ has a set of specification (SPi) that includes opinion, o_i and confidence, c_i, i.e. $SPi = (oi, ci)$. An agent's opinion, oi may conflict with another agent's opinion, o_j, or a set of other agents' opinions $\{ok ,…, ox\}$.

Definition 2. Let ai be an agent, such that $ai \in A$. Each agent in A has an Agent Confidence value (ACi) as a positive integer number that represents the level of confidence of the agent based on pre-defined factors.

Definition 3. A Conflicting Agents Set, CAS, is a set of each pair of conflicting agents.

Definition 4. For each two conflicting agents $\{(ai, aj) \in CAS\}$, their conflict strength is represented by $CSij$ with two levels of agent's confidence, High Level Confidence (HLC) and Low Level Confidence (LLC). Three situations are apparent: $ACi = ACj$, or $AC\ i > ACj$, or $ACi < ACj$.

Definition 5. For each two conflicting agents $\{(ai, aj) \in CAS\}$, we define six possible Conflict Resolution Strategies $(CRSij)$ by detecting the conflict strength $(SCij)$ towards agent confidence level (ACi, ACj).

If $AC_i = AC_j$ and both are HLC agents and SC_{ij} = high, then call Evidence Function, EF, and third party Mediator to judge CRS_{ij}= Delegation).
If $AC_i > AC_j$ and SC_{ij} = high, then (CRS_{ij}=Forcing).
If $AC_i=AC_j$ and both are HLC agents and SC_{ij}=low, then CRS_{ij}=Negotiation).
If $AC_i<AC_j$ and SC_{ij}= low then (CRS_{ij}=Submitting).
If $AC_i=AC_j$ and both are LLC agents and SC_{ij}= low, then (CRS_{ij}=Ignoring).
If $AC_i=AC_j$ and both are LLC agents and SC_{ij}=high, then (CRS_{ij}=Delegation).

5.2 The Proposed Algorithm

The agents' confidence (AC) can be used for selecting appropriate conflict resolution strategy inspired from the ***social conflict theory***. This confidence value is used by the Evaluator agent that receives all agents' opinions and makes a final decision. The benefit of this concept is that it prevents using all possible conflict resolution strategies at all conflict states, and constrained some strategies in special states. For each agent a_i in the system we define the following:

```
Define Set of agents' opinions {AO}:
If an opinion conflict is detected then Define {Conflict Resolution Strategies} as the set of all possible strategies that include{Delegation, Ignoring, Forcing, submitting, Negotiation}
Define {Conflict Strength} as the set of two levels {Strong Conflict (SC) and Weak Conflict (WC)}
Define Conflicted Agents Set (CAS);
Define{Confidence Level}as a set of two levels{High Level Confidence HLC),Low Level Confidence LLC)} for each agent in (CAS)
   Evaluate the Confidence values for each agent in (CAS);
    If Conflict Strength=SC Then
      If the conflict state is between HLC and HLC Then Return Delegation
      If the conflict state is between HLC and LLC Then Return Forcing
      If the conflict state is between LLC and LLC Then Return Delegation
    If Conflict Strength=WC Then
      If the conflict state is between LLC and LLC Then Return Ignoring
      If the conflict state is between HLC and HLC Then Return Negotiation
      If the conflict state is between HLC and LLC Then Return Submitting
```

Example: A family wants to buy three pieces of furniture, and the parents decide to consider the family members' opinions to decide what important pieces to buy. What strategies could be used if there are conflicting views?

In this case, the parents have high confidence from their experience, age and knowledge and both parents (i.e., father and mother) have almost the same level of confidence. Their children are assumed to have low level of confidence. Suppose a conflict arises between the parents' opinions, and the conflict is about buying more than two pieces of furniture. This type of conflict is considered as a Strong Conflict (both are HLC agents). From the framework, the Delegation strategy should be deployed to resolve the conflict.

If a conflict occurs between one of the parents and their son about buying one piece of furniture, the conflict is a Weak Conflict (HLC vs. LLC). Consequently, the strategy here is for the son to yield (Submitting). Table 1 summarizes the strategies

Table 2. Strategies for Conflict Resolution in Buying Furniture

Furniture	Father/Mother	Parents/Child	Child/Child
Less than two pieces (weak Conflict)	(HLC/HLC), WC **(Negotiation)**	(HLC/LLC), WC **Forcing**	(LLC/LLC), WC **Ignoring**
two pieces or more (Strong Conflict)	(HLC/HLC), SC **Delegation**	(HLC/LLC), SC **Submitting**	(LLC/LLC), SC **Delegation**

for other possibilities of conflicts. We assume that the family wants to buy three units of furniture (Table, TV, and Fan):

```
BuyTable(T) ← ChoiceTable(T)
ChoiceTable(T)←FatherChoice(T),MotherChoice(T),DaughterChoice(T,
SonChoice(T)
BuyTV (TV) ← ChoiceTV(TV)
ChoiceTV(TV)←FatherChoice(TV),MotherChoice(TV),DaughterChoice(TV),SonCh-
-oice(TV)
BuyFan (F) ← ChoiceFan(F)
ChoiceFan(F)←FatherChoice(F),MotherChoice(F),DaughterChoice(F),
SonChoice(F)
Assumptions
FatherChoice(T), ¬ MotherChoice(T)
MotheChoice(F), ¬ SonChoice(F)
Conflicts
C(FatherChoice(T))={¬ MotherChoice(T), FatherChoice(T1)/T ≠ T1
C(MotherChoice(F))= {¬ MotherChoice(F), SonChoice(F1)/F ≠ F1
```

In this example there are two conflict states: Father wants to buy (Table (T1)) but Mother Wants to buy (Table (T)), mother wants to buy (Fan (F)) but son wants to buy (Fan (F1)). The Evaluator agent checks the confidence value for all of them and the strength of each conflict. In the first state, we assume that father and mother have the same confidence level, the conflict about one piece of furniture, so the evaluator agent selects the Negotiation strategy. In the second state, Mother wants to buy (Fan (F)) but the son wants to buy (Fan (F1)). Since Mother has a confidence level higher than the son, so the strategy selected is Forcing.

6 Conclusion

We propose an approach to detect appropriate strategy for conflict resolution in multi-agent environments. We classify conflicts as Strong Conflict (SC) and Weak Conflict (WC), and categorize agents' confidence as High Level Confidence (HLC) and Low level Confidence (LLC) depending on pre-defined domain specific factors. We exploit the social conflict theory and use the confidence level to determine the best strategy for conflict resolution. A three-track algorithm between agent confidence and conflict strength is also proposed. Our future work will include evaluation of the proposed conflict resolution system and the detection of the conflicting strengths using some threshold points.

References

1. Wagner, T., Shapiro, J., Xuan, P., Lesser, V.: Multi-Level Conflict in Multi-Agent Systems. LNAI, vol. 4386. Springer, Heidelberg (2007)
2. Wang, Y., Vassileva, J.: Bayesian Network-Based Trust Model. In: Proceedings of the International Conference on Web Intelligence (WI 2003). IEEE (2003)
3. Barber, K.S., Liu, T.H., Han, D.C.: Strategic Decision-Making for Conflict Resolution in Dynamic Organized Multi-Agent Systems. A Special Issue of CERA Journal (2000)
4. Crawford, D., Bodine, R.: Conflict Resolution Education A Guide to Implementing Programs in Schools. Youth-Serving Organizations, and Community and Juvenile Justice Settings Program Report (1996)

5. Giret, A., Noriega, P.: On Grievance Protocols for Conflict Resolution in Open Multi-Agent Systems. In: Proceedings of the 44th Hawaii International Conference on System Sciences (2011)
6. Xin, G., Xiao, Y., You, H.: An Improved Dempster-Shafer Algorithm for Resolving the Conflicting Evidences. International Journal of Information Technology 11(12) (2005)
7. Rummel, R.J.: Understanding Conflict and War. The Conflict Helix, ch. 27, Conflict In The Sociocultural Field, vol. 2. Sage Publications, Beverly Hills (1976)
8. Tessier, C., Chaudron, L., Muller, H.J.: Conflict agents, Conflict management in Multi Agent System, vol. 1. Springer, Heidelberg (2000)
9. Bohinc, Š.L.: Epistemology of social work. Social Work Journal 37(6), 417–440 (1998)
10. Liu, T.H., Goel, A., Martin, C.E., Barber, K.S.: Classification and Representation of Conflict in Multi-Agent Systems. The Laboratory for Intelligent Processes and Systems, the University of Texas at Austin (1989)
11. Adler, M.R., Davis, A.B., Weihmayer, R., Worrest, R.W.: Conflict- Resolution Strategies for Nonhierarchical Distributed Agents. In: Gasser, L., Huhns, M.N. (eds.) Distributed Artificial Intelligence II, pp. 139–161. Pitman Publishing, London (1989)
12. Liu, T.H., Barber, K.S.: Selection of Conflict Resolution Strategies in Dynamically Organized Sensible Agent-based Systems. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence, Madison, p. 1193 (1998)
13. Jung, H., Tame, M.: Conflict in Agent Team, Multiagent System, Artificial Intelligent, and Simulated Organizations. LNAI, vol. 1. Springer, Heidelberg (2002)
14. Kraus, S.: Automated Negotiation and Decision Making in Multiagent Environments. Mutli-agents systems and applications. LNAI, vol. 1. Springer, Heidelberg (2002)
15. Fan, X.: Conflict Resolution with Multi-agent Argumentation. PHD thesis, University of London Imperial College of Science, Technology and Medicine Department of Computing (2012)
16. Jung, H.: Conflict Resolution Strategies and Their Performance Models for Large-Scal Multiagent Systems. PHD thesis, University of Southern California (2003)
17. Huynh, T.D., Jennings, N.R., Shadbolt, N.R.: An Integrated Trust and Reputation Model for Open Multi-agent System. In: Proceeding of the 16th European Conference on Artificial Intelligence, ECAI (2004)
18. Ramchurn, S.D., Sierra, C., Godo, L., Jenning, N.R.: A Computational Trust Model for Multi agent interactions based on confidence and reputation. In: Proceedings of 6th International Workshop of Deception, Fraud and Trust in Agent Societies, pp. 69–75 (2003)
19. Sabater, J., Sierra, C.: REGRET: A Reputation Model for Gregarious Societies. In: Proceedings of the Fifth International Conference on Autonomous Agents, Montreal, Canada, pp. 194–195. ACM Press, New York (2001)
20. Hermoso, R., Billhardt, H., Ossowski, S.: Integration Trust in Virtual Organization. In: Noriega, P., Vázquez-Salceda, J., Boella, G., Boissier, O., Dignum, V., Fornara, N., Matson, E. (eds.) COIN 2006. LNCS (LNAI), vol. 4386, pp. 19–31. Springer, Heidelberg (2007)
21. Zartman, W.: The Structuralist Dilemma in Negotiation. Research Group in International Security (1997)
22. Hazleton, M.: Conflict Management Techniques, Copyright © HumanMetrics Inc. (2013), `http://www.personalityexplorer.com/Home.aspx`

Heterogeneous Agents' Interactions in a Double Auction Virtual Stock Market

Diana Dezsi[1,2], Iulia Mărieş[2], and Florentina-Olivia Bălu[1]

[1] Dept. of High Commercial Studies, University of Geneva, Geneva, Switzerland
{diana.dezsi,florentina.balu}@unige.ch
[2] Dept. of Informatics and Economic Cybernetics, University of Economic Studies,
Bucharest, Romania
iulia.maries@hotmail.com

Abstract. The hereto paper analyses the impact of the number of heterogeneous agents in an evolutionary agent-based model of the stock market when simulated through Adaptive Modeler simulation software application. The paper compares the returns, total wealth and distributions of wealth obtained from simulating the evolutionary agent-based model with 500, 1,000, 1,500 and 2,000 agents which create a virtual stock market using a double auction trading mechanism. Within the agent-based model the population of agents is continuously adapting and evolving by using genetic programming in order to generate new agents by using the trading strategies of the best performing agents and replacing the worst performing agents in a process called breeding.

Keywords: heterogeneous agents, double auction, virtual stock market, wealth distribution.

1 Introduction

The aim of our research is to identify the main distinctions between the simulation outputs of an adaptive agent-based model for the stock market, when different numbers of agents are used to trade in the virtual stock market. In order to achieve our research aim, we use the Adaptive Modeler [1] software to simulate the adaptive agent-based model for virtual stock market generation and price forecasting of real world market-traded securities such as stocks. Thus, heterogeneous agents trade a stock floated on the stock exchange market, placing orders depending on their budget constraints and trading rules, where the virtual market is simulated as a double auction market. For the agent-based model, the software uses evolutionary computing such as the Strongly Typed Genetic Programming [2] in order to generate new agents by using the trading strategies of the best performing agents and replacing the worst performing agents in a process called breeding, thus creating adaptive, evolving and self-learning market modeling and forecasting solutions.

The agent-based models have been successfully implemented as a powerful tool in the exploration and understanding the financial markets' complexity and trading

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 205–214, 2013.

behavior, offering explanation for observed stylized facts and being able to reproduce many of them [3] [4] [5]. Arthur et al. [6] from Santa Fe Institute, Ca., USA, developed an artificial stock market which allowed for testing of agent-based models with heterogeneous agents. LeBaron [7] highlighted the main strengths and weaknesses of the Santa Fe artificial market, pointing out that learning speed of agents is very important for the model's prediction capacity.

An important development of the trading mechanism used in the agent-based models for simulating financial markets is the double auction (also called double-sided double auction, or bid-ask auction) trading mechanism implementation, the pioneers in this field being Gode and Sunders [8] [9] which introduced the zero-intelligent trader concept, further developed by Cliff [10], Gjerstad and Dickhaut [11]. Rust [12] and Phelps et al. [13] have experimented with heterogeneous agents which change their strategies during the learning process, the unprofitable strategies being replaced with the more profitable ones, thus developing adaptive models which use genetic algorithms to evolve. Walia [14] dedicated his studies to the development of the agent-based models which use genetic programming, a form of evolutionary computation similar to genetic algorithms, allowing for more flexibility and effectiveness in finding optimal solutions, programs being encoded as tree structures, thus crossover and mutation operators being applied easier. The later is similar with the learning process used in the hereto paper.

The wealth represents the total value of cash and shares that an agent holds. Many of the papers simulating agent-based models assume equal initial amount of wealth for the agents, which is an unrealistic assumption, as real markets involve heterogeneity in terms of wealth distribution of traders. According to empirical studies, the income distribution follows a power-law distribution, which we will also use in the hereto paper for the initial endowment of the population, as in [5].

The stock market involves a large number of traders and we suppose that the total number of agents N is of great impact on the output of an agent-based model simulation. In order to assess the impact of the number of agents on the market behavior, we perform several experiments in which N takes the values 500, 1,000, 1,500 and 2,000.

The organization of this paper is as follows. Section 2 presents the specifications of the adaptive agent-based model used in the simulations, Section 3 describes the datasets used in this study, while the results of the simulations that have been performed are presented in section 4, the paper ending with the conclusions and avenues for future work.

2 Evolutionary Agent-Based Model Specifications

An agent-based model is a computational model for simulating the actions and interactions of multiple agents in order to analyze the effects on a complex system as a whole, and represents a powerful tool in the understanding of markets and trading behavior. An agent-based model of a stock market consists of a population of agents (representing investors) and a price discovery and clearing mechanism (representing a virtual stock market). As regards to the stock markets, agent-based models can successfully replicate time series features like fat-tailed distributions and volatility clustering, on which standard financial models offer few explanations, as market prices derive from the interaction of a large number of heterogeneous investors with different decision making methods and different investment goals. The complex dynamics of these heterogeneous investors and the resulting price formation process require a simulation model of multiple heterogeneous agents and a virtual market. Research has shown that complex behavior can emerge from simulations of agents with relatively simple decision rules.

The evolutionary agent-based model referred to in this paper is simulated in Adaptive Modeler software, which supports up to 2,000 agents and 20,000 observations for each simulation. The model consists of a population of agents and a Virtual Market on which the agents trade the envisaged security. The agents are autonomous and heterogeneous entities representing the traders of the stock market, each having their own *wealth* (cash and shares) and their own trading strategy called the *genome*.

The general cycle of the evolutionary agent-based model used in this paper, which repeats at each of the analyzed quotes, is described as follows:

1. *Import real stock market data.*
2. *Initialization of the model*: At the initialization, the agents are endowed with an initial wealth according to a Pareto probability distribution, a well known power law distribution commonly used to describe wealth or income distributions, describing unequal distribution, where a large part of the total wealth is owned by a small percentage of individuals, which was first described by Pareto [15] . They are also given a trading rule which is called the genome, which is randomly created by taking in account the selected genes (which represent functions) using genetic programming. Broker fees are variable and are set to 0.2% of the transaction value. There is no market maker. All the parameters for each of the two models and their values are described in *Table 1*.
3. *Receive new quote bar.*
4. *Agents evaluate trading rules and place orders*: Agents get access to historical prices and evaluate their trading rules according to the genomes allocated in the initialization process, resulting in a desired position as a percentage of wealth limited by the budget constraints, and a limit price. Agents are two-way traders during the simulations, meaning that they are allowed to both sell and buy during multiple periods, and they are one-way traders during a single period (in our case a day) corresponding to an auction, as they are able to submit only one order per auction, either buy or sell. The position is generated in a random manner, while the limit price is generated after a technical analysis has been performed, according to the genome structure which represents trading functions.
5. *Virtual Stock Market clearing and forecast generation*: The Virtual Stock Market determines the clearing price using a clearing house, which is a discrete time double-sided auction mechanism in which the Virtual Stock Market collects all bids (buying orders) and asks (selling orders) submitted by the agents and then clears the market at a price where the supply quantity equals the demanded quantity, therefore the clearing price is the price for which the highest trading

volume from limit orders can be matched, thus all agents establish their final positions and cash at the same time. In case the same highest trading volume can be matched at multiple prices, then the clearing price will be the average of the lowest and the highest of those prices. Market orders have no influence on the clearing price, only executed orders do. The Virtual Stock Market also executes all executable orders, and forecasts the price for the next bar.

6. *Breeding*: During the breeding process, new agents are created from best performing agents in order to replace the worst performing agents, creating new genomes by recombining the parent genomes through a crossover operation, and creating unique genomes by mutating a part of the genome. The breeding process repeats at each bar**,** with the condition that the agents must have a minimum breeding age of 80 bars, in order to be able to assess the agents' performance.
7. *The model waits for a new quote*: If the model receives new quotes, it will repeat the process described at points 4-6. If there are no more quotes to be processed, the simulation ends.

Table 1. General settings of the models. Market and agents' parameters configuration in the simulations

Parameter Type	Parameter Name	Parameter Value
Market Parameters	No. of trading periods (bars)	20,000
	No. of agents	500/1,000/1,500/2,000
	Minimum price increment for prices generated by model	0.01
	Spread	0.01%
	Variable Broker fee	0.2%
Agent Parameters	Wealth Distribution	Pareto distribution, Pareto index 2
	Position Distribution	Gaussian distribution
	Min. position unit	20%
	Max. genome size	1000
	Max. genome depth	20
	Min. initial genome depth	2
	Max. initial genome depth	5
	Genes	CurPos, LevUnit, Rmarket, Vmarket, Long, Short, Cash, Bar, IsMon, IsTue, IsWed, IsThu, IsFri, close, bid, ask, average, min, max, >, change, +, dir, isupbar, upbars, pos, lim, Advice, and, or, not, if
	Breeding Cycle Frequency	1 bar

Table 1. *(Continued.)*

Parameter Type	Parameter Name	Parameter Value
	Minimum breeding age	80
	Initial selection: randomly select	100% of agents of minimum breeding age or older
	Parent selection	5% agents of initial selection will breed
	Mutation probability	10% per offspring

In order to obtain random seed, the Adaptive Modeler software uses the Mersenne Twister algorithm [16] to generate pseudo random number sequences for the initial creation of trading rules or genomes and for the crossover and mutation operators of the breeding process.

The interactions among agents are enabled by auctions, an approach pioneered by Vickrey [17] who saw the market institution as a game of incomplete information in which agents do not know each others' private values. Therefore, in double auctions the study aims at maximizing the social welfare and identifying how price formation develops dynamically [18]. A double auction is a trading mechanism in which buyers and sellers can enter bid or ask limit orders and accept asks or bids entered by other traders [8]. This trading mechanism was chosen to be used for the virtual market simulation in the Adaptive Modeler models because most of the stock markets are organized as double auctions. Double auction stock markets converge to the equilibrium derived by assuming that the traders are profit-maximizing Bayesian, being an example of a microeconomic system, as described in Hurwicz (1986) [19] and Smith (1982) [20]. In the double auction markets, agents introduce bid or ask orders, each order consisting of a price and quantity. The bids and asks orders received are put in the order book and an attempt is made to match them. The price of the trades arranged must lie in the bid-ask spread (interval between bid price and ask price). Furthermore, the use of double auction trading mechanism generates stochastic waiting times between two trades, as stated by Scalas [21].

The genome of the agent uses a tree composed of genes which generates the trading strategies. The initial node is called *Advice* and combines the position generated by *RndPos* function and the limit price value generated by the *Limit* function into a buy or sell order advice. The *RndPos* gene returns a desired position value ranging from -100% to 100% which is randomly generated from a uniform distribution. The *Limit* function uses simple technical indicator initially generated in a random manner from the list of functions selected to be used in the model, which develop during the breeding process, in order to generate the limit price for the buy or sell order. The buy or sell order is introduced in the market after comparing the desired position with the agent's current position and calculating the number of shares that need to be bought or sold, taking also in consideration the available cash. The trading rules of the model use historical price data as input from the Virtual Stock

Market, and return an advice consisting of a desired position, as a percentage of wealth, and an order limit price for buying or selling the security. The trading rules are implemented by genetic programming technology explained bellow. Through evolution the trading rules are set to use the input data and functions (trading strategies) that have the most predictive value.

The agents' trading rules development is implemented in the software by using the *Strongly Typed Genetic Programming* (STGP) approach, and use the input data and functions that have the most predictive value in order for the agents with poor performance to be replaced by new agents whose trading rules are created by recombining and mutating the trading rules of the agents with good performance. The STGP was introduced by Montana (2002) [2], with the scope of improving the genetic programming technique by introducing data types constraints for all the procedures, functions and variables, thus decreasing the search time and improving the generalization performance of the solution found. Therefore, the *genomes* (programs) represent the agents' trading rules and they contain *genes* (functions), thus agents trade the security on the Virtual Market based on their analysis of historical quotes.

During the breeding process, new offspring agents are created from some of the best performing agents to replace some of the worst performing agents. In order to achieve this, at every bar, agents with the highest *Breeding Fitness Return* are selected as parents, and the genomes (trading rules) of pairs of these parents are then recombined through genetic crossover to create new genomes that are given to new offspring agents. These new agents replace agents with the lowest *Replacement Fitness Return*. The fitness functions are a measurement of the agent's investment return over a certain period, therefore the *Breeding fitness return* is computed as a short term trailing return measure of the wealth over the last 80 analyzed quotes and represents the selection criterion for breeding (best agents), while the *Replacement fitness return* is computed as the average return per bar and represents the selection criterion for replacement (worst agents).

3 Data

The data used in this paper was retrieved from *www.altreva.com* and Bloomberg application and contains daily data for the S&P500 stock index, including open, high, low and close values. The analysed period is January 3rd 1950 – March 12th 2013, meaning 15,800 trading days. Such a large period of time ensures the optimisation of the learning process and an increased level of adaptability and evolution of the agents' trading rules within the model, thus generating better simulation results.

The parameters used in the model are described in Section 2, Table 1, and remain constant during the simulations, except the number of agents which will take the following values: 500, 1,000, 1,500 and 2,000. The simulations will be processed by the Adaptive Modeler software application, using a double auction trading mechanism and heterogeneous agents which interact within the Virtual Stock Market, while the population of agents adapts and evolves using genetic programming.

4 Simulation Results

The simulation results were averaged over 40 independent simulations, with the same parameters configuration values except for the number of agents, each conducted over 63 years with different initial seeds of the random number generators, to ensure that the results of the simulations are constant, allowing for an assessment of the robustness and accuracy of the simulation results. The market and agent parameters are explained in section 2 and summarized in Table 1.

In order to analyze the fit of the price forecast generated by the simulations when compared to real stock market data, the Root Mean Squared Error (RMSE) indicator was computed for each simulation and plotted in Fig. 1.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(F_{t-i} - P_{t-i})^2}$$

where $\mathrm{n} = \mathrm{t} - \mathrm{s} + 1$, when the calculation period is (s, t)

F_i represents the forecast value generated by the Virtual Stock Market for date i, P_i represents the real stock market value from the input data.

The results show that when the number of the agents who participate in the Virtual Stock Market rises, the fit of the model improves as the RMSE indicator drops. Therefore, the model which incorporates 2,000 agents is also the most reliable, consistent, robust and accurate, when compared to the other models, thus generating the best forecast for the stock price.

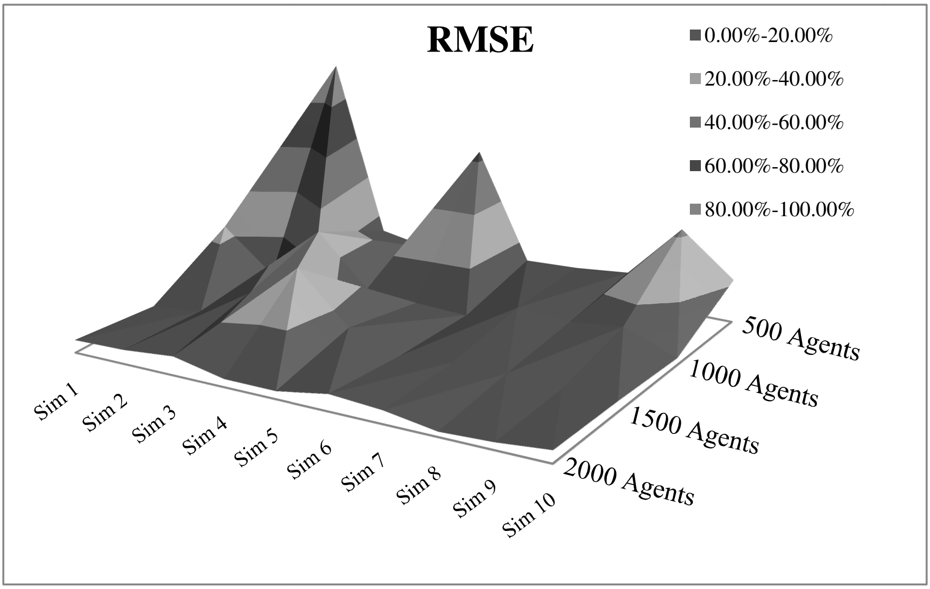


Fig. 1. RMSE - The fit of simulation results with real data when the number of agents changes

As regards to the wealth distribution of the population in the models, the 2,000 agents model preserves best the power law distribution initially allocated, while for the other models in some simulations the discrepancies between agents' wealth grow stronger, as a few number of agents gain market power, earning huge returns, as illustrated in the examples in Fig. 2.

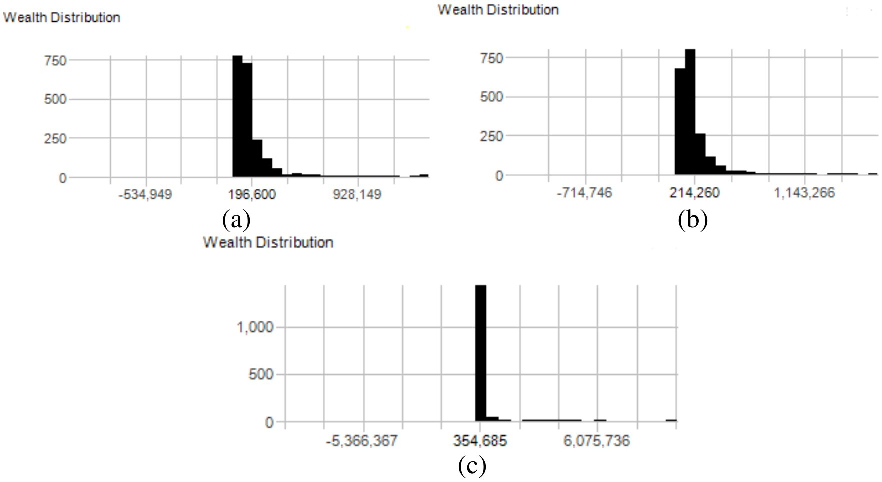


Fig. 2. Wealth distributions for: (a) power-law initial wealth distribution; (b) power-law like final wealth distribution (from 2,000 agent model simulation); (c) large discrepancy final wealth distribution (from 500 agent model simulation).

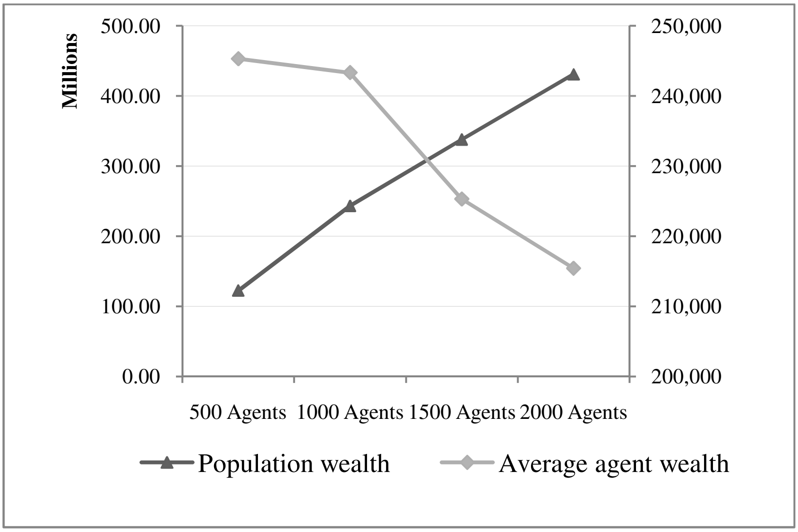


Fig. 3. Average agents' wealth versus population wealth (averaged values for each of the 500/1,000/1,500/2,000 agent model simulations)

According to the averaged results obtained from the simulations, the highest population wealth is obtained for the 2,000 agent model, although the average agent wealth resulted from the simulations is the lowest. Therefore, the best model from the point of view of social welfare is the 2,000 agent model, which is most important for market design, achieved through the double-sided auction trading mechanism and a sufficiently large number of agents.

5 Conclusions

The results following the assessment of the impact of the number of agents in the agent-based model outcome resulted from the stock market simulations have shown that the 2,000 agent based model is best suited to simulate the stock market when compared to the 500, 1,000 and 1,500 agent based model. The conclusion comes from analyzing the fit of the models in terms of forecast abilities, power-law distribution for wealth and the population's wealth versus average agent wealth.

The hereto paper brings to light the importance of the agent-based models when it comes to number of agents impact over the simulation results, taking in consideration that many of the agent-based models simulations in the past were computed over small numbers of agents, especially the experiments based on human agents interactions or human mixed with artificial agents interactions, which can be irrelevant in a stock market simulation, especially the ones involving a double-sided auction trading mechanism.

Further research should focus on increasing the number of agents to 5,000 or 10,000, but the version of software used in this paper does not allow simulations with such a large number of agents.

Acknowledgment. We are grateful for the financial support of the Scientific Exchange Programme Sciex-NMS.ch, project code 12.040, "ROU: Swiss Research Fellowships".

This work was co-financed from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007-2013, project number POSDRU/107/1.5/S/77213 „Ph.D. for a career in interdisciplinary economic research at the European standards".

References

1. Altreva, http://www.altreva.com/technology.html
2. Montana, D.: Strongly Typed Genetic Programming. Evolutionary Computation 3(2), 199–230 (2002)
3. Alfi, V., Cristelli, M., Pietronero, L., Zaccaria, A.: Minimal agent based model for financial markets I: origin and self-organization of stylized facts. The European Physical Journal 67(3), 385–397 (2009)
4. Daniel, G.: Asynchronous simulations of a limit order book. University of Manchester. Ph.D. Thesis (2006)

5. Aloud, M., Fasli, M., Tsang, E., Dupuis, A., Olsen, R.: Modelling the High-Frequency FX Market: an Agent-Based Approach. University of Essex. Technical Reports, Colchester (2013)
6. Arthur, B., Holland, J., LeBaron, B., Palmer, R., Tayler, P.: Asset Pricing Under Endogenous Expectations in an Artificial Stock Market. In: Arthur, W.B., Durlauf, S.N., Lane, D. (eds.) The Economy as an Evolving, Complex System II, pp. 15–44. Addison Wesley, Redwood City (1997)
7. LeBaron, B.: Building the Santa Fe Artificial Stock Market. Physica, 1–19 (2002)
8. Gode, D., Sunder, S.: Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality. Journal of Political Economy 101, 119–137 (1993)
9. Gode, D., Sunder, S.: Lower Bounds for Efficiency of Surplus Extraction in Double Auctions. In: The Double Auction Market: Institutions, Theories, and Evidence (1993)
10. Cliff, D.: Evolution of Market Mechanism Through a Continuous Space of Auction-Types. Technical report, Hewlett-Packard Research Laboratories, Bristol, England (2001)
11. Gjerstad, S., Dickhaut, J.: Price formation in double auctions. Games and Economic Behavior 22, 1–29 (1998)
12. Rust, J., Palmer, R., Miller, J.: A Double Auction Market for Computerized Traders, Santa Fe Institute, New Mexico (1989)
13. Phelps, S., Marcinkiewicz, M., Parsons, S., McBurney, P.: Using population-based search and evolutionary game theory to acquire better-response strategies for the double-auction market. In: Proceedings of IJCAI 2005 Workshop on Trading Agent Design and Analysis, TADA 2005 (2005)
14. Walia, V., Byde, A., Cliff, D.: Evolving Market Design in Zero-Intelligence Trader Markets. In: IEEE International Conference on E-Commerce (IEEE-CEC03), Newport Beach, CA., USA (2003)
15. Pareto, V.: Cours d'economie politique. Rouge, Lausanne (1897)
16. Matsumoto, M., Nishimura, T.: Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. ACM Trans. on Modeling and Computer Simulation 8(1), 3–30 (1998)
17. Vickrey, W.: Counterspeculation, Auctions, and Competitive Sealed Tenders. Journal of Finance 16(1), 8–37 (1961)
18. Niu, J., Parsons, S.: An Investigation Report on Auction Mechanism Design. Cornell University Library, arXiv:0904.1258 (2013)
19. Hurwitz, L.: On Informationally Decentralized Systems. In: Decision and Organization, pp. 297–336. Univ. of Minnesota Press, Minnesota (1986)
20. Smith, V.L.: Microeconomic Systems as an Experimental Science. The American Economic Review 72(5), 923–955 (1982)
21. Scalas, E.: Mixtures of Compound Poisson Processes as models of tick-by-tick financial data. Chaos, Solitons & Fractals 34, 33–40 (2007)

Knowledge-Based e-Contract Negotiation among Agents Using Semantic Web Technologies

Kalliopi Kravari, Christos Papavasileiou, and Nick Bassiliades

Dep. of Informatics, Aristotle University of Thessaloniki, GR-54124, Thessaloniki, Greece
{kkravari,cpapavas,nbassili}ATcsd.auth.gr

Abstract. E-Commerce enabled new ways of transactions. Companies and individuals negotiate and make contracts every day. Practically, contracts are agreements between parties that must be kept. These agreements affect the involved parties irretrievably. Hence, negotiating them efficiently is proved vital. To this end we propose the use of intelligent agents, which benefit from Semantic Web technologies, such as RDF and RuleML, for data and policy exchanges. Each agent encounter is characterized by the interaction or negotiation protocol and each party's strategy. This study defines a knowledge-based negotiation procedure where protocols and strategies are separated enabling reusability and thus enabling agent participation in interaction processes without the need of reprogramming. In addition, we present the integration of this methodology into a multi-agent knowledge-based framework and next a use case scenario using the contract net protocol that demonstrates the added value of the approach.

Keywords: Semantic Web, Agents, e-Contract negotiation, Reaction RuleML.

1 Introduction

The massive growth of e-Commerce [13] is indisputable, being clear that it will continue to grow. However, e-Commerce is rather complicated, since the parties involved have to collect information, negotiate and safely execute transactions. Although companies and individuals interact every day, they face difficulties in reaching agreements; namely contracts that create relations and obligations that must be kept. Negotiating efficiently such a contract is important since the decisions made during negotiation affects the parties irretrievably. To this end we propose the use of Intelligent Agents (IAs) [8]. IAs benefit from Semantic Web (SW) technologies [3], performing unsupervised complex actions on behalf of their users, reflecting their specific needs and preferences. While SW full vision may be a bit distant, there are already capabilities that can make software more interoperable and cheaper to maintain. For instance, the use of SW technologies, such as RDF and RuleML, for data and policy exchanges maximizes interoperability among parties. Hence, as IAs are gradually enriched with SW technologies their use is increasing.

Each agent is able to manage a private policy (or strategy), a set of rules representing requirements, obligations and restrictions, and personal data that meet its

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 215–224, 2013.

user's interests. Sophisticated tasks, such as negotiation and brokering services, are already carried out efficiently by IAs. On the other hand, each transaction among parties is strictly specified by an interaction protocol, a set of rules that specify among others guidelines and restrictions on the parties involved. Hence, using IAs could be the answer for a flexible but efficient (contract) agreement procedure management. However, in order to reach SW agents' maximum efficiency, a well–formed modeling framework is needed. It should be re-usable, easily comprehensible by the user (promoting agent usage) and easily analyzable by the agent (promoting automatization).

Usually both private negotiation strategies and public interaction protocols are jointly hard-coded in each agent. Undoubtedly, it is a common and convenient practice, however it is inflexible. Yet, separating policies (personal strategies) from protocols is imperative. Each agent's policy is private and any disclosure of it could lead to incalculable loss. On the other hand, each protocol should be a common resource since agents must comply with the same interaction protocol in order to interact. Hence, this study attempts to define the necessary requirements and procedures that will let agents interact without the need of reprogramming. The proposed knowledge-based approach enables agents to choose the appropriate protocol (e.g. library of re-usable protocols) and combine it with their personal strategy by using SW technology.

Hence, this article proposes the use of SW languages for expressing both protocol and strategy, in addition to separating them. Separating strategy from the protocol and automatically combining them will let agents to modify their behavior while remaining compliant to the protocol with no extra programming cost. Hence, a fully automated procedure for agent transactions (here in contracts) is presented. Nevertheless, an appropriate framework providing enough compliance with the proposed SW technologies should be used, thus, an integration of the above methodology into EMERALD [11], a knowledge-based MAS, is presented. The rest of the paper is structured as follows: Section 2 gives an overview of the approach, Section 3 briefly overviews EMERALD, while Section 4 illustrates a Contract Net use case, which better displays the potential of the approach. The paper is concluded with references to related work, conclusions and directions for future improvements.

2 Overview

E-Contracts are the most common procedures in everyday life. The main differentiation is that they are modeled, specified and executed by a software system, overcoming the delays and drawbacks of the manual process. Hence, here, we study contract protocols and more specifically the FIPA Contract Net Interaction Protocol.

2.1 FIPA Contract Net Interaction Protocol

Although, contract interaction protocols are acknowledged as vital, their modeling is a really challenging task. Yet, several domain-depended interaction protocols have already been developed. The Contract Net Protocol (CNET), for instance, is a noteworthy and probably the most widely used protocol, firstly introduced by Smith [15].

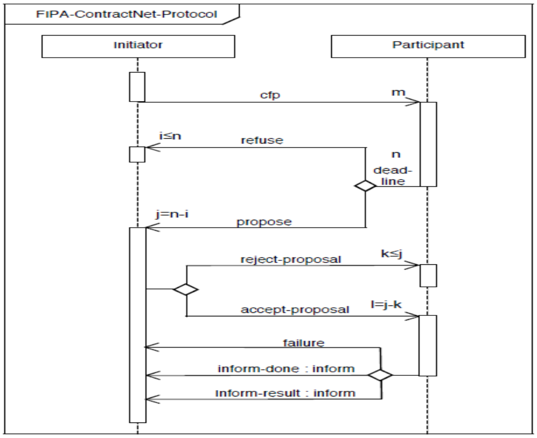


Fig. 1. FIPA Contract Net Interaction Protocol

In CNET negotiation is considered as a two-way communication in which an agent evaluates the offer of assigning a contract or receiving one from its own perspective. Since then, much research has been done, successfully resolving important issues in the field. Although CNET was proved valuable in a variety of situations, it had to be modified in order to reflect changes in agent technology. In this direction, FIPA, an IEEE Computer Society standards organization, provides among others the FIPA Contract Net Interaction Protocol [16]. This standardized protocol has been used over time as the basis for a variety of cases. According to the FIPA specification (Fig. 1) in the contract net interaction protocol, one agent (the Initiator) takes the role of manager which wishes to have some task performed by one or more other agents (the Participants) and further wishes to optimize a function that characterizes the task. This characteristic could be for instance the minimum price. For a given task, the Initiator has to send a call for proposal message communicating its request. Next, any number of the Participants may respond positively; the rest must refuse. Negotiations then continue with the Participants that accepted the call. A positive response however is not a strict acceptance but rather a counter proposal. Hence, the Initiator has to evaluate the offers and ignore the refusals. Finally, it has to accept the best offer by sending back an acceptance messages whereas reject messages should be send to the rest.

2.2 Separating Interaction Protocol from Agent Strategy

However, having an appropriate protocol is not enough, flexibility and reusability is also needed, which actually can be obtained by separating each agent's private policy from the protocol. To this end, in the proposed approach two main rulesets were defined. The first one is related to the Strategy (agent's personal policy), while the second one is related to the Protocol. The Strategy ruleset defines the agent's personal preferences whereas the Protocol ruleset mainly orchestrates message exchange among agents. It is obvious that the message exchange, defined by the protocol, is the key for a successful transaction. Hence, providing appropriate message structures is vital. To this end, Reaction RuleML was chosen for expressing both the protocol and the strategy rules [4], [7]. This rule language was chosen for two major reasons. Firstly, it is flexible in rule representation and secondly its syntax supports a message structure that can include all message modules provided by the FIPA specifications. The message structures in RuleML (Fig. 2) and FIPA are similar since they both contain predicates for *Sender*, *Receiver*, *Content (Payload)*, *Protocol* and C*onversation Identifier* (called *conversation-id* in FIPA and *oid* in RuleML).

```
<Message mode="outbound" directive="CFP″
  <oid> <!-- conversation ID--> </oid>
  <protocol> <!-- transport protocol --> </protocol>
  …
</Message>
```

Fig. 2. (Call-For-Proposal) Message structure in Reaction-RuleML syntax

Using appropriately the above message structures, agents are able to exchange from simple facts to rulebases (sets of rules). Even more useful is the fact that due to the conversation-id module, agents will be also able to get involved in longwinded and usually asynchronous communications, and thus being flexible. In this context, Reaction RuleML is expedient since it provides specific predicates (`SendMsg` and `rcvMsg`) which are appropriate for message exchange in any agent interaction. However, having both protocol and strategy rules expressed in Reaction RuleML is not enough. Agents' final behavior, the combination of protocol and strategy, should be executed in a compact way that will let them represent both their environment knowledge and their behavior patterns quite easily. To this end, the JESS execution engine and language was chosen [9]. JESS is considered as a very expressive language that can express complex logical relationships with very little code, while it is commonly used by agent programmers. Additionally, it is also FIPA compliant.

2.3 Combining Protocol with Strategy

Practically, there is a public repository where protocols are stored with their rulesets expressed in Reaction RuleML. Additionally, there are private repositories, one or more, for each agent where their personal strategy rulesets are stored, also expressed in Reaction RuleML. For instance, assume that there is a number of agents (e.g. four) interacting according to the FIPA Contract Net Interaction Protocol. One of them should be the initiator of the process and the rest (participants) will respond according to their personal intentions. Each of them will retrieve the protocol ruleset from the appropriate repository (Fig. 3), which is common for all of them, whereas they have their private repository for their strategies. In other words, the protocol will define how they will communicate whereas their strategy will define what to share. What is important here is the fission of the protocol rules to initiator and participant rules. This is essential because although there is a single protocol, it defines rules for both roles (initiator and participant) since each of them have to act from another perspective. How these sets of rules should be combined and executed is an important issue.

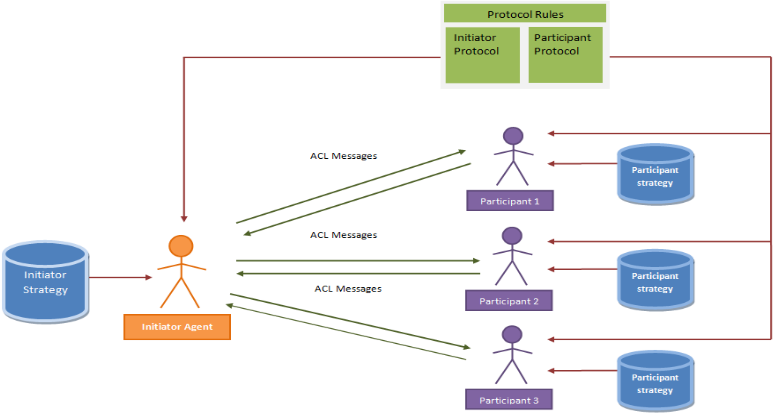


Fig. 3. Procedure overview

Fig. 4 presents the process; domain depended XSLT transformations will use both RuleML sets of rules (for protocol and strategy) and transform them to an executable ruleset (e.g. in Jess) which is fused inside each agent in order to participate in the transaction effectively. Notice that using a high-level, declarative description of the protocol and strategy no extra programming cost is needed by the agent owners.

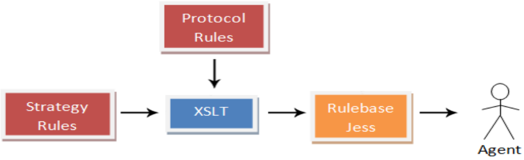


Fig. 4. Strategy – Protocol Combination

3 EMERALD

EMERALD (Fig. 5) is a multi-agent knowledge-based framework, based on SW and FIPA standards, that enables reusability and interoperability of behavior between agents. It is built on JADE [2], a reliable and widely used framework. EMERALD supported so far the implementation of various applications, like brokering and agent negotiations. This framework, in order to model and manage the parties involved in an e-Contract negotiation procedure, provides a generic, reusable agent prototype for knowledge-customizable agents (KC-Agents), consisted of an agent model (KC Model), a directory service (Advanced Yellow Pages Service) and several external Java methods (Basic Java Library). Agents that comply with this prototype are equipped with a Jess rule engine [9] and a knowledge base (KB) that contains environment knowledge (facts), behavior patterns and strategies (Jess production rules). Additionally, since trust has been recognized as a key issue in SW MASs, EMERALD adopts a variety of reputation mechanisms, both decentralized and centralized. In this study, the Jess KB is actually the agent’s data and rules that

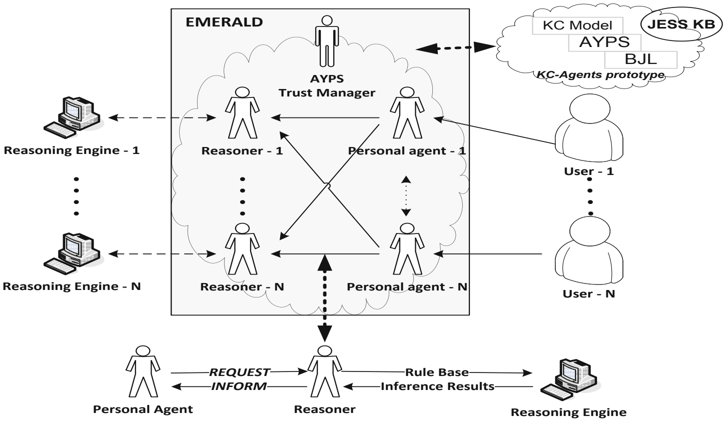


Fig. 5. EMERALD abstract architecture

comprise its policy and characterize its behavior, as described above. Using the KC-Agents prototype offers certain advantages, such as modularity, reusability, maintainability and interoperability, as opposed to having behavior hard-wired into the agent's code (e.g. in Java).

Hence, since agents do not necessarily share a common rule or logic formalism it is vital for them to find a way to exchange their position arguments seamlessly. To this end, EMERALD proposes the use of Reasoners [12], which are agents that offer reasoning services to the rest of the agent community. This approach does not rely on translation between rule formalisms, but on exchanging the results of the reasoning process of the rule base over the input data. Currently, EMERALD implements a number of Reasoners that offer services in two major reasoning paradigms: deductive rules and defeasible logic. Following the above specifications EMERALD commits to SW and FIPA standards, namely, it uses among others the RuleML language [4] since it has become a de facto standard. Additionally, it uses the RDF model [14] for data representation both for the agents' private data and the reasoning results generated during the process, as used in contract agreement interactions presented in [10].

3.1 Extending KC-Agents Prototype

EMERALD's KC-Agents prototype was extended in order to adapt to the new requirements; namely the initial separation of protocol and strategy definitions and the final combination in one Jess rulebase. So far, KC-Agents were limited in receiving one file containing both strategy and protocol; hence it was the programmer's responsibility to merge them. Each time a new behavior was needed, a new file containing both (new) strategy and protocol should be provided. Yet, a new agent model was added in the prototype based on this study. The extended KC-Agents agent derives two separated files one for the protocol and one for the strategy. Hence, whenever protocol or strategy is modified, no extra programming cost is needed. The agent will retrieve the appropriate (new) files from the corresponding repositories. The transformation and merge of them will be executed automatically. Following this approach new behaviors and protocols can be added to the private and public repositories, respectively, for future use. Agents will automatically use them when needed. In this context, the new agent model contains a function that retrieves the appropriate XSLT (from another repository) and uses it in the protocol-strategy fusion process (Fig. 5).

4 Use Case

A use case scenario based on the FIPA Contract Net Interaction Protocol is presented here in order to clarify why protocols and strategies should be separated and how an automated combination procedure will save time and programming effort. It is implemented in EMERALD and involves four parties; an initiator and three participants (Fig. 3) who comply with the new KC-Agents; hence protocol and strategy will combine automatically according to the XSLT transformation procedure. The Initiator agent is interested in the best offer for a laptop. It is aware of three potential e-shops, which are represented by the participant agents. First of all, it has to get the appropriate protocol in RuleML from the public repository. Hence, the Initiator following the protocol sends a CFP (Call-for-proposal) message containing the name of the product and a desired price to them in order to initiate the negotiation procedure. Next, it waits for their response, which could be either positive (PROPOSE) or negative (REFUSE). A positive response however is not a strict acceptance but rather a price proposal. Hence, the Initiator has to evaluate the offers and ignore the refusals. Next, it has to accept one by sending back an ACCEPT message whereas REJECT messages should be sent to the rest. The above rules are defined by the protocol since they refer to message exchange. The decision making however is defined in the private strategies.

The Initiator's strategy for instance determines the agent's main restriction; the price offered by a participant should be lower than the price the Initiator is willing to pay. That price for the Initiator is the firstly indicated price in the CFP (here 300 Euros) plus an amount (here 50 Euros). Hence, if none of the participants fulfill this restriction, the Initiator will reject all of them. On the other hand, each participant has a list (set of facts called Products) that contains their available products accompanied with the product type (e.g. laptop) and its price (e.g. 500 Euros). For instance, such a product is described in JESS syntax as: *(products (type laptop)(price 200))*. A comprehensible flow of rule activation, triggering and execution for this use case is presented in Fig. 7. The same principles can also be applied in other interaction scenarios. Using EMERALD and its KC-Agents prototype we activated the four agents defining two files, one for the protocol and one for the strategy. Following the CNET protocol a straight-forward procedure is performed from the first call to the final acceptance. Here, the transaction was successfully completed when the Initiator chose the participant called *p* by sending an *Accept Proposal* message, as presented in the EMERALD's execution diagram (Fig. 7). Furthermore, participant *p* sent back an *inform* message which according to FIPA specifications is needed in order to verify that the final decision is received.

Finally, it is worth mentioning that Reaction RuleML enables two types of rules, namely production and reactive. Production rules are used for agents' private strategy allowing them to act according to their user's will whereas reactive rules are used for the protocol (Fig. 6) allowing agents to adjust to their partner's behavior based on events related to message exchanges. Below is presented shortly a reactive (protocol) rule example in JESS syntax due to space limitations, where a participant when receives a CFP message, it posts it as a fact in the agent's internal KB. Notice how the rules interact through a predefined set of fact templates that play the role of API (here *callforp*). All RuleML files used are available at http://tinyurl.com/usecase-ruleml.

```
(defrule receive-cfp
 ACLMessage(communicative-act CFP)…(protocol fipa-contract-net))
 =>
(assert (callforp (cfp_content ?c)……(cfp_cid ?cid)))
(modify ?p_start_state (state p_check_state) (state_id ?cid)))
```

Fig. 6. Reactive Rule (Protocol): Participant receives a call

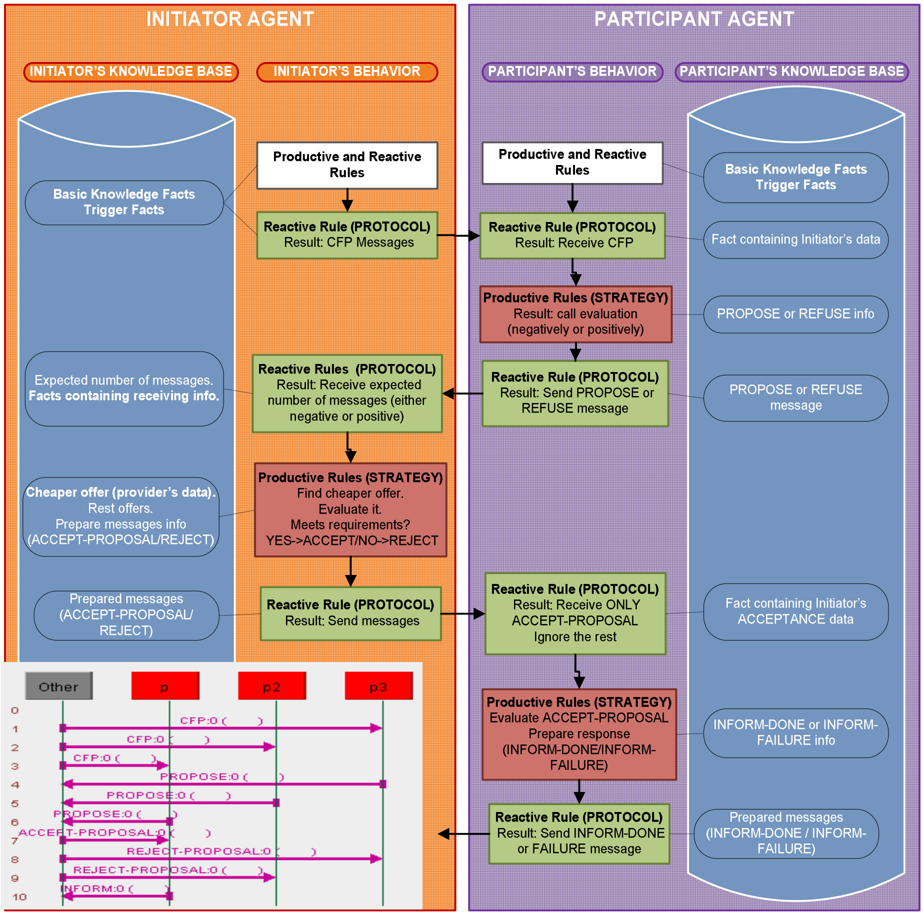

INITIATOR AGENT
INITIATOR'S KNOWLEDGE BASE
INITIATOR'S BEHAVIOR
PARTICIPANT AGENT
PARTICIPANT'S BEHAVIOR
PARTICIPANT'S KNOWLEDGE BASE
Productive and Reactive Rules
Basic Knowledge Facts Trigger Facts
Reactive Rule (PROTOCOL) Result: CFP Messages
Productive and Reactive Rules
Basic Knowledge Facts Trigger Facts
Reactive Rule (PROTOCOL) Result: Receive CFP
Fact containing Initiator's data
Productive Rules (STRATEGY) Result: call evaluation (negatively or positively)
PROPOSE or REFUSE info
Expected number of messages. Facts containing receiving info.
Reactive Rules (PROTOCOL) Result: Receive expected number of messages (either negative or positive)
Reactive Rule (PROTOCOL) Result: Send PROPOSE or REFUSE message
PROPOSE or REFUSE message
Cheaper offer (provider's data). Rest offers. Prepare messages info (ACCEPT-PROPOSAL/REJECT)
Productive Rules (STRATEGY) Find cheaper offer. Evaluate it. Meets requirements? YES->ACCEPT/NO->REJECT
Prepared messages (ACCEPT-PROPOSAL/ REJECT)
Reactive Rule (PROTOCOL) Result: Send messages
Reactive Rule (PROTOCOL) Result: Receive ONLY ACCEPT-PROPOSAL Ignore the rest
Fact containing Initiator's ACCEPTANCE data
Productive Rules (STRATEGY) Evaluate ACCEPT-PROPOSAL Prepare response (INFORM-DONE/INFORM-FAILURE)
INFORM-DONE or INFORM-FAILURE info
Reactive Rule (PROTOCOL) Result: Send INFORM-DONE or FAILURE message
Prepared messages (INFORM-DONE / INFORM-FAILURE)
Other
p
p2
p3
0
1 CFP:0 ( )
2 CFP:0 ( )
3 CFP:0 ( )
4 PROPOSE:0 ( )
5 PROPOSE:0 ( )
6 PROPOSE:0 ( )
7 ACCEPT-PROPOSAL:0 ( )
8 REJECT-PROPOSAL:0 ( )
9 REJECT-PROPOSAL:0 ( )
10 INFORM:0 ( )

Fig. 7. Scenario overview and message exchange in EMERALD

5 Related Work

A work that automates price negotiations in e-commerce transactions using a rule-based implementation is presented in [1]. It is based on JESS and the JADE framework. It concerns only price negotiation rather than a more generic protocol. Our work on the other hand concerns modeling a reusable procedure that could be used not only in this case but also in any other protocol case. Additionally, although both approaches consider IAs and FIPA standards important for maximizing automation and efficiency in users' everyday life only our approach complies with SW standards.

Another related approach is the DR-CONTRACT [6] architecture for representing and reasoning on e-Contracts in defeasible logic. The architecture captures the notions relevant to execution and performance of e-Contracts in defeasible logics. It uses a RuleML extension and RDF/XML syntax for its exported conclusions. Hence, it uses SW standards like RDF and RuleML, similarly to our approach, whereas it is an architecture focused on e-Contract procedures omitting separation of protocols and strategies, which is an important and challenging task for the field.

Concerning interoperability, Rule Responder [5] is quite similar to EMERALD. It builds a service-oriented methodology and a rule-based middleware for interchanging rules in virtual organizations. It demonstrates the interoperation of distributed platform-specific rule execution environments, with Reaction RuleML as a platform-independent rule interchange format. It has a similar view of reasoning service for agents and usage of RuleML but it is not based on FIPA specifications. In other words, it is interested in interoperability, reusability and even protocol-strategy separation, yet it doesn't support IAs but rather web services acting like agents.

6 Conclusions and Future Work

The article argued that e-Commerce met a massive growth over the past years; with e-Contracts like ordinary contracts to be the most common procedures in everyday life. However, they are rather complicated, since involved parties have to collect information, negotiate and execute transactions. We addressed this problem by using intelligent agents acting in the Semantic Web, as agents can perform the same tasks unsupervised, relieving their users of time consuming processes. To this end, this article presented a modular and reusable framework using SW technologies, such as RuleML and RDF. We defined a knowledge-based negotiation procedure where protocols and strategies are separated enabling reusability and thus enabling agent participation in interaction processes without the need of reprogramming. In addition, an integration of this methodology into a multi-agent knowledge-based framework and a use case scenario that demonstrated the added value of the approach was also presented.

As for future direction, our main interest is in extending the proposed framework to model more protocols, such as brokering, price negotiations and auctions. Our final goal is to provide a general-purpose framework for protocol-strategy separation in multi-agent environments, letting agents maximize their autonomy, flexibility and efficiency. Additionally, since agents not necessarily share the same logic or rule representation formalism, our intention is to provide a mechanism for automatic formalism transformation or interpretation.

Acknowledgments. This work is partially supported by the Greek R&D General Secretariat through a bilateral Greek-Romanian project.

References

1. Badica, C., Ganzha, M., Paprzycki, M.L.: Implementing Rule-Based Automated Price Negotiation in an Agent System. J. of Universal Computer Science 13(2), 244–266 (2007)
2. Bellifemine, F., Caire, G., Poggi, A., Rimassa, G.: JADE: A white Paper. EXP in Search of Innovation 3(3), 6–19 (2003)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American Magazine 284(5), 34–43 (2001) (revised 2008)
4. Boley, H., Paschke, A., Shafiq, O.: RuleML 1.0: The Overarching Specification of Web Rules. In: Dean, M., Hall, J., Rotolo, A., Tabet, S. (eds.) RuleML 2010. LNCS, vol. 6403, pp. 162–178. Springer, Heidelberg (2010)
5. Boley, H., Paschke, A.: Rule responder agents framework and instantiations. In: Elçi, A., Koné, M.T., Orgun, M.A. (eds.) Semantic Agent Systems. SCI, vol. 344, pp. 3–23. Springer, Heidelberg (2011)
6. Governatori, G., Hoang, D.P.: A Semantic Web Based Architecture for e-Contracts in Defeasible Logic. In: Adi, A., Stoutenburg, S., Tabet, S. (eds.) RuleML 2005. LNCS, vol. 3791, pp. 145–159. Springer, Heidelberg (2005)
7. Governatori, G., Rotolo, A.: Modelling contracts using RuleML. In: Legal Knowledge and Information Systems. Jurix 2004, pp. 141–150. IOS Press, Amsterdam (2004)
8. Hendler, J.: Agents and the Semantic Web. IEEE Intelligent Systems 16(2), 30–37 (2001)
9. JESS, the Rule Engine for the Java Platform (2008), `http://www.jessrules.com/`
10. Kravari, K., Kastori, G.-E., Bassiliades, N., Governatori, G.: A Contract Agreement Policy-Based Workflow Methodology for Agents Interacting in the Semantic Web. In: Dean, M., Hall, J., Rotolo, A., Tabet, S. (eds.) RuleML 2010. LNCS, vol. 6403, pp. 225–239. Springer, Heidelberg (2010)
11. Kravari, K., Kontopoulos, E., Bassiliades, N.: EMERALD: A multi-agent system for knowledge-based reasoning interoperability in the semantic web. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS, vol. 6040, pp. 173–182. Springer, Heidelberg (2010)
12. Kravari, K., Kontopoulos, E., Bassiliades, N.: Trusted Reasoning Services for Semantic Web Agents. Informatica: Int. J. of Computing and Informatics 34(4), 429–440 (2010)
13. Laudon, K., Traver, C.G.: E-Commerce 2012, 8th edn. Prentice Hall, New Jersey (2012)
14. Resource Description Framework (RDF): Model and Syntax Specification (2004), `http://www.w3.org/TR/PR-rdf-syntax/`
15. Smith, R.G.: The contract net protocol: high level communication and control in a distributed problem solver. IEEE Transactions on Computer 29(12), 1104–1113 (1980)
16. FIPA Communicative Act Library Specification: Fipa Contract Net Interaction Protocol Specification, version H (2003), `http://www.fipa.org/specs/`

Preliminary Experimental Results of the Multi-attribute and Logic-Based Ontology Alignment Method

Marcin Pietranik and Ngoc Thanh Nguyen

Institute of Informatics, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland
{marcin.pietranik,ngoc-thanh.nguyen}@pwr.wroc.pl

Abstract. This paper presents partial results of our novel method of ontology alignment, which is based on processing varying semantics of attributes within concepts. Following our previous publications in which we presented well grounded theoretical foundations of the alignment framework, we have created custom built implementation. This paper is an overview of the preliminary experimental results we ware able to gather. We explain the most important problems we encountered while preparing test methodology, describe used input data and give straight comparison to previously used ontology mapping verification procedures.

1 Introduction and Related Works

Ontologies can be treated as a way of expressing some part of reality ([17]), by conducting its decomposition and providing its semantic description. Their alignment is the task of designating mappings between these knowledge structures which allow convenient migration of their contents. Unlike former approaches that did not incorporate information that can be extracted from attributes of concepts in ontologies, our method is based on analyzing their semantics and the fact that these semantics may change when the same attribute is assigned to different concepts. Former solutions widely discussed in available literature were also entirely tied to OWL standard, which is the most common tool for expressing ontologies in computer readable format. It do not offer any kind of mechanism for expressing this varying meanings of attributes and furthermore, to express attributes in more details than simple *key-value* pairs.

We claim that our framework, which emphasizes explicit attributes' semantics and incorporates it in designating alignments between ontologies is consistent with the intuitive way how people see the real world. For example, consider an attribute *Weight* and its usage within two ontologies. Despite the same label (even if we discard the problem of synonymity) and the fact that in both knowledge structures they describe actual weight of some object, issues related to their straight aligning naturally appear. In one ontology its values are measured with pounds (*lb*) and in the other with kilograms (*kg*). A similar issue occurs when two attributes share the same valuation (for example, finite subset of natural numbers), but differ in the meaning (for example, one attribute refers to age and some other to length). As it will be formally explained in the next section of this paper, assigning logic sentences to the association of concepts and attributes may

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 225–234, 2013.

overcome described difficulties. For example, imagine the attribute *Address* that can be included in both concept *Person* and concept *Webpage*. Obviously, these two assignments give heterogenous meanings to the same attribute which can be easily used to distinguish them while searching for possible mappings.

In our work we have proposed the bottom-up approach to the problem. At first we have created well grounded theoretical definitions of ontologies and their components (with regard to different levels of information granularity that are available within ontologies). Then we have provided a consistent method of aligning ontologies on attribute, concept ([8]) and relation level ([9]). We approach the task of finding alignments between ontologies differently- we do not concentrate on creating a similarity (understood strictly mathematically) measure between concepts, but rather on the degree to which we can align some selected source concept to target concept. In other words - alignment represents the amount of knowledge that can be unequivocally transformed from a source concept to a target one. So far we have created a complete framework of matching ontologies on attribute, concept and relation level. The biggest difficulty that we have encountered was developing solid experimental methodology that could prove the correctness of our ideas.

The most important contributions to the topic of ontology alignment evaluation have been gathered around OAEI campaigns ([15], [16]), which are annually organized events. Their coordinators prepare broad and comprehensive test data that authors of different approaches may use to evaluate their ideas and theories. These data are build around groups of ontologies (concerning different aspects of hardnesses while designating mappings between them) from independent domains along with reference alignments between them, that have been manually composed by experts. Systems that participate in the test provide their results that are collated with reference matches and eventually standard *Precision* and *Recall* measures are calculated ([3]). Despite obvious advantages of such approach to evaluating ontology alignments, we have identified few downsides. First of all, provided benchmarks are built with OWL- that fact does not allow generic approaches to ontology alignment to use them. Secondly, the reference mappings form strict expectations concerning designating mappings. For example, that some concept A in source ontology can be aligned to the concept B in target ontology. If the concept B has a superconcept then obviously concept A can also be aligned to it. If such mapping is not present within reference alignment, then a result including such concept match is treated as an error, therefore lowering precision and recall values. Another difficulty was the fact that provided ontologies do not contain semantics of attributes (due to the fact that OWL do not provide any kind of mechanisms that allow to represent such information), so the essential data that are required by our framework. The reason why ontology alignment should be based on more generic foundations (rather than strictly OWL) are in straight correspondence with vagueness of OWL itself, which is widely discussed in [4]. Final downside are conclusions drawn from the eventual evaluation process' results. As previously mentioned, the described methodology ends with calculating *Precision* and *Recall* values. Commentaries are strictly based on these values, but we have only found one paper ([12]) that analyzed the issue deeper, so not only raw metrics but also reasons why these metrics have such values. In our opinion lack of analysis of wrong alignments (which relate to *Precision* measure) that have been designated by evaluated alignment systems and not found alignments (which relate to *Recall* measure) are indispensable for a more complete and reliable analysis of alignment method and its results.

Due to the obvious necessity of evaluating our ideas we needed to prepare other methodology that would offer both consistency with our theoretical foundations and remain in relation to OAEI campaigns.

As a prerequisite for our work we had to conduct a careful analysis of former works that had been done in the field of ontology alignment and ontology alignment evaluation. The foundations of the former topic are described in [2] which is the broad description of basic methods and techniques of aligning ontologies. Recent advances in this field has been described in [14]. This work not only overviews latest achievements, but more importantly addresses challenges that have not been yet investigated.

So how to compare our approach to methods that are based on different assumptions and input data? Furthermore, how to unequivocally state which one is better? In this paper we present preliminary results that we have gathered during the conducted experiment. We have decided to randomly pick 30 pairs of ontologies taken from OAEI campaigns and manually assign attributes and their semantics to available concepts. We are aware of possible doubts that the reader may have concerning the quality of prepared test data, but the main aim of our experiment is to show that created framework may be treated not as the straight contradiction to former solutions, but rather as a tuning tool that can be incorporated in order to improve gathered results.

The reminder of the paper is organized as follows. In Section 2, we will briefly describe our framework. Section 3 contains the detailed description of experimental methodology, along with overview of experimental environment used to conduct the test procedure. Section 4 gives a summary and shed some light on upcoming research that we plan to investigate.

2 Basic Notions and Framework Overview

We define ontology as a tuple $O = (C, R, I)$, in which C is the set of concepts, R is a set of relations between them and I is a set of instances ([10], [5]). Based on this notation our approach to ontology alignment can be defined as follows: *Having two ontologies O_1 and O_2, the task of aligning them is determining the set of triples $< c, c', M_C(c, c') >$, in which c and c' are concepts from ontologies O_1 and O_2 and the real value $M_C(c, c')$ is the degree to which concept c can be aligned to concept c'.*

The concept c from set C is defined as $c = (Id^c, A^c, V^c)$ where Id^c is a concept's label, A^c is a set of its attributes and V^c is a set of domains of attributes from A^c. This tuple can be also called *concept's structure*.

By S_A we define the set of atomic descriptions of attributes. L_s^A denotes the formal language, incorporating elements of S_A and basic logic operators $\neg, \vee, \wedge$. Semantics of attributes are defined as a partial function $S_{A,C} : A \times C \rightarrow L_s^A$, where A is the set of all possible attribute taken from the real world defined as a pair *(A, V)* in which V denotes set of valuations of attributes ([5]). Such approach allows to express varying meanings that attributes may obtain while being a part of different concepts- for example, the attribute *name* has different semantics when incorporated in the concept *Person* and concept *File*. Such approach, that provides additional descriptions of attributes gave us the possibility to formulate formal criteria for identifying relationships between attributes ([8]). These correspondences are: *equivalence* ($\equiv$), *generalization* ($\uparrow$) and *contradiction* ($\downarrow$).

Throughout this paper we will utilize the term *source ontology* and *target ontology* to designate knowledge structures that are respectively the source of data that we want to transform and the target format in which we want to express the final information.

Our attribute-based ontology alignment framework is built around three functions $M_A^{c,c'}$, M_A^c and M_C ([8]), that are used to calculate the degree to which we can align two attributes, the degree to which we can align selected attribute from source ontology and the alignment degree of two concepts. The first function is defined as follows:

$$M_A^{c,c'}(a,b) = \begin{cases} 1 & \text{if } a \equiv b \\ 1 & \text{if } a \uparrow b \text{ and not } a \equiv b \\ 1 - d_S(S_A(a,c), S_A(b,c')) & \text{otherwise} \end{cases} \tag{1}$$

We utilize the function d_S which is the distance between two semantics (two logical expressions), discussed in details in our previous work ([10]). The degree to which we can align a particular attribute from source concept to some target concept is given as:

$$M_A^c(a) = \begin{cases} \frac{1}{|Z^*|} \sum_{(a,b)\in Z^*} M_A^{c,c'} & \text{if } |Z^*| > 0 \\ M_A^{c,c'} & \text{if } |Z^*| = 0, \text{for } b = argmax_{b\in A^{c'}} M_A^{c,c'}(a,b) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Where the set Z^* is defined as $Z^* = \{(a,b) : a \in A^c, b \in A^{c'}, b = argmax_{b\in A^{c'}} M_A^{c,c'}(a,b) \wedge M_A^{c,c'}(a,b) \geq T\}$. This function, for given threshold $T \in [0,1]$, identifies *the best match* for given attribute. Having this we were able to develop the eventual definition of concept alignment degree $M_C(c,c')$. The input are two sets A^c and $A^{c'}$- respectively attributes of source concept and target concept. Initially the procedure removes unnecessary redundancy from the source concept's structure A^c (for example, dismissing equivalent attributes) and creating the working set $\overline{A_c}$. Then the formula $M_C(c,c') = \frac{\sum_{a\in\overline{A_c}} M_A^c(a)}{|\overline{A_c}|}$ is used to calculate the end degree.

Having in mind the main purpose of aligning ontologies (which comes down to answering the question about the degree to which we can align particular concept from *Source Ontology* to selected concept from *Target Ontology*) we need to point out the main differences between our approach and former ones. First of all our method is not related to any kind of similarity between concepts. As stated in [8] aforementioned base functions $M_A^{c,c'}$, M_A^c and M_C are not symmetrical, therefore our tool does not designate the closeness of two concepts, but the amount of knowledge that can be unequivocally transformed. Secondly, the foundation of this approach is built on analyzing varying semantics of attributes that describe concepts' structures. This fact mainly differentiates our approach- as can be found in [14] previously developed solutions commonly omitted this layer of information available within ontologies. This issue could be the cause of many difficulties (further presented in 3) concerning designating reliable mappings between concepts that are closely related, but differently described in terms of their names or relations that connect them. Our method, based entirely on inner structures of concepts utilize only this knowledge and does not need to process any other available data.

Because of the limited space for this paper, we cannot present complete framework in details. For a broader overview (including formal definitions, sets of postulates and comprehensive analysis) please refer to [8].

3 Experimental Results

After developing theoretical foundations we have started to work on experimental environment. On early stages of designing alignment system we have decided to implement it as a web application due to our experience in creating such solutions. Therefore, the system is written with Django web framework which is the set of tools created in Python programming language. For presentation purposes we have used HTML5/CSS3 elements and jQuery JavaScript framework. As a persistence layer we have decided to utilize MongoDB- the document oriented, schemaless database. The main reason for such choice is its open definition and possibility to store any kind of structured or unstructured data. Unlike in relational databases that force to unify stored data in fixed tables, MongoDB does not require any kind of schema. The base assumption is grouping loosely coupled data in collections, that do not force its elements to have anything in common in terms of their structures. These collections can be treated as categorization of elements present within the database. In other words, grouping them into semantically related objects (for example, collection of business cards may contain elements that have completely different structures, but they all contain some kind of contact information). Moreover, in MongoDB it is possible to dynamically change the structure of stored data during the runtime of application without any kind of modifications in the core system. In our opinion, such open approach to structuring data is well suited for creating a tool for storing and managing ontologies.

The main aim of creating the experimental environment was conducting practical verification of our method. The reason for that is the need to provide solid proof of usefulness of our approach to ontology alignment. As stated in earlier sections of this paper, our method cannot be strictly compared with solutions described in available literature. Due to the fact that the foundation of our framework is analyzing attributes and concepts structure (so the layer omitted by previous alignment tools) and furtherthe fact that our method does not imply similarities among concepts, but the degree to which they can be mapped, the strict statement that our method is better or worse is pointless.

As described in Section 1, available evaluation procedures concentrated on providing broad and complex test data with which assessed tools where confronted in terms of comparing produced mappings with reference alignments and eventual calculation of *Precision* and *Recall* measures. Such approach to the alignment evaluation assumes that used reference alignments contain all of the correct matches between ontologies and any other mappings produced by alignment solution are not valid. Such assumption is very straightforward, easy to understand and apply, but it contains hidden pitfall. For example, imagine that in the *Source Ontology* there is a concept *Article* and within *Target Ontology* there are concepts *SingleAuthorWork* and *MultiAuthorWork*. Let us also assume that within *Source Ontology* there is no knowledge concerning status of publication of articles. Obviously, the majority of information expressed in the concept *Article* may be transformed into both concepts from *Target Ontology*, but in the reference alignment (provided to test alignment methods) there is only one pair of matching concepts (e.g. *Article-SingleAuthorWork*). This assumption rejects any other mapping provided by alignment solution, therefore discarding bigger picture of ontology alignment, which is providing methods of transforming the context of source ontology into the context of target ontology. In the described situation, a possible user that is only interested in bibliographical data about certain publications, after sending such query to the knowledge base utilizing *Target Ontology* will not be presented with the complete information extracted from *Source Ontology*, but only some fragment of it. In our work we wanted to provide a solid methodology that would avoid such issues, therefore on early stages of our work we have accepted the fact that we won't be able to use formerly used experimental data.

Unfortunately, the necessity of proving the correctness of our ideas still exists. In order to unequivocally state that our framework may be useful, as previously mentioned, we have gathered and analyzed 30 pairs of ontologies taken from OAEI campaigns (from years 2009 and 2010) with their reference alignments and alignments designated by former solutions. These data are available online as OWL and plain XML files. To process them we have used Python XML parsing module and RDFlib. We have implemented a tool that processes obtained files and generates histograms of most common wrong alignments and most common not found alignments designated by systems that participated in selected campaigns.

Eventually we were able to directly input obtained summations, concepts and few attributes that sometimes occur within analyzed ontologies into the core storage unit of our system. Then, we have manually supplemented missing attributes and their semantics utilizing set S_A of atomic descriptions of attributes, which has been prepared beforehand.

Due to limited space for this paper we are not able to present created test ontologies, complete histograms and the results of the experiment that has been conducted. For illustration purposes we have hand picked some explanatory examples of concepts' alignments that has been frequently omitted by systems that participated in OAEI campaigns from selected years. The fourth column of Table 1 contains values of aforementioned histograms of not found mappings for chosen pairs of concepts.

To give an example of how concepts' structures may look like, below we present definitions of selected source and target concepts from first two columns of Table 1:

http://oaei.ontologymatching.org/2009/benchmarks/101/onto.rdf #Reference:

isbn:$id \wedge origin\_country \wedge author\_name \wedge number\_of\_edition \wedge publisher \wedge title \wedge short\_title \wedge checksum \wedge date\_of\_release \wedge is\_book$

language:$edition\_language$

series:$series\_title \wedge publisher \wedge topic\_of\_interest \wedge series\_number$

abstract:$review \wedge keywords \wedge topic\_of\_interest$

number:$number\_of\_edition$

reviewed:$quality$

volume:$journal\_number \vee magazine\_number$

Table 1. Commonly not found alignments in OAEI 2009 and 2010 campaigns

Source Concept	Target Concept	Benchmark name	Number of omissions	Alignment degree
Reference	Sqdsq	2009 254-4	14	0.941
Conference	Zqedzbx	2009 259-2	12	0.881
MotionPicture	Dscdscg	2010 248	12	0.700
Booklet	Brochure	2010 209	11	0.750
Unpublished	Zeadza	2009 202-8	9	0.906
Place	Location	2010 edas ekaw	8	0.813
ConferenceChair	GenaralChair	2010 edas sigkdd	8	0.875
PaperFullVersion	RegularPaper	2009 cmt ekaw	7	0.875
Sponsor	Sponzor	2010 iasted sigkdd	7	0.833

affiliation:$author\_origin \wedge (school\_name \vee employer\_name)$

mrNumber:$id \wedge quality \wedge short\_title \wedge abstract \wedge keywords$

keywords:$keywords \wedge topic\_of\_interest$

annote:$annotation \wedge bibliography\_entry$

size:$page\_count \wedge width \wedge height$

date:$day\_of\_release \wedge month\_of\_release \wedge year\_of\_release$

humanCreator:$author\_name$

copyright:$author\_name \wedge author\_claim$

title:$title \wedge short\_title$

issn:$id \wedge origin\_country \wedge number\_of\_edition \wedge publisher \wedge title \wedge checksum \wedge date\_of\_release \wedge is\_journal$

note:$commentary$

lccn:$id \wedge origin\_country \wedge number\_of\_edition \wedge publisher \wedge title \wedge checksum \wedge date\_of\_release \wedge released\_in\_usa$

contents:$keywords \wedge short\_title \wedge topic\_of\_interest$

http://oaei.ontologymatching.org/2009/benchmarks/101/onto.rdf #Conference:

name:$conference\_name \wedge conference\_abbrevation \wedge year\_of\_event$

shortName:$conference\_abbrevation$

issue:$((\neg journal\_number \wedge \neg series\_number \wedge number\_of\_edition) \vee (\neg journal\_number \wedge series\_number \wedge \neg number\_of\_edition) \vee (journal\_number \wedge \neg series\_number \wedge \neg number\_of\_edition)) \wedge publisher$

organizer:$organizer\_name$

location:$place \wedge venue$

http://oaei.ontologymatching.org/2010/benchmarks/101/onto.rdf #MotionPicture:

author:$director\_name$

title:$movie\_title$

cast:$actors\_set$

copyrights:$director\_name \wedge author\_claim$

date_of_release:$date\_of\_release$

music:$composer\_name$

http://oaei.ontologymatching.org/2010/benchmarks/101/onto.rdf#Booklet:

size:$page\_count \wedge width \wedge height$
rights:$author\_name \wedge author\_claim$
title:$title \wedge short\_title$
short_title: $short\_title$
author_name:$author\_name$
self_publisher: $\neg publisher \wedge \neg funder$
http://oaei.ontologymatching.org/2009/benchmarks/101/onto.rdf #Unpublished:
note:$commentary$
author:$author\_name$
title:$title \wedge short\_title$
availbility:$\neg publisher \wedge \neg editor\_name \wedge \neg is\_reviewed$
http://iasted #Sponsor:
status:$is\_funder \wedge (is\_bronze\_sponsor \vee is\_silver\_sponsor \vee is\_gold\_sponsor)$
tin:$tin\_number \wedge is\_company$
name:$company\_name$
http://oaei.ontologymatching.org/2009/benchmarks/254-4/onto.rdf#sqdsq :
isbn:$id \wedge origin\_country \wedge author\_name \wedge number\_of\_edition \wedge publisher \wedge title \wedge short\_title \wedge checksum \wedge date\_of\_release \wedge is\_book$
senum:$series\_title \wedge publisher \wedge topic\_of\_interest \wedge series\_number$
vol:$journal\_number \vee magazine\_number$
abstract:$review \wedge keywords \wedge topic\_of\_interest \wedge short\_content$
page_count:$page\_count$
bib:$bibliography\_entry$
pub:$publisher \wedge price$
mon:$month\_of\_release$
yea:$year\_of\_release$
number_of_edition:$number\_of\_edition$
qos:$quality$
issn:$id \wedge origin\_country \wedge number\_of\_edition \wedge publisher \wedge title \wedge checksum \wedge date\_of\_release \wedge is\_journal$
lccn:$id \wedge origin\_country \wedge number\_of\_edition \wedge publisher \wedge title \wedge checksum \wedge date\_of\_release \wedge released\_in\_usa$
writer:$author\_name \wedge edition\_language$
afiliations:$author\_origin$
day:$day\_of\_release$
keys:$keywords \wedge topic\_of\_interest$
http://oaei.ontologymatching.org/2010/benchmarks/248/onto.rdf#dscdscg:
vbxzvcxuyhgsrafed:$director\_name$
xvxcvsfv:$director\_name \wedge author\_claim$
xbvsdwerqerqwrt:$title \wedge short\_title$
zxcvzxv:$date\_of\_release$
xvbxzvb:$duration\_time$

The last column of Table 1 contains concepts' alignment degree that have been collected during the experiment. Our system gives clear results that can be used to designate reliable alignments between two ontologies by analyzing logic sentences assigned to associations of attributes and concepts that may include both positive and negative knowledge (so nonnegated and negated literals from set S_A). Proposed method returns values from the range [0,1] and it is able to clearly distinguish concepts that differ in their structures. What is worth mentioning is the fact that it can be used not only to select matching concepts (by confronting calculated degrees with some accepted threshold), but also to select parts of source ontologies that can be used to fill lacking pieces of target ontology. This is the biggest difference between our method and former solutions. It is dictated by the fact that the function that calculates the degree to which we can align source concept to target concept is not symmetrical, therefore it is not any kind of similarity or even metric.

In our opinion, the proposed method is closer to the intuitive way of finding matching concepts which would be used by experts designating matches. Even with semantics that are straightforward and not complicated, obtained results are good enough to claim that our method can be useful when two semantic descriptions must be matched. What is interesting is the fact that our methodology can be adapted to any kind of application that requires such matching procedure, even if it is not straightly related to the topic of ontologies.

On the concept level and when whole ontologies are going to be aligned, our method has quadratic computational complexity, referring to sizes of concepts' sets from the target and the source ontology. More importantly, processing only attributes from concepts is the only real requirement and there is no necessity to analyze any kind of knowledge that may be extracted from ontologies (such as concepts' taxonomy, their relations etc.). This makes it well suited for situations where only selected parts of ontologies need to be matched.

Due to explained differences, it cannot be strictly compared with former works, it should rather be treated as their extension or a method of adjusting already designated mappings. Conducted test procedure that has been only partially described in this paper allowed us to perform robust statistical analysis based on well established methods of experimental outcomes' interpretation taken from the field of Information Retrieval [11]. Obtained results prove the correctness and effectiveness of proposed ontology alignment framework and we will present gathered conclusions in our upcoming publications.

4 Future Works and Summary

Preliminary results presented in this paper unequivocally shows that our framework gives promising results. It is able to produce unambiguous alignments between two heterogeneous ontologies and it appears to be well suited for filtering initial set of designated mappings. We claim that it can be developed into the mature framework capable of not only finding mappings between ontologies, but also be considered as a solid theoretical and practical foundation of any ontology management application. In the future we plan to further work on implemented environment. We want to create a flexible API to provide convenient method of integrating it into any kind of application that requires matching semantic descriptions of any kind of objects (e.g. schemas of federated data warehouses). We will also address the issue of aligning ontologies on the instance level.

Acknowledgement. This research project was supported by grant No. 2011/01/N/ST6/01438 from the National Science Centre, Poland.

References

1. Danilowicz, C., Nguyen, N.T.: Consensus-based partitions in the space of ordered partitions. Pattern Recognition 21(3), 269–273 (1988)
2. Euzenat, J., Shvaiko, P.: Ontology Matching, 1st edn. Springer, Heidelberg (2007)
3. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology Alignment Evaluation Initiative: Six Years of Experience. In: Spaccapietra, S. (ed.) Journal on Data Semantics XV. LNCS, vol. 6720, pp. 158–192. Springer, Heidelberg (2011)
4. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: OWL 2: The next step for OWL. Web Semantics: Science, Services and Agents on the World Wide Web 6, 309–322 (2008)
5. Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management (Advanced Information and Knowledge Processing). Springer (2008)
6. Nguyen, N.T.: Processing Inconsistency of Knowledge in Determining Knowledge of a Collective. Cybernetics and Systems 40(8), 670–688 (2009)
7. Meilicke, C., García-Castro, R., Freitas, F., van Hage, W.R., Montiel-Ponsoda, E., Ribeiro de Azevedo, R., Stuckenschmidt, H., Sváb-Zamazal, O., Svátek, V., Tamilin, A., Trojahn, C., Wang, S.: MultiFarm: A benchmark for multilingual ontology matching. Web Semantics: Science, Services and Agents on the World Wide Web 15, 62–68 (2012)
8. Pietranik, M., Nguyen, N.T.: A Method for Ontology Alignment Based on Semantics of Attributes. Cybernetics and Systems Cybernetics and Systems 43(4), 319–339 (2012)
9. Pietranik, M., Nguyen, N.T.: Ontology Relation Alignment Based on Attribute Semantics. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part II. LNCS, vol. 7654, pp. 49–58. Springer, Heidelberg (2012)
10. Pietranik, M., Nguyen, N.T.: Semantic Distance Measure between Ontology Concept's Attributes. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part I. LNCS, vol. 6881, pp. 210–219. Springer, Heidelberg (2011)
11. Van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworth-Heinemann, Newton (1979)
12. Sabou, M., d'Aquin, M., Motta, E.: Exploring the semantic web as background knowledge for ontology matching. In: Spaccapietra, S., Pan, J.Z., Thiran, P., Halpin, T., Staab, S., Svatek, V., Shvaiko, P., Roddick, J. (eds.) Journal on Data Semantics XI. LNCS, vol. 5383, pp. 156–190. Springer, Heidelberg (2008)
13. Scharffe, F., Euzenat, J.: Linked Data Meets Ontology Matching - Enhancing Data Linking through Ontology Alignments. In: KEOD 2011, pp. 279–284 (2011)
14. Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEEE Trans. Knowl. Data Eng. 25(1), 158–176 (2013)
15. Shvaiko, P., Euzenat, J., Giunchiglia, F., Stuckenschmidt, H., Fridman Noy, N., Rosenthal, A. (eds.): Proceedings of the 4th International Workshop on Ontology Matching (OM 2009), collocated with the 8th International Semantic Web Conference (ISWC 2009), Chantilly, USA, October 25. CEUR Workshop Proceedings 551, CEUR-WS.org (2009)
16. Shvaiko, P., Euzenat, J., Giunchiglia, F., Stuckenschmidt, H., Mao, M., Cruz, I.: Proceedings of the 5th International Workshop on Ontology Matching (2010), http://ceur-ws.org/Vol-689/
17. Staab, S., Studer, R.: Handbook on Ontologies, 2nd edn., XIX, 811 p. 121 illus., Hardcover. Springer (2009) ISBN: 978-3-540-70999-2

Intuitionistic Fuzzy Control Based on Association Rules

Ion Iancu, Mihai Gabroveanu, and Mirel Cosulschi

Department of Computer Science, University of Craiova, 13 A.I. Cuza Street,
Craiova, 200585, Romania
i_iancu@yahoo.com, {mihaiug,mirelc}@central.ucv.ro

Abstract. The task of the standard Mamdani fuzzy logic controller is to find a crisp control action from the fuzzy rule-base and from a set of crisp inputs. In this paper we modify this controller in order to work with intuitionistic fuzzy sets and to activate a set of rules having the same conclusion. Usually, the inference rules used in a fuzzy logic controller are given by a domain expert; in our system, these rules are automatically induced as fuzzy association rules starting from a training set. The fuzzy confidence value associated with each rule is used to obtain the fuzzy set inferred by our system.

Keywords: fuzzy logic controller, intuitionistic fuzzy sets, t-norm, implication, association rules.

1 Introduction

The knowledge base of a rule-based system may contain imprecisions which appear in the description of the rules given by the expert. Because such an inference can not be made by the methods which use multi-valued logic, Zadeh gave a theory of approximate reasoning [16] that is the deduction of imprecise conclusions from a set of imprecise premises. The theory of approximate reasoning is based on fuzzy logic inference processes. An important part of fuzzy reasoning is represented by Fuzzy Logic Control (FLC), which is very useful when the needed models are not known or when they are too complex for analysis with conventional quantitative techniques. A standard FLC system consists of four major parts: Fuzzification interface, Fuzzy rule-base, Fuzzy inference machine and Defuzzification interface.

In this paper we modify the standard Mamdani controller in order to work with intuitionistic fuzzy sets which represent a natural generalization of usual fuzzy sets; such a set is characterized by two functions representing the degree of membership and the degree of non-membership and is a better model for representing the imperfect knowledge. In this way, we obtain an Intuitionistic Fuzzy Logic Controller (IFLC). Usually, the inference rules used by a FLC are provided by a domain expert. In our case, these rules are induced as fuzzy association rules using a training set and the system activates only a set of rules having the same conclusion (or consequence).

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 235–244, 2013.

The proposed IFLC system works as follows:

1. generates, from training data, the rules used by inference engine;
2. computes the firing level of each rule, corresponding to input data;
3. computes the matching value of the input data with each set of rules having the same conclusion;
4. selects the best set of rules used in the inference process;
5. computes the fuzzy set that represents the conclusion inferred and its corresponding crisp value obtained by defuzzification process.

The rest of paper is organized as follows: Section 2 presents the basic concepts with references to our approach. In Section 3 it is described the proposed IFLC system. The Section 4 contains an example about presented system and the last section discusses conclusions and future works.

2 Basic Concepts

2.1 Intuitionistic Fuzzy Sets

The knowledge base of a rule-based system may contain imperfect information which is inherent in the description of the rules given by the expert. This information may be incomplete, imprecise, fragmentary, not fully reliable, vague, contradictory or deficient in some other ways. In these cases, some difficulties appear: the difficulty of representing the deduction rules expressed, generally, by means of natural language and the difficulty of utilization of these rules when the observed facts do not match the condition expressed in the premise of the rule, but are not too different from them. Nowadays, we can handle much of this information's imperfection using fuzzy logic ([6]). A membership function of a classical fuzzy set assigns to each element from the universe of discourse a number from the unit interval to indicate the membership degree to the set under consideration. The non-membership degree is just natural defined as the complement to 1 of the membership degree. However, a human being who expresses the membership degree of a given element in a fuzzy set very often does not express corresponding degree of non membership as the complement to 1. This reflects a well known psychological fact that the linguistic negation not always identifies with logical negation. Thus Atanassov ([3]) introduced the concept of an intuitionistic fuzzy set characterized by two functions expressing the degree of belongingness and the degree of non belongingness, respectively. Atanassov's intuitionistic fuzzy sets (see [3], [4]), can be viewed as a tool that may help better model imperfect information, especially under imperfectly defined facts and imprecise knowledge.

Let X denote a universe of discourse. Then an intuitionistic fuzzy set A in X (see [3], [4]) is a set of ordered triples

$$A = \{(x, \mu_A(x), \nu_A(x)) | x \in X\}$$

where $\mu_A, \nu_A : X \to [0, 1]$ satisfy $0 \leq \mu_A(x) + \nu_A(x) \leq 1, \forall x \in X$. For each x the numbers $\mu_A(x)$ and $\nu_A(x)$ represent the degree of membership and the degree of non-membership of the element $x \in X$ to $A \subset X$, respectively. For each element $x \in X$ we can compute the, so-called, *intuitionistic fuzzy index* (or *hesitation* [14]) of x in A defined as follows

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x).$$

Of course, a fuzzy set is a particular case of the intuitionistic fuzzy set with $\nu_A(x) = 1 - \mu_A(x)$. The working with intuitionistic fuzzy sets instead of fuzzy sets imply the adding of another degree of freedom (ν_A or π_A) to μ_A. The intuitionistic fuzzy sets offer a new possibility to represent imperfect knowledge and, therefore, to describe in a more adequate way many real problems. Such problems appear when we face with human opinions involving two or more answer of the type ([13]): "Yes", "Not", "I do not know", "I am not sure", etc.

Voting can be a good example of such a situation, as human voters may be divided into three groups of those who: "Vote for", "Vote against", "Abstain or Give invalid votes".

This third group is of great interest (from the point of view of customer behavior analysis, for instance) because people from this undecided group after proper enhancement (e. g. different marketing activities) can finally become sure, i. e. become persons voting for or against (customers wishing to buy products advertised).

The intersection and the union of two intuitionistic fuzzy sets in X,

$$A = \{(x, \mu_A(x), \nu_A(x) | x \in X)\} \text{ and } B = \{(x, \mu_B(x), \nu_B(x) | x \in X)\},$$

are given by

$$(A \cap B)(x) = \{(x, min(\mu_A(x), \mu_B(x)), max(\nu_A(x), \nu_B(x)))\}$$

and

$$(A \cup B)(x) = \{(x, max(\mu_A(x), \mu_B(x)), min(\nu_A(x), \nu_B(x)))\}.$$

The majority of applications operate with intuitionistic fuzzy numbers; we choose to work with trapezoidal intuitionistic fuzzy numbers.

Definition 1. *A trapezoidal intuitionistic fuzzy number A with parameters $b_1 \leq a_1 \leq b_2 \leq a_2 \leq a_3 \leq b_3 \leq a_4 \leq b_4$ and denoted by*

$$A = (b_1, a_1, b_2, a_2, a_3, b_3, a_4, b_4)$$

is given such as:

$$\mu_A(x) = \begin{cases} 0 \ \ if \ \ x < a_1 \\ \dfrac{x - a_1}{a_2 - a_1} \ \ if \ \ a_1 \leq x \leq a_2 \\ 1 \ \ if \ \ a_2 \leq x \leq a_3 \\ \dfrac{x - a_4}{a_3 - a_4} \ \ if \ \ a_3 \leq x \leq a_4 \\ 0 \ \ for \ \ a_4 < x \end{cases} \qquad \nu_A(x) = \begin{cases} 1 \ \ if \ \ x < b_1 \\ \dfrac{x - b_2}{b_1 - b_2} \ \ if \ \ b_1 \leq x \leq b_2 \\ 0 \ \ if \ \ b_2 \leq x \leq b_3 \\ \dfrac{x - b_3}{b_4 - b_3} \ \ if \ \ b_3 \leq x \leq b_4 \\ 1 \ \ if \ \ b_4 < x \end{cases}$$

The rule

$$if\ X\ is\ A\ then\ Y\ is\ B$$

is represented by its conditional possibility distribution ([17], [16]) $\pi_{Y/X}$:

$$\pi_{Y/X}(v,u) = \mu_A(u) \to \mu_B(v),\ \forall u \in U,\ \forall v \in V$$

where $\to$ is an implication operator, μ_A and μ_B are the membership functions of the fuzzy sets A and B, respectively, U is the domain of X and V is the domain of Y. One of the most important implication is Lukasiewicz implication ([5]), $I_L(x,y) = \min(1 - x + y, 1)$.

Another notion used in this paper is t-norm.

Definition 2. *A function $T : [0,1]^2 \to [0,1]$ is a triangular norm (in short, $t-$norm) iff it is commutative, associative, non-decreasing and $T(x,1) = x,\ \forall x \in [0,1]$.*

2.2 Fuzzy Association Rules

Mining of association rules represents one of the most important task in data mining. An association rule describes an interesting relationship among different attributes. The task of discovering boolean association rules was introduced by Agrawal in [2]. Fuzzy association rules can handle both quantitative and categorical data and are expressed in linguistic terms, which are more natural and understandable for human beings.

The basic problem of finding fuzzy association rules was introduced in [12]. Let $\mathcal{DB} = \{t_1, \ldots, t_n\}$ be a database characterized by a set $\mathcal{I} = \{i_1, \ldots, i_m\}$ of categorical or quantitative attributes (items). For each attribute i_k, $(k = 1, \ldots, m)$, we will consider $n(k)$ associated fuzzy sets. Let $F_{i_k} = \{F^1_{i_k}, \ldots, F^{n(k)}_{i_k}\}$ be the set of all these fuzzy sets. Typically, a domain expert provides the set of fuzzy sets associated with attributes, along with their corresponding membership functions. We denote with $\mu_{F^j_{i_k}}$ the membership functions of fuzzy sets $F^j_{i_k}$.

We call **fuzzy itemset** the tuple $\langle X, F_X \rangle$, where $X \subseteq \mathcal{I}$ is a set of attributes and F_X is a set of fuzzy sets associated with attributes from X.

Definition 3. *A **fuzzy association rule** is an implication of the following form:*

$$if\, X\ is\ F_X\ then\, Y\ is\ F_Y$$

where $X = \{x_1, \ldots, x_p\}$ and $Y = \{y_1, \ldots, y_q\}$ are disjoint subset of $\mathcal{I}$, and $F_X = \{a_1, \ldots, a_p\}$ and $F_Y = \{b_1, \ldots, b_q\}$ are fuzzy sets related to attributes from X and Y respectively. We denote this rule with:

$$\langle X, F_X \rangle \Rightarrow \langle Y, F_Y \rangle$$

An example of a fuzzy association rule is the following:

"IF *Age* is *young* and *Income* is *high* THEN *Cars* is *many*"

Here, $X = \{Age, Income\}$, $Y = \{Cars\}$, $F_X = \{young, high\}$, $F_Y = \{many\}$ and the rule can be represented as:

$$\langle\{Age, Income\}, \{young, high\}\rangle \Rightarrow \langle\{Cars\}, \{many\}\rangle$$

In order to express the quality of a fuzzy association rule two quality measures, fuzzy support and fuzzy confidence, have been proposed in [12].

Definition 4 (Itemset fuzzy support value). *The* ***fuzzy support value*** *of itemset fuzzy itemset $\langle X, F_X\rangle$ in $\mathcal{DB}$ is:*

$$FS_{\langle X,F_X\rangle} = \frac{\sum_{t_i\in\mathcal{DB}}\prod_{x_j\in X}\alpha_{a_j}(t_i[x_j])}{|\mathcal{DB}|} \tag{1}$$

where

$$\alpha_{a_j}(t_i[x_j]) = \begin{cases}\mu_{a_j}(t_i[x_j]), & \textit{if } \mu_{a_j}(t_i[x_j]) \geq \omega\\ 0, & \textit{otherwise}\end{cases} \tag{2}$$

and ω is a user specified minimum threshold for the membership function. Thus, the values of membership functions less than this minimum threshold, ω, are ignored.

Definition 5 (Rule fuzzy support value). *Let $\langle X, F_X\rangle \Rightarrow \langle Y, F_Y\rangle$ be a fuzzy association rule. The* ***fuzzy support value of the rule*** *is defined as fuzzy support value of the itemset $\langle\{X, Y\}, \{F_X, F_Y\}\rangle$:*

$$FS_{\langle X,F_X\rangle\Rightarrow\langle Y,F_Y\rangle} = FS_{\langle\{X,Y\},\{F_X,F_Y\}\rangle}$$

Definition 6 (Rule Fuzzy Confidence). *Let $\langle X, F_X\rangle \Rightarrow \langle Y, F_Y\rangle$, a fuzzy association rule. The* ***fuzzy confidence value*** *of the rule is defined as:*

$$FC_{\langle X,F_X\rangle\Rightarrow\langle Y,F_Y\rangle} = \frac{FS_{\langle Z,F_Z\rangle}}{FS_{\langle X,F_X\rangle}}$$

where $Z = \{X, Y\}$ and $F_Z = \{F_X, F_Y\}$.

A fuzzy association rule is considered as *interesting* if it has enough support and high confidence value.

3 Proposed Intuitionistic Fuzzy System

According to the structure of an FLC, an IFLC requires the following operations: fuzzification, reasoning and defuzzification.

3.1 Fuzzification and Firing Levels

A fuzzification operator transforms crisp data into fuzzy sets. For instance, $x_0 \in U$ is fuzzified into $\overline{x}_0$ according to the relations:

$$\mu_{\overline{x}_0}(x) = \begin{cases}1 & if \quad x = x_0\\ 0 & otherwise\end{cases} \qquad \nu_{\overline{x}_0}(x) = \begin{cases}0 & if \quad x = x_0\\ 1 & otherwise\end{cases}$$

The $\mu-$firing level and $\nu-$firing level of an intuitionistic fuzzy set A and a crisp value x_0 as input are $\mu_A(x_0)$ and $\nu_A(x_0)$, respectively.

3.2 Reasoning

In order to perform reasoning a set of rules are necessary. Typically rules for fuzzy logic controllers appear in if-then form and are obtained from the knowledge of experts and operators. As a result, the rules are limited, subjective and inaccurate.

In our system, these rules are automatically induced as fuzzy association rules starting from a training set. We can use any algorithm for mining fuzzy association rules (see [7], [8], [9], [10]) to induce fuzzy association rules. We consider that the training set is described as a set de transactions $\mathcal{DB} = \{t_1, \ldots, t_n\}$ characterized by a set $\mathcal{I} = \{i_1, \ldots, i_m\}$ of attributes. These attributes are represented by the input and output variables of fuzzy logic controller. For each attribute (variable) i_k, $k = 1, \ldots, m$, we consider $n(k)$ linguistic values represented as fuzzy sets.

We generate only rules with input attributes in premise and output attributes in conclusion. The generated rules has the following form:

$$R_i : if\ X_1\ is\ A_i^1\ and\ ...\ and\ X_r\ is\ A_i^r\ then\ Y\ is\ C_i : \quad (FS_i, FC_i) \tag{3}$$

where $X_j, j \in \{1, 2, ..., r\}$ represent the input variables, Y is an output variable, $A_i^j, j \in \{1, 2, ..., r\}$ and C_i are linguistic values associated with X_j and Y respectively, and (FS_i, FC_i) are the fuzzy support and the fuzzy confidence of the rule.

For a given rule R_i the input data $x = \{x_1, \ldots, x_r\}$ generates the $\mu-$firing level μ_i, the $\nu-$firing level ν_i and the intuitionistic fuzzy index π_i, computed as

$$\mu_i = T(\mu_i^1, ..., \mu_i^r), \quad \nu_i = T(\nu_i^1, ..., \nu_i^r), \quad \pi_i = T(\pi_i^1, ..., \pi_i^r) \tag{4}$$

where T is a t-norm while μ_i^j is the $\mu-$firing level, ν_i^j is the $\nu-$firing level and π_i^j is the intuitionistic fuzzy index, respectively, for $A_i^j, j \in \{1, 2, ..., r\}$.

We partition the generated rules in subsets with rules having the same conclusion:

$$\mathcal{R}(C) = \begin{cases} R_1 : if\ X_1^1\ is\ A_1^1\ and\ ...\ and\ X_1^{r_1}\ is\ A_1^{r_1}\ then\ Y\ is\ C : (FS_1, FC_1) \\ \ldots \\ R_P : if\ X_P^1\ is\ A_P^1\ and\ ...\ and\ X_P^{r_P}\ is\ A_P^{r_P}\ then\ Y\ is\ C : (FS_P, FC_P) \end{cases} \tag{5}$$

For each rule subset $\mathcal{R}(C) = \{R_1, \ldots, \mathcal{R}_P\}$ we compute the matching value of input data x as follows:

$$MR(C) = \frac{\sum_{i=1}^{P} \alpha_i * FC_i}{\sum_{i=1}^{P} FC_i} \tag{6}$$

where α_i is the firing level of the rule R_i, computed as

$$\alpha_i = (1 - \pi_i)\mu_i + \pi_i \nu_i.$$

In order to compute the output of IFLC for the input x, the system selects the subset of rules $\mathcal{R}(C)$ having the maximum matching measure. This subset will be identified as a single rule having the conclusion C and the firing levels l_μ and

l_ν, where $l_\mu = \dfrac{\sum_{i=1}^{P} \mu_i * FC_i}{\sum_{i=1}^{P} FC_i}$ and $l_\nu = \dfrac{\sum_{i=1}^{P} \nu_i * FC_i}{\sum_{i=1}^{P} FC_i}$. Using the Lukasiewicz implication, we get two conclusions:

- the $\mu-$conclusion C^μ, which is the traditional output of the rule computed using the membership function and the firing level μ

$$\mu_{C^\mu}(v) = I(l_\mu, \mu_C(v)), \forall v \in V.$$

- the $\nu-$conclusion C^ν, which is the output of the rule computed using the non-membership function and the firing level ν

$$\mu_{C^\nu}(v) = I(l_\nu, \mu_C(v)), \forall v \in V.$$

The crisp value y_0 associated to a inferred conclusion C' is obtained by defuzzification method.

3.3 Defuzzification

The $\mu-$conclusion and the $\nu-$conclusion inferred using the set of rules $\mathcal{R}(C)$ are given by the following expressions:

$$\mu_{C^\mu}(x) = \begin{cases} 1 - l_\mu \quad if \quad x \le a_1 \\ \dfrac{x - a_1}{a_2 - a_1} + 1 - l_\mu \quad if \quad x \in [a_1, l_\mu(a_2 - a_1) + a_1] \\ 1 \quad if \quad x \in [l_\mu(a_2 - a_1) + a_1, l_\mu(a_3 - a_4) + a_4] \\ \dfrac{x - a_4}{a_3 - a_4} + 1 - l_\mu \quad if \quad x \in [l_\mu(a_3 - a_4) + a_4, a_4] \\ 1 - l_\mu \quad if \quad a_4 \le x \end{cases}$$

$$\mu_{C^\nu}(x) = \begin{cases} 1 \quad if \quad x \le l_\nu(b_1 - b_2) + b_2 \\ \dfrac{x - b_2}{b_1 - b_2} + 1 - l_\nu \quad if \quad x \in [l_\nu(b_1 - b_2) + b_2, b_2] \\ 1 - l_\nu \quad if \quad x \in [b_2, b_3] \\ \dfrac{x - b_3}{b_4 - b_3} + 1 - l_\nu \quad if \quad x \in [b_3, l_\nu(b_4 - b_3) + b_3] \\ 1 \quad if \quad x \ge l_\nu(b_4 - b - 3) + b_3 \end{cases}$$

where l_μ and l_ν are the firing levels defined in the previous section.

The conclusions C^μ and C^ν are defuzzified into y^μ and y^ν, respectively, using the Middle-of-Maxima and Middle-of-Minima, respectively; the defuzzified output corresponding to conclusion C of the set $\mathcal{R}(C)$ is computed as a linear combination $y = (1 - \pi)y^\mu + \pi y^\nu$, where π is the intuitionistic fuzzy index associated with the set $\mathcal{R}(C)$, computed as $\pi = \dfrac{\sum_{i=1}^{P} \pi_i * FC_i}{\sum_{i=1}^{P} FC_i}$.

For $\pi = 0$ the IFLC system reduces to the system from [11].

4 A Case Study

In order to show how the proposed system works, we consider an example inspired from [1] concerning washing machines. We consider an IFLC system with two inputs and one output. The input variables are *degree-of-dirt* (DD) and *type-of-dirt* (TD); the output variable is *washing-time* (WT). We consider the universes of discourse $[0, 100]$ for the input variables and $[0, 60]$ for the output variable.

For the input variable DD we can take into consideration the following three linguistic variables:

$$F_{DD} = \{Small, Medium, Large\}.$$

Similarly, let

$$F_{TD} = \{VeryNotGreasy, NotGreasy, Medium, Greasy, VeryGreasy\}$$

be the set of linguistic variables associated with the input variable TD.

For the output variable WT we consider the following set of linguistic variables

$$F_{WT} = \{VeryShort, Short, Medium, Long, VeryLong\}.$$

The membership and the non-membership functions for all linguistic variables are defined as trapezoidal intuitionistic fuzzy numbers.

In order to extract the rules used by the inference engine of IFLC, we use a modified implementation of Fuzzy Apriori-T algorithm [7]. This algorithm runs on a training dataset of the form (DD, TD, WT) and keep only rules with support and confidence greater than or equal to the minimum support threshold and minimum confidence threshold respectively.

The Table 1 contains the fuzzy association rules obtained applying the Fuzzy Apriori-T algorithm on training dataset. Partitioning this set of rules, we get the following subsets having the same conclusion:

Table 1. Fuzzy Association Rules

ID	Rule	Confidence
R_1	*If DD is Large and TD is Greasy then WT is VeryLong*	91,27%
R_2	*If DD is Medium and TD is Greasy then WT is Long*	92,06%
R_3	*If DD is Small and TD is Greasy then WT is Long*	91,83%
R_4	*If DD is Large and TD is Medium then WT is Long*	83,54%
R_5	*If DD is Medium and TD is Medium then WT is Medium*	84,52%
R_6	*If DD is Small and TD is Medium then WT is Medium*	92,11%
R_7	*If DD is Large and TD is NotGreasy then WT is Medium*	92,89%
R_8	*If DD is Medium and TD is NotGreasy then WT is Short*	94,94%
R_9	*If DD is Small and TD is NotGreasy then WT is VeryShort*	74,65%

$\mathcal{R}(VeryLong) = \{R_1\}$, $\mathcal{R}(Long) = \{R_2, R_3, R_4\}$, $\mathcal{R}(Medium) = \{R_5, R_6, R_7\}$, $\mathcal{R}(Short) = \{R_8\}$ and $\mathcal{R}(VeryShort) = \{R_9\}$.

Now, we have inference rules for our IFLC system. Let consider that we want to compute the output for the following input data $x = (79, 62)$. Using the t-norm Product $T(x, y) = xy$, we compute for every rule R_i the values: the $\mu-$firing level μ_i, the $\nu-$firing level ν_i, the fuzzy index π_i and the firing index α_i.

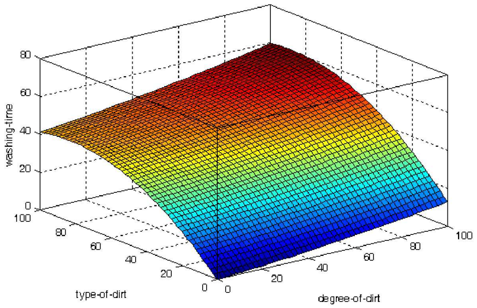


Fig. 1. Input/output response surfaces for IFLC system

After this step, for each partition of rules we compute the matching value of the input data x using the formula (6) and select the rule set having the maximum matching measure, $\mathcal{R}(Long)$.

For the set $\mathcal{R}(Long)$, the conclusions $Long^{\mu}$ and $Long^{\nu}$ are defuzzified into $y^{\mu} = 44.3229$ and $y^{\nu} = 44.7509$, respectively. Because the fuzzy index for the set $\mathcal{R}(Long)$ is $\pi = 0.10483398$, the conclusion inferred by the IFLC system is $y = 44.3678$. If we use the classic FLC system [11] the inferred conclusion gives $y = 43.17632$. The response surfaces given by IFLC system is presented in Figure 1. In order to evaluate the performance of the IFLC system we made a comparison with classical FLC system for various inputs, which highlighted the performance of the proposed algorithm compared to the classic version. For sample inputs: $(13, 82)$, $(15, 80)$, $(11, 78)$, $(16, 85)$, $(10, 75)$, $(12, 79)$, $(14, 76)$, $(13, 84)$, $(11, 82)$, $(77, 47)$, $(73, 45)$, $(74, 42)$, $(75, 45)$, $(78, 46)$, the dispersions of the results are $D_1^2 = 23.5308 * 10^{-3}$ for FLC system and $D_2^2 = 23.1986 * 10^{-3}$ for IFLC system. Because $D_2^2 < D_1^2$ it means that the intuitionistic fuzzy system is more robust than the classical. Other statistics highlight this property. For instance, the amplitude (difference between extreme values) is 0.42923 for FLC system and 0.42115 for IFLC system; the coefficient of variation (standard deviation relative to the arithmetic mean) is $3.5519 * 10^{-3}$ for FLC system and $3.4371 * 10^{-3}$ for IFLC system.

5 Conclusion

This paper presents a fuzzy controller model of Mamdani type working with intuitionistic fuzzy sets; thus this controller is an extension of the model from [15],[11]. While the standard Mamdani controller activates a set of rules with different conclusions, our model activates a set of rules having the same conclusion; thus we obtain a fuzzy set as output, which can be defuzzified in order to obtain a crisp value. Moreover, the rules used by Fuzzy Inference Engine are generated using Data Mining techniques and taking into account the rule fuzzy confidence. The sample explained in Section 4 can be used as a model for various real-world applications. In the future we intend to extend this version in order to work with crisp data, intervals and/or linguistic terms as inputs.

References

1. Agarwal, M.: Fuzzy Logic Control of Washing Machines, http://softcomputing.tripod.com/sample_termpaper.pdf
2. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
3. Atanassov, K.: Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 20, 87–96 (1986)
4. Atanassov, K.: Intuitionistic Fuzzy Sets: Theory and Applications. Physica-Verlag (1999)
5. Baczynski, M., Jayaram, B.: Fuzzy implications. STUDFUZZ, vol. 231. Springer, Berlin (2008)
6. Bellman, R.E., Zadeh, L.A.: Decision making in a fuzzy environment. Management Sciences 17, 141–164 (1970)
7. Coenen, F.: The LUCS-KDD Fuzzy Apriori-T Software, Department of Computer Science, The University of Liverpool, UK (2008), http://www.csc.liv.ac.uk/~frans/KDD/Software/Apriori_TFP/aprioriTFP.html
8. Gabroveanu, M.: Mining association rules. In: Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches, ch. XXVIII, pp. 647–673. IGI-Global, USA (2009)
9. Gyenesei, A.: Mining Weighted Association Rules for Fuzzy Quantitative Items. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 416–423. Springer, Heidelberg (2000)
10. Hong, T.-P., Kuo, C.-S., Chi, S.-C., Wang, S.-L.: Mining Fuzzy Rules from Quantitative Data Based on the ApriotiTid Algorithm. In: SAC 2000: Proceedings of the, ACM Symposium on Applied Computing, pp. 534–536. ACM Press, Como (2000)
11. Iancu, I., Gabroveanu, M.: Fuzzy logic controller based on association rules. Annals of the University of Craiova - Mathematics and Computer Science Series 37(3), 12–21 (2010)
12. Kuok, C., Fu, A., Wong, M.H.: Mining fuzzy association rules in databases. SIGMOD Rec. 27(1), 41–46 (1998)
13. Szmidt, E., Kacprzyk, J.: Analysis of agreement in a group of experts via distances between intuitionistic preferences. In: Proceedings of IPMU 2002, Universite Savoie, France, pp. 1859–1865 (2002)
14. Szmidt, E., Kacprzyk, J.: Intuitionistic fuzzy sets in some medical applications. Notes of IFS 7(4), 58–64 (2001)
15. Tajbakhsh, A., Rahmati, M., Mirzaei, A.: Intrusion detection using fuzzy association rules. Applied Soft Computing 9(2), 462–469 (2009)
16. Zadeh, L.A.: A theory of approximate reasoning. In: Machine Intelligence, pp. 149–194. John Wiley & Sons, New York (1979)
17. Zadeh, L.A.: Fuzzy sets as a basis for a theory of a possibility. Fuzzy Sets and Systems 1, 2–28 (1978)

Joining Data Mining with a Knowledge-Based System for Efficient Personalized Speech Therapy

Mirela Danubianu

Stefan cel Mare, University of Suceava
mdanub@eed.usv.ro

Abstract. This paper to present an approach in which data mining technique could contribute to the enrichment of knowledge base of expert system embedded in TERAPERS.- an intelligent system which aims to assist the personalized therapy of speech disorders. Since March 2008 the system is used by the therapists from Regional Speech Therapy Center of Suceava, and it provides a data set which can serve as a foundation for data mining operations. Patterns resulted from these operations may complete the set of rules which form the knowledge base of expert system in TERAPERS.

Keywords: expert system, data mining, classification rules.

1 Introduction

Knowledge-based systems are valuable tools in decision-making process. They aim to solve problems within a certain domain using artificial intelligence, and in order to do that, they contain two main components: a knowledge base and an inference engine. The most frequent knowledge based systems are rule-based expert systems.

The greatest problem for the development of an efficient system is to build a good knowledge base. In order to do that, knowledge acquisition is a sensible problem. Obviously each domain provide several sources of knowledge, as specialized literature or human experts, but the modern achievements in data processing area could offer new possibilities to enrich a knowledge base. One of these modern achievements is data mining. As part of the complex process of knowledge discovery in databases, data mining tries to find useful patterns in large amounts of data, which have no apparent relationship between them. Such a pattern can take the form of IF-THEN rules.

This paper aims to present an initial research regarding some ways in which one of the data mining methods- classification- could contribute to enrich the knowledge base of the expert system embedded in TERAPERS in order to improve the personalized therapy of dyslalia.

2 Knowledge-Based Systems

Knowledge-based systems, known also as expert systems, are computer programs that incorporate the knowledge and analytical skills of human experts in certain area.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 245–254, 2013.

Turban defines such system as: “a computer program that simulates the judgment and behavior of a human or an organization that has expert knowledge and experience in a particular field” [1]. Having a complex architecture, as presented in Figure 1, an expert system includes the following components: the knowledge base which contains the knowledge of the human experts of a specific area; the facts base, which embodies data from a problem to be resolved as well as the facts, resulted from the reasoning made by inference engine over the knowledge base, and the inference engine that is the module that performs the transformation. It starts from facts, activates the correspondent knowledge from the knowledge base and builds the reasoning which leads to new facts. Other components are: the explanation module which has the role to present in accessible forms the justification of the reasoning made by the inference engine; the knowledge acquisition module which transform the human expert’s knowledge in the appropriate form for the system use and the user graphic interface (GUI) which allows to the users to access the expert system.

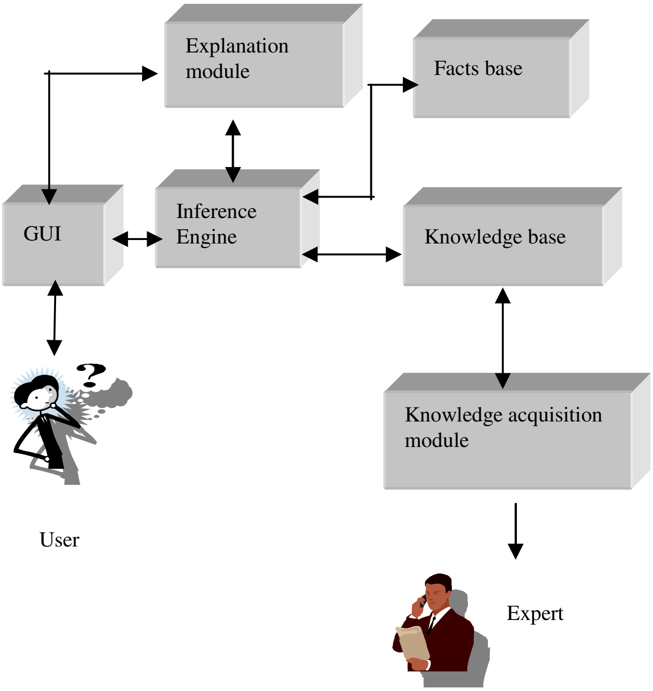


Fig. 1. Knowledge-based system architecture

The most used technique of knowledge representation is rule-based. A rule has the form “***IF condition THEN action***”. In a rule-based expert system, the domain knowledge is represented as sets of rules that are checked over a collection of facts or knowledge about the current situation. When the IF section of a rule is satisfied by the facts, the action specified by the THEN section is performed. The IF section of a rule is compared with the facts and the rules whose IF section matched the facts are executed. This action may modify the set of facts in the knowledge base.

The great challenge for the development of an efficient expert system is to build a good knowledge base. This means a knowledge base that has to be complete, consistent, coherent and non-redundant. In order to do that, knowledge acquisition is one of the main problems. Obviously each domain provide several sources of knowledge, as specialized literature or human experts, but the modern achievements in data processing area could offer new possibilities to enrich a knowledge base.

3 Knowledge Discovery in Databases and Data Mining

Defined as “the process of identifying valid, novel, potentially useful, and understandable patterns in data” [2], knowledge discovery in databases (KDD) is the result of the technological progress associated with the increased need for valuable and new knowledge. Over time, several models for KDD process have been proposed, but the most known is the industrial model – CRISP-DM, presented in Figure 2. Accordingly to this model, KDD process consists of the following six steps: business understanding, data understanding, data preparation, modeling, evaluation of the model and deployment [3].

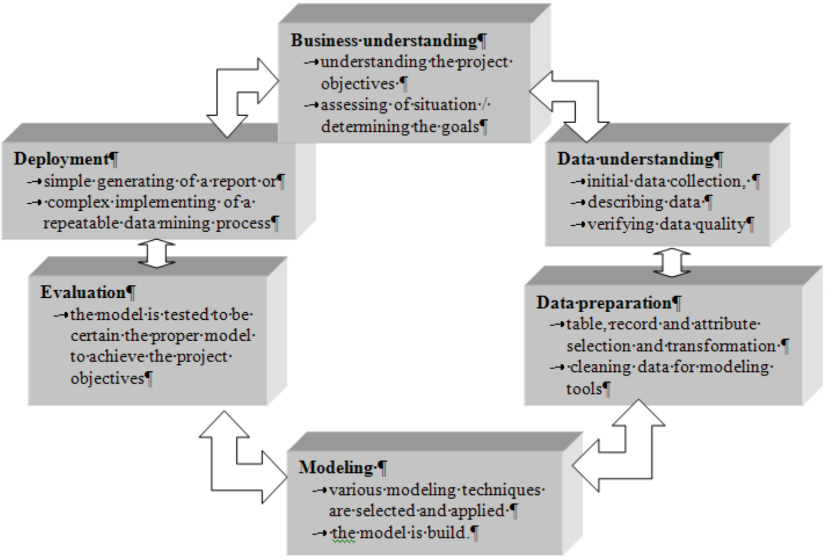

Business understanding¶
→understanding the project objectives ¶
→assessing of situation / determining the goals¶
Data understanding¶
→initial data collection, ¶
→describing data ¶
→verifying data quality¶
Data preparation¶
→table, record and attribute selection and transformation ¶
→ cleaning data for modeling tools¶
Modeling ¶
→various modeling techniques are selected and applied ¶
→ the model is build.¶
Evaluation¶
→the model is tested to be certain the proper model to achieve the project objectives¶
Deployment¶
→simple generating of a report or¶
→ complex implementing of a repeatable data mining process¶

Fig. 2. CRISP-DM Model

Defined as the semiautomatic extraction of knowledge from huge volumes of data, data mining correspond to the modeling step in the knowledge discovery in databases process. It involves the application of intelligent methods in order to discover new and interesting patterns from large volumes of data, and it may facilitate the discovery from apparently unrelated data, relationships that can anticipate future problems or might solve the studied problems. Data mining applications may solve two kinds of problems: prediction and knowledge discovery. Prediction is considered the main goal of data mining, but it is often preceded by description.

For each of these problems it is indicated to use some associated methods. For prediction we can use classification or regression while for knowledge discovery we can use deviation detection, clustering, association rules, database segmentation, sequence analysis or visualization [4].

3.1 Data Mining Methods

Placed in the category of supervised learning methods, *classification* is a two-phase process. First a model that describes a set of predetermined classes or concepts is built. Each case is assumed to belong to a predefined class, as determined by one of its attributes, called the class label attribute. In the second step, the model is used for classification, but before that, it is necessary to estimate the predictive accuracy.

Whereas classification determines the set membership of the samples, the prediction of continuous values can be modeled by *regression.* In this case model design consists of finding a structure for it, on computing an optimal value for its parameters and assessing the model quality.

Clustering, often referred to as unsupervised learning, involve a process that discovers structures in data without any supervision. As the name clustering implies, unsupervised algorithm is capable of discovering structures on its own by exploiting similarities or differences between individual data points on a data set.

Association rules mining is an important data mining method that aims to find interesting dependencies in large sets of data items. Interesting associations between data items can often lead to information used for decision making.

Classification

Classification is the process by which a set of common properties of objects in a database are identified, and based on these similarities objects are allocated to different classes, using a model.

Classification's objective is to analyze a training data set and develop an accurate description or model for each class using the available features. These class descriptions are then used to classify future test data from the database or to develop a description (so called "classification rules") for each class in the database.

In practice, one of the most frequently used method is classification based on decision tree [5] [6], because it is suitable to approximate a target function with discrete values. Once built, the decision tree can predict a new data instance, by following a path that starts from the root to a leaf node.

One of the advantages of decision trees lies in the fact that they can easily translate into a set of ***IF -THEN*** rules easier to understand. For each path, from the root node to a leaf node, a rule is created and each pair of attribute values along the path forms a conjunction.

4 Speech Disorders and Speech Therapy

A speech disorder may be defined as a problem of fluency, voice, and/or how a person utters speech sounds. It may have different causes, from organic to functional, neurological or psycho-social causes. The prevention and treatment of speech disorders is a complex issue, stirring the interest of speech therapists, as well as of those asked to contribute to the children's language education. Early treatment of speech impairments ensures improved efficiency, as psycholinguistic automatisms are not consolidated in young children, and, through adequate educational interventions, they can be replaced by correct speech acts. Differential diagnosis will decide upon the therapy for correcting language, as psycho diagnosis allows an adequate therapeutic program, and the elaboration of a prognosis regarding the evolution of the child, along with the therapeutic process. The therapy has to be adapted to each language therapist, to each particular case, to the child’s learning rhythm and style, as well as to the level of the impairment.

Dyslalia is articulation disorder that consists in difficulties with the way sounds are formed and strung together. These are usually characterized by omitting, distorting a sound or substituting one sound for another. The key issues in dyslalia therapy are shown in Figure 3.

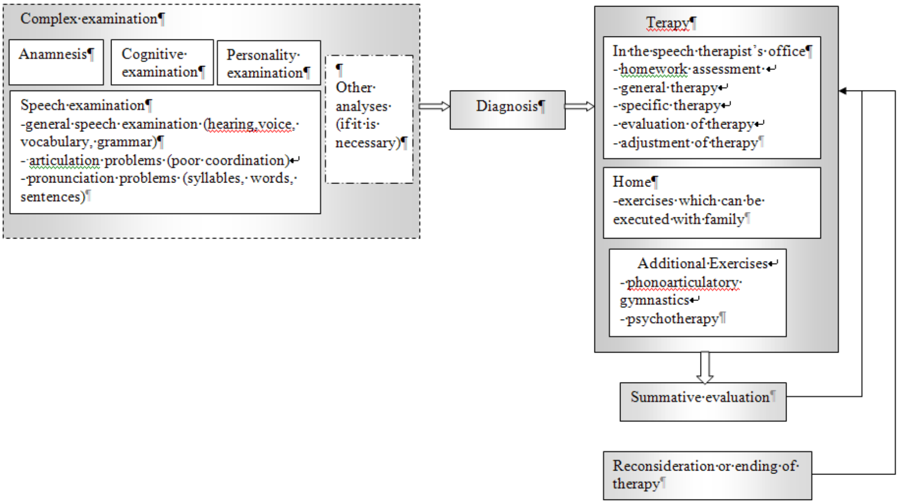

Complex examination
Anamnesis
Cognitive examination
Personality examination
Speech examination
-general speech examination (hearing,voice, vocabulary, grammar)
- articulation problems (poor coordination)
-pronunciation problems (syllables, words, sentences)
Other analyses (if it is necessary)
Diagnosis
Terapy
In the speech therapist’s office
-homework assessment
-general therapy
-specific therapy
-evaluation of therapy
-adjustment of therapy
Home
-exercises which can be executed with family
Additional Exercises
-phonoarticulatory gymnastics
-psychotherapy
Summative evaluation
Reconsideration or ending of therapy

Fig. 3. Key issues in speech therapy [7]

5 TERAPERS

TERAPERS is a computer-based speech therapy system that aims to assist the personalized therapy of dyslalia. This is based on the interactions between six functional blocks [8]: child, speech therapist, lab monitor program, expert system, 3D model, and child monitor program, as is shown in Figure 4.

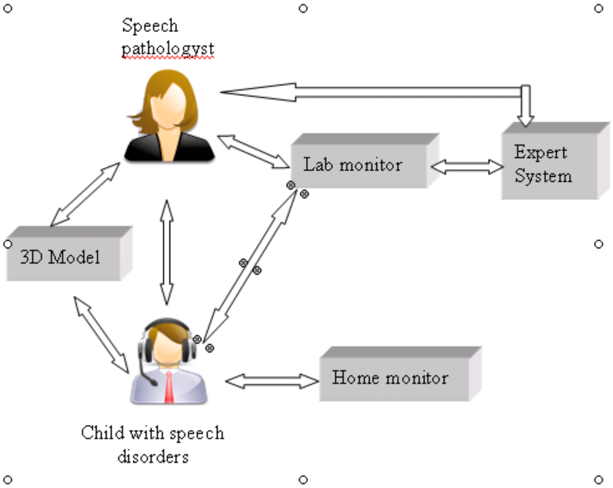


Fig. 4. The intelligent system's blocks relationships

The lab monitor program allows the introduction of a complex examination's information and offers the possibility of making periodically records with the child's speech. The child receives an instant audio feedback and he can see the history of his audio recordings. The role of home monitor program is to create a virtual interface between teacher and child (home speech therapy). This component is implemented both for PC and PDA. It can run exercises in a game manner, can offer feedback and can perform statistics base on current subject scores. The monitor program performs homework transmission to the child PC or PDA. Later, when the child comes back, he can receive the activity report. 3D model provide viewing of the correct positioning of language, lips and teeth for each sound. The professor will analyze the images offered by the 3D model and can correct some of the mistakes. Expert system, if activated, makes suggestions regarding some training parameters like session frequency, length and content (exercises) according with some input variables. If the teacher observes erroneous conclusions, he can view the inferential route and can change the knowledge base. The expert takes the data input from the monitor program and generates, upon request, sets of personalized exercises. Lab monitor program is an interface between the speech therapist and other components like data base, expert system and child monitor program. At this level, speech therapist can collect both textual and audio information regarding each child, can administrate exercises and can manage all therapy aspects: selection of children, scheduling for therapy, offer all statistical reports that are required [8].

The system contains two main components: an intelligent system installed on each speech therapist's office computer and a mobile system used as a friend of child therapy. The two systems are connected.

5.1 Expert System Embedded in TERAPERS

The expert system is based on a therapy guide, written in a natural language. This guide formalized in knowledge base consists of [10]:

- the muscular of phonon-articulator system development methods (e.g. setting up exercises for cheeks, lips and tong);
- the rhythm of respiration controlling methods (e.g. supervised inspiration and expiration from the temporal and intensity standpoint);
- the phonomatic hear development methods (e.g. the onomatopoeic pronunciation, rhythmic pronunciation exercises, distinguish along the paronyms);
- the method for the sound consolidation (e.g. the pronunciation sound of direct, inverse and complex syllable, of words, of paronyms, etc);
- the sound's utilization in complex contexts (e.g. sentence, short stories, poems, riddles).

In this project we have used an expert system for therapy of dyslalic children. It aims to create a better model for speech therapist decisions. We use a rule-based expert system which has two major advantages. First is, that usually that kind of systems nor requires a large training set, and since the expert thinking is explicitly spelled out, we know how he thinks about the problem. Regarding that, it has the disadvantage that the knowledge acquisition phase may be difficult.

Another advantage of rule-based expert systems is the potential ability of rule-based expert systems to learn by creation of new rules and addition of new data to the expert knowledge data base.

We implement over 150 rules for control various aspects of personalized therapy (19 variables presented in Figure 5). These rules are currently validated by speech therapists and can be modified in a distributed manner.

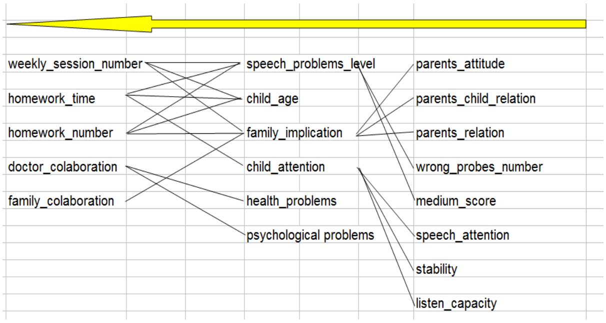


Fig. 5. Variables used for expert system [8]

6 Data Mining and Expert System in TERAPERS

An overview of the data stored in TERAPERS reveals that there are a lot of variables that are considered for the design and implementation of personalized therapy.

As we have noted above, speech therapy is a complex process. Adapting the therapy programs involves a complex examination of children and recording of relevant data relating to personal and family anamnesis. Complex examination of how the children articulate the phonemes in various constructions allows a diagnosis and classification in a given category of severity. Anamnesis data collected may provide information relative to various causes that may negatively influence the normal development of the language. It contains historical data and data provided by the cognitive and personality examination [4]. On provide to the applied personalized therapy programs data such as number of sessions/week, exercises for each phase of therapy and the changes of the original program according to the patient evolution. In addition, the report downloaded from the mobile device collects data on the efforts of child self-employment. These data refer to the exercises done, the number of repetitious for each of these exercises and the results obtained. [4] The tracking of child's progress materializes data which indicate the moment of assessing the child and his status at that time. All these data are stored in a relational database, composed of 60 tables. [4] Data collected by the TERAPERS system together with data from other sources (eg demographic data, medical or psychological research) is the set of raw data that will be the subject of data mining. One of the methods we have tested is classification which places the people with different speech impairments in predefined classes. Thus it is possible to track the size and structure of various groups. We use classification which is based on the information contained in many predictor variables, such as personal or familial anamnesis data or related to lifestyle, to join the patients with different segments. For this purpose we have considered a data set containing 102 features, and around 300 cases. These features relate to data anamnesis, data from complex examination and results from evaluations occasioned by the end of each stage of therapy and at the end of the whole process.

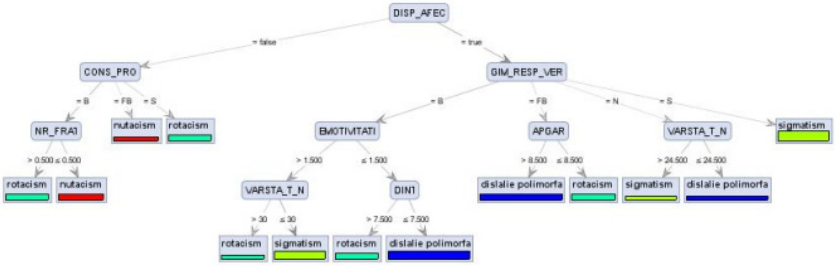
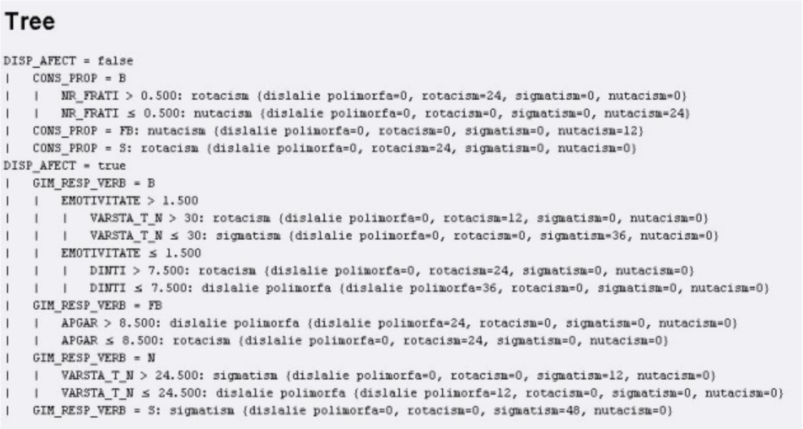


Fig. 6. Decision tree built on TERAPERS data

Several experiments were conducted using RapidMiner5 [9]. For each of them a label class was set, and based on available data, decision tree were built. We used decision trees because these allow a natural transposition into a minimal set of rules having as consequent the class label. Such a decision tree, having as class label the "*diagnosis*" attribute is shown in Figure 6.

Subsequently, decision trees were converted into sets of rules, as shown in Figure 7.



Tree

```
DISP_AFECT = false
|   CONS_PROP = B
|   |   NR_FRATI > 0.500: rotacism {dislalie polimorfa=0, rotacism=24, sigmatism=0, nutacism=0}
|   |   NR_FRATI ≤ 0.500: nutacism {dislalie polimorfa=0, rotacism=0, sigmatism=0, nutacism=24}
|   CONS_PROP = FB: nutacism {dislalie polimorfa=0, rotacism=0, sigmatism=0, nutacism=12}
|   CONS_PROP = S: rotacism {dislalie polimorfa=0, rotacism=24, sigmatism=0, nutacism=0}
DISP_AFECT = true
|   GIM_RESP_VERB = B
|   |   EMOTIVITATE > 1.500
|   |   |   VARSTA_T_N > 30: rotacism {dislalie polimorfa=0, rotacism=12, sigmatism=0, nutacism=0}
|   |   |   VARSTA_T_N ≤ 30: sigmatism {dislalie polimorfa=0, rotacism=0, sigmatism=36, nutacism=0}
|   |   EMOTIVITATE ≤ 1.500
|   |   |   DINTI > 7.500: rotacism {dislalie polimorfa=0, rotacism=24, sigmatism=0, nutacism=0}
|   |   |   DINTI ≤ 7.500: dislalie polimorfa {dislalie polimorfa=36, rotacism=0, sigmatism=0, nutacism=0}
|   GIM_RESP_VERB = FB
|   |   APGAR > 8.500: dislalie polimorfa {dislalie polimorfa=24, rotacism=0, sigmatism=0, nutacism=0}
|   |   APGAR ≤ 8.500: rotacism {dislalie polimorfa=0, rotacism=24, sigmatism=0, nutacism=0}
|   GIM_RESP_VERB = N
|   |   VARSTA_T_N > 24.500: sigmatism {dislalie polimorfa=0, rotacism=0, sigmatism=12, nutacism=0}
|   |   VARSTA_T_N ≤ 24.500: dislalie polimorfa {dislalie polimorfa=12, rotacism=0, sigmatism=0, nutacism=0}
|   GIM_RESP_VERB = S: sigmatism {dislalie polimorfa=0, rotacism=0, sigmatism=48, nutacism=0}
```

Fig. 7. Set of rules resulted from a decision tree

In our experiments we used as class labels, on the one hand the diagnosis, and on the other hand the results of assessments made by experts in various moments of therapy.

To correctly classify children' states at the end of each phase of therapy we considered as predictive attributes along with anamnesis data and those that are focused on therapeutic schema designed such as: the number of sessions / week, the exercises performed, the results of independent work. As mentioned earlier, both TERAPERS' expert system and classification models use IF-THEN rules. However, among these rules is a big difference. Rules that form the knowledge base of the expert system have the form:

$$\text{IF expr THEN action} \tag{1}$$

In case the IF expression is TRUE it triggers the execution of an action.
The rules for a classification model obtained from a data mining process have the form:

$$\text{IF expr THEN class} \tag{2}$$

If the IF expression is TRUE, the case is placed in a certain class. For a potential inclusion in the knowledge base of the expert system, these rules must undergo a post-processing step in which, if possible, the class to be associated with a particular action. Finally, the IF expression will be associated with an action. Following our experiments and the evaluation by therapists of pattern achieved, the set of variables considered for rules construction was completed with variable such as “child's emotivity” and “affective disposition” that can influence, for example, the number of wrong probes in the complex evaluation or the child's attention. The set of possible actions, established by the speech therapy experts, was also supplemented with some attempts to suggest a sequence and structure of exercises to be done, to lead to an ameliorated state of the child at the end of different stages of therapy. Until now the knowledge base of the expert system was increased by around 15%. Now, it is necessary to make a practice validation for new rules.

7 Conclusion and Future Work

In this paper we present a means by which using data mining techniques, we were able to complete the rule base of an expert system embedded in a speech therapy assistant system. TERAPERS system was developed within the Center for Computer Research in the University "Stefan cel Mare", and starting March 2008 it is used by the therapists from Regional Speech Therapy Center of Suceava. Data collected in the exploitation were the foundation for the application of several data mining techniques. Rules-based classification has proven to be useful both for tracking the size and structure of various groups of patients by placing the people with different speech impairments in predefined classes and to enrich the knowledge base of the expert system embedded in TERAPERS. Next period, a practice validation of the new rules will be made, and with the accumulation of new data we will try to refine the existing rules or to discover others rules through data mining.

References

1. Turban, E.: Expert System and Applied Artificial Intelligence. Macmillan Publishing Company, New York (1992)
2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Cambridge (1996)
3. Wirth, R., Hipp, J.: CRIPS-DM Towards a standard process model for data mining. In: Proceedings of the Fourth International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, pp. 29–39 (2000)
4. Danubianu, M., Pentiuc, S.G., Socaciu, T.: Towards the Optimized Personalized Therapy of Speech Disorders by Data Mining Techniques. In: The Fourth International Multi Conference on Computing in the Global Information Technology, ICCGI 2009, August 23-29, vol. CD, pp. 1–6 (2009)
5. Quinlan, J.R.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
6. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
7. Danubianu, M., Tobolcea, I.: Using data mining approach for personalizing the therapy of dysalia. In: E-Health and Bioengineering Conference (EHB), pp. 1–4 (2011)
8. Danubianu, M., Pentiuc, S.G., Schipor, O., Nestor, M., Ungurean, I., Schipor, D.M.: TERAPERS - Intelligent Solution for Personalized Therapy of Speech Disorders. International Journal on Advances in Life Science 1(1), 26–35 (2009) ISSN: 1942-2660
9. Mierswa, I., Wurst, M., Klingenberg, R., Scholz, M., Evler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 935–940 (2006)
10. Pentiuc, S.G., Schipor, O.A., Danubianu, M., Schipr, M.D.: Therapy of Dyslalia Affecting Pre-Scholars. In: Proceedings of Third European Conference on the Use of Modern Communication Technologies - ECUMICT, Gent, Belgium (2008)

On Interacting with Collective Knowledge of Group Facilitation

Constantin-Bala Zamfirescu[1] and Ciprian Candea[2]

[1] "Lucian Blaga" University of Sibiu, Faculty of Engineering, Department of Computer Science and Automatic Control, Emil Cioran 4, 69121 Sibiu, Romania
zbc@acm.org
[2] Ropardo SRL, Research and Development Department, Reconstructiei 2A
550129, Sibiu, Romania
ciprian.candea@ropardo.ro

Abstract. Group decision process design is a well-known class of ill-structured, dynamic, and going-concerns problem. The paper presents a human-computer interaction engineering approach to design a software prototype that provides personalized, contextual and actionable recommendations for this problem. The approach emphasizes the computational aspects of collective intelligence to structure these recommendations based on the collective knowledge that reflects not only the design space *per se*, but the collective experience in exploiting it as well. It is demonstrated by: 1) detailing the engineering issues of an implemented prototype for the group decision process design; and 2) explaining its functionalities through a representative set of interaction scenarios.

Keywords: design space exploration, stigmergic systems, human-computer interaction, and collective knowledge.

1 Introduction

Group Decision Support System (GDSS) is one of the most complex collaborative working environments in terms of its design space and extensive application domains. Theoretically, as an open composition of highly configurable collaborative tools, it requires a high level of expertise for an effective exploration of the available design space. Several field studies of GDSS research [1, 2] acknowledge the tight correlation between the optimal design of a **group decision process** (GDP) and the expert knowledge to arrange the collaborative tools. If such knowledge may be incrementally developed in the form of "corporate knowledge" by exploring the design space for a GDSS that features a closed and well-known set of collaborative tools, the complexity of the **GDP design** (GDPD) space becomes intractable in the context of increased tendency to adopt cloud-based GDSS solutions with an open and dynamic array of composite tools [3-4]. s

From the knowledge engineering stance, this move towards an order of magnitude larger design space is leading inevitably to a diachronic perspective over the GDSS

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 255–265, 2013.

capability to exploit the collective knowledge which may be recorded and preserved from a group to another. As envisaged in the general context of societal-scale GDSS [5, 6], the "Business-as-a-Service" cloud model may bring together a very disperse set of collective knowledge (i.e. GDP business models, patterns of interactions) that may facilitate the exploration and exploitation of the GDPD space.

In contrast with the traditional centralized approach of developing expert systems that explicitly codifies the experts' knowledge for training and/or guiding inexperienced users [7] (i.e. "facilitation in a box"), in our previous research we investigated the possibility to externalize and support, from a diachronic stance, the effective use of collaborative facilitation knowledge with self-development capabilities [8-10]. Similar with the most flourishing collaborative working environments of the last decade (e.g., Google, eBay, Amazon, Wikipedia), we follow the swarming models of computation to develop the collaborative knowledge for the effective exploration of the GDPD space. In these models humans are externalizing their knowledge in a shared environment and exploit them by employing some stigmergic coordination mechanisms that prevent their cognitive load to be exposed to the complexity of an open and dynamic exploration space [11]. In this context, stigmergy brings its intrinsic capability to produce complex knowledge structures without need for any arrangement, organization or direct communication [12], minimizing the frictions encountered in social communication (i.e. negotiation, agreements, etc.).

Conceptually, this collective knowledge structure may be viewed as a **collective mental map** (CMM) [13] of the GDPD space. In this complex conceptual space, the users' aptitude to reason within this space becomes an essential capability of an effective exploitation of this knowledge. As noted by Parunak et. al. [14], through iterative queries reformulations a CMM may provide personalized, contextual, actionable recommendations based on the past collective experiences of its users in the identification, representation and manipulation of partial knowledge.

In this context, the paper presents a human-computer interaction engineering approach to design a software prototype that provides personalized, contextual and actionable recommendations for the GDPD. The form of these recommendations is contextually structured from the collective knowledge that reflects not only the design space *per se*, but the collective experience in exploiting it as well. Consequently, the remaining part of this paper is organized as follows. The next section briefly presents the previous work the contribution of this paper is relaying on. It basically includes the proof of concept for the proposed approach in agent-based social simulation experiments. The third section highlights the engineering issues of designing the software prototype: the interaction framework with the CMM and the key computational elements required to provide practical recommendations for GDPD. Through an exhaustive set of interaction scenarios with the CMM, the functionalities of this prototype are exemplified in the forth section and concluded in the last one.

2 Background Work

Due to its abstraction power, we have used the CMM concept [13] to store the collective knowledge and actions for the GDPD space [8], [10]. It presumes a composition of component models (specialized knowledge, represented as node) linked in a weighted graph, signifying a navigation map that supports the design effort to reach any component model from any place of the CMM where it is needed.

The **component model** (CM) commonly stands for the possible states of the design problem, while the links are the possible design actions that guide the decision from one state to another. In the GDSS research domain, the CM is the smallest piece of knowledge for GDPD. Due to its inherent complexity, the relevant structure for a CM in GDPD is an open problem (e.g., "ThinkLet" [15], "session topic" [16], or a "collaborative activity" in a business process modeling notation [4]) that is irrelevant herein. These CMs belong to one of the five possible types identified in collaboration engineering [15] (i.e. converge, diverge, organize, evaluate, and negotiate) and reflect some general patterns of transforming the collective knowledge.

A link between two CMs is reflecting the users' evaluation of the performance for a possible move from one problem state to another relative to a problem type. The relation to the problem type is achieved using user-defined tags for the performances associated with a link. The tags serve as dynamic navigation links in the GDPD space, a common approach in programming the collective intelligence [17]. Note that the knowledge encoded in the CMM for the GDPD space are created, structured and refined dynamically by all the users who are using them.

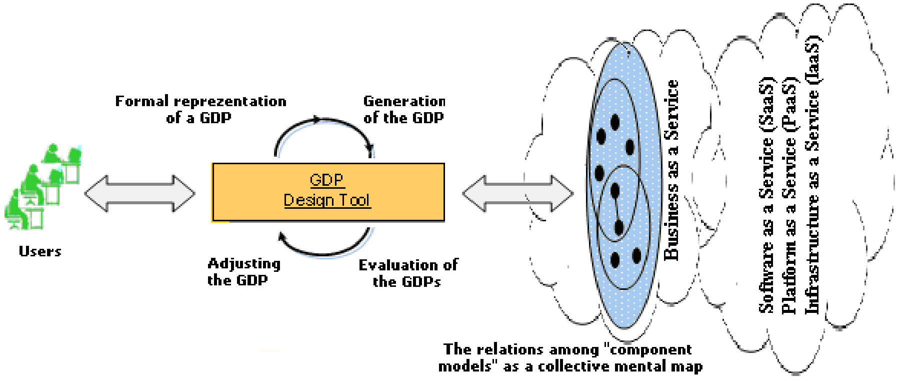

Formal reprezentation of a GDP
Generation of the GDP
Users
GDP Design Tool
Adjusting the GDP
Evaluation of the GDPs
Business as a Service
Software as a Service (SaaS)
Platform as a Service (PaaS)
Infrastructure as a Service (IaaS)
The relations among "component models" as a collective mental map

Fig. 1. The interaction with the CMM of GDPD space

This knowledge model for the GDPD space has been previously investigated through different use scenarios in an agent-based simulation environment. Thus, we have shown that [8-10]: 1) the stigmergic coordination mechanisms permit the exploitation of some contextual factors (i.e. problem type frequency and diversity, users' experience with the GDSS) to guide the GDPD; 2) the effectiveness of these mechanisms depends on how the GDP model is structured on different levels of abstractions and incorporates execution uncertainties; 3) the information used in the stigmergic coordination mechanisms implicitly lead to the possibility of offering contextual recommendations. On this theoretical basis, we have further developed a general interface model and a prototype that facilitates the GDPD [18]. In this model, the interaction with the CMM is considered an iterative process that requires the formal representation, generation, evaluation and adjustment of subsequent designs (Fig. 1). From the engineering stance, this process presumes the development of an application that supports the interaction with the CMM by: 1) visually representing a GDP model, and 2) filtering these representations in relation to user's intentional stance.

3 Engineering Issues

3.1 The Interaction Framework with the CMM

The dialog between the users and the CMM for GDPD is inspired from the **shared-plans theory** (SPT) [19], [20]. According to the SPT, the structure of a dialog includes a description of the actions and plans that underlay the collaboration. The theory is used as foundation to inform the engineering of the GDPD tool and not to implement it in its formal way as in [20]. Contrary to the classical planning theories, SPT introduces the representation of the intentional context in which actions are executed. To define a shared plan, the theory is using simple concepts, i.e. *plan*, *receipt* and *action*.

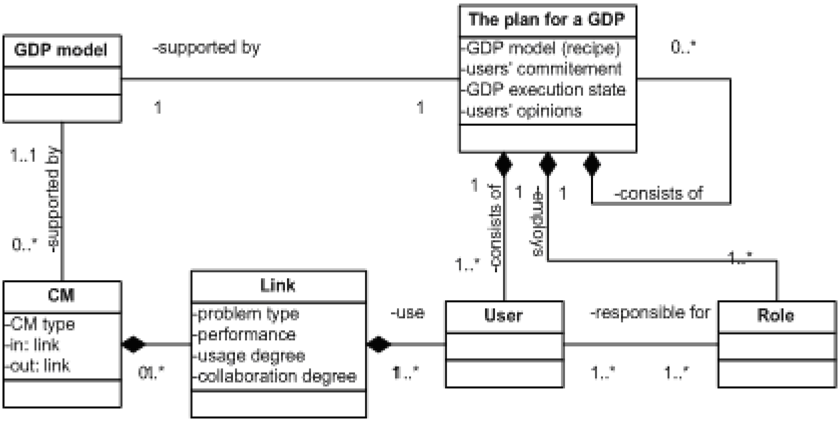


Fig. 2. A simplified data structure for the GDPD

If we consider the GDPD as a collaborative *plan* (Fig. 2), its description presumes the execution of some core activities (i.e. model identification, users' selection, monitor the GDP implementation, and record the users' views and commitments relative to its completion) which are further used to structure the interaction between the users and the GDPD tool. By completing these activities, a plan will reflect the users' mental state in designing and executing a complex action from the GDP model. Consequently, a complex *action* corresponds to the sub-problems in which the initial one can be decomposed. For every complex action there may be one or more implementation methods, called *recipes,* which assume the identification of composite CMs. Thus, to complete a complex action, the users involved in GDPD should identify the recipes that lead to the realization of the action. Even if there may be many recipes for the implementation of an action, only one of them can be selected and executed. This method of hierarchical decomposition of the problem is similar to hierarchical tasks network [21] and can be described in business process modeling notations [22].

All these GDPD activities are realized within the CMM conceptual space depicted in Fig. 2 with the corresponding data structures (i.e. CMs, Link and User). In this way, the CMM codifies correlated knowledge among users and CMs, reflecting the users' evaluation of CM's performance (a node in the graph) relative to a problem type. The performances are stored for each "problem type" in a variable associated with each tagged link from the graph.

3.2 Some Computational Aspects of Collaborative Design

Collective Performance. The performance recorded in the links between the CMs describes the CMM state over time (Fig. 2). In our case, we have applied a simple weighted additive rule for the aggregation of performances [8]: $CP_{jk}(CM_k,t) = CP_{jk}(CM_k,t\text{-}1) + IP_{jk}(CM_k)/w,$ where: t is the temporal component of the model, incremented for each successive design activity; k is the CM's identification index; $IP_{jk}(CM_k)$ is the user's individual evaluation of the performance for the k-th CM assessed from the side of CM j at moment t; $CP_{jk}(CM_k,t)$ and $CP_{jk}(CM_k,t\text{-}1)$ are the new and previous values of the **collective performance** (CP) stored on the link between the CMs j and k for a problem type (Fig. 2); and w is a tuning parameter, given by the implementation engineer, to weight the impact of the last performance assessment.

Subjective Performance. Since a CM's performance is influenced by the group composition as well, the CP for the group of users involved in a GDPD $(CP^{<Group>})$ may be computed as a linear function dependent by: 1) the CP of a CM evaluated by the entire users community; 2) the usage degree $(UD^{<Group>})$ of a CM by the group; and 3) the collaboration degree $(CD^{<Group>})$ among the group's members relative to a CM. Thus, the subjective performance of the k-th CM evaluated from the side of CM j is: $CP^{<Group>}{}_{jk}\ (CM_k) = \lambda_1 CP_{jk}(CM_k) + \lambda_2 UD^{<Group>}(CM_k) + \lambda_3 CD^{<Group>}(CM_k)$, where: $\lambda_1, \lambda_2, \lambda_3$ are tuning parameters (implementation context dependent), to reflect the impact of users experience with the GDSS (λ_2) and prior collaborations (λ_3) in relation to CP (λ_1); $UD^{<Group>}$ is the average use of a CM by each member of the group; and for n users that compose the group:

$$CD^{<Group>}(CM_k) = \sum_{\substack{u,v=1 \\ u \neq v}}^{n} nc_{uv}(CM_k) \Big/ n(n-1)/2 \tag{1}$$

is the average of previous collaborative use of CM_k by the users u and v (nc_{uv}). Note that all these data are recorded in the CMM (Fig 2).

Selection Entropy. The self-organization of relations between the CMs after successive performance evaluations over time entails for a rational behavior during the GDPD a "decrease of freedom" (i.e. lessening the probability distribution of preferences among alternative CMs) in following the possible suggestions provided by the system. Consequently, for a problem type, this corresponds to the Shannon normalized entropy [23, 24], where the preference for an alternative CM is computed by normalizing its related performance to the aggregation of performances for all the available alternatives. Note that for simplicity reasons, the performance of a CM stands for the $CP^{<Group>}$ as computed above. Thus, the entropy associated with the selection of a certain CM is a measure of cognitive complexity for the GDPD, being a local metrics that can be contextually computed for each CM selection activity.

Spatial Entropy. In addition to selection entropy, [23-24] introduced the notion of spatial entropy as a global measure to show how exhaustively/properly the problem space has been explored. Consequently, it is a reference measure that can be used to assess the trust in the preference for a CM. Note that GDPD is a dynamic design space, where CMs may appear/disappear over time, and consequently undermined by the risk to fall into local minima due to an improper exploration of the design space. For the GDPD, the spatial entropy was computed by substituting the classical probability associated with the preference for a CM with the ratio of problem types that used that CM in the GDPD to the total number of exploitable CMs. Thus, by assessing the selection and spatial entropies the suggestions provided by the system may be correlated with the following phase in which the CMM is structured [23]: *bootstrapping* (the exploration phase, when the preferences have no relevance), *development* (the structure formation phase, when the preferences have limited relevance) and *maintenance* (the exploitation phase, when the preferences have adequate relevance).

4 Interactions Scenarios with the CMM

This section details the functionalities for providing contextual, actionable recommendations based on the knowledge recorded in the CMM. As described in the previous section, GDPD is blend of mutually supporting activities (e.g., interpretation and exchange of information; creation and modification of the design alternatives; selection of feasible modeling alternatives; implementation and monitoring the GDP, etc.) that can be reduced to a subset of *elementary processes*: selecting the model, detailing the model and delegation of responsibilities. Note that these alternatives are not compulsory actions; the user can choose any of them during the GDPD space exploration.

4.1 An Example of Representing a GDP Design Alternatives as a Shared-Plan

In Fig. 3 is shown an example of structuring the GDP according to the SPT, where the CMs are the basic collaborative activities that may be used to construct the complex action associated with the group-decision problem of *"portfolio selection"*. The recipe of this complex action (or the GDP model) can be achieved by implementing one of the two alternatives of structuring the GDP: *"rational assessment"* or *"social assessment"*. These actions are complex and, consequently, they need to be subsequently decomposed into actions reflecting the sequence of CM types that compose the model. For example, the *"rational assessment"* complex action may be further decomposed into: *"criteria generation"* (to generate the evaluation criteria) → *"criteria selection"* (to select the most relevant ones) → *"weights assessment"* (to establish the criteria's weight) → *"project evaluation"* (to evaluate the projects against the relevant criteria) → *"project selection"* (to select the relevant projects). These final actions become executable when the corresponding CM or the recipe that implements the CM type is identified. For instance, the complex action *"criteria generation"* has CM1 as implementation recipe, while *"criteria selection"* CM2. Similarly with the situation when a problem may be modeled and abstracted in alternative ways (i.e. *"rational assessment"* and *"social assessment"*), there may be various implementation options for an action (i.e. *"weights assessment"* can be implemented either by the recipe provided by the CM3 or CM4).

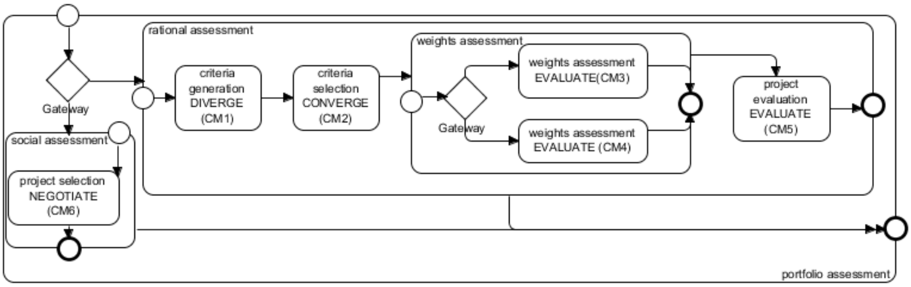

rational assessment
weights assessment
Gateway
criteria generation DIVERGE (CM1)
criteria selection CONVERGE (CM2)
weights assessment EVALUATE(CM3)
Gateway
weights assessment EVALUATE (CM4)
project evaluation EVALUATE (CM5)
social assessment
project selection NEGOTIATE (CM6)
portfolio assessment

Fig. 3. The GDPD for a "portfolio selection" problem in business process modeling notation

4.2 Model Selection

Selecting a GDP model implies the activities for the identification of a complete or partial model for a problem type. Simply, these activities involve the selection of a predefined model of a GDP. There are many possible interaction scenarios here, depending on the way in which the problem has been decomposed into sub-problems.

The Model Has Not Been Structured. If the GDP model is independent on other decisions, the selection corresponds to the classical situation of GDSS use when the facilitator selects a predefined model for the GDPD without any need to structure it. Nevertheless, the user interaction with the CMM is not limited to the selection of a complete model. The selection of a GDP model involves a dialogue with the user who generates, evaluates and modifies the GDP model.

For the *"portfolio assessment"* example, if we imagine the user is trying to identify a GDP model by querying the information recorded in tags (e.g., "assessment" or "rational assessment") there will be a list with many possible design models that corresponds to the keywords used to describe the problem. After selecting one of them, the system automatically generates all the design solutions for further investigation (Fig.

4). The solution space is a synthetic representation automatically generated and includes all the CMs and their relationships as they were experienced by users in solving this problem. Note that the generated solutions space can be very complex if the system has been used for very different problems types. To minimize this solutions space at a manageable level, the knowledge recorded in the CMM allows the implementation of several filtering options, such as to restrict the design space to those CMs experienced by the user in the past (i.e. the individual mental map) or those that provide a performance over a user-defined threshold. Moreover, the combination of local and spatial entropy (computed as mentioned in 3.2) may provide additional information as regards the level of trust in the performance of a CM as it has been evaluated in earlier uses of the system by its users. In Fig 4, these levels of trust (i.e. bootstrapping, development and maintenance; see 3.2) leave the users the entire responsibility to explore or exploit the design space. Consequently, some solutions may be relatively relevant and requires an assessment from the user side. In this case the evaluation entails the removal of irrelevant parts from the generated solutions.

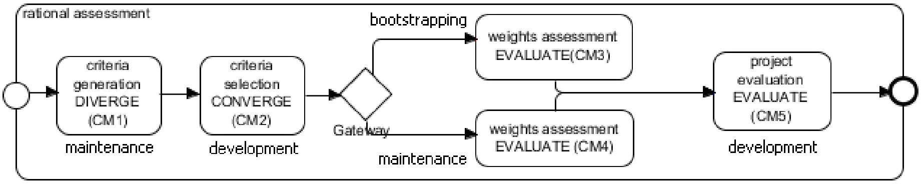

rational assessment
criteria generation DIVERGE (CM1)
maintenance
criteria selection CONVERGE (CM2)
development
Gateway
bootstrapping
weights assessment EVALUATE(CM3)
maintenance
weights assessment EVALUATE (CM4)
project evaluation EVALUATE (CM5)
development

Fig. 4. An example of the solution space generated from the CMM

The Model has been Structured. In the previous example, we considered the model selection is independent of the context in which the decision is made. Even for relatively simple models, this strategy leads to complex representations of the solutions space which are difficult to understand and manipulate. This is due to the impossibility of restricting the solution space relative to user's intentional stance. In the example from Fig. 3 there are only two alternatives for structuring the GDP model: *"social assessment"* and *"rational assessment"*. Both alternatives can be combined and decomposed themselves. For example, a *"rational assessment"* of the feasible projects may be followed by a *"social assessment"*, or a *"rational assessment"* can start immediately with the evaluation phase when the criteria and their corresponding weights are fixed or have been formerly concluded. Under these circumstances, when the solution space is very complex, the user is expected to firstly decompose the *"portfolio assessment"* problem into *"social or rational assessment"*, secondly decompose the *"rational assessment"* in *"criteria generation"* and *"selection"*, *"weights assessment"* and *"projects evaluation"* and so on (see Fig. 3). In this case, we are able to filter the identification of the GDP model depending on the context in which the query is made (inside a given abstraction, i.e. *"rational assessment"*) by adding implicitly the relevant tags that capture the user's intentional stance. On the other hand, the same problem may have models with different performances depending on the modeling context (e.g., *"social assessment"* may have as modeling alternatives CMs with different performances if it is executed after a *"rational assessment"* or not).

4.3 Model Elaboration

In contrast with the identification of a predefined solution, the elaboration of a GDP model corresponds to the design activities required to elaborate (create from scratch) a solution by defining: abstraction levels for the GDPD, CM types and CMs. Note that the identification and elaboration of a GDP model are essentially complementary processes that are combined during the design activities. For instance, one can choose a predefined pattern for the GDPD of a sub-problem, while another may develop a particular design solution. Choosing one of or combining them reflects the user's attitude and experience in GDPD (i.e. conservative or creative).

In Fig. 5 is shown a possible interactions flow with the GDPD tool for the *"portfolio assessment"* problem. We assume the user starts with a formal representation of the model by defining the CM types that compose the GDP (i.e. diverge → converge → evaluate → evaluate), and after that he/she is trying to generate the model by identifying the corresponding CMs. The system may support in this case both alternatives to choose a CM: either focalized on the structure and content of the results, obtained after the execution of each CM (upstream navigation in the CMM), or conversely, when the user is accounting for the available knowledge inside the group and consequently explores the possibilities to transform the obtained results (downstream navigation in the CMM). Obviously, in the context of a real GDPD, the user employs in a creative way both strategies.

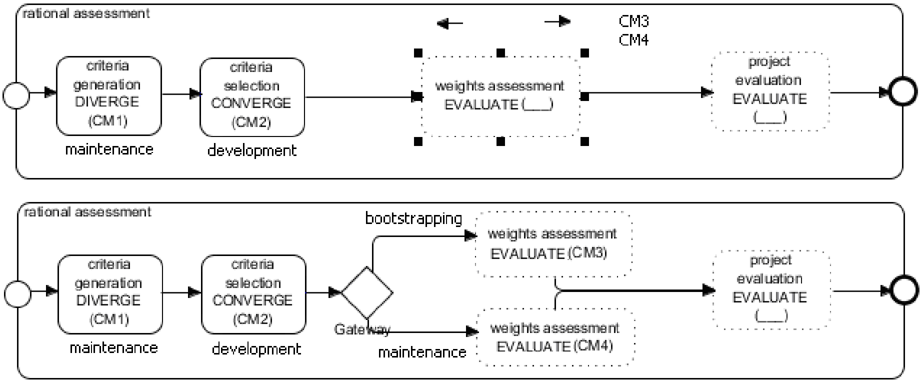

rational assessment
CM3
CM4
criteria generation DIVERGE (CM1)
maintenance
criteria selection CONVERGE (CM2)
development
weights assessment EVALUATE (___)
project evaluation EVALUATE (___)
rational assessment
bootstrapping
criteria generation DIVERGE (CM1)
maintenance
criteria selection CONVERGE (CM2)
development
Gateway
weights assessment EVALUATE (CM3)
weights assessment EVALUATE (CM4)
maintenance
project evaluation EVALUATE (___)

Fig. 5. An example of interacting with the CMM during the model elaboration

The implementation of this mechanism is an immediate consequence of the user's navigation in the CMM structure of the GDPD space. Unlike the case when the GDP is found as a result of an identification process, in this situation the suggestions offered by the system consider the outward performances from the selected CMs. Thus, the user can choose one or more of the suggested alternatives for modeling the GDP (e.g., in Fig. 5 the user has selected both modeling suggestions for implementing the CM type *"evaluation"*: CM3 and CM4) which will be reflected in the final model through the "xor" node. Note that this process is iterative, wherein the user generates, evaluates and adjusts the GDP model according to his objectives of using the GDSS.

4.4 Delegation of Responsibilities

Delegation of responsibilities may be performed at different levels of GDPD, from a user to an entire group. As mentioned, the CMM records data related to users' experience and their degree of past collaborations at the CMs level. Consequently, the major aim of this process is to identify the users who are able to design the GDP and/or guarantee a certain level of collaboration performance according to social sciences theories (i.e. adaptive structuration theory, activity theory, etc.). The computational elements described in section 3.2, aim at identifying the users who are indirectly maximizing the degree of sharing a common mental model for the GDPD. This means that the initiator of a group decision should start with a formal representation of the GDPD. The designed model will form the basis for the identification of users who are maximizing its performance as a result of considering their past experience ($UD^{<Group>}$) and collaborations ($CD^{<Group>}$).

5 Conclusions

In this paper we investigated the multiple ways of interacting, during the GDPD, with a complex knowledge structure (i.e. CMM) of a well-known class of ill-structured, dynamic, and going-concerns problem. Even if the GDPD is a blend of mutually supporting activities (e.g., interpretation and exchange of information; creation and modification of design alternatives; selection of feasible modeling alternatives; implementation and monitoring the GDP, etc.), it can be reduced to a subset of elementary processes (i.e. selecting the model, detailing the model and delegation of responsibilities) that can be guided through personalized, contextual and actionable recommendations. Through a representative set of interaction scenarios with the CMM, we have shown how these recommendations may be computed from some basic data recorded in the CMM. Clearly all the theoretical issues presented in this paper need to be further investigated in real world settings, where many practical concerns related to its deployment, acceptability and usability require an adequate answer.

References

1. Vreede, G.J., Boonstra, J.A., Niederman, F.: What Is Effective GSS Facilitation? A Qualitative Inquiry Into Participants' Perceptions. In: 35th Hawaiian Int. Conf. on System Sciences. IEEE Press, Los Alamitos (2002)
2. Lagroue, H.: The effectiveness of virtual facilitation in supporting GDSS appropriation and structured group decision making. Doctoral dissertation, Louisiana University (2006)
3. Sallam, R.L.: Who's who in collaborative decision-making. Research Note No. G00214928. Gartner, Stamford (2011)
4. Heiko, T.: Cloud-Based Collaborative Decision Making. J. of Decision Support System Technology 4(4), 39–59 (2012)
5. Turoff, M., Hiltz, S., Cho, H., Li, Z., Wang, Y.: Social decision support system. In: 35th Hawaiian Int. Conf. on System Sciences. IEEE Press, Los Alamitos (2002)

6. Rodriguez, M.A., Steinbock, D.J.: Group Holographic Modeling for Societal-Scale Decision-Making Systems. In: N. American Assoc. for Computational Social and Organizational Science Conf., Pittsburgh (2004)
7. Briggs, R., Kolfschoten, G., Vreede, G.J., Albrecht, C., Lukosch, S.: Facilitator in a Box. Information Systems, 1–10 (2010)
8. Zamfirescu, C.B., Duta, L., Candea, C.: Implementing the "Design for Emergence" Principle in GDSS. In: Frontiers in AI and Applications, vol. 212, pp. 61–72. IOS Press (2010)
9. Zamfirescu, C.B., Duta, L., Iantovics, B.: On Investigating the Cognitive Complexity of Designing the GDP. Studies in Informatics and Control 19, 263–270 (2010)
10. Zamfirescu, C.B., Filip, F.G.: Swarming models for facilitating collaborative decisions. International J. of Computers, Communications & Control 1, 1841–1844 (2010)
11. Parunak, V.D., Brueckner, S.A.: The Cognitive Aptitude of Swarming Agents (2009), https://activewiki.net/download/attachments/6258699/CASA.pdf
12. Rosen, D., Suthers, D.D.: Stigmergy and collaboration: Tracing the contingencies of mediated interaction. In: 44th Hawaiian Int. Conf. on System Sciences. IEEE Press (2011)
13. Heylighen, F.: Collective Intelligence and its Implementation on the Web: algorithms to develop a collective mental map. J. Comput. and Math. Org. Theory 5, 253–280 (1999)
14. Parunak, H.V.D., Brueckner, S.A., Matthews, R., Sauter, J.: Swarming methods for geospatial reasoning. Int. J. of Geographical Information Science 20(9), 945–964 (2006)
15. Briggs, R.O., Vreede, G.J., Nunamaker Jr., J.F.: Collaboration Engineering with ThinkLets to Pursue Sustained Success with GSS. J. of Management Inf. Systems 19, 31–63 (2003)
16. Ender, G.: OpenSpace-Online: State-of-the-art Real-Time Conferencing (2009), http://www.openspace-online.com/OpenSpace-Online_eBook_en.pdf
17. Satnam, A.: Collective intelligence in action. Manning Publications (2008)
18. Zamfirescu, C.B., Candea, C.: A Stigmergic Guiding System to Facilitate the Group Decision Process. In: 28th IEEE Int. Conf. on Data Engineering, pp. 98–102. IEEE Press, Washington (2012)
19. Grosz, B., Kraus, S.: Collaborative plans for complex group action. Artificial Intelligence 86, 269–357 (1996)
20. Grosz, B., Kraus, S.: The Evolution of Shared Plans. In: Rao, A., Wooldridge, M. (eds.) Foundations of Rational Agency, pp. 227–262. Kluwer Academic Press (1999)
21. Ghallab, M., Nau, D., Traverso, P.: Hierarchical Task Network Planning. Automated Planning: Theory and Practice. Morgan Kaufmann (2004)
22. Object Management Group: Business Process Model and Notation (BPMN) Version 2.0. Document No. formal/2011-01-03 (2011)
23. Guerin, S., Kunkle, D.: Emergence of constraint in self-organizing systems. Nonlinear Dynamics, Psychology, and Life Sciences 8(2), 131–146 (2004)
24. Parunak, V.H.D., Brukner, S.: Entropy and Self-Organization in Multi-Agent Systems. In: Fifth Int. Conf. on Autonomous Agents, Montreal, pp. 124–130 (2001)

Automatic Lexical Alignment between Syntactically Weak Related Languages. Application for English and Romanian

Mihaela Colhon*

Departament of Computer Science, Faculty of Exact Sciences, Craiova România
mcolhon@inf.ucv.ro

Abstract. In this paper we describe an alignment system that takes English-Romanian parallel sentences (*bitexts*) and aligns them at their content-word level. A syntactic feature approach combined with a dictionary lookup is used as primary technique to perform word alignments. Other used methods take into account local word grouping or the nearest aligned neighbors approach to filter between many-to-many word alignments. Building an alignment system at the word level, one can use it in the creation of new resources, for example collections of parallel sequences of texts in the two languages based on which translation schemes could be learned.

Keywords: Machine Translation, Alignment Scores, Word Aligner.

1 Introduction

Parallel texts form a very valuable resource for several Natural Language Processing difficult tasks, such as Machine Translation [15], Cross Language Information Retrieval or Word Sense Disambiguation [5]. While resources like bilingual dictionaries and parallel grammars help to improve the quality of Machine Translation (MT) systems, by aligning two texts in different languages at various levels (i.e. documents, sections, paragraphs, sentences and words) we can help the creation of new parallel lexical resources [6].

The sentence alignment process maps sentences in the Source Language to their translation in the Target Language. A precise evaluation of the techniques would be most valuable [8]. Yet, opposed to Speech Recognition, Information Retrieval or Message Understanding systems, there is no established framework for

* The author M. Colhon has been funded for this research by the strategic grant POSDRU/89/1.5/S/61968, Project ID 61986 (2009), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007-2013. Also, the author would like to thank the Natural Language Group of Faculty of Computer Science, Al.I. Cuza University of Iaşi, Romania and to "Mihai Drăgănescu" Research Institute for Artificial Intelligence team for providing the resources upon which this study was developed.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 266–275, 2013.

the evaluation of parallel text alignment [14]. Recent work deal with mining parallel sentences from parallel corpora or even non-parallel but comparable corpora (such as [7] and [10]). The system presented in [7] uses a dictionary to translate some of the words of the source sentence, and then uses these translations to query a database for finding translation candidates. *LEXACC* - Lucene-Based Parallel Sentence Extractor from Comparable Corpora [10] was developed to work on comparable corpora and obtained state-of-the-art results in comparison with established approaches [10].

The motivation for our work comes from the fact that word-aligned parallel corpora do not generally exist. This paper presents a method for identifying corresponding words in parallel texts, a task known as *word alignment*. The implemented approach focuses on the content words of the two languages and on some functional words such as numerals or conjunctions. The algorithm makes use of an English-Romanian dictionary and exploits phrase annotation in order to find, for a given source word for which alignments have to be determined the translation equivalent in the target language[1]. As input, it requires only a bilingual corpus pre-aligned at the sentence level and with POS data annotations for the words of both languages.

The results of this study could be used in order to derive alignments at the sentence level that work on comparable corpora. More precisely, as it is stated in [10], content-word alignments could be used to make a Boolean query which is then passed to a search engine to find a list of translation candidates for sentence alignment. Also, building this alignment system, one can use it in the creation of new resources, for example collections of parallel sequences of texts in the two languages based on which translation schemes could be learned.

The word aligner presented in this paper takes as input a pair of aligned sentences and aligns words of the two sentences. The alignments are determined based on an English-Romanian dictionary sorted by translation scores that was generated with the TREQ module [12]. In the alignment technique we do not use non-lexical measures, such as sentence length information, but we made use of alignment distances between semantically related words. We implemented several translation features that are already settled in the literature such as *POS affinity* [11], *Locality* [13] or *Relative position/Distortion* [9]. The evaluations show comparable results with other word-aligners developed for the same pair of languages. YAWA [13] is a three stage lexical aligner that uses bilingual translation lexicons and phrase boundaries detection for the sentences in both languages in order to align words of a given bitext. The three phases of the YAWA aligner are: *Content Words Alignment*, *Chunks Alignment* and *Dealing with sequences of unaligned words*.

MEBA [13] uses an iterative algorithm that takes advantage of all preprocessing phases needed by YAWA. As an improvement with respect to the

[1] Because this alignment study is developed in view of a MT development, we use terminology from the MT applications, such as source word/language (the word/language that is to be translated) and target word/language (the word/language in which translation is made).

YAWA aligner, in MEBA each of the iterations can be configured to align different categories of tokens (named entities, dates and numbers, content words, functional words, punctuation) in decreasing order of statistical evidence. In MEBA are implemented several translation features such as *Similarity Measure*, *Obliqueness*, *Locality* and *Collocation detection* for both languages.

Our word-aligner does not request pre-processing phases over the two languages. For the source language (English, in our case) it necessitates a good syntactic parser. Then, all the possible alignments between English and Romanian words, as given by an English-Romanian translation dictionary, are marked. Then some filtering methods are implemented in order to select the correct alignment(s) from the set of initial possible ones. As it is shown in the Evaluation section, for an English-Romanian corpus consisting of 1400 parallel sentences, the precision of the proposed word aligner are comparable with YAWA and MEBA processions.

1.1 The Corpus

The alignment algorithm was implemented to work on an English-Romanian corpus, in which sentences are aligned one-to-one and the words are annotated with morpho-syntactic data. We found that the English-Romanian parallel corpus developed based on the English and Romanian parts of the Acquis-Communitaire corpus is a perfect resource for the algorithm requirements. All the words of this bilingual corpus are annotated with lemmas, morphosyntactic information (gender, number, person and case) and Part of Speech markers. The tagsets used to annotate the words of the English-Romanian corpus come from MULTEXT-East morphosyntactic specifications, 3rd Version .

As it was already specified, in the presented approach, a bilingual dictionary is used to build possible word-pairs translation. The used dictionary does not necessary pairs the base forms of English words with the base forms of their corresponding Romanian words. As English language has a poor morphology, the base form of its words results immediately. But, for Romanian words a morphological tool [3] to find Romanian words lemma is needed.

The inflection simplicity in English makes that the majority of researchers in the field of computational linguistics neglect the inflection morphology [1]. This is not the case for Romanian where the need for suitable computational models of morphology asked for a Romanian morphological dictionary[2], resource that is currently available thanks to the efforts of the Natural Language Processing Group of Faculty of Computer Science from Al. I. Cuza University of Iaşi.

For this version, the aligner takes into account only some special words, named in the followings as *anchor words*. The group of anchor words includes all content words of the two languages and some functional words (such as numerals and conjunctions). All the other functional words that could occur in the sentences are discarded.

[2] http://instrumente.infoiasi.ro/WebPosRo/resources/posDictRoDiacr.txt

Also considered are n-grams in both languages as the word-aligner does not take into account only 1:1 word alignments, but the general case of $m_1 : m_2$ alignments, for m_1, $m_2 \geq 1$. No patterns are used to identify n-grams. The only condition imposed is that they should be contiguous sequences of words.

2 The Word Aligner

The word aligner takes as input a pair of aligned sentences and finds the best translation correspondences for a certain category of words, named *anchor words* which include all the content words of the two languages such as nouns, verbs (non-auxiliary forms), adjectives and adverbs and some functional words such as numerals, conjunctions and all the abbreviations[3]. The translation correspondents for the English words are primarily obtained with a *Dictionary Lookup Approach.* Others approaches are used for filtering in case of multiple possible translations and are grouped under the name of *Filtering Methods for Multiple Alignments.* The alignments for each English word take into account the words full form and their part-of-speech (shortly, POS) data. Frequently, the translation equivalents have the same POS but relying on such a restriction would seriously affect the alignment recall [12]. For this reason, our alignment model makes use of some translation compatibilities rules for pairs of source and target words based on their POS data.

Remark 1. As the translation scores are determined based on a bilingual translation lexicon, this implies that these measures depend greatly on the linguistic resource quality and completeness [10].

Nevertheless, the translation dictionary was obtained by running GIZA++ on the English and Romanian parts of JRC-Acquis corpus - the same corpus upon which this word-aligner algorithm was tested. Thus, for the presented study, we consider that the dictionary is properly constructed in order to cover the lexical constructions upon which the algorithm is running.

Each pair of parallel sentences (s, t) is represented as: $s = (w_1^s, \dots, w_n^s)$ and $t = (w_1^t, \dots, w_m^t)$ where $w_1^s, \dots, w_n^s$ are the anchor words that appear in the source sentence s and $w_1^t, \dots, w_m^t$ are the anchor words of the target sentence t, in the orders they occur in the sentence $n, m > 0$.

For two parallel sentences (s, t), the probability score that the English word w_i^s from the sentence s is translated with the Romanian word w_j^t in the sentence t will be noted in what follows by $P_{(s,t)}(w_j^t | w_i^s) \in [0, 1]$.

The case of at least two consecutive Romanian translations $w_{j_1}^t, \dots, w_{j_k}^t$ for a single English word w_i^s will be noted by $P_{(s,t)}(w_{j_1}^t, \dots, w_{j_k}^t | w_i^s)$, $k \geq 2$.

[3] We choose to include these kind of functional words in the alignment process because usually these words have 1:1 translation alignments and thus they could help finding the correct alignment for the neighbor content words with multiple words translations

Proposition 1. For two parallel sentences $s = (w_1^s, \ldots, w_n^s)$ and $t = (w_1^t, \ldots, w_m^t)$ given as lists of anchor words, we note the set of the alignments by:

$$A \subseteq (w_1^s, \ldots, w_n^s) \times (w_1^t, \ldots, w_m^t)$$

We note the list of the source words aligned with at least one target word by $pr_{(w_1^s,\ldots,w_n^s)}A$. We have $pr_{(w_1^s,\ldots,w_n^s)}A = \{w_i^s \mid \exists(w_i^s, w_j^t) \in A, 1 \leq j \leq m\}_{1 \leq i \leq n}$.

For any pair of parallel sentences (s,t), the translation candidates w_j^t for a source word w^s are determined based on the *translation probability scores* taken from the dictionary as follows.

Proposition 2. For each w_j^t an alignment candidate for an English word w^s, we note $(w^s, w_j^t) \in A$ if and only if there is an entry (w^s, w^t) in the dictionary such that score$(w^s, w^t) > 0.001$[4] and $lemma(w^t) = lemma(w_j^t)$.

In order to set the correct alignments for a particular English content word w^s, we do not select the best 1:1 translation from a set of possible translations $w_1^t, \ldots, w_k^t$ such that the translation probability score is maximum. Instead, we chose to give an advantage to those w_j^t that exactly match with a form given in the dictionary as translation equivalent for w^s.

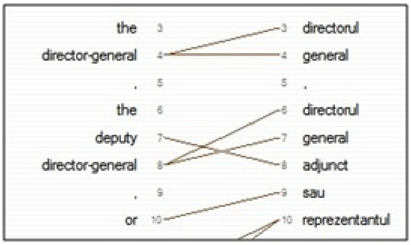


Fig. 1. Multiple alignments for a single English word

Using the introduced notations we define the score for an English word w^s to be translated by w^t, in the context given by the parallel sentences pair (s,t) as follows:

$$P_{(s,t)}(w^t \mid w^s) = \begin{cases} 0, & \text{if score}(w^s, w^{t'}) < 0.001, \\ 1, & \text{if } w^t = w^{t'}, \\ \text{score}(w^s, w^{t'}), & \text{if lemma}(w^t)\text{=lemma}(w^{t'}). \end{cases}$$

where $w^{t'}$ is a translation candidate for w^s as it results from the bilingual dictionary.

[4] In this alignment algorithm, we do not align words for which the translation score is smaller than 0.001, but for further refinements we consider to take into considerations also these cases but only for unaligned pairs of words.

The aligner has to take into account the cases in which an unique English word is aligned with more than one Romanian word. For example, in Figure 1, the English word "director-general" is considered aligned with the Romanian words "directorul" and "general". The score in case of multiple Romanian translations $w^t_{j_1}, \ldots, w^t_{j_k}$ for a single English word w^s_i is calculated as follows:

$$P_{(s,t)}(w^t_{j_1}, \ldots, w^t_{j_k} \mid w^s) = max\{P_{(s,t)}(w^t_{j_1} \mid w^s), \ldots, P_{(s,t)}(w^t_{j_k} \mid w^s)\}$$

After attaching scores, the possible translations for a source word are passed to the *Filtering Methods for Multiple Alignments* ($FMMA$) in order to extract the correct alignment(s). The implemented filtering methods are described in the next section.

3 Filtering Methods for Multiple Alignments

Even if the two sentences are treated from the point of view of their lists of anchor words, word-ordering inside the sentence is very important when the algorithm has to choose an alignment against another (the so-called *Distorsion Locality property* [9]).

As it is said in [4], word alignments should benefit from phrase information as these kind of syntactical information can provide contextual translation information. This property is referred to in the literature as *locality* property [11]. Thus, when we have to determine the correct alignment for a given English word w_i, the algorithm first checks the alignments of the neighbour anchor words w_j within the limits given by the syntactic phrases that include w_i.

$FMMA_1$*:* **Scenario:** This filtering is applied for source words w^s which has more than one possible alignment.
Method: The right alignment from a set of multiple possible alignments is considered to be the one that is at a minimum distance with respect to the alignments of the closest anchor word from of w^s.

In Figure 2 it is shown a successfully application of $FMMA_1$. It is the case of the possible alignments: ("sole", "doar"), ("sole", "unica") $\in A$ and the algorithm correctly removes from the set A the alignment ("sole", "doar").

$FMMA_2$ **Scenario:** This filtering is applied in the case of two different source words that share the same possible alignment.
Method: If two source words w^s and $w^{s'}$ have the same possible alignment: $(w^s, w^t), (w^{s'}, w^t) \in A$ and (w^s, w^t) is the only possible alignment for w^s with a greater score than the one corresponding to $w^{s'}$, that is $P(w^t \mid w^s) > P(w^t \mid w^{s'})$ then the alignment $(w^{s'}, w^t)$ is removed from A.

In Figure 3 is shown a case in which $FMMA_2$ was applied successfully on the set of possible alignments A as it results from the dictionary:
("may", "Poate") $\in A$, P("Poate" | "may") $= 1$ because the dictionary contains the entry ("may", "poate")

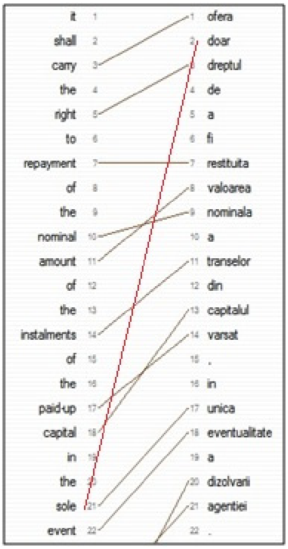


Fig. 2. $FMMA_1$ correctly removes an alignment

(“may”, “facă”) $\in A$, P(“fac” | “may”) = 0.002149 because the dictionary contains the entry (“may”, “face”), lemma(“facă”)=lemma(“face”)= “face”

(“also”, “de_asemenea”) $\in A$, P(“de_asemenea” | “also”) = 1 because the dictionary contains the entry (“also”, “de_asemenea”)

(“also”, “Poate”) $\in A$, P(“Poate” | “also”) = 0.0015749 because the dictionary contains the entry (“also”, “pot”), lemma(“poate”) = lemma(pot) = “putea”

(“raise”, “fac”) $\in A$, P(“facă” | “raise”) = 1 because the dictionary contains the entry (“raise”, “facă”)

The following alignments were removed by $FMMA_2$:

(“may”, “facă”) was removed from A because of the alignment (“raise”, “facă”)

(“also”, “Poate”) was removed from A because of the alignment (“may”, “Poate”)

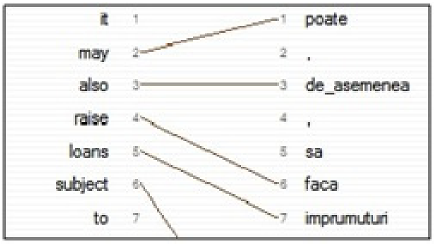


Fig. 3. $FMMA_2$ finds the right alignment for “may”, “also” and “raise”

In the case we still have more than one possible alignment for a given English word w^s after applying $FMMA_1$ and $FMMA_2$ conditions, the bounds limits given by the *nearest aligned word preceding it* and *the nearest aligned word following it* are used in order to find the right alignment. This approach is determined by the well-known fact that words of a sentence do not translate independently of each other.

$FMMA_3$ **Scenario:** This filtering is applied for source anchor words that have more than one possible alignment, for which the filtering based on the closest anchor word alignment (the $FMMA_1$ filter) could not be applied.
Method: For the alignment algorithm we have to consider the extreme sparseness of inversion, i.e., for example, sequences a, b being reversed as b, a inside two parallel sentences during translation. The natural word order of the considered pair of languages could determine crossing word alignment links as it is exemplified in [2]. Thus we have to consider some alignments that exceed some of the previous or of the next neighbours anchor words but we have to filter away the wrong alignments that do not match with any of the neighbour's alignments.

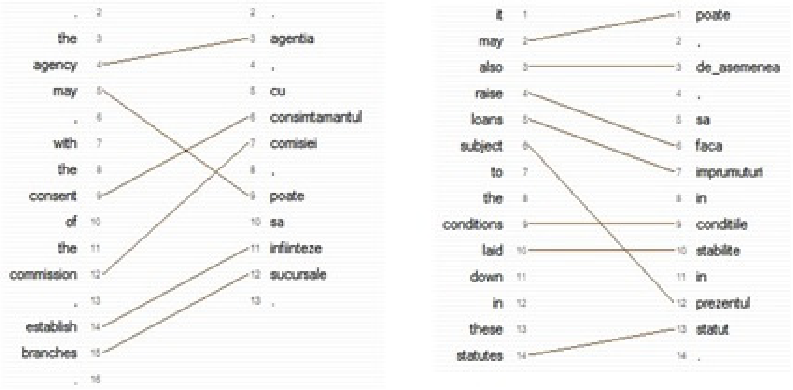


Fig. 4. Alignments of the source words that exceed the alignments of the neighbors

In Figure 4 it is exemplified a correct alignment (left) and a wrong alignment (right) both of them exceeding some of the neighbour alignments. After applying all the filtering conditions on the set of possible alignments A acquired from the dictionary, we get the final set of alignments for the considered pair of parallel sentences and we note this set by A_{final}. We have:

$$A_{final} = FMMA_3(FMMA_2(FMMA_1(A)))$$

Following the translation similarity measure for two sentences introduced in [10], we define a measure that takes into account only the anchor-word alignments between two parallel sentences (s,t):

$$P(s,t) = \sum_{w_i^s \in pr_1(A_{final})} P_{(s,t)}(w_1^t, \ldots, w_m^t \mid w_i^s)$$

where the two sentence are reduced to their anchor-words: $s = (w_1^s, \ldots, w_n^s)$ and $t = (w_1^t, \ldots, w_m^t)$.

4 Evaluation

Like in most Natural Language Processing or Pattern Recognition systems, the performance of the proposed word aligner is measured in terms of *Precision*, *Recall* and *F-measure*. In this specic context, these concepts can be simply explained as follows: given a corpus of parallel sentences, C, there is a number C_{total} of correct alignments between them. After aligning the texts with the word aligner, $Match(C)$ represents the total number of correct correspondences found by the aligner in C while $Unmatch(C)$ represents the number of incorrect correspondences found by the aligner in the corpus C. The precision $P(C)$ and recall $R(C)$ are then given by: $P(C) = Match(C)/(Match(C) + Unmatch(C))$ and $R(C) = Match(C)/C_{total}$. The *F-measure* is the harmonic mean of Precision and Recall: *F-measure* $= 2 * P * R/(P + R)$

Table 1. Evaluation Results

No. sentences	C_{total}	$Match(C)$	$Unmatch(C)$	$P(C)$	$R(C)$	*F-measure*
435	4638	4287	422	0.9103	0.9243	0.9172
1400	15112	14084	1305	0.9151	0.9319	0.9234

Table 1 presents the results obtained, first on a smaller set of sentences (435 sentences), and then on a bigger corpus consisting of 1400 sentences. As it can be seen, there are small differences in terms of Precision and Recall for the two test sets which means that the algorithm is stable on bigger data. Even if the scores are below the ones obtained in a similar scenario by YAWA (Precision: 0.9850) and MEBA (Precision: 0.9408), we consider that the performance of the proposed algorithm is good, taking into account that it uses only preprocessing tools for Romanian that are currently available in the scientific community (more precisely, it needs only POS annotations for the Romanian texts).

5 Conclusions and Future Work

This paper describes a dictionary-based parallel sentence word aligner. The aligner considers only the anchor words (all the content words and some functional words of the considered languages) of two parallel sentences. Information about the sentence's length is not used, as this type of approach is considered knowledge-poor.

In the future, if we have a pair of parallel sentences (s, t) such as $0.8 < P(s, t) < 1$, then the algorithm has to explore all possible combinations of source and target unaligned words in order to identify new possible alignment pairs.

References

1. Alhazov, A., Boian, E., Cojocaru, S., Rogozhin, Y.: Modelling Inflections in Romanian Language by P Systems with String Replication. The Computer Science Journal of Moldova 17(2), 160–178 (2009)
2. Colhon, M.: Language Engineering for Syntactic Knowledge Transfer. Computer Science and Information Systems Journal (ComSIS) 9(3), 1231–1248 (2012)
3. Cristea, D., Simionescu, R., Haja, G.: Reconstructing the Diachronic Morphology of Romanian from Dictionary Citations. In: Proceedings of Conference on Language Resources and Evaluation, LREC 2012 (2012)
4. Holmqvist, M.: Heuristic word alignment with parallel phrases. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation, LREC 2010 (2010)
5. Ma, X.: Champollion: A Robust Parallel Text Sentence Aligner. In: 5th International Conference on Language Resources and Evaluation, LREC 2006, pp. 489–492 (2006)
6. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MITPress, Cambridge (2003)
7. Munteanu, D., Marcu, D.: Improving Machine Translation Performance by Exploiting Comparable Corpora. Computational Linguistics 31(4), 477–504 (2005)
8. Santos, A.: A survey on parallel corpora alignment. In: MI-Star 2011, Braga, Portugal (2011)
9. Ştefănescu, D.: Intelligent Information Extraction of Multilingual Corpora. PhD Thesis. Romanian Academy. Research Institute for Artificial Intelligence (2010)
10. Ştefănescu, D., Ion, R., Hunsicker, S.: Hybrid Parallel Sentence Mining from Comparable Corpora. In: Proceedings of the 16th Conference of the European Association for Machine Translation, EAMT 2012, pp. 137–144 (2012)
11. Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D.: Combined word alignments. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, pp. 107–110. Association for Computational Linguistics (June 2005)
12. Tufiş, D.: From Word Alignment to Word Senses, via Multilingual Wordnets. The Computer Science Journal of Moldova - CSJM 14(1), 3–33 (2006)
13. Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D.: Improved Lexical Alignment by Combining Multiple Reified Alignments. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics - EACL 2006, Trento, Italy, pp. 153–160. Association for Computational Linguistics (April 2006) ISBN 1-9324-32-61-2
14. Véronis, J., Langlais, P.: Evaluation of parallel text alignment systems. The *ARCADE* project. In: Véronis, J. (ed.) Parallel Text Processing, pp. 369–388. Kluwer Academic Publishers, The Netherlands (2000)
15. Yamada, K., Knight, K.: A syntax-based statistical translation model. In: 39th Meeting of the Association for Computational Linguistics ACL 2001, Toulouse, France, pp. 523-530 (2001)

Monitoring and Predicting Journalistic Profiles

Daniela Gîfu[1] and Dan Cristea[1,2]

[1] "Alexandru Ioan Cuza" University, Faculty of Computer Science,
16, General Berthelot St., 700483, Iaşi
{daniela.gifu,dcristea}@info.uaic.ro
[2] Institute for Theoretical Computer Science, Romanian Academy - Iaşi branch,
2, T. Codrescu St., 700481, Iaşi

Abstract. The paper develops a pilot study aiming at finding a methodology for identifying features of journalistic writing profiles. The study is based on capturing dominant discursive tonalities, knowing that journalistic discourse entails public legitimacy. We made use of a series of natural language processing tools as preliminary steps in revealing three egocentric journalistic identities. Based on quantitative analysis, such as syntactic and lexical-semantic, we put in evidence qualitative pragmatic aspects. The final goal of this study is to configure a tool for automatic analysis of journalistic profiles. We have concentrated on three top Romanian journalists, whose newspaper publications have been monitored over a period of three months and semi-automatically analyzed. As a result, a database with their profiles is collected and interpreted. Such a tool could be of interest to mass-media, but also to specialists in communication and public relations, to political parties and to the public opinion, in general.

Keywords: journalism, lexicon, syntactic analysis, semantic analysis, egocentrism features.

1 Introduction

The automatic identification of the features of journalistic profiles, although in the attention of researches in Natural Language Processing, is currently a problem only partly solved. We rely on the human capacity to assign significances[1][21] of words, accepting the argument that the human beings have the potential to understand the natural language. That is why, current methods for identification of discursive tonalities, by using statistical or symbolic algorithms, approximate the human ability to classify, to identify the author's position and the nature of his opinions. The interest is focused on identifying the linguistic modalities to persuade or seduce an auditorium about the truth of the speaker/writer's ideas. A well-crafted argument in the architecture of a discourse can reveal the intellectual nature of a journalist which has as principal purpose to serve editorial policies through his articles.

A technique for profiling the author of a text is given by the use of linguistic markers that can reveal much more about the author's personality, often, more than he

[1] Here, the significance concept is synonymous with understanding.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 276–285, 2013.

reveals by other ways (e.g. in an interview). This approach emphasizes, in fact, the importance of using a natural language processing system able to extract basic linguistics features on a large amount of texts, that can be arranged as a collection of pragmatic knowledge in order to inventory the features of journalistic profiles.

Such a methodology could be useful for communication and public relations specialists who build, suggest, etc. discursive structures of different public actors that represent them (press, politicians, economists, secret services and so on). One of the reasons for the effort to represent mentally journalistic portraits (identifying the discursive nature of a journalist) is important in improving communication. Another reason is that it can help to improve the editorial policies and thus to offer a better adequacy of the journalistic thesis to the expectations of the community and the needs they represent.

Section 2 presents the state of the art. Section 3 depicts a new vision about linguistics boundaries between *I* (*we*) and *you*, then in section 4, after a short description of the corpus analyzed, we present the methodology applied in identifying the *egocentrism* features of the journalist profile. Finally, Section 5 presents some conclusions and directions for the future work.

2 State of the Art

Our study combines tools that enable classification of the texts [7] and automatic recovery methods of text information monitored applicable pragmalinguistics studies that is used in journalistic language in order to identify features of journalistic profiles [15], [22]. Of major importance are some researches based on the collection of new media texts, used in identifying characteristics underlying implementation of text classifiers [11], [17], [13]. A comparative analysis of *n* texts signed by *n* different journalists provides the criteria or heuristics which can be implemented in a platform for natural language processing for characterization of each document and each signatory of text. When two or more collections of texts, each having a single author are compared, it can be determined statistically, to which typology it belongs. These criteria of differentiation can be: the use of longer sentences, with specific ways of structuring the sentence/paragraph; the use of special means to divide words at the end of the row, using the abbreviations (with points between each letter, with no points, all letters are uppercase), the use of particular expressions or phrases with a significantly higher frequency than the common language, using words and/or phrases in another language, use of quotes; using a specific number/relationship of words (adjectives, adverbs) or preferences for specific classes of pronouns; an author prefer some specific morphological variations and spellings of words (replacing diacritics with similar groups of letters); the use of punctuation, or emoticons that become elements of nonverbal or paraverbal language etc.

Simple features such as the (punctuation, functional words, and n-grams) are too common to be used in order to characterize the style of an author. Some researchers use the frequency of part-of-speech (POS) tags [14] to classify user profiles from a transcribed dialogical relationship. Features like: average number of words per sentence, average number of letters in a word and use of punctuation marks are often encountered in the studies leaned towards the profile identification of authors [4].

In this context, it is important to remember the lexical aspects and types of discourse (including the choice of words, the type of participation in communicative act) for the identification of the discursive features [6]. A number of linguistic and pragmatic studies aiming the identification of characteristics of a particular type of speech, or an author [10] make reference to: argumentation [3], preference for the personal pronouns (first person, singular and plural and second person plural), exclamations and rhetorical questions, etc. Some platforms are known as to man-machine mediated interaction [2], [12], based on automatic learning to enhance the skills of detection for attributes in the identification of authors profiles for certain texts [20].

One aspect in our study is the profiling of a discursive style starting from the type of opinion and how it is expressed, manual annotation being a first step in this sense. In the production of the text to which we make reference, a journalist is expressing her/his opinions about a particular subject and interacts with an imaginary audience. Content analysis is used in many applications to capture potential conflicts or to detect discursively the various opinions of the signatory [8]. For the Romanian language it was introduced a similar system of annotation [18], which does not exclude, however, subjective aspects. These approaches use specific dictionaries [9] or emotional intensity [5].

By making these choices we can trace the socio-pragma-linguistics profile of the enunciator, otherwise a part of our purpose, correlating certain computational techniques with pragmalinguistics theories.

3 Pragmalinguistics Boundaries

If social boundaries are strategies of social policy, by extrapolation, pragmalinguistics boundaries are strategies of discursive intentionality. These expression processes are involved in the construction of the identity of the speaker, being it wilfully, coerced or involuntary. We have introduced here the *boundary* concept just to clarify that we are not talking of limit[2] but the separation of the two protagonists, which define an act of communication (transmitter and receiver). The representation of the communication between the two parts of this split communication space is here decisive: "boundaries are not air tight, are never occlusive, but more or less fluid, moving and permeable [19]. The dividing line [1], which separates "I"/"we" from "you", is vital for their identity. For example, the relation "I-we" allows the definition of membership of the journalist to the editorial team, known as the signers of the articles of the print media apt their speech to the political editorial which they represent.

Pragmalinguistics boundaries become tools that make it possible to identify inclusions and potential exclusions from the editorial board (we exemplify with the journalist Ion Cristoiu, who, for a while, was one of the most vocal opponents of media that he serves now). In Figure 1, we have sketched the process of communication initiated by a journalist: the publication – represented by “I”/"we" – is addressed to the public opinion – "you". The linguistic boundaries, in fact pragmalinguistics, are distinguishing traits that the addressee of the article would

[2] Limitations in the communication sciences, known as communicational barriers, are a series of obstacles that arise between transmitter and receiver and that reduce the efficiency of messages sent within a process of communication.

want to identify in the language of the journalist, but which can also be a barrier for deciphering the correct message. This may be due either to a wilful intent on the part of the media player (the journalist or the publication, through its policies) or an unconscious intention (the use of a journalistic code, understood by the receiver[3]). The intentioned discursive purpose when using "I" is a modality of shaping a pragma-discursive dimension.

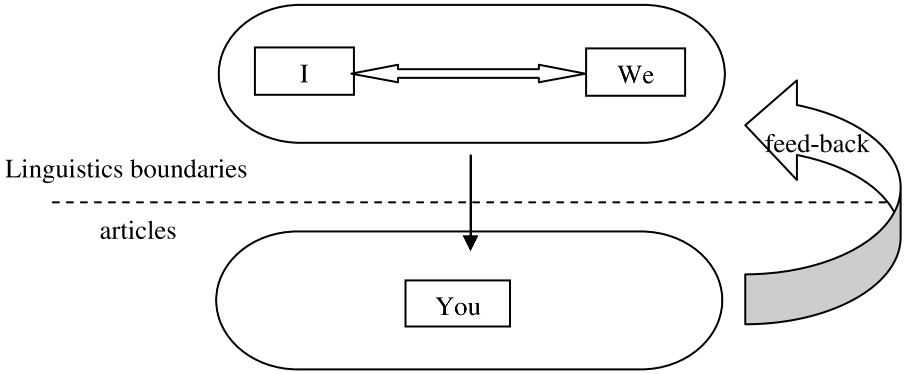


Fig. 1. Communicational process ("I" ±"We" and "You")

The use of "I"/"we", on one hand, and "you", on the other, are signs that can help in the process of classification of an egocentric typology. In this study we restrain at doing a simple statistics that would subsequently help in advancing a more rigorous investigation over the distance that the writer interposes between himself (or his publication) and his readers.

4 A Case Study

The method and instruments for processing natural language used in this paper confirm the premise that *investigating journalists' identity in pragmalinguistics terms can be captured based on statistics depicted from a corpus of texts*. Since language suffers a perpetual metamorphosis, and the one of the virtual media is even more rapidly changing, it would be good if the corpus would be acquired over different periods of time. Although we believe that the proposed approach has an important degree of generality, allowing for inclusion of more investigation axis, we are also aware that the analyzed corpus in this initial research is still insufficient for drawing crisp egocentrism classification conclusions. What we propose, therefore, is merely a methodology for typifying journalistic styles, focused on egocentrism elements. The exercise is drawn around some of the Romanian editorialists with great outlet to the public.

[3] For instance, when you write in other language than that spoken by the receiver.

4.1 The Corpus

For the elaboration of preliminary conclusions on the configuration process of the journalist identity, we collected, stored and processed 1,463 relevant articles (summing up 28,879 words), published by three journalists: Cristian Tudor Popescu (CTP), Marius Tucă (MT), Ion Cristoiu (IC), during January – March 2013 by three important Romanian newspapers having similar profiles[4], but usually displaying totally disjoint opinions and journalistic styles on any topic.

4.2 Methodology

We present briefly the accomplished steps:

- by attentive reading, we identified 3 typology of egocentrism journalistic, that can be called: *egocentric (self-)ironic*, *egocentric puffy*, and *egocentric all-knowing*.
- we established a number of features (belonging to the syntactic, lexical-semantic and pragmalinguistics levels of analysis) that are, more or less, subject to automatic extraction: the use of the personal pronouns, familiarity in communication, ironies, punctuation, etc.; the semantic classes of being `rational` and `emotional` (with their sub-classes), `nationalism`, `sexual`; comments that are aggressive, etc.; number of enumerations/article and the expressions with emotional content;
- we have tagged all texts for POS in order to highlight the personal pronouns first person[5], singular and plural, but also second-person, plural. All sentences containing personal pronoun forms have also been semantically analyzed (using the DAT software) and pragmatically (qualitative analysis).

4.3 The Pragmalinguistics Analysis

In order to proceed with the syntactic analysis, the text bodies were annotated with syntactic information, in XML. Two sources of information have been used, involving manual (200 units[6] for each journalist monitored) and automatic annotation (350 units – for the first, 210 – for the second, and 483 – for the third). The manually annotated segments (see Table 1) included the interrogative and exclamatory count, and all pronoun forms, first person, singular and plural, but, also, second person, plural (nominative, dative, and accusative cases). After highlighting the characteristics of the specified syntax, we were interested to see in what phrase structures are the personal pronouns used, operation performed using POS tagging [16].

[4] These are national dailies of general information, tabloids with a circulation of tens of thousands of copies per edition, each. The newspapers were monitored on their websites: *Evenimentul zilei* – www.evz.ro, *Gândul* – www.gandul.info, *Jurnalul național* – `www.jurnalul.ro`

[5] Here we have also reserved the lemma *subsemnatul* (undersigned), which replaces the pronoun "I" in the nominative case.

[6] Sentences.

We can conclude about the syntactic structure (Table 1) for each journalist as follows:

- the first type of journalist, CTP, prefers medium length texts (approx. 22 sentences/article). His interrogative (50) and exclamatory (32) sentences have a rhetorical purpose; the audience can or should be able to reflect to CTP's opinions.

Certain pragmalinguistics boundaries can be identified in the speech of this journalist:

1. personal pronoun, first-person, singular - evokes in the sentence, through personal pronouns "I", entities present implicitly or explicitly in the universal speech, the communication situation being defining (e.g. *because I want to contribute with what I can to bring viewers in the Romanian theatre*).
2. personal pronoun, first-person, plural - empathises with public opinion (e.g. *The funny parliamentary Becali is invited several times a day to televisions by us in order for him to make a mockery of the rule of law, common sense and human dignity, to insult women, to curse men*).
3. personal pronoun, second-person, plural – assuming the role of very good connoisseur of governmental management (e.g. *The reasoning seems wrong for you, dear readers, sick or healthy?*).

Table 1. Syntactic descriptors used in this research

Descriptors	CTP	MT	IC
I, sg. (*eu, îmi, mi, m, mine, mă*)	29	25	118
I., pl (*noi, ne, ni*)	3	14	30
II, pl. (*voi, vouă, vă, vi, v*)	2	25	5
Sentences	22	7	43
Exclamatory	22	18	99
Interrogative	50	34	62

- the second journalist investigated, MT, is very expeditious (approx. 7 sentences/article), but often the receptor must find the answers to his interrogations (34). The reader can discover also some aggressive exclamatory sentences (18).

From the perspective of pragmalinguistics boundaries, this journalist uses:

1. personal pronouns, first-person, singular – in contexts with nostalgic tint, sometimes anxious (e.g. *I love break-ups over the shoulder*).
2. personal pronoun, first-person, plural – emphasizes, easy to recreate, in this sense, the name of the daily newspaper that he leads and his quality in the editorial. (e.g. *Contacted by the National Journal, Alina Alexandra Mihai, winner of the Giumbix contest, told us how her idea came out…*).
3. personal pronoun, second-person, plural – the clear delimitation, even pornographic, from others (e.g. *It fills your screen with boobs and butts / sex is on money, do not hurry,/don't masturbate, because you will use the credit card.*)

- the third type of journalist, IC, writes very long articles (approx. 43 sentences/article), because he insists in details in any political subject (demonstrating

to be a mature political analyst). It is noted that he uses the personal pronoun "I" more often than the other two journalists, denoting a higher concern towards himself. Here are a few examples:
1. the personal pronoun, first person, singular – emphasizing his experience as a journalist, but also his ideological position: *In 2003, I was, as a director, at the forefront of the Realitatea TV, shepherded by Silviu Prigoană.*
2. the personal pronoun, first person, plural – underlines the verticality (positioning himself in the good camp) in relation to the wicked. Here's a snippet of speech: *The distinguished are left with the millions, and we, the rest, with the honour.*
3. the personal pronoun, second person, plural – clear induction in the eyes of public opinion to the camp of morality he makes part of. For instance: *Put yourselves in our job and you are going to be fine!*

4.4 Lexical-Semantic Analysis

The corpus was processed with the DAT[7] tool (a platform for lexical-semantic analysis of public discourse). To identify the predominant tonalities in the discourses of each journalist, we included in this case study only 15th semantic classes, arranged hierarchically: `positive` (with 3 subclasses: `spectacular, firmness` and `moderation), rational` (with 5 subclasses: `uncertain`, `inhibition`, `intuition`, `certain`, and `determine`), and `negative` (with 3 subclasses `anger, anxiety`, `sadness`), and `sexual.`

When an occurrence belonging to a lower level class is detected in the input file, all counters in the hierarchy from that class to the root will be incremented.

Bellow, we have the results outputted by DAT (Fig. 2), when analysing the streams of textual data for each semantic class. The 3 profiles analysed (CTP, MT and IC) can be interpreted as follows:

- the discourse o the first type of journalist, `CTP`, is predominantly negative in emotional tonality (class `negative`), in two different intensities (classes `anger` and `sadness`). In general, he prefers ironical expresses (for instance: *His work is a systematic and tenacious huge collection of kitsch and clichés*). Although he is a well known journalist, sometimes he mentions that he is a specialist in the cinema art. (e.g. *I do it as a cinema-goer*). Also, he prefers to use expressions in the English language (for instance, *Who`s that stumblin`around in the dark? State your business or prepare to get winged! or Auf wiedersehen, Bullseye! or Alexandre Dumas is black and so on*), but in an ironic way. We will call this type, *egocentric (self-)ironic.*
- the second type, `MT`, is a dynamic guy (class `positive`) and prefers metaphorical languages (class `spectacular`) but, often, in a pornographic tone (*don't masturbate, because you will use the credit card, / Look only at the nipples, like some teenagers*). Most of the times, he has in attention the themes presented in his show, "The MT Show". He takes every opportunity to promote their own journalistic press

[7] DAT (Discourse Analysis Tool) has some similarities with LIWC (Linguistic Inquire and Word Count), used during the American presidential elections in 2008 [Pennebaker, J. W.]. The Romanian lexicon resourcing DAT contains a collection of over 9,500 entries (lemmas).

trust (for instance: *a new TV rubric in the MT Show*). He writes short, confidently in himself (class `certain`), often in verse, because he enjoys being a bohemian wistful (e.g. *How many question marks/sending a riot lifestyle/of a desertion from proper decorum/things/people,/ to be able to respond.*). We will call this type, *egocentric puffy*.

- the third type, `IC`, has a rational discourse (class `rational`) is very convinced of his ideas. He prefers long texts to explain in a determined way (class `determine`) the chosen subject (especially, with political flavour). He is an old fashioned journalist; his articles have a title, a short resume and a long body). Actually, all his texts appear under the heading *Romania's C.*

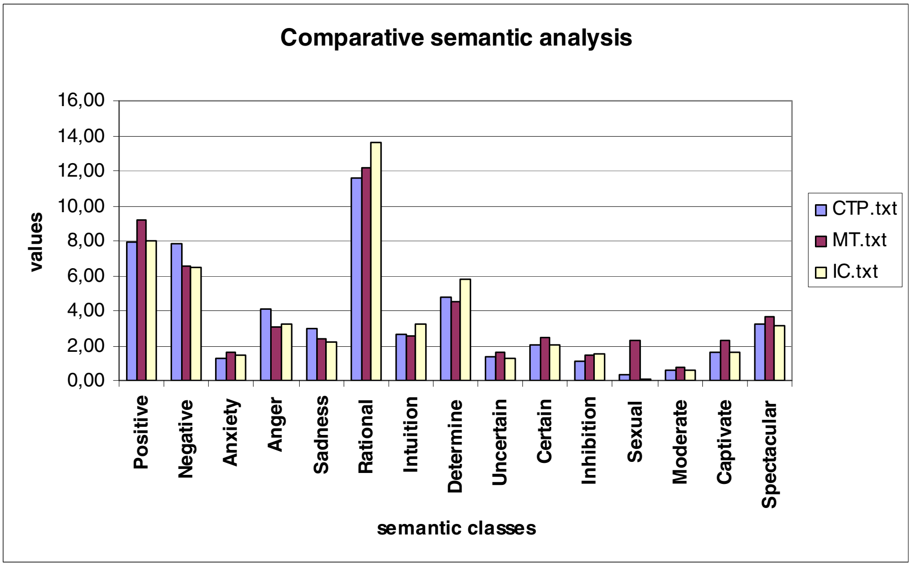


Fig. 2. The comparative semantic analysis for journalistic articles

He has an indulgent point of view over the President (for instance, *these days that were spent, a few things have made me realize that the president was right*), even if not long time ago he had a different opinion. Now he attacks the Ponta's Parliament, situated on the opposite side of the president (e.g. *I noticed first of all Victor Ponta's inability to overcome the posture of politician opposition, essentially babble*"). We will call this type *egocentric all-knowing*.

5 Conclusions

The discursive-*egocentric* differences, found in this study, can be attributed partly to idiosyncratic rhetorical styles, and partially, to the ethics the authored adhered to (editorial policy applied in the public space). Of course, in the current context, ethics is confused with a specific ideology that belongs to a specific editorial group. No less important in identifying the pragmalinguistic characteristics is the cultural universe and the generation to which belongs the signatory of a press article.

We believe that the findings revealed in the present study may lay the basis for the delineation of a journalistic identity that brings in the space of the Romanian journalism critics an expansion of the possibilities of public discourse analysis by computer mediated techniques.

We are aware that the corpus of manually annotated texts is still in an early phase and this study should be understood only as allowing to perform a pilot study towards a statistical investigation on a larger corpus, that would be used in a process of automatic learning, such that, in the future, the machine be capable of efficient automatic annotation. Right now we are testing the feasibility of using our natural language processing instruments in the automatic analysis of journalistic texts. Not less important if to reveal new significant features of discourse which, on one hand, could be automatically extracted from the text and, on the other, are useful for our goals. It is also important to perform tests that would reveal to what extend the authorship types identified (3 – after the present study, maybe more – after a more rigorous one) could be clearly delimited by statistical means.

Another focus of attention in our future research is towards supervising the public, as commenter of the journalistic texts, an aspect which has not received much attention to date. In doing all these, we look also to adapt our instruments to other languages as well, which, among other things, would allow us to compare the results obtained in the Department against those published elsewhere.

Acknowledgments. In order to perform this research the first author received financial support from the POSDRU/89/1.5/S/63663 grant. We are grateful to Radu Simionescu, from the NLP-Group@UAIC-FII, for POS-tagging the Romanian corpus.

References

1. Barth, F.: Ethnic Group and Boundaries, Bergen, Oslo, Universitetsforlaget (1969)
2. Denis, A.L., Quignard, M., Freard, D., Detienne, F., Baker, M., Barcellini, F.: Détection de conflits dans les communautés épistémiques en ligne? In: TALN 2012, Grenoble, France (2012)
3. Ducrot, O., Anscombre, J.-C.: L'argumentation dans la langue, Mardaga (1983)
4. Forsyth, E., Martell, C.: Lexical and Discourse Analysis of Online Chat Dialog. In: International Conference on Semantic Computing (2009)
5. Eensoo, E., Valette, M.: Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. In: Proceedings of TALN 2012, Grenoble, France (2012)
6. Garera, N., Yarowsky, D.: Modelling Latent Biographic Attributes in Conversational Genres. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, August 2-7, pp. 710–718 (2009)
7. Gîfu, D., Cristea, D.: Multi-dimensional analysis of political language. In: Park, J.J.(J.H.), Leung, V., Shon, T., Wang, C. (eds.) Future Information Technology, Application, and Service: FutureTech 2012, vol. 1, pp. 213–221. Springer, Netherlands (2012)

8. Grivel, L., Bousquet, O.: A discourse analysis methodology based on semantic principles - an application to brands, journalists and consumers discourses. Journal of Intelligence Studies in Business 1, 76–86 (2011)
9. Iftene, A., Rotaru, A.: User Profile Modelling in eLearning using Sentiment Extraction from Text. In: Research in Computing Science, Special issue: Natural Language Processing and its Applications, Mexico, vol. 46, pp. 267–278 (2010)
10. Kerbrat-Orecchioni, C.: Analyse des conversations et négociations conversationnelles. In: Grosjean, M., Mondada, L., (eds.) La Négociation au Travail, Lyon, PUL/ARCI, pp. 17–41 (2004)
11. Lin, J.: Automatic Author Profiling of Online Chat Logs, M.S. Thesis, Naval Postgraduate School, Monterey (2007)
12. Lortal, G., Todirascu-Courtier, A., Lewkowicz, M.: AnT&CoW:Share, Classify and Elaborate Documents by means of Annotation. Journal of Digital Information Management 6(1), 61–70 (2008)
13. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count. In: LIWC 2001. Erlbaum Publishers, Mahwah (2001)
14. Portele, T.: Data-driven Classification of Linguistic Styles. In: Spoken Dialogues. COLING (The 19th International Conference on Computational Linguistics) (2002)
15. Schiaffino, S., Amandi, A.: Intelligent user profiling. In: Bramer, M. (ed.) Artificial Intelligence. LNCS (LNAI), vol. 5640, pp. 193–216. Springer, Heidelberg (2009)
16. Simionescu, R.: POS-tagger hibrid. Dissertation at the "Alexandru Ioan Cuza" Universitatea of Iaşi (2011)
17. Stark, A., Dürscheid, C.: SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In: Thurlow, C., Mroczek, K. (eds.) Digital Discourse. Language in the New Media, pp. 299–320. Oxford University Press, Oxford (2011)
18. Tufiş, D., Ştefănescu, D.: A Differential Semantics Approach to the Annotation of Synsets in WordNet. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC). ELRA, Malta (May 2010)
19. Wimmer, A.: The Making and Unmaking of Ethnic Boundaries: A Multilevel Process Theory. American Journal of Sociology 113(4), 970–1022 (2008)
20. White, B.Y., Frederiksen, J.R.: Inquiry, modelling, and meta-cognition: Making science accessible to students. Cognition and Instruction 16, 3–118 (1998)
21. Wittgenstein, L.: Cercetări filozofice, trad. de Mircea Dumitru şi Mircea Flonta Ed. Humanitas, 117 (2004)
22. Zukerman, I., Albrecht, D.: Predictive Statistical Models for User Modeling. User Modeling and User-Adapted Interaction 11(1-2), 5–18 (2001)

Towards the Lexicon-Based Sentiment Analysis of Polish Texts: Polarity Lexicon

Konstanty Haniewicz[1], Wojciech Rutkowski[2],
Magdalena Adamczyk[3], and Monika Kaczmarek[1]

[1] Department of Information Systems, Faculty of Informatics and Electronic Economy, Poznań University of Economics
{konstanty.haniewicz,monika.kaczmarek}@ue.poznan.pl
[2] Ciber Poland
wojciech.rutkowski@ciber.com
[3] Department of Modern Languages, University of Zielona Góra
m.adamczyk@wh.uz.zgora.pl

Abstract. Due to the increasing amount of information available on the Web, sentiment analysis aiming at an automatic identification of the emotional load of texts is growing in importance. The aim of our research is to devise a reliable method for analsing sentiment in Polish texts, which requires developing adequate polarity lexical resources. In this paper, we discuss a method of building a fine-grained polarity lexicon for Polish based on custom-built review corpora. The compiled lexicon is subsequently tested in the field of sentiment analysis reaching the accuracy level of up to 79%.

Keywords: sentiment analysis, polarity lexicon for Polish, NLP.

1 Introduction and Motivation

Emotions are an important component of texts published in the global Web. The goal of sentiment analysis it to assign to each text "a value expressing an emotional attitude: positive, negative, neutral, objective or bipolar" [22]. Most studies on sentiment analysis use textual material in English as the source of data. However, more and more resources developed specifically for English in order to extract subjective information or identify the polarity of opinions are now successfully being applied to other languages, such as Chinese [25]. Some, though altogether few, attempts of this sort have also been made for the Polish language (see [9]).

Both Polish and English are Indo-European languages; yet, the former is a member of the Balto-Slavic, and the latter of the Germanic language family. The structure of a sentence in Polish is not as constrained as in English and has a freer word order, i.e. parts of speech may be relatively freely rearranged without changing their meaning. As opposed to English, Polish is also a highly inflectional language with robust morphology. Because of these fundamental differences, we deem the task of sentiment classification of the Polish texts more challenging.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 286–295, 2013.

Irrespective of the examined language, there are two main approaches to sentiment analysis, namely the lexicon-based approach and the supervised learning approach. If the former is to be used, the required lexical resources, i.e. dictionaries of sentiment words providing for each item its sentiment score in a given domain, need to be available. While there are numerous polarity resources for the English language (e.g. [3]), those for Polish are scarce and, to the best of our knowledge, none of them is capable of fully supporting sentiment analysis, as shown in the related work section.

The main objective of our work is to build, in an automated manner, a polarity lexicon, i.e. a polarity semantic network, for the Polish language. The quality of the devised system depends strongly on the quality of the developed resources. More specifically, if the designed polarity semantic network does not cover the analysed language model sufficiently well, the results of sentiment analysis will be far from satisfactory. The aim of the study is to advance the understanding of sentiment analysis for the Polish language, thus pointing to its peculiarity, and to contribute to the development of relevant resources for the Polish language that could also be made available to the entire community.

The study follows the pro-active research path based on the design oriented research paradigm [16] and the design science paradigm [6, 7]. The adopted procedure consisted of the following three stages. First, the concept building phase took place, which resulted in establishing a solid theoretical underpining of the work presented in the subsequent sections. The next step was the approach building which involved devising a method of creating a polarity semantic network based on the pre-established theoretical concepts (importantly, the method development drew on a number of the existing approaches to building lexical resources). The last step was concerned with creating the semantic network containing 70000 concepts and running the experiments designed to test the quality of the developed artefact.

The paper is structured as follows. It opens with a brief overview of the relevant research in the domain of interest strictly related to our work. The two subsequent sections are devoted to demonstrating, respectively, the method used for developing the polarity semantic network and the conducted experiments. The paper concludes with final remarks.

2 Related Work

Sentiment analysis refers to "computational treatment of opinion, sentiment and subjectivity in text" [18]. There are two main approaches to sentiment classification, i.e. the supervised learning approach and the lexicon-based approach. Most of the machine learning based approaches use a text classification framework and consist of the following stages: tokenizing, filtering/stemming, training a classifier and, finally, applying the model. The most frequently adopted machine learning approach is the Support Vector Machines with n-gram features trained on a large set of texts with known polarities (usually positive or negative) (e.g. [19] [17]).

In turn, the lexicon-based methods make use of the existing lexical resources. Importantly, the latter are one of the main sources of linguistic information for Natural Language Processing and related areas [20]. The polarity lexical resources differ widely in complexity. They may take the form of simple lists of positive or negative words, as well as that of more complex semantic nets [13]. As word sentiments are domain dependent [23], no general-purpose polarity lexicon is capable of performing reliably for every topic [12].

Creating a polarity lexicon manually is a labour-intensive and error-prone task; in addition, providing a sufficiently widespread coverage of the lexicon is doubtful [12]. Therefore, there are a number of research initiatives aimed at creating polarity lexicons in an automated manner, using either a supervised (e.g. [24]) or unsupervised approach (e.g. [5]). A lot of methods involve utilising other lexical resources, such as thesauri and lexicons, frequently presented in the form of semantic networks. What is more, researchers often use web-documents in order to create polarity lexicons (e.g. [24]). Such lexicons are not limited to any specific word classes and additionally contain slang and multi-word expressions. Experiments have shown that, when compared to lexicons based, for instance, on WordNet [15], they make it possible to achieve a higher performance level in the polarity classification task. Furthermore, there are a few studies aiming at compiling polarity lexicons from corpora for languages other than English which employ automated methods (e.g. [8], [14]). The experiments using those lexicons show about 80% precision with respect to positive and negative polarity on adjectives and adjective phrases.

As far as the English language is concerned, numerous resources are available which successfully cater for the needs of sentiment analysis. Some of them include: SentiWordNet [3] (providing for each synonym a set of three numerical scores, describing respectively how objective (or neutral), positive and negative the terms contained in a synonym set are), WordNet-Affec [21] and ANEW [1]. To the best of our knowledge, there is no polarity lexicon for the Polish language that is readily available to the public; neither have we observed any initiatives to compile one. However, there are some general resources, such as Slowosiec (plWordNet)[1], the biggest Polish wordnet containing 94523 synsets, 133071 lexical units and almost 150000 lexical relations, which may be utilised in the process of developing polarity lexical resources. In addition, a set of tools for language processing (e.g. Morfologik, a morphological analyser) can be used.

In opposition to the few already developed approaches to investigating the emotional overtones of the Polish texts ([2], [9]), the authors of this work adopt the lexicon-based approach and attempt to create the polarity lexical resources for Polish with a view to making them widely available to the public. The accuracy of the constructed polarity semantic network is tested by using it to perform the sentiment classification task.

The mechanism used to create the above mentioned resources as well as the underlying assumptions are presented in the next section of this paper.

[1] `http://nlp.pwr.wroc.pl/en/tools-and-resources/slowosiec`

3 Design of the Mechanism

3.1 Polarity Semantic Network: Main Characteristics

The aim of the study is to provide a dedicated knowledge representation structure suitable for sentiment analysis. After a carefully conducted examination of the available resources, the decision was taken to create a semantic network, whose core is built in an automated manner, based on a number of easily accessible resources, such as dictionaries, thesauri and word lists compiled from the existing open source projects. The main effort while building the semantic network went into formulating a broad set of extraction rules. These rules were applied to the available dictionaries and other resources (such as Wikipedia) in order to identify the relations defined there. Without the structures inherent in dictionaries and other available resources, building a semantic network would be far more difficult. The constructed semantic network, which draws heavily on such structures, was reconfigured and refined thanks to a collective effort of the authors and other volunteers.

The resulting semantic network stores data on over 70000 concepts. While at the moment it is smaller than its alternatives, it has a considerable potential for growth and contains the data unavailable in other known resources. The authors intend to extend the network to over 140000 concepts along with a considerable number of categorised proper names, with a special focus on names connected with brands, organisations and various products.

The structure of the semantic network is designed to store basic data on relationships between individual concepts, such as synonyms, antonyms, hypernyms and homonyms. In addition, every term has a set of additional features, where the most prominent one is the vector storing the data on the sentiment of a concept. In contrast to the existing research, the authors decided to extend sentiment analysis to all available parts of speech. Therefore, if a concept is deemed significant in a given domain, its sentiment analysis score is stored in the semantic network along with the data on the relevant domain. Thus, the polarity semantic network set up in our work may be defined as follows:

The proposed polarity semantic network is a domain-aware sentiment lexicon Ls. Each element in Ls is associated with pairs (s_i, d_j)*, where* s_i *is the sentiment score for a given ith element in a domain* d_j*.*

Although determining the domain of the analysed textual data demands extra effort, the authors strongly believe that it is worthwhile from the language processing point of view.

The main challenge while establishing a domain is to provide a considerable amount of input data that serves as a springboard for a further, fully automated, sentiment identification process. The data that extends the semantic network is gathered from a fully automated analysis of the available corpora. The main tools for the examination of sentiment in textual data include: i) Bayes classifiers, ii) Maximum Entropy, and iii) Support Vector Machines. The result of the application of these tools is a set of terms that can be incorporated directly into the available semantic network. The incorporated terms are those that are most distinctive with respect to their domain and the actual data (originating from users' reviews).

In addition to the previously discussed aspect, the network contains data on the frequency of any given concept as culled from the reference corpus. This is an important enhancement, as it was observed that many semantic networks suffer from over-specializing when using a network in generalisation or summarising tasks. Providing frequency counts makes it possible for the applied algorithms to promote more popular terms, thus making the generalised or abstracted text more accessible to a human user. While this may not be so important in crude categorisation tasks, it is crucial when evaluating the results of various algorithms by humans.

The constructed semantic network enhanced with the discussed elements is a basic tool used in a number of experiments which aim at offering a solution that is capable of determining the sentiment of a given textual data in an on-line manner. By 'on-line' the authors understand a solution that gives the answer within milliseconds after providing input data.

The following subsections are devoted to a close description of the corpora used in the experiments and additional steps needed to devise the postulated data structure and functionality.

3.2 Data Sets

The data used in sentiment analysis was culled from a number of portals that provide reviews on various goods and services. The total number of the gathered reviews amounted to 356275. Every review, apart from the textual content, was accompanied by a mark provided by its author. The authors of this study made a fundamental assumption about the relation between the provided mark and the content of a given document. It is important to notice that the available scales were different for various sources. It was decided that all the available scales had to be normalised to a scale of 0 (lowest sentiment) to 10 (highest sentiment).

The portals of interest were specialised opinion portals that gathered reviews of particular retail services and products. The majority of opinions offered by the portals are considered to be above average in terms of the score provided by the users. After a careful study of a random sample, it was concluded that a high score is correlated with a high sentiment of the entire review.

The corpus had to be properly balanced to include an equal number of positive and negative reviews. This was done to avoid the predominance of the terms associated with a positive sentiment over the negatively charged ones. A set of sample data is given in Table 1. It has to be emphasised that these are examples of highly positive and highly negative terms.

The data was captured from the following portals: `opineo.pl`, `wizaz.pl`, `opiniuj.pl`, `mgsm.pl`, `cokupic.pl`, `wakacje.pl`. The majority of opinions found there were civil even when a low mark was given by a user. This is a challenge for researchers, as they have to discover the ways of expressing sentiment also in cases were language used to express opinion contains profanity.

Table 1. Sample highly significant positive and negative Polish terms (with approximate English interpretations)

Positive sentiment	Negative sentiment	Neutral sentiment
błyskawiczny [swift]	niesolidny [unreliable]	fajnie [fine]
przystępny [accessible]	nierealny [unreal]	powalać [strike (down)]
znakomity [excellent]	zszargać [bedraggle]	podbić [conquer]
różnorodność [diversity]	przeterminować [expire]	dostateczny [sufficient]
profeska [professional (colloq.)]	ignorancja [ignorance]	spisywać [write down]

3.3 Data Set Preparation

In order to allow for a fair calculation of the sentiment of various concepts, the prepared data set was built with two equally numerous subsets of the original corpus. This was motivated by the fact that a positive sentiment was far more prevalent in the gathered resources (over 95% of the collected reviews were positive). After the preparation of the target sub-corpus, it was divided into two parts. 90% was used to evaluate the sentiment of concepts and the remaining 10% was used as an input for experiments. The total number of the reviews in the balanced corpus amounted to 34265.

As Polish is a highly inflectional language, a procedure was devised that uses Morfologik to prepare a surrogate of an opinion that contains only the basic forms of concepts available in the prepared semantic network. At this stage all stop-words were removed.

The prepared surrogates served as an input for sentiment evaluation. The mean number of traits to be analysed for the balanced sub-corpus equalled 16.8 terms per surrogate.

3.4 Sentiment Calculation

The sentiment for the balanced corpus was calculated with the aid of a specially developed tool that uses Bayes classifiers and SVM in order to define which terms in the corpus are significant for sentiment analysis. The aim of implementing two techniques was to compare the results, yet their effectiveness proved not to diverge from those reported in the literature.

The sentiment value for a given term that was a candidate for inclusion in the semantic network was a number between 0 (lowest sentiment) and 10 (highest sentiment). The output values based on the training set were assigned to the terms in the semantic network to test its applicability.

It is important to notice that an alternative method of sentiment score storage proposed by Liu [10] is also valuable and was used in previous research [4]. This method involves the attachment of a sentiment vector that can store two or three scalars pointing to the inclination of a term towards a positive, negative or neutral sentiment, depending on its context. This is important, as terms are often polysemous and their neighbourhood can define the right scalar to apply in a given context. The total number of the terms selected by the sentiment identification tool amounted to 6685.

4 Experiments

4.1 Experiment Setup

In order to test whether the terms provided by the sentiment identification tool are the appropriate sentiment markers, a test run was carried out. The control sample was composed of 3246 reviews that were not used while establishing potential sentiment markers. A half of the corpus comprised reviews with low scores. The domain of the reviews was the same as the one discussed in the previous section (i.e. products, services, retail reviews).

The reviews were transformed into surrogates with the previously used tools. Out of the 3222 reviews, 2426 surrogates had a sufficient number of terms to calculate a prognosis of the sentiment associated with a given review surrogate. The success rate for those reviews was **78.93%**.

4.2 Discussion of the Results

It is believed that a bigger, balanced corpus should provide more sentiment markers and allow for an even higher success rate. The whole procedure must be repeated for other domains in order to achieve new levels of versatility when applying a semantic network to sentiment analysis. Additional challenges arise from the following traits of human language:

Polysemy and Homonymy. On the level of words alone, and strictly speaking their semantic structure, the highly related phenomena of homonymy and polysemy should not pass unnoticed. Formally, having their component parts identical in sound and spelling, they are indistinguishable. Semantically, while they both rest on double/multiple meanings (in which, in a de-contextualized environment, they may contribute to ambiguity), they mark two distinct types of distance between the meanings they involve.

Idioms/Idiomatic Expressions and Semantic Prosody. On the level of word combinations, the semantic phenomena that should be taken into account include the following: (a) idioms/idiomatic expressions and (b) prosody. In the former case the unique property of semantic structure is non-compositionality, which means that the idiomatic meaning of word strings cannot be derived from the meanings of their component parts (yet, depending on the level of idiomaticity, it can be more or less transparent).

The above mentioned semantic prosody, a relatively new concept introduced in 1993 by Louw [11], is concerned with the fact that the collocates of lexical items are not only random, unrelated words but also semantically defined word classes, which have either a positive or a negative meaning. The prosody of a given word, therefore, depends on the overtones carried by its collocates, and so the verb *powodować* [cause], for instance, can have a negative prosody since it tends to collocate with negatively loaded words, e.g. problem [problem], *wypadek* [accident], *śmierć* [death].

Humour and Irony. Shifting the focus to a deliberate exploitation of both semantic and pragmatic meanings for extra-linguistic purposes, there are at least two phenomena that have to be reckoned with when interpreting the data, namely humour, involving a skilful manipulation of (double/multiple) meanings, and irony, exploiting the interplay of the types of meanings. When it comes to humour, it may have some bearing on the results of the experiment insofar as it depends for its existence on the aforementioned semantic processes of polysemy and homonymy in adjectives and adverbs.

In turn, irony may skew the results as it rests on a calculated mismatch between the literal (explicit/overt) meaning of an utterance, be it a single word or an entire sentence, and the implied (implicit/covert) meaning intended by a speaker.

Grammatical and Lexical Negation. Last but not least, when faced with the data such as those used in the experiment, it needs to be born in mind that there exist structures in language, on the level of grammar and lexis alike, which are capable of producing the exact opposites of the meanings of lexical items. The most straightforward example would be a direct negation in the form of the word *nie* 'not' but less immediately obvious instances thereof were also found in the data gathered for the experiment. These included formulations such as *daleki od (zadowalającego)* [far from (satisfactory)], *mało (ciekawy)* [little (interesting)], *trudno nazwać (pomocnym)* [difficult to call (helpful)] and constructions marking contrast (e.g. *podczas gdy* [whilst], *chociaż/choć* [although], *wprawdzie*, *ale* [but]). The validation proved that a semantic network for Polish extended with domain based sentiment markers is a valuable asset in natural language processing tasks.

5 Conclusions

In this paper we presented a method of building a fine-grained polarity lexicon for Polish on the basis of custom-built review corpora. Moreover, we have tested the obtained lexicon in the field of sentiment analysis. The results of polarity identification are quite promising, with the accuracy rates reaching up to **79%**.

We strongly believe that the created polarity lexicon can be applied without additional automatic or manual processing. However, further enhancements ragarding the number of sentiment markers and supported domains are envisioned.

The future work is intended to be concerned with the development of additional techniques to improve the polarity lexicon in terms of its size and scope. Some effort will also go into introducing an extra dimension nie 'not' in terms of sentiment markers. The authors believe that a homogenised value for every concept in the lexicon reflecting the global sentiment of a concept can be valuable.

Moreover, having accomplished that, the authors would like to further extend the already mastered tools and algorithms in order to offer a coherent solution capable of analysing textual data in terms of local sentiment (at the level of sentence or paragraph, depending on the available context). What is more, the proposed polarity semantic network has a potential for a substantial growth. The growth is secured by the availability of various open access resources (thesauri and dictionaries). In addition, reorganisation and refinement of the discussed lexicon is an interesting challenge that is also going to be addressed in future research.

References

1. Bradley, M.M., Lang, P.J.: Affective norms for english words (anew): Instruction manual and affective ratings. Psychology, Technical(C-1) (1999)
2. Buczynski, A., Wawer, A.: Shallow Parsing in Sentiment Analysis of Product Reviews. In: Proceedings of the LREC 2008 Workshop on Partial Parsing: Between Chunking and Deep Parsing, pp. 14–18. ELRA, Marrakech (2008)
3. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), pp. 417–422 (2006)
4. Haniewicz, K., Rutkowski, W., Adamczyk, M.: Linguistically Aware Semantic Network for Automated Information Tracking. In: Proceedings of the Eighth International Conference on Signal-Image Technology and Internet-Based Systems, SITIS, pp. 503–509. IEEE Computer Society, Washington, DC (2012)
5. Hassan, A., Radev, D.: Identifying text polarity using random walks. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 395–403. Association for Computational Linguistics, Stroudsburg (2010)
6. Hevner, A.R.: The three cycle view of design science research. SJIS 19(2), 87–92 (2007)
7. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. Management Information Systems Quarterly 28(1), 75–106 (2004)
8. Kaji, N., Kitsuregawa, M.: Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In: Proceedings of EMNLP-CoNLL, pp. 1075–1083. Prague (2007)
9. Kowalska, K., Cai, D., Wade, S.: Sentiment Analysis of Polish Texts. International Journal of Computer and Communication Engineering 1(1), 39–42 (2012)
10. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2012)
11. Louw, B.: Irony in the Text or Insincerity in the Writer?: The Diagnostic Value of Semantic Prosodies. In: Baker, M., Francis, G., Toginini-Bognelli, E. (eds.) Text and Technology: In Honour of John Sinclair, pp. 157–176. Benjamins, Amsterdam (1993)
12. Lu, Y., Castellanos, M., Dayal, U., Zhai, C.: Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 347–356. ACM, New York (2011)
13. Maks, I., Vossen, P.: Different Approaches to Automatic Polarity Annotation at Synset Level. In: Proceedings of the First International ESSLLI Workshop on Lexical Resources, pp. 63–71. Publ. online (2011)
14. Maks, I., Vossen, P.: Building a Fine-grained Subjectivity Lexicon from a Web Corpus. In: Chair, N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp. 3070–3076. European Language Resources Association (ELRA), Istanbul (2012)

15. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM 38, 39–41 (1995)
16. Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., et al.: Memorandum on design-oriented information systems research. EJIS 20, 7–10 (2011)
17. Paltoglou, G., Thelwall, M.: A study of Information Retrieval weighting schemes for sentiment analysis. Association for Computational Linguistics, pp. 1386–1395 (2010)
18. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2(1-2), 1–135 (2008)
19. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, vol. 10, pp. 79–86. Association for Computational Linguistics, Stroudsburg (2002)
20. Sagot, B.: Introduction. In: Proceedings of WoLeR 2011, the 1st International Workshop on Lexical Resources, p. 6. Ljubljana (2011)
21. Strapparava, C., Valitutti, A.: WordNet-Affect: An Affective Extension of WordNet. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, vol. 4, pp. 1083–1086. ELRA, Lisboa (2004)
22. Tromp, E., Pechenizkiy, M.: Senticorr: Multilingual sentiment analysis of personal correspondence. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, ICDMW 2011, pp. 1247–1250. IEEE Computer Society, Washington, DC (2011)
23. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. Inf. Syst. 21(4), 315–346 (2003)
24. Velikovich, L., Blair-Goldensohn, S., Hannan, K., McDonald, R.: The viability of web-derived polarity lexicons. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 777–785. Association for Computational Linguistics, Los Angeles (2010)
25. Zhang, C., Zeng, D., Li, J., Wang, F.-Y., Zuo, W.: Sentiment analysis of chinese documents: From sentence to document level. J. Am. Soc. Inf. Sci. Technol. 60(12), 2474–2487 (2009)

Bringing Common Sense to WordNet with a Word Game

Jacek Rzeniewicz and Julian Szymański

Department of Computer Systems Architecture,
Gdańsk University of Technology, Poland
julian.szymanski@eti.pg.gda.pl, jrzeniewicz@gmail.com

Abstract. We present a tool for common sense knowledge acquisition in form of a twenty questions game. The described approach uses WordNet dictionary, which rich taxonomy allows to keep cognitive economy and accelerate knowledge propagation, although sometimes inferences made on hierarchical relations result in noise. We extend the dictionary with common sense assertions acquired during the games played with humans. The facts added to the knowledge base use eight new relation types. After 700 games played in the animals domain the average number of assertions per concept raised over 89%. Evaluation of the newly acquired facts indicates high quality of knowledge captured using proposed approach.

Keywords: common sense knowledge acquisition, WordNet, word games.

1 Introduction

Common sense knowledge are facts and rules that are obvious for humans, such as that father is always older than his child, that windows are transparent or that burned skin hurts. Those facts and rules allow to perform reasoning, which is necessary for deep text mining, natural language processing, object recognition, context–aware interfaces and more AI problems. Over the last decades several significant projects have been launched in attempt to capture those elementary facts about the world into machine readable dictionaries. The most important are WordNet [1], CyC [2], SUMO [3] and ConceptNet [4]. The latter is the most advanced: initially assembled from English sentences acquired by the Open Mind Common Sense project [5], later became multilingual, combining knowledge from other sources, including WordNet, DBPedia [6] and Verbosity game [7]. However, the fact remains that handcrafted knowledge bases, reliable as they are, stay relatively small and expensive in development, while automatic methods tend to produce noise assertions.

Several successful crowdsourcing experiments in last years like reCAPTCHA [8] and Foldit [9] have shown the great potential of human–based computation. Verbosity have generated a significant amount of statements in a short time. 200 thousands assertions acquired from players were included into Open Mind Common Sense data set in 2009 [10]. We consider getting the Internet users directly

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 296–305, 2013.

involved in the process of common sense facts acquisition and verification as the most promising idea in this field. Well designed interaction with humans can provide both high quality of the statements and reasonable volume of assertions.

We created an on–line questions game Piclick[1] for common sense facts acquisition. This is an implementation of twenty questions game, where one person thinks of a concept while the other asks him a series of *yes/no* questions and attempts to guess what his partner thinks of. Here our system is in role of the guesser, in each step providing players with 3 unanswered questions to choose from. Every given answer maps to a fact in the knowledge base, and that fact's confidence depends on coherence among players. Only assertions that multiple players agreed upon yield 100% certain facts. A core characteristic of the system is that its knowledge base extends the WordNet dictionary. Having started from the sole Wordnet, the system was acquiring facts and evolving towards a rich common sense knowledge base. Utilizing the great amount of concepts defined in the dictionary allowed to only learn new features about them rather than learning a whole knowledge base from scratch. Also making use of WordNet's numerous *IS_A* relations makes new features propagate faster in the knowledge base. Finally, the result of the training process is a common sense knowledge base where most of the facts can be expressed using WordNet synsets and a few new relations.

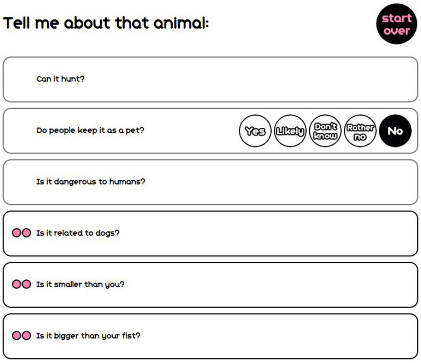
The structure of the rest of this paper is as follows: In section 2 we describe the Piclick game with the algorithms behind. Section 3 reports the results we achieved acquiring common sense facts with the game. Discussion and our future plans conclude the paper in section 4.

2 The Piclick Game

Every week, the time people spend playing video games totals to over 3 billion hours [11]. Players perform great number of tasks that require intelligence and intuition, although the results are of no value in the real–world. In the past few years, many researchers and game designers got involved in attempts to make use of the players' computing power to solve real–world problems [9, 11–13]. A few of those games with a purpose [13] have proven a great potential of crowdsourcing difficult problems to video games' players. Some of them were successfully employed to gather common sense facts [7] and also to verify existing assertions [14].

We propose a system to extend and improve WordNet using twenty questions game (Figure 1). This is a word game where one player thinks of a concept and the other one (in this case, our server) asks a series of yes/no questions regarding that concept. Once collected a few answers, the latter guesses what the concept his partner was thinking of was. Upon each game's completion, some knowledge can be added to the system. This is not limited to the user's answers during

[1] `https://kask.eti.pg.gda.pl/pinqee/game`



Tell me about that animal:

Can it hunt?

Do people keep it as a pet?

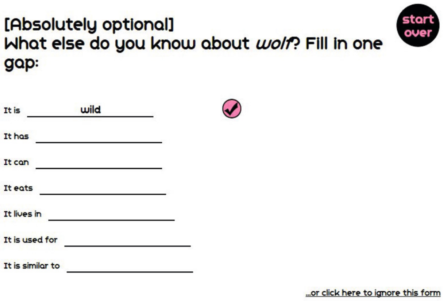
Is it dangerous to humans?

Is it related to dogs?

Is it smaller than you?

Is it bigger than your fist?

Fig. 1. The Piclick game's interface



[Absolutely optional]
What else do you know about *wolf*? Fill in one gap:

It is wild

It has

It can

It eats

It lives in

It is used for

It is similar to

...or click here to ignore this form

Fig. 2. Templates let users enter new concepts into the system

the game. Having cleared up what the subject of the game was, the system prompts the player to answer yet another question: *what do you know about [concept]?* User can reply to that by submitting a fact in a form of an object–relation–object triple. Similarly as in Verbosity, we provide users with sentence templates containing blank fields for the feature object (Figure 2). Previously we experimented asking users to fill in two blank fields for concept and relation, although for many players it was not clear which parts of the sentence they were expected to enter in those separate fields. During the tests in animals domain, templates for 7 relations were made available: *is* (meant to link with adjectives as opposed to IS_A), *can*, *lives_in*, *eats*, *used_for*, *similar_to*, *bigger_than* and *smaller_than*.

There is another games' side effect except for learning facts about concepts. For each question, so called *comprehension rate* is computed as a fraction of *don't know* answers to all answers. This rate is taken into account when selecting questions in later games, but more universally, it is an indicator of how easy to understand a concepts is. For example, comprehension rate equals 0 for *does it have peristome?* and 0.97 for *does it have bones?*. Those values, treated as a measure of concepts' comprehension, stratify otherwise flat WordNet.

In Verbosity, Narrators could send bogus hints to Guessers in order to complete the game faster, e.g. by spelling the answer in separate clues or sending look–alike or sound–alike words [10]. In Piclick players interact with the system, so there is little chance of such cooperative cheating. Players verify each others assertions as they answer the same questions. Contradicting statements either loose the strength of connection between concept and a feature or make it effectively disappear. Such design is to assure good quality of the facts in the knowledge base.

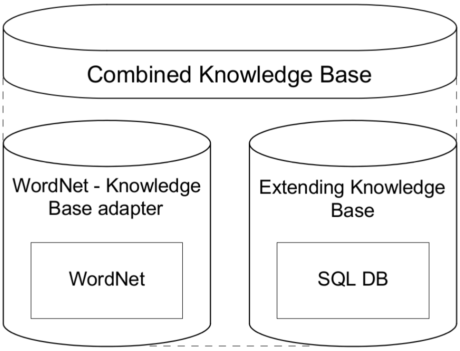


Fig. 3. Knowledge base structure

2.1 Knowledge Base

Local WordNet interface [15] fed with WordNet 3.0 files was wrapped by our knowledge base adapter. All the information gathered from users during the games is stored in a separate SQL knowledge base. Those two are aggregated and combined by an external instance, as shown in Figure 3. Such approach is very flexible since the acquired knowledge is kept separate as a stand alone and logically coherent knowledge base that can be used on its own. Decoupled knowledge base architecture allows to introduce modifications to the knowledge combining logic. Also, throughout the whole lifetime of the system a mapping between WordNet 3.0 and the whole knowledge base is maintained.

The knowledge base is structured in form of a semantic network built up from the fuzzy triples of knowledge expressed in the *vw*ORF notation [16]. Figure 4 depicts an example assertion expressed this way. Here the fact that birds have wings was associated with parameters support $w = 0.85$ and confidence $v = 0.97$. This means that a bird most usually has a wing and that this assertion is almost sure. The *vw*ORF notation is capable of expressing elementary sentences. More complex assertions, e.g. that birds usually have two wings, cannot be expressed this way, but this could be enabled by allowing a whole triple on the "object" side of another triple.

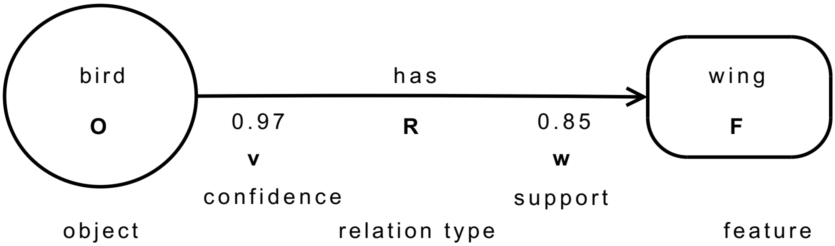


Fig. 4. Example assertion in the vwORF notation

Each triple in the knowledge base is referenced by at least one assertion, an entity representing an atomic piece of information (e.g. a single "click" user does in the game or a statement from WordNet). Every assertion is assigned a weight, it also carries a support value analogue to the triple's w. For a given triple, its parameters w and v functionally depend on referencing assertions, namely on their weighted average (w) and standard deviation (v). When all the individual assertions indicate equal feature support, the standard deviation is 0 and the confidence value equals 1. The confidence decreases as standard deviation goes up for incoherent assertions; in a boundary case of opposite assertions (such of support values 1 and -1), standard deviation approaches 1 and confidence tends to 0.

The weights were set equal for both types of assertions: coming from players' "clicks" and those reflecting WordNet's relations. This way, if a single user contradicts a statement acquired from WordNet, appropriate triple is given 0 confidence. At least two contradicts are required to invert the support of a triple sourced from WordNet.

2.2 The Search Algorithm

We develop a semantic search algorithm [17] that's objective is to minimize the number of questions required to identify the concept user has in mind. Semantic network representation is easily browsable for humans, but only little amount of the overall information stored in the network is stated explicitly. For the sake of computations, the knowledge is translated into the form of a semantic space. The semantic space is a high dimensional space where one dimension is associated with a unique pair (*relation*, *feature*). Those pairs are referred to as questions. Each concept in the semantic space is represented by a Concept Description Vector (CDV) [16]. Every element e of each CDV holds two real numbers e_w and e_v denoting respectively support of the connection and the confidence of knowledge. Those values are copied from corresponding triple, if such exists, or inherited from another vector, provided that there is a positive support path with relation *IS_A* going to that vector.

The search algorithm operates on the semantic space constructed from the semantic network. Over the consecutive iterations, a subspace $O(ANSV)$ of candidate vectors is maintained. Any CDV can enter or leave $O(ANSV)$ in between two iterations. The answers that user gives build up a vector ANSV. Each time a new answer appears, similarity s between ANSV and every CDV is computed:

$$s = \frac{ANSV \cdot CDV}{|ANSV| \cdot \sqrt{\sum_{q:ANSV[q]\neq 0} CDV\,[q]^2}} \tag{1}$$

where $CDV[q]$ denotes the confidence and support product of CDV's element for dimension q.

In each iteration system computes questions that are the most informative across the candidates subspace $O\left(ANSV\right)$. Any CDV is included into $O\left(ANSV\right)$ with probability p:

$$p = \begin{cases} 0, & \text{if } s \leq s_l \\ \frac{1}{2} - \frac{1}{2}\cos\left(\pi\frac{s-s_l}{s_u-s_l}\right) & \text{if } s_l < s < s_u \\ 1, & \text{if } s \geq s_u \end{cases} \quad (2)$$

where:

s_u – safety threshold set to $s_{max} - 0.1$

s_l – rejection threshold set to $s_{max} - 0.4 + 0.03 \cdot k$

s_{max} – current iteration highest similarity

k – number of affirmative answers so far

Once the subspace $O\left(ANSV\right)$ is defined, each question q is assigned information gain IG:

$$IG = -\frac{\omega_\oslash}{\omega} \cdot \left(\frac{\omega_\ominus}{\omega_\oslash}\log_2\frac{\omega_\ominus}{\omega_\oslash} + \frac{\omega_\oplus}{\omega_\oslash}\log_2\frac{\omega_\oplus}{\omega_\oslash}\right) \quad (3)$$

where:

ω – weight of all CDV in $O\left(ANSV\right)$

$\omega_\oslash$ – weight of all CDV such that $CDV\left[q\right] \neq 0$

$\omega_\oplus$ – weight of all CDV such that $CDV\left[q\right] > 0$

$\omega_\ominus$ – weight of all CDV such that $CDV\left[q\right] < 0$

and weight of CDV is equal to its similarity to ANSV. Questions are then ordered according to a rank R:

$$R = \frac{IG + 3 \cdot CR}{4} \quad (4)$$

where CR denotes question's comprehension rate, the fraction of answers *don't know* to all answers or 1 when the question was answered less than 2 times. In the set of the highest rank questions we compute correlation between pairs and filter out duplicates (e.g. one of the questions *is it a mammal?* and *does it have hair?* would be removed). In the pair of correlated questions the one with lower R value is left out. Then the player is prompted to answer one of the 3 questions of the highest R after the filtering.

2.3 Word Sense Disambiguation

In the last stage of the game users are prompted to fill in sentence templates in order to convey additional facts regarding the game's subject. Those sentences then need to be expressed in a form of triples of knowledge, so that they are populated in the knowledge base. In order to achieve this, the feature typed in by the user must be mapped to an appropriate knowledge base concept. The implemented word sense disambiguation method makes use of concepts' descriptions sourced from WordNet glosses.

For each sentence template there is a part of speech (POS) constraint defined as well as a set of keywords that are likely to occur in the right concept's gloss. During the disambiguation all the concepts linked with the specified word are retrieved. First, the concepts marked with different POS are filtered out. Then, concepts' descriptions are searched for the keywords. The one containing the biggest number of keywords is selected. For example, the template *it has* _________ *color* was assigned an adjective POS constraint and keywords [*color, colour*]. Whenever player fills in this template – typing for example *blue* – all adjectives linked to the word *blue* are retrieved. There are 8 of them, but only one contains the word *color* in its gloss.

3 Results

The knowledge base itself is represented as a semantic network. Due to feature inheritance, full information is explicitly stated only in semantic space, where concepts are represented as vectors of features. Therefore the amount of knowledge gained by the system was evaluated by comparing the semantic space reflecting the sole WordNet knowledge to the space after the games were played. Since the games were only conducted in the animals domain, the following statistical information is identically restricted.

The following results come from a state of semantic space captured after 700 games in animals domain. Table 1 presents numbers of vectors and questions (dimensions) in the semantic space before and after the games were conducted. Number of vectors did not change as a result of the games, since all the animals that players thought of were already defined in WordNet and represented in the knowledge base. Semantic space dimensionality grew up as users were adding new features by filling the templates at the final stage of the game.

Vectors in the semantic space hold support and confidence for each dimension (feature). It is reasonable and convenient to take product of those values when considering concept's relation to a given feature. When this product's module is greater than 0.2, we consider it an assertion. Tables 2 and 3 present average number of assertions per concept available in the semantic space before and after the games were conducted. Average number of facts per concept raised from 24.6 to 46.5, which is over 89%. More than half of the facts based on new relation types are expressed with relations IS and CAN. Figure 5 presents how the number of assertions changed in function of the games played. In the beginning there were no negative facts in the semantic space since there aren't any in WordNet.

We performed a test to evaluate the quality of the newly acquired knowledge. Facts about 50 random animals that occurred as game subjects were evaluated by playing the game one more time. All those games were conducted in such a way that the first listed question was always the one that got answered, and the answer *I don't know* was never clicked. Any dubious questions were checked in online encyclopedias. Out of those 50 games 29 ended successfully, and features of those animals that were not recognized were inspected manually. We found the

Table 1. Size of the semantic space before and after the games

	before	after
# vectors	4017	4017
# questions	1059	1410

Table 2. Total and average (per concept) numbers of assertions in the semantic space before the games

		assertions	
		total	avg.
relations	IS_A	34878	8.7
	HAS_A	63870	15.9
	all	100023	24.6

Table 3. Total and average (per concept) numbers of assertions in the semantic space after the games

		assertions			
		affirmative		all	
		total	avg.	total	avg.
relations	IS_A	36123	9.0	36123	9.0
	HAS_A	68922	17.2	90471	22.5
	IS	3682	0.9	16288	4.1
	CAN	5537	1.4	13662	3.4
	LIVES_IN	3336	0.8	7224	1.8
	EATS	1357	0.3	3389	0.8
	USED_FOR	132	0	496	0.1
	SIMILAR_TO	2182	0.5	6711	1.7
	BIGGER_THAN	3174	0.8	5877	1.5
	SMALLER_THAN	3849	1	4845	1.2
	all	128294	31.9	185086	46.1

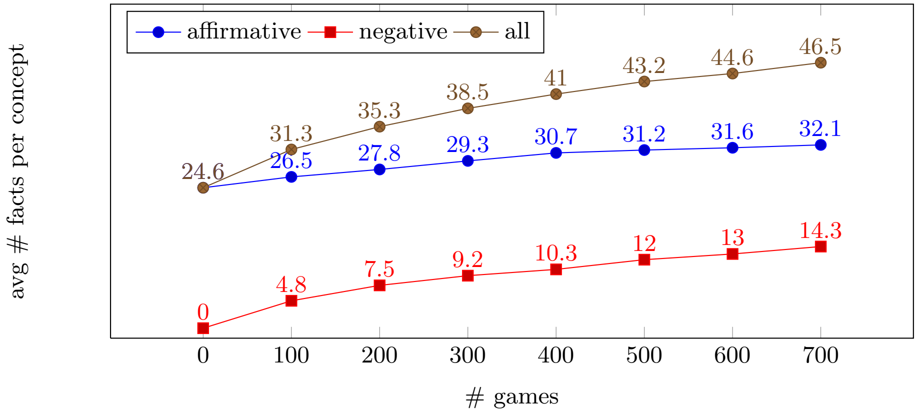


Fig. 5. Average number of affirmative, negative and total facts depending on number of games played

quality of the facts learned by the system satisfactory and learned that the main failure reason was lack of distinctive features rather than wrong facts acquired by the system. For example, there was no feature differentiating between hippo and rhinoceros. Without such feature, the system could not learn to tell these animals apart regardless on the number of training games. We also noticed that relatively formulated features only introduce noise: *it can run very fast* turned out to be useless, but *it can run faster than a domestic cat* gives much better results.

4 Future Plans

700 games played so far provided the system with significant number of new facts - average number of facts per concept raised over 89%. However, much more games are needed until the knowledge base stratifies by facts' confidence. Only then it will be possible to evaluate the most confident new facts in terms of uniqueness and accuracy. The number and quality of those facts will be the project's actual contribution. Fortunately, the game got enough attention to say that such evaluation will be possible within the next months.

Results obtained after the game was released to public indicate that minor moderation must be run in order to make optimal use of players' efforts. It is often necessary to rephrase newly acquired questions so that they become more clear or to merge duplicate features. Sometimes players make mistakes verifying results of the games. Also they are allowed to enter new features themselves, so it happens that noise objects appear in the knowledge base. More management tools should be implemented to minimize noise acquired by the system.

So far, the game has been collecting new facts within the animals domain. We will certainly attempt to make a selection of domains available in order to make the gameplay more attractive and acquire facts in other fields. Also different variants of the game may be investigated, e.g. we can let the player ask questions or assign him/her a concept and encourage to finish the game in minimal number of steps.

As soon as reasonable number of games is completed in various domains, we will be able to make them available online, in a similar fashion to ConceptNet's web interface. Synchronization with WordNet synsets will assure those facts be highly usable.

Acknowledgments. The work has been supported by the Polish Ministry of Science and Higher Education under research grant N N516 432338.

References

1. Miller, G.A., Beckitch, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to wordnet: An on-line lexical database (1993)
2. Lenat, D.B., Guha, R.V., Pittman, K., Pratt, D., Shepherd, M.: Cyc: toward programs with common sense. Communications of the ACM 33(8), 30–49 (1990)
3. Niles, I., Pease, A.: Towards a standard upper ontology, pp. 2–9. ACM Press (2001)
4. Speer, R., Havasi, C.: Representing general relational knowledge in conceptnet 5. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA), Istanbul (May 2012)
5. Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T., Zhu, W.L.: Open mind common sense: Knowledge acquisition from the general public, pp. 1223–1237. Springer (2002)
6. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Web Semant. 7(3), 154–165 (2009)

7. von Ahn, L., Kedia, M., Blum, M.: Verbosity: a game for collecting common-sense facts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2006, pp. 75–78. ACM, New York (2006)
8. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science 321(5895), 1465–1468 (2008)
9. Eiben, C.B., Siegel, J.B., Bale, J.B., Cooper, S., Khatib, F., Shen, B.W., Players, F., Stoddard, B.L., Popovic, Z., Baker, D.: Increased diels-alderase activity through backbone remodeling guided by foldit players. electronically (February 2012), `http://dx.doi.org/10.1038/nbt.2109`
10. Speer, R., Havasi, C., Surana, H.: Using verbosity: Common sense data from games with a purpose (2010)
11. McGonigal, J.: Reality Is Broken: Why Games Make Us Better and How They Can Change the World (2011)
12. Riedel-Kruse, I.H., Chung, A.M., Dura, B., Hamilton, A.L., Lee, B.C.: Design, engineering and utility of biotic games. Lab Chip 11, 14–22 (2011)
13. von Ahn, L.: Games with a purpose. IEEE Computer Magazine 39(6), 92–94 (2006)
14. Herdagdelen, A., Baroni, M.: The concept game: Better commonsense knowledge extraction by combining text mining and a game with a purpose (2010)
15. Gerber, M.: C# API for WordNet 3.0. (2012), `https://ptl.sys.virginia.edu/ptl/members/matthew-gerber/software#wordnet` (accessed April 10, 2012)
16. Szymański, J., Duch, W.: Information retrieval with semantic memory model. Cognitive Systems Research 14, 84–100 (2012)
17. Rzeniewicz, J., Szymański, J., Duch, W.: Adaptive algorithm for interactive question-based search. In: Shi, Z., Leake, D., Vadera, S. (eds.) Intelligent Information Processing VI. IFIP AICT, vol. 385, pp. 186–195. Springer, Heidelberg (2012)

Improvement of Imperfect String Matching Based on Asymmetric n-Grams

Julian Szymański and Tomasz Boiński

Department of Computer Architecture,
Gdańsk University of Technology, Poland
{julian.szymanski,tomasz.boinski}@eti.pg.gda.pl

Abstract. Typical approaches to string comparing treats them as either different or identical without taking into account the possibility of misspelling of the word. In this article we present an approach we used for improvement of imperfect string matching that allows one to reconstruct potential string distortions. The proposed method increases the quality of imperfect string matching, allowing the lookup of misspelled words without significant impact on computational effectiveness. The paper presents the proposed method, experimental data sets and obtained results of comparison to state of the art methods.

Keywords: imperfect string matching, information retrieval, recognition memory.

1 Introduction

The transformation performed by recognition memory from the space of perceptions into a space of inner representations, allows one to put on the perceived things structures that organize the perception. This mechanism is also used on the linguistic level where particular words should be selected from the stream of perception, which is a prerequisite of language understanding. The recognition memory works for all different modes of perception. On a visual level it reconstructs elements of the images, and an on audio channel it allows to identify words.

In this article we focus on the aspect of recognition memory which allows the correction of string characters using the algorithm performing efficient imperfect string matching. It should be stressed that this is only one of the tasks that recognition memory can perform, and its implementation can serve as an elementary block of a cognitive system. The task of string recognition (or matching the string to the one stored in the dictionary) is a very important element in information retrieval [1] systems so it will be analysed also from that perspective. The task of imperfect string matching can also be applied in spell checking, thus making its implementation very useful.

String comparison performed by machines is typically a binary process: two literals are the same or not. The computers usually do not consider the strings as similar, but only the same or different. While we are processing the natural language in written form, where word misspellings are common, the selection of words, non-words and their correspondences become an important issue. It should be noticed that selecting a subset of words from the set of all possible character combinations is significant.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 306–315, 2013.

In retrieval systems, selecting the exact matches of the particular user query is not sufficient, due to possible misspellings and also different inflections (not considering semantics of the utterances). Other uses of presented methods include sequence analysis systems such as comparing genomes [2].

2 Imperfect Matching Methods

Most of the approaches to imperfect string matching are based on metrics that describe distances between two strings. The most intuitive was introduced by Hamming [3]. The Hamming distance for two words of the same length is the number of places where they differ. For text correction this measure is not very useful as it takes into account only one possible typing error – change of a letter.

An improvement of the string matching algorithm has been proposed by Levenshtein [4]. The measure calculates a minimal number of operations, such as letter removal or inserting that leads to transforming one string to another. Two same strings will have a Levenshtein distance equal to 0. If they are different, minimal distance will be 1 and the maximal will be the length of the longer string. The algorithm in its basic form is a bit slow but it has been optimized and according to Exorbyte it allows one to check 2.5 mln words in 10 seconds [5] and in that form it is used eg. in *Yahoo!* search engine.

The Levenshtein distance has one serious drawback. It does not take into account operation of changing two letters. Research indicates that this type of error is one of the common misspellings that humans do (80 %) [6]. The measure was adopted to take into account this additional operation. The modified distance is named Damerau-Levenshtein. Despite the adding of one additional operation that does not seem to be complicated, it strongly influence the computational effectiveness of the algorithm.

Aforementioned methods for string comparison are basic approaches for imperfect string matching [7]. Their advantage is not very complicated implementations. The Hamming distance is fast but it is only limited to strings of the same length. The Damerau-Levenshtein measure is known to have approximately 80% efficiency in misspellings correction. The drawback of these methods is their limitation to the selected number of basic operations on strings that not cover all possible errors humans can do. E.g. semantically the Polish word *pies* (dog) is closer to the misspelled word *pias* than *wies* (countryside), but using presented distances they have the same distance because they have the same difference – a letter change.

The problems with measures based on editing distances have been compensated for by using extended representations of the strings. One of the most popular approaches are n-grams that are widely used in sequence analysis. An n-gram is an n–element subsequence taken from a given sequence. Usually the n-grams are created by dividing the original sequence on the pieces having equal n length that may overlap. The method is similar to the approach used in Support Vector Machines where the increase of dimensions with a particular kernel allows one to separate objects that are not separable in higher dimensions. The imperfect string matching based on n-grams representation usually gives more precise results than the former presented approaches based on editing distances [8].

The n-gram approach also allows one to save disk memory. The methods based on editing distances require you to store a whole dictionary in memory. The n-grams approach can optimize memory usage by creating the all n-grams dictionary. It should be noticed that the size of the dictionary is proportional to n. If n is considerably smaller than average string length, it will take much less space. To optimize the n-grams approach and to create effective word representation it is required to store only references to the proper n-grams that store only reference i.d. instead of whole strings, which is known to be more efficient [8].

To compare the n-grams representations, usually two approaches are used: slower, based on Marcov models that allows capturing the order of n-grams, and a faster method that is based on comparison of sets, where order is omitted. Due to the effectiveness of n-grams representations usage of Tversky index (known for set-theoretic approach for string comparing) provide similar results to Marcov models [8].

The methods presented above are widely used. Among them open-source approaches for string correction should be mentioned, namely Hunspell, Aspell, Ispell [9]. They can be used both as stand alone applications, or by third party software. Ispell is based on Damerau-Levenshtein metrics where the distance is not higher then 1 and it does not take into account spelling rules. Its successor Aspell is optimized for over 70 languages. It is also based on Damerau-Levenshtein metrics but contains optimisations and extensions for improving the speed of comparisons by using cached list of elements called soundslike. The Aspell has been integrated with many applications such as Notepad++, the older version of Opera, Gedit, AbiWord. Hunspell is a spelling corrector and morphological analyser. It has been enriched with an approach based on n-grams. It has been integrated with Google Chrome, LibreOffice, LyX, Mozilla Firefox, Opera 10+.

3 Extension of n-Grams: bi2quadro-Grams

In our approach we use a different method of string representation than that typically used with n-grams. We performed series of experiments comparing different combinations of n-grams used for representation as well as a source word (given for assessment) and target words (words stored in the dictionary). The experiments have shown that much better results can be achieved if we use different representations of words in the dictionary and source words. Typical approaches use the same n for source and target word, typically $n = 2$. In our approach the dictionary words are represented with quadro-grams and the source word is represented with bi-grams. As representations of source and target words are different the key issue is the way of their comparison. We can not compare source and target n-grams directly, as their n numbers are different. Thus, the method for comparing different n-grams lengths is crucial here.

3.1 Different n-Grams Length

Having different n-grams length the calculation of the similarity metric is performed within different dimensions. Thus, both n-grams – source and target should be mapped into the same dimension, in other words their number should be equal. The number ng of n-grams generated for a given string equals $ng = length(string) - (n - 1)$.

The solution for performing comparison of n-grams with different lengths was to add additional blank letters to the representation with bigger n, and perform the similarity evaluation after mapping the bi-gram onto fragments of higher n-grams. In that way each of bi-grams have their corresponding place in higher n-gram representation. In our case we test tri- and quadro-grams.

Tri-gram representation is extended by adding one blank letter at the end of the last gram, quadro-grams are extended by adding blank letter at the end of the last gram and blank letter at the beginning of the first one. The example of partition of the strings for comparing tri- and quadro-grams with bi-grams has been shown in Table 1.

Table 1. Example for dividing the string on bi/tri/quadro-grams

string	n-grams								
klawiatura	kl	la	aw	wi	ia	at	tu	ur	ra
klawiatura	kla	law	awi	wia	iat	atu	tur	ura	ra_
klawiatura	_kla	klaw	lawi	awia	wiat	iatu	atur	tura	ura_

3.2 Calculating n-Grams Similarity

The similarity between two strings can be expressed as a number of matched n-grams. This value is in the range $[0, source\_bi\text{-}grams\_number]$. After normalization the strings similarity represented with n-grams is expressed using Formula 1.

$$similarity = \frac{number\_of\_matched\_bi\text{-}grams}{total\_number\_of\_bi\text{-}grams}, \tag{1}$$

In the results we presented we use distance between two strings that is calculated as $distance = 1 - similarity$.

The similarity between two n-grams having different n can be calculated by counting the number of n-grams from the smaller n representation in the higher one. However, this approach gives good results only if the source and target words are the same. If in the source word some errors appear, such as insertion of an additional letter, or shifting of letters, the approach gave unintuitive results.

The examples in Tables 2 and 3 show such cases when a letter is removed or inserted. The examples also show that different similarity values can be achieved for different n-gram sizes – the last column shows the number of n-grams matched.

Table 2. Letter removal – comparison of bi-, tri- and quadro-grams

n-gram type	string	n-grams							matched n-grams
source string	podstwka	po	od	ds	st	tw	wk	ka	7
bi-grams	podstawka	po	od	ds	st	ta	aw	wk	4
tri-grams	podstawka	pod	ods	dst	sta	taw	awk	wka	6
quadro-grams	podstawka	_pod	pods	odst	dsta	staw	tawk	awka	6

Table 3. Letter insertion – comparison of bi-, tri- and quadro-grams

n-gram type	string	n-grams							matched n-grams
source string	tabelica	ta	ab	be	el	li	ic	ca	7
bi-grams	tablica	ta	ab	bl	li	ic	ca	a_	2
tri-grams	tablica	tab	abl	bli	lic	ica	ca_	a__	2
quadro-grams	tablica	_tab	tabl	abli	blic	lica	ica_	ca__	5

What can be seen from the examples provided in Tables 2 and 3 in the case of letter removal, using both tri-grams and quadro-grams representations for target word gave identical results. In the case of letter insertion, using quadro-grams gave better results than using bi- or tri-grams. However not all bi-grams of the source word could be matched.

As the comparison of n-grams should take into account different possibilities of spelling errors, we identify different types of matching:

- exact match: $ab \rightarrow abcd$.
- shifted match $ab \rightarrow acbd$.
- commutative match $ab \rightarrow cabd$.
- reverse match.: $ab \rightarrow bacd$.
- adding additional letter $ab \rightarrow abcx$

The introduction of those matchings increases efficiency of the algorithm. E.g. in case of letter removal the number of matched bi-grams increases (Table 4).

Table 4. Letter removal – comparison of bi-, tri- and quadro-grams

n-gram type	string	n-grams							matched n-grams
source string	podstwka	po	od	ds	st	tw	wk	ka	7
bi-grams	podstawka	po	od	ds	st	ta	aw	wk	4
tri-grams	podstawka	pod	ods	dst	sta	taw	awk	wka	7
quadro-grams	podstawka	_pod	pods	odst	dsta	staw	tawk	awka	7

Table 5. Weights for different types of matchings

Type of matching	Symbol	Weight
exact match	ab	$1,0$
shifted match	bc, da	$0,8$
commutative match	ac, db	$0,8$
reverse match	ba, cb, ad, ca, bd	$0,1$
adding additional letter	x	$0,7$

The introduction of different types of matching allows us to add weights to the similarity measure that focus the similarity measure on the particular aspect of the matching. Without the weights the results can be misleading. E.g. distance between “podstwka” and “podstawka” will be 0 which is obviously not true. Thus each type of the aforementioned matchings was assigned a weight as presented in Table 5.

Taking weights into account gives more intuitive results. The example for a word with a removed letter is presented in Table 6. Type, number of found corresponding n-grams, and distance between the source and target word is presented in Table 7.

In all cases, results obtained by using tri- and quadro-grams were better than when using only bi-grams. For exact, shifted, commutative and reverse matchings the results obtained for tri- and quadro-grams were identical. When the source word had an additional letter inserted, the usage of quadro-grams gave better results than bi- and tri-grams representations. Such an example is presented in Table 8. Type, number of found corresponding n-grams, and distance between the source and target word is presented in Table 9.

The examples presented above indicate that the standard n-grams approach allows you to precisely compare only identical words. This method does not take into account possible spelling errors. Introduction of weighted types of matchings allows for recognition of misspelled words based on their similarity to the correct ones. Thus, despite

Table 6. Letter removal – example including weights

n-gram type	string	n-grams							matched n-grams
source word	podstwka	po	od	ds	st	tw	wk	ka	7
bi-grams	podstawka	po ab	od ab	ds ab	st ab	ta -	aw -	wk -	4
tri-grams	podstawka	pod ab	ods ab	dst ab	sta ab	taw ac	awk bc	wka bc	7
quadro-grams	podstawka	_pod ab	pods ab	odst ab	dsta ab	staw ac	tawk bc	awka bc	7

Table 7. Letter removal – distance including weights

n-gram type	string	number of bi-grams	distance
source word	podstwka	7	–
bi-grams	podstawka	$4ab$	$1 - 4 \cdot 1/7 \approx 0,43$
tri-grams	podstawka	$4ab$, $1ac$, $2bc$	$1 - (4 \cdot 1 + 1 \cdot 0,8 + 2 \cdot 0,8)/7 \approx 0,09$
quadro-grams	podstawka	$4ab$, $1ac$, $2bc$	$1 - (4 \cdot 1 + 1 \cdot 0,8 + 2 \cdot 0,8)/7 \approx 0,09$

Table 8. Letter insertion – example with weights

n-gram type	string	n-grams							matched n-grams
source word	tabelica	ta	ab	be	el	li	ic	ca	7
bi-grams	tablica	ta ab	ab ab	bl x	li x	ic -	ca -	a_ -	2
tri-grams	tablica	tab ab	abl ab	bli x	lic x	ica -	ca_ -	a__ -	2
quadro-grams	tablica	_tab ab	tabl ab	abli x	blic x	lica da	ica_ da	ca__ da	5

Table 9. Letter insertion – distance with weights

n-gram type	string	number of bi-grams	distance
source word	tabelica	7	–
bi-grams	tablica	$2ab$, $2x$	$1-(2\cdot 1+2\cdot 0,7)/7 \approx 0,51$
tri-grams	tablica	$2ab$, $2x$	$1-(2\cdot 1+2\cdot 0,7)/7 \approx 0,51$
quadro-grams	tablica	$2ab$, $3da$, $2x$	$1-(2\cdot 1+3\cdot 0,8+2\cdot 0,7)/7 \approx 0,17$

having to map different dimensions of n-grams, obtained results show that by the introduction of handling for spelling errors, quality of matching increases.

3.3 Different Word Lengths

Bi-grams matching allows string comparison when for each bi-gram of source word we have a corresponding bi-gram available for the target word. If a target word is shorter than the source one, some of bi-grams cannot be matched. In that case last grams from the target word will not be compared, and thus they do not influence the results. It results in the cases where comparing the word $klawiatora$ with $klawiatura$ or $klawiaturaaaaa$ gives the same result which is not acceptable.

To compensate for this we added a penalty for a difference in word length between target and source one. The penalty is calculated as presented by Formula 2. It is then added to the distance, thus increasing it for every additional letter.

$$penalty = \frac{length\_of\_source\_word - length\_of\_target\_word}{10} \tag{2}$$

The difference between lengths has been divided by 10 to be normalized into $[0, 1]$. We assume that the difference in lengths of source and target words is smaller than 10. The penalty allows one to eliminate the words that are similar to the source one in the n-grams located at the beginning, but which are longer.

4 The Experiments

For the experiments presented in this paper two open-source dictionaries were used:

- LibreOffice dictionary – 50 911 words,
- Open Source English words dictionary of 153 222 words [10]

They contains together 162 463 different words, as 41 670 were found in both dictionaries. From those dictionaries a corpora containing 1000 correct words. This corpora was then modified to include 10%, 20%, ... , 100% of misspelled words. As a set of target words we use three publicly available dictionaries: *wikipedia*, *aspell* and *trec*.

Five different approaches to correct distorted words were tested and compared with bi2quadro measure. The results for 1000-word corpora have been presented in Figure 1.

The quality of the algorithm depends on the degree of correctness of the corpora. For corpora that included only properly spelled words for all cases the proposed measure

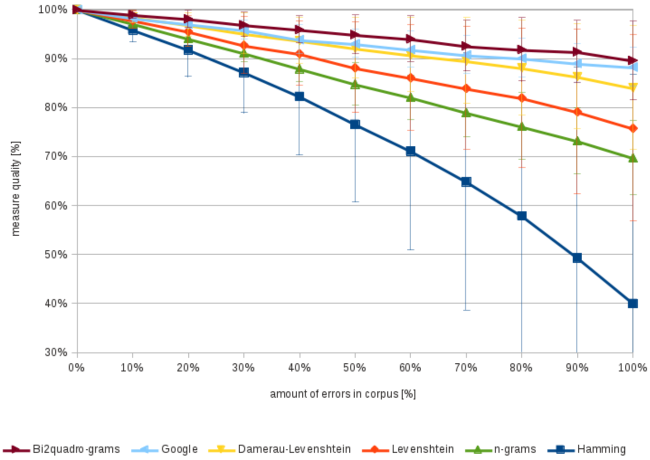


Fig. 1. Quality tests for different approaches to imperfect matching

reaches 100% of correctness which indicates that it does not introduce noise to the dictionary. In all other cases, where test corpora contained erroneous words, our proposed measure achieved better results than other measures. It also behaves better than the one used by Google, that has been tested by providing sample strings into the Google search box and comparing the recommendation of correction returned by search engine.

The proposed measure was also tested using the *wikipedia*, *aspell* and *trec* corpora. The results have been computed using an F-measure and presented in Figure 2. Once again, bi2quadro-grams measure achieved better results than other widely used measures obtaining an average score around 90%.

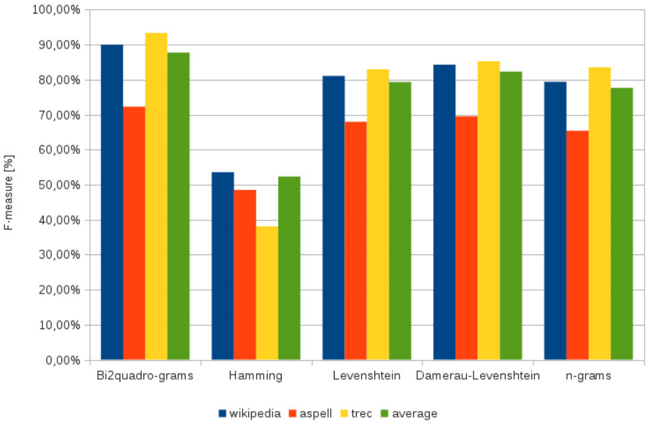


Fig. 2. Quality tests for different measures using *wikipedia*, *aspell* and *trec* corpora

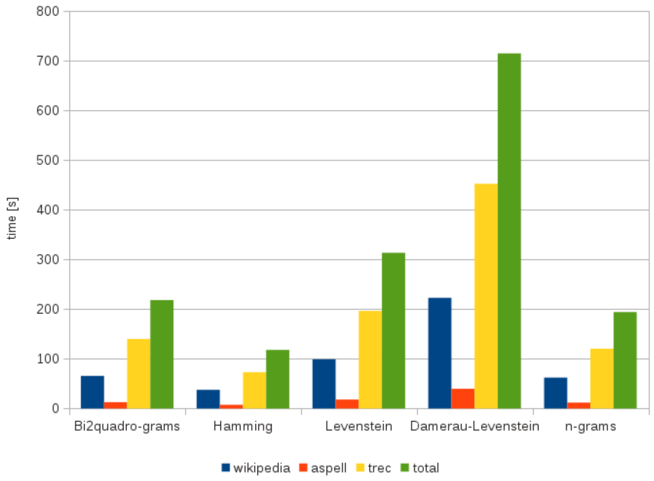


Fig. 3. Performance tests for different measures using *wikipedia*, *aspell* and *trec* corpora

The proposed approach is also competitive in terms of performance. As in other approaches the time needed to perform word lookup is directly related to the size of the corpora. The proposed measure is only slightly slower than the standard n-grams approach, about 2-times slower than Hamming measure and faster than Levenshtein and Damerau-Levenshtein approach while yielding better results (Figure 3).

5 Conclusions

The proposed method of bi2quadro-grams achieves better results than other widely used approaches. The proposed measure achieved almost 90% accuracy, outperforming other known metrics like Hamming, Levenshtein or n-grams measures. The proposed metric also outperforms widely used dictionaries found in Microsoft Word in terms of time. For the dictionary of 162 463 words, the lookup took only 10 ms, whereas Word needed over 15 ms. Bi2quadro-grams also usually produces only one word as a result. Most of the widely-used dictionaries present lengthy lists of words for the user to choose from. As such, the proposed metric could be used in any word lookup and correction applications, yielding better results than most of the available solutions while preserving the computational time.

The presented approach we plan to use in our search engine `http://kask.eti.pg.gda.pl/BetterSearch` aiming at improving information retrieval from Wikipedia. The module that allows you to correct the keywords entered by the user should improve precision of the search. Beside information retrieval, this system offers additional functionalities that allows one to go beyond keyword-based search [11].

Acknowledgements. This work has been supported by the National Centre for Research and Development (NCBiR) under research Grant No. SP/I/1/77065/1 SYNAT: ”Establishment of the universal, open, hosting and communication, repository platform for network resources of knowledge to be used by science, education and open knowledge society”.

References

1. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
2. Saxena, S., Jónsson, Z., Dutta, A.: Small rnas with imperfect match to endogenous mrna repress translation. Journal of Biological Chemistry 278, 44312–44319 (2003)
3. Hamming, R.: Error detecting and error correcting codes. Bell System Technical Journal 29, 147–160 (1950)
4. Lcvenshtcin, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics-Doklady 10 (1966)
5. Sulzberger, C.: Efficient implementation of the levenshtein-algorithm (2009), `http://www.levenshtein.net/` (February 28, 2012)
6. Damerau, F.J.: A technique for computer detection and correction of spelling errors. Commun. ACM 7, 171–176 (1964)
7. Hall, P., Dowling, G.: Approximate string matching. ACM Computing Surveys (CSUR) 12, 381–402 (1980)
8. Navarro, G., Baeza-Yates, R., Sutinen, E., Tarhio, J.: Indexing methods for approximate string matching. IEEE Data Engineering Bulletin 24, 19–27 (2001)
9. Atkinson, K.: Gnu aspell (2011), `http://aspell.net/` (March 07, 2012)
10. WinEdt: Winedt dictionaries - english (uk) (2010), `tug.ctan.org/tex-archive/systems/win32/winedt/dict/uk.zip` (March 14, 2012)
11. Deptuła, M., Szymański, J., Krawczyk, H.: Interactive information search in text data collections. In: Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., Niezgódka, M. (eds.) Intell. Tools for Building a Scientific Information. SCI, vol. 467, pp. 25–40. Springer, Heidelberg (2013)

WikiDetect: Automatic Vandalism Detection for Wikipedia Using Linguistic Features

Dan Cioiu and Traian Rebedea

University Politehnica of Bucharest, Department of Computer Science and Engineering,
313 Splaiul Independetei, Bucharest, Romania
dan.cioiu@gmail.com, traian.rebedea@cs.pub.ro

Abstract. Vandalism of the content has always been one of the greatest problems for Wikipedia, yet only few completely automatic solutions for solving it have been developed so far. Volunteers still spend large amounts of time correcting vandalized page edits, instead of using this time to improve the quality of the content of articles. The purpose of this paper is to introduce a new vandalism detection system, that only uses natural language processing and machine learning techniques. The system has been evaluated on a corpus of real vandalized data in order to test its performance and justify the design choices. The same expert annotated wikitext, extracted from the encyclopedia's database, is used to evaluate different vandalism detection algorithms. The paper presents a critical analysis of the obtained results, comparing them to existing solutions, and suggests different statistical classification methods that bring several improvements to the task at hand.

Keywords: Vandalism Detection, Wikipedia, Natural Language Processing, Classification.

1 Introduction

Wikipedia is an online collaborative encyclopedia, consisting of articles that can be edited by anybody. Updating an article usually increases its value by quality additions or objective corrections. Sometimes, though, Wikipedia's openness eases malicious editors to negatively modify or spam an article or even to plainly delete content.

The modification of a Wikipedia article is called an edit and its state, form and content at a certain moment in time are referred to as a revision. Wikipedia keeps a complete log of all information on the revisions, edit authors and articles evolution.

This paper describes a software system capable of analyzing the text of an article and identifying the attributes that are able to discriminate between malicious edits and constructive and well-intended ones. Wikipedia currently employs some official techniques that address this issue, such as human tracking of articles leading to manual edit reverts, or automated programs specialized in vandalism detection.

Vandalism on Wikipedia can be difficult to define, but it is generally described as a malicious action coming from a person whose goal is to inflict a negative impact on the article and the reader community. Such actions are usually concealed, ironic, or

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 316–325, 2013.

destructive. Some well-intended edits can also have a negative impact, but they shouldn't be thought of as acts of vandalism [1].

An effective way to counter vandalism was introduced by the administrators of Wikipedia through the addition of bots which respond to human commands, safely and rapidly reverting malicious edits. They report their actions and detect users that hide their true IP addresses using proxies.

Our goal is to develop and test a new solution of automatic vandalism detection on Wikipedia. It should be able to process articles extracted from a real source of articles and produce results that are measurable and close to, if not better, than the ones of current solutions.

The task consists in the separation of a data set into two categories: clean (or well-intended) edits and vandalized (or malicious) edits. We must define rules that accurately determine if a revision is the result of vandalism and implement a system that splits the articles into the two categories quickly and efficiently. To solve the data separation problem we will use a machine learning approach, categorizing revisions using several classification algorithms into one of the two classes.

The input data is a set of edits. Each contains information about two consecutive revisions of the same article: the one before and the one after the edit was performed. Given a set of *N* functions, an edit is transformed into a vector with *N*-elements after applying each function on its content.

The paper continues with a brief outline of the previous results obtained using machine learning approaches for the task at hand. Then, section 3 contains the most important details of the implemented system – *WikiDetect*. The most consistent part of the paper is section 4 that explains the evaluation methodology and the subsequent refinements of the system that resulted from this evaluation. The paper ends with a summary of the most important findings and conclusions.

2 Previous Research

Vandalism detection by means of classification algorithms was also addressed by Potthast et al. [1] using logistic regression. Their approach increased the *F*-Measure performance of existent solutions by 49% and achieved 83% precision at 77% recall.

Adler et al. [2] used a decision tree-based classifier. The target variable is the analyzed edit and the input variables are computed based on past versions of the article, user profile and revision comments. A total of 15 decision nodes are applied, including the date and time of the revision, comment length, past and current text histogram. ADTree algorithm, proposed by Freund and Mason [3], was chosen to classify the data and achieved 36.9% precision at 77.1% recall and area under ROC curve of 90.4%.

Harpalani et al. [4] used a compound solution to detect vandalism on Wikipedia articles. The NBTree model is a hybrid between decision trees and classic Bayesian classifiers. Again, an 8-node tree based on 11 characteristics is built. Tree branches determine the path to a leaf after node decisions are applied. When the leaf is reached, a Naïve Bayes algorithm is run and thus, the data is double classified. Experiments revealed 61.5% precision at 25.2% recall and area under ROC curve of 88.7%.

Several attempts were made to incorporate external user reputation tools, such as WikiTrust (www.wikitrust.net). It can be used to extract information on the revision history of certain users or in-time edit statistics over different geographic locations. Using this information usually increases the performance of the vandalism detection systems. For example, in terms of area under ROC curve, West and Lee [5] achieved 95% and Adler et al. [2] achieved 93.4%.

3 System Description

We have approached the vandalism detection task by designing a system capable of processing an input set of data and specifying a set of vandalism verdicts as outputs. Moreover, we have tried to improve the results of the application without using any information about the reputation of the users related to the edit, thus only using information about the edit itself and the article, mainly its text and history of edits.

For evaluation purposes, we partition the set of inputs into training and test data. The detection system follows a machine learning pattern by computing a feature vector for each revision and, based on the training data (that provides correct verdicts for the contained edits), attempts to establish a correspondence between vector values and each of the two categories, regular and vandalized. This is achieved by means of classification algorithms applied to the vectors. We compute several standard information retrieval measures such as precision, recall, and area under ROC curve to evaluate the system's performance and quantify the overall efficiency of the learning mechanisms.

WikiDetect uses for training and evaluation a large corpus provided by the *Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse* (http://pan.webis.de). In the training stage we process the corpus from PAN 2010 and the final tests are run on the data from both 2010 and 2011.

The 2010 *PAN* corpus is made up of over 32,000 training revisions – with membership classes determined manually by authorized annotators –, and of over 130,000 revisions with unknown class membership that must be determined. All data was extracted from Wikipedia and hold information on revisions pre- and post-edit contents, comments, date and user name or IP address [6].

Upon analyzing Wikipedia's policy page on vandalism [7], which describes different types of malicious content changes, we considered a set of 28 independent features as being relevant in the process of vandalism detection. These are: length of the comment associated to the edit, ratio between current and past revision lengths (in characters), length of longest sequence of identical consecutive characters, length of longest word from the article, or edit comment, number of non-alpha characters from added HTML comments, number of suspicious (but non-vulgar) words added to the article, HTML comments, or edit comments, number of new blocks of HTML comments, whether or not the edit is a single-word update, number of consecutive character sequences added, article image, mediawiki tags, or references count variation, whether the user is anonymous, number of vulgar words added to the text, or hidden in HTML comments, ratio between vulgar words and total number of words in the edit comments, number of added pronouns, variation of vulgar words, and pronouns, ratio between number of added uppercase letters and total added letters in the article body, ratio between number of added uppercase letters and total number of characters in the edit comments, number of identical character sequences added, similarity measure between added text and initial text, and between deleted text and resulted text, computed using text vector cosine coefficients, semantic similarity measure between added text and initial text, and between deleted text and resulted text.

WikiDetect uses Java CSV library [8] to read data from the CSV files inside the corpus. Diff, Match and Patch library [9] is then used to detect the differences between the texts of the two edits. We have also integrated Apache Lucene [10] into WikiDetect, as it provides indexing and query mechanisms for large corpora of texts, along with other components such as spellcheckers or result ranking.

The words from WikiDetect's input data are stemmed and added to an index. When querying the index, the ranking system outputs the similarity measure between the query string and matched document. The system uses this feature to determine how similar the new text is compared to the old text. Lucene uses a complex score ranking, which is based on cosine similarity.

WordNet [11] is a lexical collection of nouns, verbs, adjectives and adverbs for the English language. Each word belongs to a dictionary-like set of cognitive synonyms. WS4J [12] is a Java library that implements several word similarity algorithms, working with the WordNet database. WikiDetect uses WS4J to determine the semantic relevance of the added and deleted text against the initial and, respectively, final state of the revision using the similarity measure proposed by Jiang and Conrath [13]. To measure the similarity between two words, we create a chain of words from one end and progress to the other moving across a minimum number of synonym sets. The length of the resulted chain will determine the similarity score. Between two texts, the similarity is the sum of single-highest similarities between each new word and each of the words already in the text.

Queries performed with Lucene and WordNet quantify the quality of the added or deleted text within the revision's context. After it is determined, we can tell whether the user was malicious or added value to the article, using an estimation of the semantic similarity score between the edits.

Data classification is then performed using the Weka data mining framework (http://www.cs.waikato.ac.nz/ml/weka/). We have evaluated the system by running the data against the following classification algorithms: ADTree, NBTree, BayesNet, BayesLogisticRegression and SVM. All but the latter are bundled with Weka. SVM support is provided by LibSVM [14] – a library of SVM tools, and SVM Weka [15] – a Weka interface for LibSVM. We chose to apply these classification algorithms because they were used in the previously described papers and achieved the best results.

4 System Evaluation

4.1 Preliminary Experiments

A series of classification experiments were carried out to analyze the impact of elements such as feature selection, parameter tweaking, or text parsing methods on the system's performance and efficiency. This section describes the major steps we have undertaken and which led to the final training results.

Table 1. Results of the preliminary experiments: unnormalized data, balanced classes

BayesianLogisticRegression – could not be applied						
BayesNet (85.785% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.889	0.241	0.92	0.889	0.905	0.913	0.0
0.759	0.111	0.686	0.759	0.721	0.913	1.0
Weighted avg.						
0.858	0.21	0.864	0.858	0.86	0.913	---
NBTree (85.7749% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.892	0.249	0.918	0.892	0.905	0.912	0.0
0.751	0.108	0.689	0.751	0.718	0.912	1.0
Weighted avg.						
0.858	0.215	0.863	0.858	0.86	0.912	---
SVM (34.4977% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.2	0.2	0.758	0.2	0.316	0.5	0.0
0.8	0.8	0.242	0.8	0.371	0.5	1.0
Weighted avg.						
0.345	0.345	0.633	0.345	0.330	0.5	---
ADTree (86.0575% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.901	0.265	0.914	0.901	0.907	0.908	0.0
0.735	0.099	0.702	0.735	0.718	0.908	1.0
Weighted avg.						
0.861	0.225	0.863	0.861	0.862	0.908	---

Table 2. Results after the second run: well-spread unnormalized data, balanced classes

BayesianLogisticRegression (77.052% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.99	0.919	0.772	0.99	0.867	0.536	0.0
0.081	0.01	0.728	0.081	0.145	0.536	1.0
Weighted avg.						
0.771	0.699	0.761	0.771	0.693	0.536	---
BayesNet (84.5634% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.881	0.265	0.913	0.881	0.896	0.899	0.0
0.735	0.119	0.663	0.735	0.697	0.899	1.0
Weighted avg.						
0.846	0.23	0.852	0.846	0.848	0.899	---
NBTree (84.9571% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.896	0.295	0.905	0.896	0.9	0.895	0.0
0.705	0.104	0.683	0.705	0.694	0.895	1.0
Weighted avg.						
0.85	0.249	0.851	0.85	0.85	0.895	---
SVM – takes too long to process the data						
ADTree (85.0177% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.898	0.299	0.904	0.898	0.901	0.899	0.0
0.701	0.102	0.686	0.701	0.694	0.899	1.0
Weighted avg.						
0.85	0.251	0.851	0.85	0.851	0.899	---

For the first full experiment we computed the features mentioned in the preceding chapter, except the more complex ones that use Lucene and WordNet, on all revisions from the training corpus. We introduced a filter that discards 75% of all regular edits because the two classes are unbalanced – the ratio of regular to vandalized articles was 9:1. Therefore, all the classifiers are trained on 10,000 instances. The results of the first experiments are depicted in Table 1.

It can be seen that, due to the fact that the instance attributes were not normalized, the BayesLogisticRegression classifier could not be applied. Also, SVM shows poor results because the value vectors are not fully compatible with its operation; to maximize efficiency, SVM should rely on data that is well-spread in a certain [minValue, maxValue] domain. Precision and recall with ADTree and NBTree are similar and exceed 70%.

To improve these results, we ran another experiment where we modified the methods that output values observed to be close to certain extremities. Some features were computed relative to total values and generated many 0s and 1s, so we chose to use absolute values that are not influenced by the initial assumptions. Also, we filtered out intervals like *[minValue, x]* and *[x, maxValue]*, where *x* is a distinguishable pole describing mostly regular edits. All numbers from such intervals were replaced by *minValue* and, respectively, *maxValue*. The obtained results are summarized in Table 2.

Table 3. Results after the third run: well-spread normalized data, balanced classes

BayesianLogisticRegression (79.5457% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.839	0.341	0.885	0.839	0.862	0.749	0.0
0.659	0.161	0.566	0.659	0.609	0.749	1.0
Weighted avg.						
0.795	0.298	0.808	0.795	0.8	0.749	---
BayesNet (84.5634% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.881	0.265	0.913	0.881	0.896	0.899	0.0
0.735	0.119	0.663	0.735	0.697	0.899	1.0
Weighted avg.						
0.846	0.23	0.852	0.846	0.848	0.899	---
NBTree (85.0076% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.897	0.297	0.905	0.897	0.901	0.896	0.0
0.703	0.103	0.685	0.703	0.694	0.896	1.0
Weighted avg.						
0.85	0.25	0.851	0.850	0.851	0.896	---
SVM (79.6164% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.836	0.33	0.888	0.836	0.862	0.753	0.0
0.67	0.164	0.566	0.67	0.614	0.753	1.0
Weighted avg.						
0.796	0.289	0.811	0.796	0.802	0.753	---
ADTree (85.0177% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.898	0.299	0.904	0.898	0.901	0.899	0.0
0.701	0.102	0.686	0.701	0.694	0.899	1.0
Weighted avg.						
0.85	0.251	0.851	0.85	0.851	0.899	---

The second experiment showed noticeable improvements – ADTree and NBTree performance has risen to 85% precision and recall, and to an area under ROC curve of just under 90%. Still, SVM needs a long time to produce results and BayesianLogisticRegression produces poor results. To correctly run these two classifiers we must normalize the feature attributes. For the third experiment – see Table 3 – we eliminated the empirical normalization and introduced the following Weka filter for normalization: weka.Filters.Unsupervised.Attribute.Normalize.

The normalization procedure clearly has a positive effect, especially on the results produced by the SVM classifier. It builds an N-dimensional hyperplane, and the normalization guarantees that all dimensions are equally important for final instance verdicts and that they are independent in the process of marking which edits are vandalized and which are not.

4.2 Final Experiment

The results of the previous experiments rely on a subset of features. For the final one, we add the four characteristics that form the base of our semantic analysis of the revisions: the Lucene and WordNet similarity scores for text addition and removal.

Table 4. Final training results: semantic analysis included

BayesianLogisticRegression (84.5028% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.935	0.438	0.87	0.935	0.902	0.749	0.0
0.562	0.065	0.735	0.562	0.637	0.749	1.0
Weighted avg.						
0.845	0.348	0.837	0.845	0.838	0.749	---
BayesNet (85.3508% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.879	0.227	0.924	0.879	0.901	0.908	0.0
0.773	0.121	0.671	0.773	0.718	0.908	1.0
Weighted avg.						
0.854	0.201	0.863	0.854	0.857	0.908	---
NBTree (86.9965% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.920	0.288	0.909	0.920	0.915	0.907	0.0
0.712	0.080	0.740	0.712	0.726	0.907	1.0
Weighted avg.						
0.87	0.238	0.868	0.870	0.869	0.907	---
SVM (84.2908% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.932	0.436	0.870	0.932	0.9	0.748	0.0
0.564	0.068	0.725	0.564	0.634	0.748	1.0
Weighted avg.						
0.843	0.347	0.835	0.843	0.836	0.748	---
ADTree (86.4917% correctly identified)						
TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.927	0.331	0.898	0.927	0.912	0.911	0.0
0.669	0.073	0.746	0.669	0.705	0.911	1.0
Weighted avg.						
0.865	0.268	0.861	0.865	0.862	0.911	---

We believe that the final results of the training phase, depicted in Table 4, should be a combination (such as voting) of all the five classification algorithms. To reach the best combination, we worked on increasing each algorithm's individual performance without limiting to the ones that would singularly perform better.

We consider a set of feature selection search methods, including Greedy, Genetic, Best first, or Scatter search. We apply them to our attributes list using Weka, in order to define a subset of relevant features. As a result of this analysis, we can state that the most defining characteristics of a vandalized Wikipedia revision are: the flag that indicates whether the user is anonymous, Lucene's similarity score for text addition, the number of suspicious phrases, the length of the longest added word, the number of added characters, the ratio between the number of added uppercase letters and total added letters and variation of pronouns in the article body.

4.3 Evaluation on the Test Corpus

The WikiDetect system was evaluated on the test corpora from the *PAN* workshops, both from 2010 and 2011. For this, we built a composite classification model (using voting) by joining the five training sub-models.

Table 4 showed that the algorithms achieve (relatively) distinct performance indicators. Before, they were independently run on the training data and output a verdict of class membership between 0% (clean edit) and 100% (vandalized edit). We joined the five sets of verdicts into a custom corpus and applied a new ADTree classifier for voting, in order to obtain a single set of verdicts.

We calculated the area under the ROC curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC) for class *1.0* (class of vandalized edits). We achieved 0.893 and, respectively, 0.482 for the corpus of 2010 and 0.889 and, respectively, 0.530 for the corpus of 2011. Tables 5 and 6 show a comparison of WikiDetect to the top performing systems developed for *PAN-10* [6] and *PAN-11* [16].

Table 5. WikiDetect compared to the best solutions from PAN-10

ROC-AUC	PR-AUC	Detector
0.92236	0.66522	Mola Velasco [17]
0.90351	0.49263	Adler et al. [2]
0.89856	0.44756	Javanmardi [18]
0.89377	0.56213	Chichkov [19]
0.89375	**0.48294**	**WikiDetect**
0.87990	0.41365	Seaward [20]
0.87669	0.42203	Hagedus et al. [21]
0.85875	0.41498	Harpalani et al. [4]

Table 6. WikiDetect compared to the best solutios from PAN-11

ROC-AUC	PR-AUC	Detector
0.95313	0.82230	West şi Lee [5]
0.88898	**0.53001**	**WikiDetect**
0.82963	0.42464	Drăguşanu et al. [22]

5 Summary and Conclusions

This paper described the steps towards the design and development of a solution for automatic vandalism detection on Wikipedia using only natural language processing and machine learning, without taking into account user profiles and reputation. Several classification algorithms were used and a detailed comparison and analysis of their performance was carried out.

Tests show that we achieved good results and the best classifier is an ADTree (area under ROC curve - 0.91). We chose to run multiple classifiers and observe their verdicts in order to build a final model that would inherit all their best features and reach an overall better performance. However, after independent runs of single classifiers certain vandalism patterns still remain undetected.

Our solution considers the motivations of malicious users and official vandalism categories described by Wikipedia. The fact that this information was not meant to be used in automatic detection software makes their successful integration even more rewarding.

WikiDetect also succeeds in applying syntactic and semantic analysis to test contents. The latter proved to be decisive in measuring the quality of user contribution to the article.

At the end, we mention that the integration of user history and reputation into the process of vandalism detection will further improve the results. This could be made using either the revision history or geographic reputation. Semantic analysis can be improved even further, by comparing user contribution to paragraph and article topics. Advanced pragmatic analysis should also improve the detection rate of vandalism.

Acknowledgement. The research presented in this paper was partially supported by project No. 264207, ERRIC-Empowering Romanian Research on Intelligent Information Technologies (FP7-REGPOT-2010-1).

References

1. Potthast, M., Stein, B., Gerling, R.: Automatic Vandalism Detection in Wikipedia. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 663–668. Springer, Heidelberg (2008)
2. Adler, B.T., de Alfaro, L., Pye, I.: Detecting Wikipedia Vandalism using WikiTrust. In: Proceedings of the 2010 Conference on Multilingual and Multimodal Information Access Evaluation (2010)

3. Freund, Y., Mason, L.: The Alternating Decision Tree Algorithm. In: Proceedings of the 16th International Conference on Machine Learning (1999)
4. Harpalani, M., Phumprao, T., Bassi, M., Hart, M., Johnson, R.: Wiki Vandalysis - Wikipedia Vandalism Analysis. In: Proceedings of the 2010 Conference on Multilingual and Multimodal Information Access Evaluation (2010)
5. West, A.G., Kannan, S., Lee, I.: Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata. In: Proceedings of the Third European Workshop on System Security EUROSEC (2010)
6. Potthast, M.: Crowdsourcing a Wikipedia Vandalism Corpus. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2010)
7. Wikipedia vandalism policy, `http://en.wikipedia.org/wiki/Wikipedia:Vandalism`
8. Java CSV library, `http://sourceforge.net/projects/javacsv/`
9. Diff, Match and Patch library, `http://code.google.com/p/google-diff-match-patch`
10. Apache Lucene, `http://lucene.apache.org/core/`
11. WordNet – a lexical database for English, `http://wordnet.princeton.edu`
12. WS4J library, `http://ws4j.googlecode.com`
13. Jiang, J., Conrath, D.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: Proceedings of International Conference Research on Computational Linguistics, ROCLING X (1997)
14. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
15. SVM Weka, `http://cns.bu.edu/~gsc/CN710/pmwiki.php?n=Main.SVMWeka`
16. PAN 2011 conference website, `http://www.webis.de/research/events/pan-11`
17. Mola Velasco, S.M.: Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals. Notebook Papers of CLEF 2010 LABs and Workshops (2010)
18. Javanmardi, S.: Vandalism detection in Wikipedia: a high-performing, feature-rich model and its reduction through Lasso. In: Proceedings of the 7th International Symposium on Wikis and Open Collaboration (2011)
19. Chichkov, D.: Submission to the 1st International Competition on Wikipedia Vandalism Detection. SC Software Inc. (2010)
20. Seaward, L.: Submission to the 1st International Competition on Wikipedia Vandalism Detection. Universtiy of Ottawa (2010)
21. Hegedus, I., Ormándi, R., Farkas, R., Jelasity, M.: Novel Balanced Feature Representation for Wikipedia Vandalism Detection Task: Lab Report for PAN at CLEF 2010 (2010)
22. Drăguşanu, C.-A., Cufliuc, M., Iftene, A.: Detecting Wikipedia Vandalism using Machine Learning. Notebook Paper for the CLEF 2011 LABs Workshop (2011)

An Improved Application Level Multi-path Streaming Algorithm for HBDP Networks

Dan Schrager and Florin Radulescu

University *Politehnica* of Bucharest, Romania
Faculty of Automatic Control and Computer Science

Abstract. Data intensive computing applications which require access to geographically distributed data are increasingly important in the grid and cloud computing field. We designed a novel approach to the task of transferring bulk data between remote hosts, based on the fork-join queueing model, by combining, at application level, the use of classical pipes as a mean of serializing data locally with multiple parallel TCP connections which ensure high bandwidth throughput. Our new experiments in 1 Gbps and up to 9 Gbps high bandwidth networks have proven the validity of our improved multi-path transmission algorithms. Both simple file transfers and arbitrary data streaming were done effectively, efficiently, in parallel, between pairs of distributed processes connected via extended fast pipes.

Keywords: fork-join queues, pipe programming paradigm, multi-path.

1 Introduction

Our research is placed in the context of grid and cloud computing where the ability to transfer large amounts of data over wide area networks is increasingly important. For example, the Atlas experiment [2] is transferring 10 PB of data a year; or the VIRGO collaboration [18] that has storage needs in the order of many hundreds of TB, and very low latency requirements.

The task of moving large files among distant hosts is usually handled by using modified enhanced versions of TCP [10] (e.g. SACK [15], Vegas [12], HighSpeed [17], FAST [4], MulTCP [9]), non-TCP protocols (e.g. XCP [5]), or dedicated applications and libraries which either use networking parallelism (e.g. XFTP [14], PSockets [8], GridFTP [19]) or UDP [11] as a transport means (e.g. RBUDP [7], SABUL [20]). However, TCP modifications are backward compatible but typically require long standardization times and are difficult to disseminate; non-TCP protocols require usually expensive changes in routers; libraries have to be included in new applications, while existing programs may be too specialized.

In a related paper [6] we have presented our novel approach based on a pipeline mechanism, of high speed and capacity, which interconnects, at application level, pairs of data producer/consumer processes running on different hosts in a WAN network. It is worth emphasizing the generality and elegance of the approach which

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 326–335, 2013.

manages to exploit, through encapsulation, the pipe programming paradigm - specific to modern operating systems (Unix) - based on reuse of existing system components, as opposed to building new monolithic applications.

In terms of its ability to use efficiently multiple paths of different bandwidth, our application resembles to Multipath TCP [1] which is however implemented in the operating system (Linux) kernel and so suffers, as the majority of TCP extensions, from difficulties of dissemination and standardization.

Compared to the class of applications or library functions for networking transfers using multiple concurrent TCP connections, which is a member of, our application differs from previous achievements by targeting high rate streaming and therefore not being limited to file transfers of known size.

In this paper we extend our initial research with experiments made in HBDP networks of 1 and up to 9 Gbps. We were able to evaluate the streaming performance of our transfer algorithms and we have also brought a number of amendments to our multi-path shaping techniques meant to increase the overall throughput and transmission stability.

The following section presents the fork-join queueing model that is underlying our research. The main improvements added to our weighted transmission algorithms are included in Section 3. Section 4 presents methodology and experimental results and their analysis is done in Section 5. Finally, conclusions and future research directions are presented in Section 6.

2 The Fork-Join Queue Model

The architecture of our system, presented in detail in [6] and here in Fig. 1, is based on the model of synchronized queues, also known as fork-join queues [13], specific to parallel and distributed processing in general. Thus, data read by the parent process from a pipeline, which makes a single point of interface with the underlying (Unix-type) OS, is split, in order, among N child processes, via a shared memory segment, containing FIFO networking data buffers of size B (fork phase). The join happens actually at the receiver, with a symmetrical structure, where data is recomposed, in the same order, from its parts, and leaves the system through another writing pipeline.

The performance of our application is determined by its sojourn time, defined as time elapsed between the fork and join moments, or, in other terms, by its global transmission rate R.

Evidently, it depends on individual TCP transmission rates, r_i, with $i \in [1, N]$.

Denote $r_{min} = \min\limits_{i=1,N}\{r_i\}$, $r_{max} = \max\limits_{i=1,N}\{r_i\}$ and let $t_i = \frac{B_i}{r_i}$ with $B_i = B$, $t_{max} = \frac{B}{r_{min}}$ and $t_{min} = \frac{B}{r_{max}}$.

Then given the in order transmission and streaming synchronization constraints:

$$R = \frac{\sum\limits_{i=1}^{N} B_i}{t_{max}} = \frac{NB}{\frac{B}{r_{min}}} = Nr_{min} \tag{1}$$

The above equation poses a problem for unequal transmission rates, so we decided to adapt the utilization degree of our buffers depending on measured throughput, according to formula:

$$B_i = B\frac{t_{min}}{t_i} = B\frac{r_i}{r_{max}} \tag{2}$$

Since now:

$$t_{max} = \max_{i=1,N}\{t_i\} = \max_{i=1,N}\{\frac{B_i}{r_i}\} = \frac{B}{r_{max}} \tag{3}$$

equation (1) becomes:

$$R = \frac{\sum_{i=1}^{N} B_i}{t_{max}} = \frac{\frac{B}{r_{max}}\sum_{i=1}^{N} r_i}{\frac{B}{r_{max}}} = \sum_{i=1}^{N} r_i \tag{4}$$

which makes an ideal approximation of R as sum of all transmission rates, provided that the individual measured times t_i are accurate. Next section will detail a few improvements we brought to our initial [6] shaping algorithms dealing with a better estimation of individual sojourn times.

3 Parallel Transmission Algorithms Improvements

We kept unchanged the overall structure of our transmission algorithms detailed in [6]. We remind it here briefly, as being formed out of:

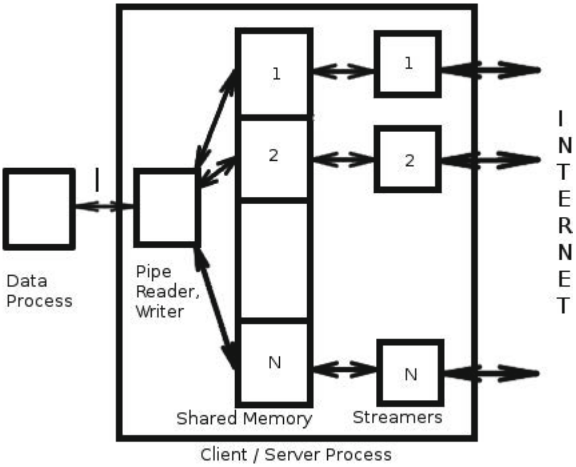


Fig. 1. Architecture of Parallel Streaming System

- an infinite transmission loop with independent synchronization regions at each individual transmission/reception buffer level.
- an agreed data streaming ordered over N parallel TCP connections.

It is seen as a global fork-join queue which interconnects with speed, through a generalized long and large pipeline, pairs of remote data producer/consumer processes.

To accommodate multi-path support at application level we superposed transmission shaping strategies, implemented, for simplity and efficiency, exclusively at the sender side, and having three main phases:

- measurement of per connection/queue sojourn times, $\{t_i\}$, when all buffers are of equal size, followed by adjustment (weighting) of buffer sizes according to formula (2) above.
- a short waiting period to attain steady-state TCP transmission.
- a feedback loop, determining eventual reshaping when some paths change or/and transmission rates are throttled dynamically, in time.

These strategies are meant to obtain the shortest global sojourn time, setting the overall throughput closer to that computed by equation (4).

To better support a larger class of network bandwidths, we identified, during experiments presented in the next section, a few improvements to our initial multi-path algorithms, leading to an improved performance and overall stability.

3.1 A More Precise Sojourn Times Measurement

We averaged the transmission times over a number of steps varying inversely proportional to N (but no less than 2 steps), instead of a fixed number of steps before. We used the mean and standard deviation of the N times as indicators whether the weighted buffer times are better than those initially measured with equally sized buffers: when both mean and standard deviation were bigger, we decided to continue with the equally sized buffers; when only the mean was bigger, we triggered an early reshaping, without entering the feedback loop at all. We also considered an alarm mechanism, able to trigger a clocked reshaping, correcting rare cases when the global transmission rate was wrongly driven by the slowest connection.

These strategies helped with reducing rate measurement errors, due to the characteristically uneven TCP rate over time.

3.2 Stability Improvements

We managed to reduce unnecessary reshaping phases, and got better stability and throughput, by considering only significant (more than 40%) increases in the mean sojourn times as a reason for triggering reshaping, discounting decreases. In the same time, we employed a moving reference to the feedback loop - a rolling average (over last three values) of the mean sojourn times, instead of the fixed initial mean value of them.

4 Methodology and Experimental Results

We have expanded our 100 Mbps bounded network initial measurements, reported earlier in [6], over two new classes of networks: an ad-hoc 1 Gbps link between two regular computers, and an IPoIB Infiniband network, with a maximum bandwidth of about 9 Gbps, composed of powerful Opteron nodes, each having 12 cores with hyper-threading enabled (total 24 processors) running at 3 GHz, and 32 GB of installed RAM.

We have used the same application, bbftpPRO [3], where our simple striping algorithms, alongside streaming with optional multi-path shaping are implemented in an uniform manner. This has guaranteed equal TCP configurations (*cwnd*, buffer sizes) and instrumentation over all the experiments we performed. Thus we are able to study the behavior of our improved algorithms in more demanding conditions. For example, we reached a maximum transfer rate above 14 Gbps when using the loopback (local) interface of nodes belonging to the fast cluster, during memory-to-memory streaming experiments.

4.1 File Striping vs. File Streaming Performance Evaluation

Firstly, we evaluated the impact of synchronizations imposed by the fast streaming model as compared to plain file striping. Since striping requires transfer of files of known size, and to eliminate latency caused by hard disk drives, we used in memory (*tmpfs*) files of size adequate to network throughput levels.

Figure 2 presents the average rate measurements during transfers of a 400 MB memory file, using 1 to 14 parallel TCP connections, in the 1 Gbps network, in both striping and streaming modes.

The average throughput obtained while striping and streaming a 10 GB memory file in the 9 Gbps network is shown in Fig. 3, where the number of parallel TCP connections varies between 1 and 24.

4.2 Unweighted vs. Weighted Streaming Performance Evaluation

Secondly, we compared the average throughput of the weighted streaming algorithm with its unweighted form. We used a special testbed made out of a number

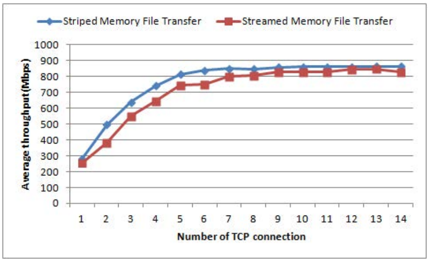


Fig. 2. File striping vs. file streaming, 1 Gbps network

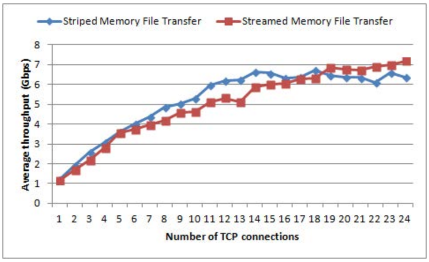


Fig. 3. File striping vs. file streaming, 9 Gbps network

of independent channels with different transfer rates, overlapping the underlying physical network. We configured our network topology using the traffic control program *tc*. We set three channels in the 1 Gbps network, of 100, 200, and 300 Mbps maximum throughput, and similarly, another three channels of 1, 2, and 3 Gbps maximum rate, in the 9 Gbps network.

Static Mode. Our static tests were of type memory-to-memory and lasted for 60 seconds each, so to achieve TCP stability. We varied the number of concurrent TCP connections, their distribution per channel and the number of channels used in the same time. We allocated 1 connection per channel in the 1 Gbps network, in a round-robin mode, until all connections got alloted. Same allocation strategy was used in the 9 Gbps network, except that the number of connections per channel was 2, at each step. For example, out of a total of 3 connections, 2 went to the 200 Mbps channel and 1 to the 100 Mbps one; out of 8 connections, 4 went to the 3 Gbps channel, 2 to the 2 Gbps channel, and another 2 to the 1 Gbps one, etc.

In Fig. 4 we present the average throughput for two, 200 and 100 Mbps, channels, and 2 to 8 parallel TCP connections - in the 1 Gbps network. Figure 5 shows the average throughput for three, 300, 200, and 100 Mbps, channels, and 3 to 9 parallel TCP connections - in the same network.

The 9 Gbps network case is presented in Fig. 6 - average throughput for two, 2 and 1 Gbps, channels with 4 to 24 TCP connections, and in Fig. 7 - average throughput for three, 3, 2, and 1 Gbps, channels with 6 to 24 TCP connections.

Dynamic Mode. We were also interested to evaluate the dynamic response of our streaming (multi-path) algorithms when the maximum throughput of the participant connections fluctuates. Thus, we reconfigured alternatively, every 20 seconds, the maximum rate of individual connections, by associating them with channels having different bandwidth, in a continous memory-to-memory transfer loop.

We captured networking snapshots of the *System Monitor* program, for experiments in the 1 Gbps network. Figure 8 shows 8 unweighted parallel connections,

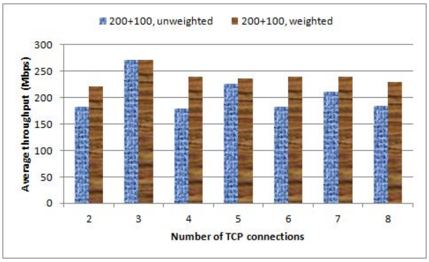


Fig. 4. Static mode, two channels, 200+100 Mbps

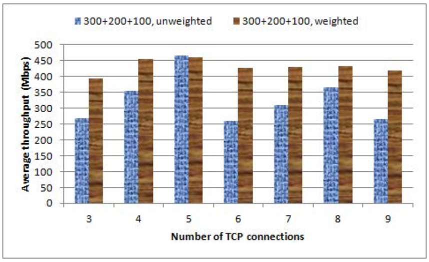


Fig. 5. Static mode, three channels, 300+200+100 Mbps

4 in a 100 Mbps channel, and the other 4 in a 200 and then 300 Mbps channel, alternatively. Snapshot in Fig. 9 illustrates the weighted transmission, in the same context.

In the 9 Gbps network we instrumented our dynamic experiments by intermediary piping the transferred data through the *pmr* program and collecting its instantaneous rate reports once every second. Figure 10 shows both the weighted and unweighted transmission of 12 parallel connections, 6 in a 1 Gbps channel, and the other 6 in a 2 and then 3 Gbps channel, alternatively.

5 Analysis of Experimental Results

From a quantitative point of view, when transmission does not require shaping but only extra synchronisations imposed by the fast streaming mechanisms, both striping and streaming seem equally rapid. This is apparent from Fig. 2 and 3 where the average total throughput increases approximately linearly with the number of parallel connections until reaching a saturation value of about 80%

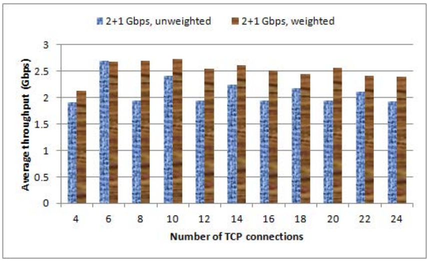


Fig. 6. Static mode, two channels, 2+1 Gbps

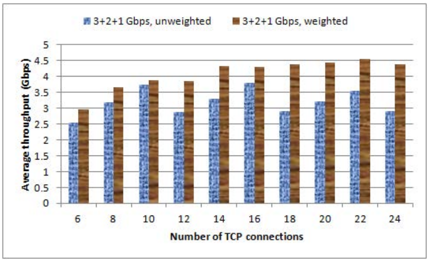
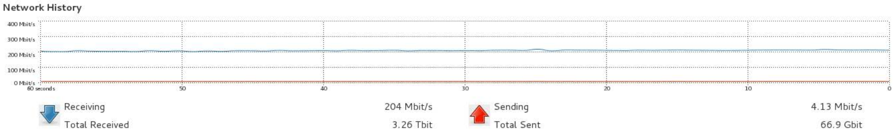
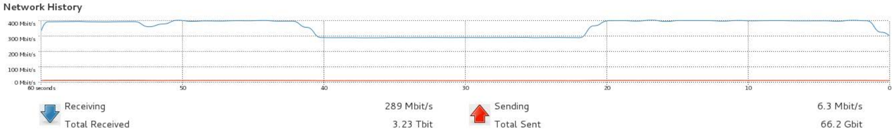


Fig. 7. Static mode, three channels, 3+2+1 Gbps



Receiving 204 Mbit/s
Total Received 3.26 Tbit
Sending 4.13 Mbit/s
Total Sent 66.9 Gbit

Fig. 8. Dynamic streaming, 8 TCP connections, unweighted 100+200|300 Mbps



Receiving 289 Mbit/s
Total Received 3.23 Tbit
Sending 6.3 Mbit/s
Total Sent 66.2 Gbit

Fig. 9. Dynamic streaming, 8 TCP connections, weighted 100+200|300 Mbps

of the available bandwidth, for both algorithms as well. At a closer examination, striping appears to have an early advantage over streaming, because every microsecond counts in such fast networks, and there is an unavoidable synchronization overhead in the case of streaming, absent for (independent) striping. However, when transmission parameters are well tuned (e.g. increased buffer sizes) and for a larger number of parallel connections, the average throughput becomes mostly equal, therefore streaming penalty is negligible.

5.1 Shaping Analysis

In any case, multi-path weighted (shaped) streaming holds an overall qualitative and quantitative edge over its unweighted mode, as demonstrated in the multichanneled network topology experiments. Both static and dynamic performances show that the weighted transmission algorithm uses better the available bandwidth and responds promptly to alteration of bandwidth over time, relative to the unweighted one.

Static Analysis. Thus, in the static 1 Gbps network case: data presented in Fig. 4 shows, for two channels (of 200 and 100 Mbps), that the average throughput is 16.6% higher, with a peak of +32.7% for 4 concurrent connections, when using the shaping algorithm. The utilization degree of the available bandwidth (equal to 300 Mbps) averages 79.9% in the weighted mode, while only 68.5% in the unweighted mode.

Data shown in Fig. 5, for three channels (of 300, 200, and 100 Mbps), computes an average throughput 31.8% higher, peaking at +64.5% for 6 parallel connections, when shaping is used. The utilization degree of the available bandwidth (equal to 600 Mbps) averages 71.8% in the weighted mode, while only 54.5% in the unweighted mode.

The static 9 Gbps network case: data presented in Fig. 6 shows, for two channels (of 2 and 1 Gbps), that the average throughput is 19.4% higher, with a peak of +39.1% for 8 connections, when shaping. The utilization degree of the available bandwidth (equal to 3 Gbps) averages 84% in the weighted mode, while only 70.3% in the unweighted mode.

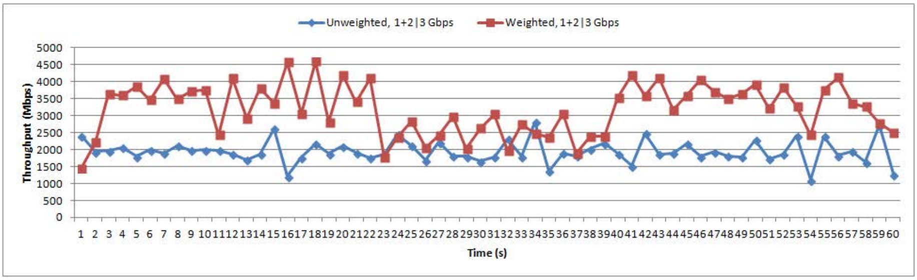


Fig. 10. Dynamic streaming, 12 TCP connections

Data shown in Fig. 7, for three channels (of 3, 2, and 1 Gbps), computes an average throughput 27.2% higher, peaking at +50.5% for 18 connections, when shaping is in effect. The utilization degree of the available bandwidth (equal to 6 Gbps) averages 67.9% in the weighted mode, while only 53.4% in the unweighted mode.

It is noteworthy, from a qualitative standpoint, the clear distinction between the rates achieved by the weighted algorithm as compared to the unweighted one. While the former maintains approximately the same higher throughput regardless of TCP connections distribution to channels, the latter is obviously affected by it, hence the sawtooth shape of its throughput line.

Dynamic Analysis. In the absence of shaping, throughput stays unmodified, at around 200 Mbps as shown in Fig. 8 for the 1 Gbps network experiments, and at about 1.95 Gbps as shown in Fig. 10 for the 9 Gbps network. These values correspond to the rate of the slowest connection, multiplied by the number of connections, as predicted by theory.

In contrast, the weighted algorithm is sensible to the ±50% bandwidth variation happening every 20 seconds. Figure 9 shows, for the 1 Gbps network case, two average throughput levels, one just below 300 Mbps and the other just below 400 Mbps, both approximating the bandwidth sum of the used channels, as expectedly. In Fig. 10, for the 9 Gbps network case, the two average throughput levels are about 2.62 Gbps and 3.48 Gbps, both representing about 87% of the channels' bandwidth sum, in their respective states.

Finally, these results also demonstrate that the improvements made to our multi-path algorithms, described in Section 3 above, have indeed contributed to a better overall transmission stability.

6 Conclusions and Future Work

This paper has presented the behavior of our improved data streaming algorithms over an extended range of networks capable of high bandwidth of 1 and up to 9 Gbps. We managed to preserve the advantages of our model which generalizes the pipe programming paradigm between fast data reading/writing processes distributed over the Internet.

The experiments and results described and analyzed in the previous sections demonstrate that our application is capable of streaming data at very high rates with only a small penalty due to synchronization constraints. We also showed that our improved algorithms handle the multi-path transmission case efficiently, in both static and dynamic regimes, yielding better throughput and stability.

Based on these results, and those presented earlier in [6], we are confident that the proposed model and the new fast transfer algorithms meet the modern requirements of today's research domains interested in transmitting massive bulk data elegantly, efficiently, effectively and at parallel speeds.

6.1 Future Research Directions

We consider further improving the stability and efficiency of our algorithms, and also automating transmission parameter choosing, like optimal number of parallel connections or buffer sizes. We appreciate that the expected diversification and continuous development of networks in the future would certainly benefit from further research extending our current work.

References

1. Ford, A., Raiciu, C., Handley, M., Bonaventure, O.: TCP Extensions for Multipath Operation with Multiple Addresses draft-ietf-mptcp-multiaddressed-09. IETF draft (June 2012)
2. ATLAS Experiment, `http://www.atlas.ch/`
3. bbftpPRO, `http://bbftppro.myftp.org/`
4. Jin, C., Wei, D., Low, S.: FAST TCP: Motivation, Architecture, Algorithms, Performance. In: Proceedings of IEEE Infocom, Hong Kong (March 2004)
5. Katabi, D.: Decoupling Congestion Control and Bandwidth Allocation Policy with Application to High Bandwidth-Delay Product Networks. Ph.D. dissertation, Massachusetts Institute of Technology (March 2003)
6. Schrager, D., Radulescu, F.: Efficient Algorithms for Fast Data Transfers Using Long and Large Pipes in WAN Networks. In: 19th International Conference on Control Systems and Computer Science (CSCS-19) (May 2013)
7. He, E., Leigh, J., Yu, O., DeFanti, T.: Reliable Blast UDP: Predictable High Performance Bulk Data Transfer. In: Proceedings of IEEE International Conference on Cluster Computing (2002)
8. Sivakumar, H., Bailey, S., Grossman, R.L.: PSockets: The case for application-level network striping for data intensive applications using high speed wide area networks. In: Proceedings of the IEEE/ACM SC2000 Conference (2000)
9. Crowcroft, J., Oechslin, P.: Differentiated End-to-End Internet Services Using a Weighted Proportional Fair Sharing TCP. ACM SIGCOMM Computer Communication Review 28(3), 53–69 (1998)
10. Postel, J.: Transmission Control Protocol. RFC 793 (September 1981)
11. Postel, J.: User Datagram Protocol. RFC 768 (August 1980)
12. Brakmo, L., Peterson, L.: TCP Vegas: End to End Congestion Avoidance on a Global Internet. IEEE Journal on Selected Areas in Communications 13(8) (October 1995)
13. Flatto, L., Hahn, S.: Two parallel queues created by arrivals with two demands. SIAM J. Appl. Math. 44(5), 1041–1053 (1984)
14. Allman, M., Ostermann, S., Kruse, H.: Data Transfer Efficiency over Satellite Circuits Using A Multi-Socket Extension to the File Transfer Protocol (FTP). In: Proceedings of the ACTS Results Conference, NASA Lewis Research (1995)
15. Mathis, M., Mahdavi, J., Floyd, S., Romanow, A.: TCP Selective Acknowledgment Options. RFC 2018 (October 1996)
16. pmr, `http://zakalwe.fi/~shd/foss/pmr/`
17. Floyd, S.: HighSpeed TCP for Large Congestion Windows. RFC 3649 (December 2003)
18. VIRGO Detector, `https://wwwcascina.virgo.infn.it/`
19. Allcock, W., Bester, J., Bresnahan, J., Chervenak, A., Liming, L., Tuecke, S.: GridFTP: Protocol Extensions to FTP for the Grid. Global Grid Forum Recommendation Document GFD.20 (2005)
20. Gu, Y., Grossman, R.: SABUL: A Transport Protocol for Grid Computing. Journal of Grid Computing 1(4), 377–386 (2003)

A Novel Approach to Automated Cell Counting Based on a Difference of Convex Functions Algorithm (DCA)

Le Thi Hoai An[1,2], Le Minh Tam[3], and Nguyen Thi Bich Thuy[4]

[1]Laboratory of Theoretical and Applied Computer Science (LITA)
UFR MIM, University of Lorraine,
Ile du Saulcy, 57045 Metz, France
[2]Lorraine Research Laboratory in Computer Science and its Applications
CNRS UMR 7503, University of Lorraine, 54506 Nancy, France
[3] DES en Cardiologie et maladies vasculaires, Université de Paris VI, France
[4] Faculty of Mathematics Mechanics Informatics, HUS-VNU, Vietnam
hoai-an.le-thi@univ-lorraine.fr,
minh-tam.le@inserm.fr,
thi-bich-thuy.nguyen9@etu.univ-lorraine.fr

Abstract. Cytological analysis, specially the cell counting, is an important element in the diagnosis of many diseases. Cell segmentation, the major phase of cell counting procedure, was basically performed by intensity thresholding, feature detection, morphological filtering, region accumulation and deformable model fitting. We present in this paper an automatic method for cell counting with segmentation based on Feature Weighted Fuzzy Clustering via a Difference of Convex functions with Optimization Algorithm called DCA. This new application of our method can give promising results compared to the traditional manual analysis despite the very high cell density.

Keywords: Image Segmentation, Cell counting, Fuzzy Clustering, DC programming, DCA.

1 Introduction

Cytological analysis is an important element in the diagnosis of several diseases. The traditional method for an expert to achieve the differential counting is very tedious and time-consuming. Counting should be automated but it can become a complicated process.

Some examples of common techniques used in cell segmentation are thresholding [11], cell modeling[11], filtering, mathematical morphology [1], watershed clustering [6] and fuzzy sets [14]. Each algorithm is ultimately a combination of segmentation methods for adaptation to cell types. The development of an efficient segmentation algorithm, the main step of cell counting, constitutes a challenge for researchers in this domain.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 336–345, 2013.

Segmentation methods can be classified in four categories: methods based on pixels, on areas, on contours and on physical model for image formation. Fuzzy C-Means (FCM) clustering introduced by Bezdek in 1981[2], is a most widely used fuzzy clustering method. Recently, Hichem and al. proposed a FCM using features weighted algorithm [4], called SCAD (Simultaneous Clustering and Attribute Discrimination) to solve the problem. Furthermore, using feature weight can get more meaningful clusters in the problem of segmenting colour image [4].

The problem FCM using features weighted can be stated as follows. Let $\mathcal{X} := \{x_1, x_2, ..., x_n\}$ be a data set of n entities with m attributes and the known number of clusters k $(2 \leq k \leq n)$. Denote by Λ a $k \times m$ matrix defined as $\Lambda = (\lambda_{l,i})$ where $\lambda_{l,i}$ defines the relevance of i-th feature to the cluster C_l. $W = (w_{j,l}) \in \mathbb{R}^{n \times k}$ with $j = 1, \ldots, n$ and $l = 1, \ldots, k$ called the *fuzzy partition* matrix in which each element $w_{j,l}$ indicates the membership degree of each point x_j in the cluster C_l (the probability that a point x_j belongs to the cluster C_l).

We are regrouping the set $\mathcal{X}$ into k clusters in order to minimize the sum of squared distances from the entities to the centroid of their cluster. The dissimilarity measure is defined by m weighted attributes. Then a straightforward formulation of the clustering using weighted dissimilarity measures is $(\mu, \beta$ are exponents greater than 1):

$$\begin{cases} \min F(W, Z, \Lambda) := \sum\limits_{l=1}^{k} \sum\limits_{j=1}^{n} \sum\limits_{i=1}^{m} w_{jl}^{\mu} \lambda_{li}^{\beta} (z_{li} - x_{ji})^2 \\ s.t: \sum\limits_{l=1}^{k} w_{jl} = 1, j = 1..n, \\ \quad \sum\limits_{i=1}^{m} \lambda_{li} = 1, l = 1..k, \\ \quad w_{jl} \in \ [0,1], j = 1..n, l = 1..k, \\ \quad \lambda_{li} \in [0,1], l = 1..k, i = 1..m. \end{cases} \tag{1}$$

Problem (1) is very difficult due to the nonconvexity of the objective function. Moreover, in real applications this is a very large scale problem (high dimension and large data set, i.e. n and m are very large), that is why global optimization approaches such as Branch & Bound, Cutting plane algorithms etc. cannot be used. In [4], the authors proposed **SCAD** algorithm to solve the problem (1).

In this paper we investigate a novel approach in DC programming framework to solve Problem (1). DC programming and DCA is an innovative approach in nonconvex programming which were introduced by Pham Dinh Tao in a preliminary form in 1985 and have been extensively developed since 1994 by Le Thi Hoai An and Pham Dinh Tao. They become now classic and increasingly popular. DCA has been successfully applied to several areas, in particular for Machine Learning. This motivates us to use DCA for solving the hard problem(1) in the context of cell counting.

DC programming and DCA address a general DC program of the form

$$\alpha = \inf\{\mathcal{F}(x) := g(x) - h(x) \ : \ x \in \mathbb{R}^n\}, \quad (P_{dc})$$

where g and h are lower semi-continuous proper convex functions on $\mathbb{R}^n$. Such a function $\mathcal{F}$ is called a DC function, and $g - h$ is called DC decomposition of $\mathcal{F}$ while the convex functions g and h are called DC components of $\mathcal{F}$. Based on the DC duality and the local optimality conditions, the idea behind DCA is simple : each iteration r of DCA approximates the concave part $-h$ by its affine majorization (that corresponds to taking $y^r \in \partial h(x^r)$) and minimizes the resulting convex function

$$\min\{g(x) - h(x^r) - \langle x - x^r, y^r\rangle : x \in \mathbb{R}^n\} \quad (P_r)$$

to obtain x^{r+1}.

The construction of DCA involves DC components g and h but not the function $\mathcal{F}$ itself. Hence, for a DC program, each DC decomposition corresponds to a different version of DCA. Since a DC function $\mathcal{F}$ has an infinite number of DC decompositions which have crucial impacts on the qualities (speed of convergence, robustness, efficiency, globality of computed solutions,...) of DCA, the search of a "good" DC decomposition is important from an algorithmic point of views. This fact is crucial for nonconvex nonsmooth programs for which DCA is one of the rare effective algorithms. How to develop an efficient algorithm based on the generic DCA scheme for a practical problem is thus a sensible question to be studied. It would be wrong to think that using DCA for efficiently solving a practical problem is a simple procedure, since the general DCA scheme is rather a philosophy than an algorithm. There is not only one DCA but one family of DCAs for a considered problem. The design of an efficient DCA for a concrete problem is an art which should be based on theoretical tools and on its special structure.

The DCA scheme used in this paper is very simple and inexpensive, because it is based on a suitable DC decomposition of the objective function. Then we developed a combination of the mathematical morphology and segmentation method based in DCA to different cells counting problems. Results of cell counting of SCAD and our method are compared subsequently with the manual analysis, considered as the reference.

In Section 2 we will present in detail our algorithm for image segmentation and some morphological operations. The computational results will be presented on Section 3.

2 Methodology

First, we will describe briefly the segmentation phase using DCA. For more detail, the readers are refereed to [10].

2.1 DCA Approach for Solving Problem (1)

Let $\alpha_i := \min_{j=1,...,n} x_{j,i}$, $\gamma_i := \max_{j=1,...,n} x_{j,i}.$, $z_l \in \mathcal{T}_l := \Pi_{i=1}^m [\alpha_i, \gamma_i]$ for all $l = 1, ..., k$, and $Z \in \mathcal{T} := \Pi_{l=1}^k \mathcal{T}_l$. Let Δ_l (resp. $\mathcal{C}_j$) be the $(m-1)$-simplex in $\mathbb{R}^m$(resp. $(k-1)$-simplex in $\mathbb{R}^k$), for each $l \in \{1, ..., k\}$ (resp. for each $j \in \{1, ..., n\}$), defined by:

$$\begin{aligned} \Delta_l &:= \left\{ \Lambda_l := (\lambda_{l,i})_i \in [0,1]^m : \sum_{i=1}^{m} \lambda_{l,i} = 1 \right\}; \\ \mathcal{C}_j &:= \left\{ W_j := (w_{j,l})_l \in [0,1]^k : \sum_{l=1}^{k} w_{j,l} = 1 \right\}, \end{aligned}$$

and $\mathcal{C} := \Pi_{j=1}^n \, \mathcal{C}_j, \mathcal{T} := \Pi_{l=1}^k \, \mathcal{T}_l, \Delta := \Pi_{l=1}^k \Delta_l$.

Then the problem (1) can be rewritten as:

$$\min \left\{ F(W, Z, \Lambda) : (W, Z, \Lambda) \in (\mathcal{C} \times \mathcal{T} \times \Delta) \right\}. \tag{2}$$

A DC formulation of (1) can be

$$\min \left\{ F(W, Z, \Lambda) := G(W, Z, \Lambda) - H(W, Z, \Lambda) : (W, Z, \Lambda) \in (\mathcal{C} \times \mathcal{T} \times \Delta) \right\}, \tag{3}$$

where

$$H(W, Z, \Lambda) := \sum_{l=1}^{k} \sum_{j=1}^{n} \sum_{i=1}^{m} \left[\frac{\rho}{2} \left(w_{jl}^2 + z_{li}^2 + \lambda_{li}^2 \right) - w_{jl}^{\mu} \lambda_{li}^{\beta} (z_{li} - x_{ji})^2 \right] \tag{4}$$

and

$$G(W, Z, \Lambda) := \frac{\rho}{2} \sum_{l=1}^{k} \sum_{j=1}^{n} \sum_{i=1}^{m} \left(w_{jl}^2 + z_{li}^2 + \lambda_{li}^2 \right)$$

are convex functions on $(\mathcal{C} \times \mathcal{T} \times \Delta)$ with $\rho := \max \ \{\mu(\mu-1)\delta^2 + 2\mu\delta + \beta\mu\delta^2; 2\mu\delta + 2 + 2\beta\delta; \beta\mu\delta^2 + 2\beta\delta + \beta(\beta-1)\delta^2\}$, $\delta = \gamma - \alpha$, $\mu > 1$, $\beta > 1$.

For designing a DCA applied to (3), we first need to compute $(\bar{W}^r, \bar{Z}^r, \bar{\Lambda}^r) \in \partial H(W^r, Z^r, \Lambda^r)$ (a subgradient of H at W^r, Z^r and then solve the convex program

$$\min \left\{ \frac{\rho}{2} \sum_{l=1}^{k} \sum_{j=1}^{n} \sum_{i=1}^{m} \left(w_{jl}^2 + z_{li}^2 + \lambda_{li}^2 \right) - \langle (W, Z, \Lambda), (\bar{W}^r, \bar{Z}^r, \bar{\Lambda}^r) \rangle : (W, Z, \Lambda) \in (\mathcal{C} \times \mathcal{T} \times \Delta) \right\}. \tag{5}$$

The function H is differentiable and its gradient at the point (W^r, Z^r, Λ^r) is given by:

$$\begin{aligned} \bar{W}^r &= \nabla_W H(W, Z, \Lambda) = \left(m\rho w_{jl} - \sum_{i=1}^{m} \mu w_{jl}^{\mu-1} \lambda_{li}^{\beta} (z_{li} - x_{ji})^2 \right)_{j=1..n}^{l=1..k}, \\ \bar{Z}^r &= \nabla_Z H(W, Z, \Lambda) = \left(n\rho z_{li} - \sum_{j=1}^{n} 2 w_{jl}^{\mu} \lambda_{li}^{\beta} (z_{li} - x_{ji}) \right)_{l=1..k}^{i=1..m}, \\ \bar{\Lambda}^r &= \nabla_\Lambda H(W, Z, \Lambda) = \left(n\rho \lambda_{li} - \sum_{j=1}^{n} \beta w_{jl}^{\mu} \lambda_{li}^{\beta-1} (z_{li} - x_{ji})^2 \right)_{l=1..k}^{i=1..m}. \end{aligned} \tag{6}$$

The solution of the auxiliary problem (5) is explicitly computed as (Proj stands for the projection)

$$\begin{aligned}
&(W^{r+1})_j = \mathrm{Proj}_{\mathcal{C}_j}\left(\tfrac{1}{m\rho}(\bar{W}^r)_j\right) \; j=1,...n;\\
&(Z^{r+1})_{l,i} = \mathrm{Proj}_{[\alpha_i,\gamma_i]}\left(\tfrac{1}{n\rho}(\bar{Z}^r)_{li}\right) \; l=1,..,k, i=1,...m;\\
&(\Lambda^{r+1})_l = \mathrm{Proj}_{\Delta_l}\left(\tfrac{1}{n\rho}(\bar{\Lambda}^r)_l\right) \; l=1,...k.
\end{aligned} \tag{7}$$

The computation of $(Z^{r+1})_{l,i}$ is given by

$$Z^{r+1}_{l,i} = \begin{cases} \frac{1}{n\rho}\bar{Z}^r_{li} \text{ if } n\rho\alpha_i \le \bar{Z}^r_{li} \le n\rho\gamma_i, \\ \alpha_i \text{ if } \bar{Z}^r_{li} < n\rho\alpha_i, \\ \gamma_i \text{ if } \bar{Z}^r_{li} > n\rho\gamma_i \end{cases} \qquad \forall i=1,\ldots,m, l=1,\ldots,k. \tag{8}$$

For computing the projection of points onto a simplex $\Delta_l/\mathcal{C}_j$, some efficient algorithms are available among them we use the very inexpensive algorithm developed in [5].

Hence, DCA applied on (3) can be described as follows.

DCA-SI: DCA applied to (3)

- **Initialization:** Choose W^0, Z^0 and Λ^0. Let $\epsilon > 0$ be sufficiently small, $r = 0$.
- **Repeat**
  - ○ Compute $(\bar{W}^r, \bar{Z}^r, \bar{\Lambda}^r)$ via (6).
  - ○ Compute $(W^{r+1}, Z^{r+1}, \Lambda^{r+1})$ via (7).
  - ○ $r = r + 1$
- **Until** $\|(W^{r+1}, Z^{r+1}, \Lambda^{r+1}) - (W^r, Z^r, \Lambda^r)\| \le \epsilon$ or $|F(W^{r+1}, Z^{r+1}, \Lambda^{r+1}) - F(W^r, Z^r, \Lambda^r)| \le \epsilon$.

2.2 Morphological Operations

Mathematical morphology was first introduced by Georges Matheron and Jean Serra [13][15]. The basic operations of mathematical morphology are the dilation, erosion, closure and disconnection. The function of dilation is increasing the image while erosion makes it less. Closure operation helps to close the inner hole region and eliminate the bays along the border area and disconnection is used to gets rid of small fragments, protruding regions near its borders. Based on these operations, various morphological operations were developed.

After segmentation phase, almost cells are defined in the binary image, but some noncellular particles were also present. The binary image is thus further processed to remove the objects that do not correspond to the cells of interest by applying some morphological operations: region filling for achieving solid particles, filtering for removing small artifacts, etc.

Showed in Fig.1 are the results of some morphological operations. Fig.1(b), the ImFill(BW1,'holes') function in Matlab was used to fill holes in the input image. In Fig.1(c), BwAreaOpen(bw,p) helps to remove from a binary image all connected objects that have fewer than p pixels. The value of p is chosen large enough so that it can eliminate the wrong objects but still remains the correct cells.

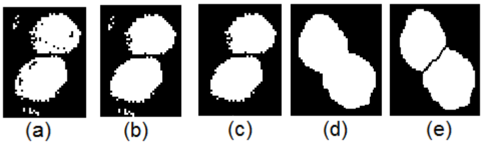


Fig. 1. Some morphological operations: (a): after segmentation, (b): fill holes, (c): remove small objects, (d): overlapping cells, (e): separated cells by watershed

In the final binary image there were still overlapping or touching cells. A further procedure using watershed was thus applied to separate these cells before counting. The idea of watershed comes from an example about the great divide, this particular line which separates the U.S.A. into two regions. A drop of water falling on one side of this line flows down until it reaches the Atlantic Ocean, whereas a drop falling on the other side flows down to the Pacific Ocean. As we shall see in further detail later, this great divide constitutes a typical example of a watershed line. The two regions it separates are called the catchment basins of the Atlantic and the Pacific Oceans, respectively. The two Oceans are the minima associated with these catchment basins [12].

In Fig.1(d),(e) is an demo of watershed using.

3 Computational Experiments and Results

Our algorithm were implemented in the Visual C++ 2008 combined with Matlab R2007a, and were performed on a PC Intel i5 CPU650, 3.2 GHz of 4GB RAM. Images for experiment are Oligodendrocyte cells (one type of brain cells)(image **I1**), Oligodendrogliomas [16](image **I2**), Mouse liver cells [17] (image **I3**).

As the same way in [4], we map each pixel to an 8-dimensional feature vector consisting of three colors, three texture features and the two positions of pixels. The three color features are L^*a^*b coordinates of the color image. The three

Table 1. Cell counting results

Image	Size	DCA-SI	SCAD	Threshold	Manual
I1	1360×1024	434	477	434	430
I2	800 ×533	248	476	473	295
I3	400 ×292	126	133	80	123

texture features (polarity, anisotropy and contrast (cf. [3,4]) are computed as follows. First, the image $I(x,y)$ is convolved with Gaussian smoothing kernels $G_\delta(x,y)$ of several scales δ: $M_\delta(x,y) = G_\delta(x,y) \otimes (\varDelta I(x,y))(\varDelta I(x,y))^t$.

- The polarity is defined by $p = |E_+ - E_-|/(E_+ - E_-)$, where E_+ and E_- represent, respectively, the number of gradient vectors in the matrix Gauss kernels $G_\delta(x,y)$ of scale δ at the pixel (x,y) on the positive and negative sides of the dominant orientation. For each pixel, an optimal scale value is selected such that it corresponds to the value where polarity stabilizes with respect to scale. Let p^* be the polarity at the selected scale.
- The anisotropy is computed by $a = 1 - \lambda_2/\lambda_1$, where λ_1, λ_2 are the eigenvalues of $M_\delta(x,y)$ at the selected scale.
- The texture contrast is defined as $c = 2(\sqrt{\lambda_1 + \lambda_2})^3$.

The results of **DCA-SI** are compared with the results of **SCAD** ([4]) and thresholding methods. For both algorithms **DCA-SI** and **SCAD**, the parameter β is chosen in the interval $[1.0, 4.0]$ while the parameter of μ is taken in $[1.5, 2.5]$.

Fig.2 shows some segmentation results in comparing **DCA-SI** and **SCAD** and thresholding. In image **I1**, segmentation results of all three algorithm were

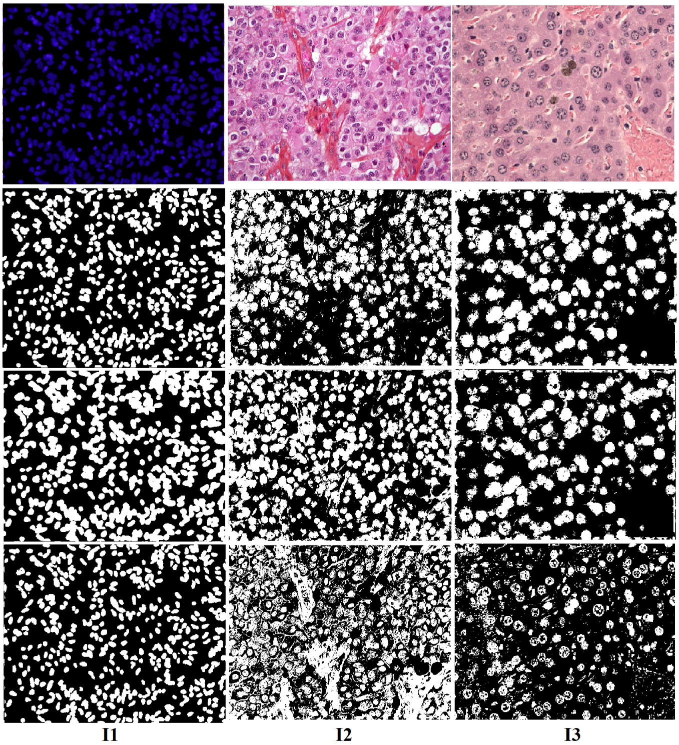


Fig. 2. The rows show respectively: Original images, images segmented by **DCA-SI**, **SCAD**, thresholding

good as cells were very clear and separable from the background. However in image **I2**, thresholding can not detect the cells with the blood as cells' colour and environment were close. Consequently, some mistakes were found in the segmentation result in row 4. while **DCA-SI** can eliminate the blood out (row 2). In **SCAD** result (row 3) it was difficult to count in the following phase because of existence of some blur region. In image **I3**, we obtained cells with much holes and fragments as the thresholding method separated cells based on pixels' intensity. In this case, **DCA-SI** and **SCAD** which used more information of a pixel (texture and position feature) gave the cells more solid and clear. Subsequently, cells segmented by **DCA-SI** were more sharpen and smooth than **SCAD**.

The counting results of three algorithms on three images are reported in the Table 1 below.

This table showed that **DCA-SI** which have better results on segmentation phase can give more exact results in the counting phase.

Fig. 3 contains the counting results on images **I1**, **I2**, **I3** using **DCA-SI** segmentation method combined with some morphological operations above.

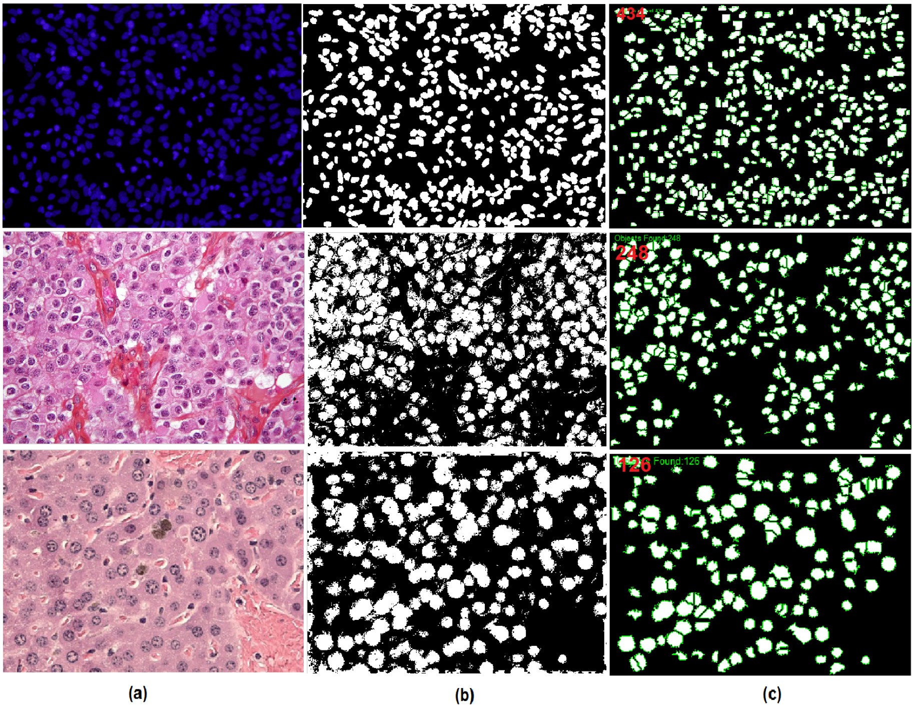


Fig. 3. Original(a), Segmented by **DCA-SI** (b) and Count result(c) images

4 Conclusion

We have presented an automatic segmentation technique for microscope cell images which is an important step in cell counting problem. The proposed segmentation technique, based on Feature Weighted Fuzzy Clustering via DCA, was evaluated by comparing with two others methods frequently used in cell counting problem. In our experiments, the counting results given by DCA are better than those of SCAD and thresholding. DCA method appears to be an effective segmentation technique in complex large-scale histological images and could be applied in routine practice.

References

1. Anoraganingrum, D.: Cell Segmentation with Median Filter and Mathematical MorphologyOperation. In: Proc. Intl. Conf. on Image Anal. and Proc., pp. 1043–1046 (1999)
2. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithm. Plenum Press, New York (1981)
3. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Color and Texture-Based Image Segmentation Using EM and Its Application to Content-Based Image Retrieval. In: Proceedings of the Sixth International Conference on Computer Vision, January 4-7, pp. 675–682 (1998)
4. Hichem, F., Olfa, N.: Unsupervised learning of prototypes and attribute weights. Pattern Recognition 37(3), 567–581 (2004)
5. Joaquim, J.J., Raydan, M., Rosa, S.S., Santos, S.A.: On the solution of the symmetric eigenvalue complementarity problem by the spectral projected gradient algorithm. Numerical Algorithms 47(4), 391–407 (2008)
6. Jiang, K., Liao, Q., Dai, S.: A Novel White Blood Cell Segmentation Scheme Using Scale-Space Filtering and Watershed Clustering. In: Proc. 2nd Intl Conf on Machine Learning and Cybern, pp. 2820–2825 (2003)
7. Le Thi, H.A., Le Hoai, M., Pham Dinh, T.: Fuzzy clustering based on nonconvex optimisation approaches using difference of convex (DC) functions algorithms. Journal of Advances in Data Analysis and Classification 2, 1–20 (2007)
8. Le Thi, H.A., Le Hoai, M., Nguyen Trong, P., Pham Dinh, T.: Noisy image segmentation by a robust clustering algorithm based on DC programming and DCA. In: Perner, P. (ed.) ICDM 2008. LNCS (LNAI), vol. 5077, pp. 72–86. Springer, Heidelberg (2008)
9. Le Thi, H.A., Pham Dinh, T.: Minimum Sum-of-Squares Clustering by DC Programming and DCA. In: Huang, D.-S., Jo, K.-H., Lee, H.-H., Kang, H.-J., Bevilacqua, V. (eds.) ICIC 2009. LNCS, vol. 5755, pp. 327–340. Springer, Heidelberg (2009)
10. Le Hoai, M., Nguyen, T.B.T., Ta, M.T., Le Thi, H.A.: Image Segmentation via Feature Weighted Fuzzy Clustering by a DCA based algorithm. In: Nguyen, N.T., van Do, T., Thi, H.A. (eds.) ICCSAMA 2013. SCI, vol. 479, pp. 53–63. Springer, Heidelberg (2013)
11. Liao, Q., Deng, Y.: An Accurate Segmentation Method for White Blood Cell Images. In: IEEE Intl Sym on Biomedical Imaging, pp. 245–248 (2002)

12. Vincent, L., Soille, P.: Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(6) (June 1991)
13. Matheron, G.: Random Sets and Integral Geometry. Wiley, New York (1975)
14. Theera-Umpon, N.: White Blood Cell Segmentation and Classification in Microscopic Bone Marrow Images. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3614, pp. 787–796. Springer, Heidelberg (2005)
15. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, New York (1983)
16. `http://foodmedicaleponyms.blogspot.fr/2011/03/fried-egg-like-cells.html` (accessed on March 2013)
17. `http://nanotechweb.org/cws/article/tech/33794` (accessed on March 2013)

An Effective Ant Colony Optimization Algorithm for the Minimum Sum Coloring Problem

Sidi Mohamed Douiri and Souad Elbernoussi

Research Laboratory Mathematics, Computing and Applications, Faculty of Sciences,
University of Mohammed V-Agdal, Rabat, Morocco
douirisidimohamed@hotmail.fr

Abstract. Ant colony optimization is a collective problem solving approach that simulates the foraging behavior of ants. It is a class of metaheuristics which made a success in NP-hard combinational optimization problems. In this paper we study the minimum sum coloring problem (MSCP), which is an NP-hard problem derived from the graph coloring problem (GCP). The goal of this problem is to minimize the sum of colors used in a graph. We propose for this problem a method based on ant colony optimization, which we tested on several benchmark graphs from the usual literature. By comparing the test results with those found in the literature, we demonstrate the effectiveness of the proposed method.

Keywords: minimum sum coloring, upper bounds, ant colony optimization.

1 Introduction

Given an undirected graph $G = (V, E)$, the graph coloring problem (GCP) requires to assign a color to each vertex such that no two adjacent vertices bear the same color. The challenge is to find the smallest number of colors. Finding the smallest number of different colors used for a valid coloring, called the chromatic number $\chi(G)$ of a graph G, is an NP-hard problem [1]. This problem often arises in several real applications, including frequency assignment [2], scheduling [3], timetabling [4], register allocation [5], etc.

A few exact approaches have been proposed in the literature, all of which can only be applied to small graphs [6]. Several metaheuristics have been applied to the graph coloring problem, e.g. tabu search [7], simulated annealing [8], ant colony optimization [9], genetic algorithm [10], etc.

In this work we are interested in the minimum sum coloring problem (MSCP). The aim is to minimize the sum of colors assigned to vertices. MSCP was introduced for the first time by Kubicka et al. who demonstrated its NP-completeness [11]. In 2007, Kokosinski et al. [13] gave upper and lower bounds by using a parallel genetic algorithm based on GPX and CEX crossover with a number of iterations from 5000 to 10000. In 2010, Moukrim et al. presented a lower bound

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 346–355, 2013.

for MSCP by studying several approaches based on extraction of partial graphs [14]. They gave a coloring for the complementary graph using a greedy algorithm MRLF that has complexity $O(n^3)$ [15], and they were able to improve the results given by Kokosinski et al. In 2011, Douiri and Elbernoussi provided upper bounds by using a genetic algorithm where the initial population is given by applying the DBG algorithm [16], Douiri and Elbernoussi proposed a heuristic to approximate a upper bound of MSCP, based on the extraction of maximal independent set [17], they also used a hybridization of Max-Min Ant System (MMAS) and a local heuristic [20]. In [21] the authors try to give a lower bound for MSCP by looking for a decomposition into cliques of the graph, based on the metaheuristic of ant colony optimization.

The rest of this paper is structured as follows. Section 2 briefly describes the theoretical background of the graph coloring problem (GCP) and the minimum sum coloring problem (MSCP). The principles of the ant colony optimization (ACO) algorithm are presented in Section 3. Our ACO based algorithm for MSCP will be described in detail in Section 4. Section 5 reports computational experiments, and also contains a brief discussion of the obtained results. Finally, conclusions and future work are presented in the last section.

2 Preliminary Definitions

We consider an undirected graph $G = (V, E)$, where V is the set of vertices and E denotes the set of edges. A k-coloring of such a graph $G = (V, E)$ is a mapping $c : V \longrightarrow \{1, ..., k\}$ such that $\forall (i, j) \in E \Rightarrow c(i) \neq c(j)$. A k-coloring of G is a partition of V into k independent sets. The smallest number of different colors used for a valid coloring is called the chromatic number, and denoted by $\chi(G)$.

The minimum sum coloring problem consists in finding a valid coloring so that $\sum_{v \in V} c(v)$ is minimal. This sum is denoted by $\sum(G) = min_c \sum_{v \in V} c(v)$. The smallest number of colors used to color G in the MSCP problem is called the strength of G and denoted $s(G)$. It is observed in [11] that even for trees whose chromatic number is 2, it is sometimes necessary to use more than 2 colors to find the optimal solution to MSCP, the tree in Figure 1 illustrates well this fact.

If we decompose the graph G as independent subsets $V_1, V_2, ..., V_k$, then we obtain a valid k-coloring by assigning to each subset V_i the color i, $1 \leq i \leq k$ and $\sum(G) = \sum_{i \in (1,...,k)} i. \mid V_i \mid$, with $\mid V_1 \mid \geq \mid V_2 \mid \geq .... \geq \mid V_k \mid$.

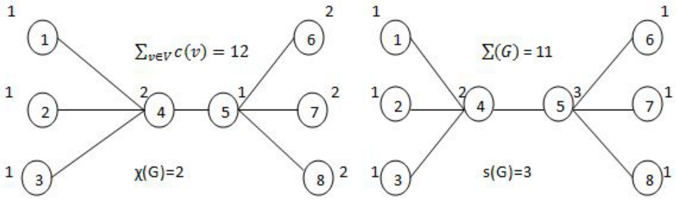


Fig. 1. Optimal solution for MSCP requires 3

3 Ant Colony Optimization (ACO)

Ant colony optimization (ACO) is a population-based metaheuristic initiated by Dorigo [18] in 1992. It's a cooperative search algorithm inspired by the behavior of real ants. ACO algorithm was firstly used for solving traveling salesman problem (TSP) [19], and then has been successfully applied to a large number of difficult problems like the routing in telecommunication networks, quadratic assignment problem (QAP) [22], graph coloring problems [9], etc.

The main idea is that it is indirect local communication among the individuals of a population of artificial ants. They deposit a substance, called pheromone on their path and form a pheromone trail, which enables them to find shortest paths between their nest and food sources. The pheromone decays over time, so that the path information is progressively lost if the pheromone is not reinforced by other ants, resulting in much less pheromone on less visited paths see Figure 2. Additionally a certain amount of pheromone evaporates this evaporation can be adjusted with a parameter called evaporation rate ρ. At each solution construction step, an ant has to decide to which neighboring vertex to move. This decision is made probabilistically based on pheromone values and some heuristic informations.

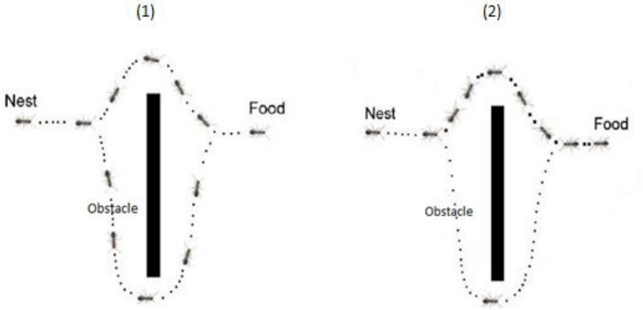


Fig. 2. How real ants find a shortest path. (1)The probability that ants take the short or the long path to the food source is same. (2)The probability to take again the short path is higher.

4 Resolution Approach

We consider p =(c(1),c(2),.....,c(V)) as solution given by an ant corresponding to an assignment of k colors to all the vertices of the graph G, where V is the vertices number of graph G. Our algorithm finds both the sum of minimum graph coloring $\sum(G)$ and the strength $s(G)$. The first step consists in initializing the tracks of pheromone, thereafter, in each cycle, every ant constructs a solution (coloring) and the pheromone trails are updated. The algorithm stops when it reaches the maximum number of cycles.

Each time, the ant chooses an i-th vertex, and we look for a set C of candidate colors for this vertex by using a constructive algorithm. The i-th vertex receives the color l from the set C with the probability:

$$p(i \leftarrow l) = \frac{\tau_i(l)^\alpha \eta_i^\beta \mu_i^\gamma(l)}{\sum\limits_{l \in C} \tau_i(l)^\alpha} \tag{1}$$

$\alpha > 0$, $\beta > 0$ and γ are the relative influence of every variable, their choice is determined experimentally, and $\tau_i(l)$ is the amount of virtual pheromone deposed on color l (color of i-th vertex), $\eta_i = \frac{1}{nclr}$ where $nclr$ is the current number of colors used before coloring the i-th vertex. $\mu_i(l) = \frac{1}{sum(l)}$ where $sum(l)$ is the sum coloring found after have to assign the color l to i-th vertex.

The construction of the candidate set of colors to color an i-th vertex is based on the connections of i-th vertex and the current number of colors used. In the Figure 3(a) the vertex $\{4\}$ is related to the vertex $\{3\}$. Therefore the candidate set of colors is $C = \{1, 3\}$, the vertex $\{4\}$ receives a color of C with a probability which depends of τ_4 and η_4. For the graph 3(b) the vertex $\{3\}$ is related to the vertices $\{1, 2\}$. The vertex $\{3\}$ can take neither the color 1 nor the color 2 then $C = \emptyset$, and then the vertex $\{3\}$ receives the color $(nclr + 1) = 3$, where nclr is the current color before coloring the i-th vertex.

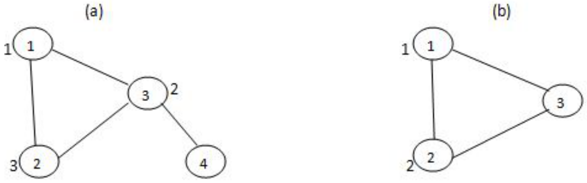


Fig. 3. Construction of the set C

4.1 Pheromone Update Rule

Local Update: Once the ant s generates a solution, the pheromone quantities deposited by each ant on each color is updated as follows:

$$\Delta\tau_i(l) \leftarrow \Delta\tau_i(l) + \frac{L_{loc}}{nclr_s + sum_s} \tag{2}$$

and

$$\tau_i(l) \leftarrow (1 - \rho_{loc}).\tau_i(l) + \Delta\tau_i(l) \tag{3}$$

where

. $\Delta\tau_i(l)$ is the addition of pheromone by an ant s on the color l assigned to the i-th vertex.

- L_{loc} is a constant.
- ρ_{loc} evaporation factor$(0 < \rho_{loc} < 1)$.
- $nclr_s$ the number of colors given by an ant s.
- sum_s the sum coloring given by an ant s

Global Update: At each iteration T, after all ants have completed their solutions, pheromone evaporation on all colors chooses is triggered, and then according to (4) each ant s deposits a quantity of pheromone $\Delta\tau_i(l)$ on each color that it has used.

$$\Delta\tau_i(l) \leftarrow \Delta\tau_i(l) + \frac{L_{gl}}{somclr_{best}} \tag{4}$$

where

- $\Delta\tau_i(l)$ is the addition of pheromone on colors chosen for each vertex of the best coloring in the iteration T.
- L_{gl} is a constant.
- $somclr_{best}$ the smallest sum coloring found by all ants in each iteration T.

A global update of the pheromone quantities of colors given by the best coloring is implemented with another constant factor ρ_{gl} $(0 < \rho_{gl} < 1)$ by the following rule:

$$\tau_i(l) \leftarrow (1 - \rho_{gl}).\tau_i(l) + \Delta\tau_i(l) \tag{5}$$

For each iteration, we seek the ant which gives a minimum sum coloring, and we add thereafter a quantity of pheromone on the colors chosen for each vertex of a graph,if we find a small value of the sum coloring we add then a large quantity of pheromone (4) and (5), and consequently we favor more these colors for a future choice. So this global update helps ants to find a minimum sum coloring.

4.2 Algorithm

1. Initialize pheromone trails.
2. While $T \leq T_{max}$.
3. For each ant s from 1 to nbAnts, construct a solution.
4. For each vertex $i \in V$.
5. Apply the constructive algorithm, and give the set of candidate colors C.
6. Choose a color in C for i-th vertex with the probability (1).
7. End for.
8. The local update of the quantities of pheromone, of every ant using relations (2) and (3).
9. End for.
10. The global update of the quantities of pheromone, after each iteration T using the formula (4) and (5).
11. $T \leftarrow T + 1$.
12. End while.

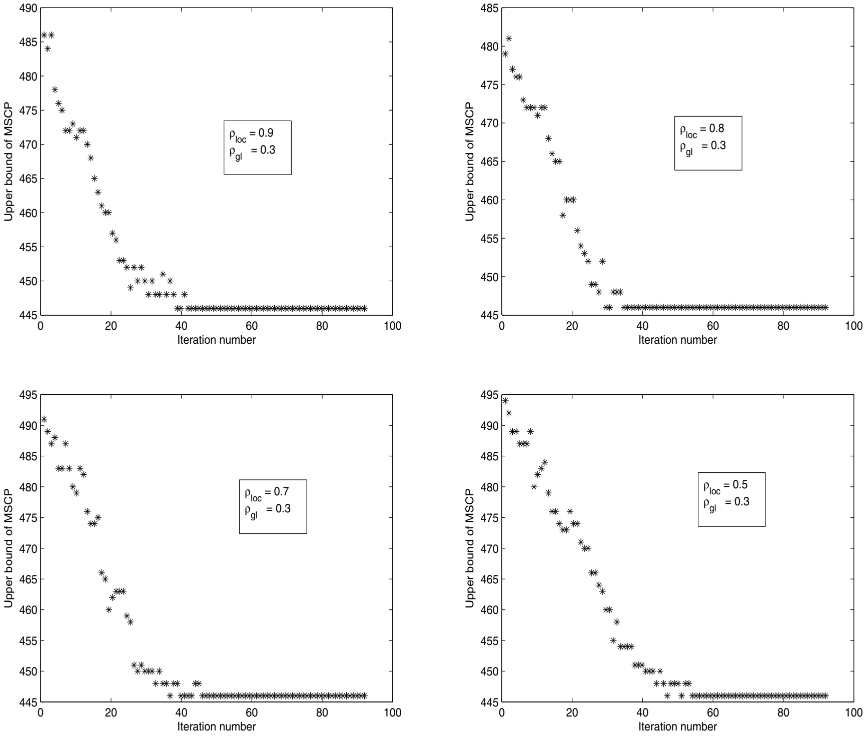


Fig. 4. Comparison of the solutions convergence with different ρ_{loc} for games120

5 Experimental Results

In this section we report the computational results obtained by our ACO algorithm for the upper bound of MSCP, the numerical results obtained are shown in Table 1, for a complete description of the instances see [23]. For each graph, we indicate the number of vertices $\mid V \mid$, the number of edges $\mid E \mid$, the chromatic number $\chi(G)$, as well as the upper bound given by a parallel genetic algorithm UB_{Kok} and the strength $s(G)_{Kok}$ given in [13]. UB is the best upper bound obtained by MRLF [15], UB_{ACO} the solution found by our ACO algorithm and $s_{ACO}(G)$ the strength of G. LB_{Ant} the lower bound given in [20]. The last column gives the success rate we run our algorithm on each graph 10 times.

A small population size may lead to converge slowly due to the lack of communication information between individuals, for each graph, we set the parameter with population size nAnts to 30. The parameters α, β and γ were tested several values between 1 to 10, and we finally chose the best experimental parameters $\alpha = 8$, $\beta = 4$ and $\gamma = 3$. The same way was used for determining the best values for the evaporation rate ρ and for parameters L_{loc}, L_{gl}. Figures 4 and 5 show the evolution of upper bound versus the iterations number for

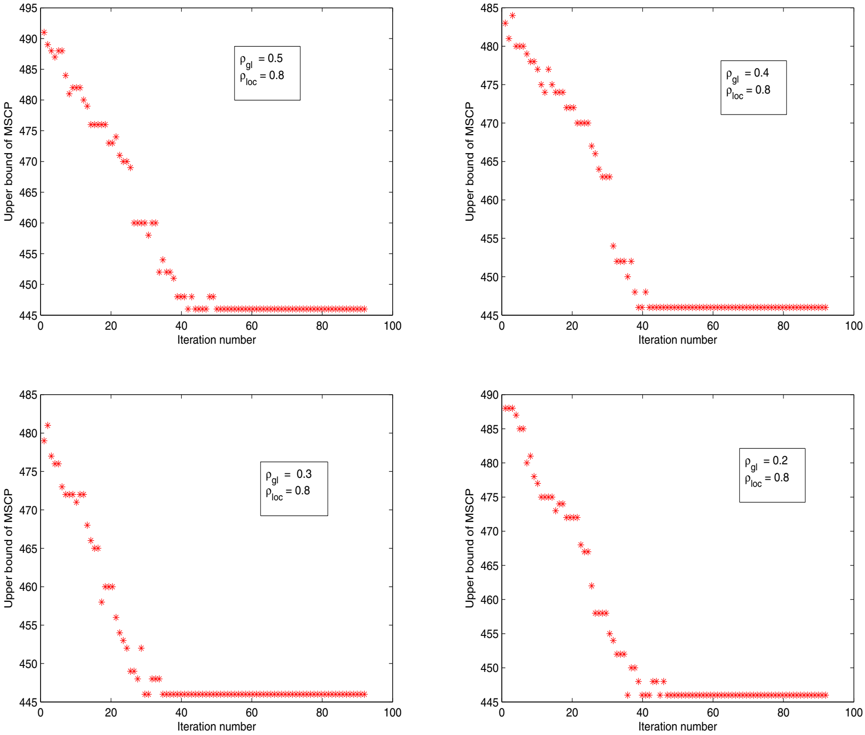


Fig. 5. Comparison of the solutions convergence with different ρ_{gl} for games120

different values of ρ_{loc} and ρ_{gl}. we tested four values for $\rho_{loc} = 0.5, 0.7, 0.8, 0.9$ and $\rho_{gl} = 0.2, 0.3, 0.4, 0.5$ on a graph named games120, we noted that for the values $\rho_{loc} = 0.8$ and $\rho_{gl} = 0.3$ our algorithm provides a good solution with a small iterations number. L_{loc} and L_{gl} were tested by different values from 10 to 20, and for each instance the best value was taken. The maximum iterations number T_{max} is taken between 60 and 200.

Most of these instances have been easily resolved, the only result that we have not improved is the upper bound of the instance queen8.8 compared to that given by Kokosinski et al. in [13].

6 Conclusion

In this paper we have presented the sum coloring problem in graphs, the main contribution of the work is the development of upper bound for the minimum sum coloring problem (MSCP) adapting ant colony optimization algorithm to this problem. Experimental results on a set of graphs prove the efficiency of our new approach. As future work, we intend to combine our ACO algorithm with greedy algorithms to improve its performances.

Table 1. Computational results

Graph	$\mid V \mid$	$\mid E \mid$	$\chi(G)$	LB_{Ant}	UB_{Kok}	$s(G)_{Kok}$	UB	UB_{ACO}	$s_{ACO}(G)$	succ
anna	138	493	11	272	281	11	277	277	11	7/10
david	87	406	11	234	243	11	241	241	11	9/10
huck	74	301	11	243	243	11	244	243	11	10/10
jean	80	254	10	216	218	10	217	217	11	8/10
queen5.5	25	160	5	75	75	5	75	75	5	10/10
queen6.6	36	290	7	126	138	8	138	138	8	10/10
queen7.7	49	476	7	196	196	7	196	196	7	10/10
queen8.8	64	728	9	288	**302**	10	303	303	10	9/10
miles250	128	387	8	316	347	8	334	334	10	10/10
miles500	128	1170	20	677	762	20	715	715	22	8/10
games120	120	638	9	442	460	9	446	446	9	9/10
myciel3	11	20	4	16	21	4	21	21	4	10/10
myciel4	23	71	5	34	45	5	45	45	5	10/10
myciel5	47	236	6	70	93	6	93	93	6	10/10
myciel6	95	755	7	142	189	7	189	189	7	10/10
myciel7	191	2360	8	286	382	8	381	381	8	10/10
fpsol2.i.1	496	11654	65	2590	*	*	*	3405	65	8/10
inithx.i.1	864	18707	54	2801	*	*	*	3679	54	4/10
mug88-1	88	146	4	163	*	*	*	190	4	10/10
mug88-25	88	146	4	162	*	*	*	187	4	10/10
mug100-1	100	166	4	187	*	*	*	211	4	10/10
mug100-25	100	166	4	185	*	*	*	214	4	10/10
2-Inser 3	37	72	4	55	*	*	*	62	4	10/10
3-Inser 3	56	110	4	84	*	*	*	92	4	10/10
zeroin.i.2	211	3541	30	1003	*	*	*	1013	30	5/10
zeroin.i.3	206	3540	30	997	*	*	*	1007	30	7/10
dsjc125.1	125	736	*	*	*	*	352	**346**	8	5/10
dsjc125.5	125	3891	*	*	*	*	1141	**1102**	20	4/10
dsjc125.9	125	6961	*	*	*	*	2653	**2586**	46	6/10
dsjc250.1	250	3218	*	*	*	*	1068	**1048**	9	4/10
dsjc250.5	250	15668	*	*	*	*	3658	**3620**	32	6/10
dsjc250.9	250	27897	*	*	*	*	8942	**8831**	74	4/10
dsjc500.1	500	12458	*	*	*	*	3229	**3094**	14	3/10
dsjc500.5	500	62624	*	*	*	*	12717	**11987**	53	3/10
dsjc500.9	500	112437	*	*	*	*	32713	**29914**	130	2/10
dsjc1000.1	1000	49629	*	*	*	*	10276	**10292**	22	4/10
dsjc1000.5	1000	249826	*	*	*	*	45408	**44708**	89	2/10
dsjc1000.9	1000	449449	*	*	*	*	119111	**115137**	234	3/10

References

1. Garey, M.R., Johnson, D.S.: Computers and Intractability: a Guide to the Theory of NP-Completeness. W.H. Freeman, NewYork (1979)
2. Gamst, A.: Some lower bounds for a class of frequency assignment problems. IEEE Transactions on Vehicular Technology 35, 8–14 (1986)
3. Leighton, F.T.: A graph coloring algorithm for large sceduling problems. Journal of Research of the National Bureau of Standards 84(6), 489–503 (1979)
4. de Werra, D.: An introduction to timetabling. European Journal of Operational Research 19, 151–162 (1985)
5. Chow, F.C., Hennessy, J.L.: The priority-based coloring approach to register allocation. ACM Transactions on Programming Languages and Systems 12, 501–536 (1990)
6. Mehrotra, A., Trick, M.A.: A column generation approach for exact graph coloring. INFORMS J. Comput. 8, 344–354 (1996)
7. Hertz, A., de Werra, D.: Using Tabu Search for Graph Coloring. Journal of Computing 39, 345–351 (1987)
8. Johnson, D.S., Aragon, C.R., McGeoch, L.A., Schevon, C.: Optimization by Simulated Annealing: An Experimental Evaluation: Part II, Graph Coloring and Number Partitioning. Operations Research 39, 378–406 (1991)
9. Costa, D., Hertz, A.: Ants Can Color Graphs. Journal of the Operational Research Society 48, 295–305 (1997)
10. Fleurent, C., Ferland, J.: Genetic and hybrid Algorithms for Graph Coloring Problem. Annals of Operations Research 63, 437–464 (1996)
11. Kubicka, E., Schwenk, A.J.: An introduction to chromatic sums. In: Proceedings of the ACM Computer Science Conference, pp. 39–45 (1989)
12. Kroon, L.G., Sen, A., Deng, H., Roy, A.: The optimal cost chromatic partition problem for trees and interval graphs. In: D'Amore, F., Marchetti-Spaccamela, A., Franciosa, P.G. (eds.) WG 1996. LNCS, vol. 1197, pp. 279–292. Springer, Heidelberg (1997)
13. Kokosiński, Z., Kwarciany, K.: On sum coloring of graphs with parallel genetic algorithms. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4431, pp. 211–219. Springer, Heidelberg (2007)
14. Moukrim, A., Sghiouer, K., Lucet, C., Li, Y.: Lower Bounds for the Minimal Sum Coloring Problem. Electronic Notes in Discrete Mathematics 36, 663–670 (2010)
15. Li, Y., Lucet, C., Moukrim, A., Sghiouer, K.: Greedy Algorithms for the Minimum Sum Coloring Problem. In: International Workshop: Logistics and Transport (2009)
16. Douiri, S.M., Elbernoussi, S.: New algorithm for the sum coloring problem. Int. J. Contemp. Math. Sciences 6(10), 453–463 (2011)
17. Douiri, S.M., Elbernoussi, S.: A New Heuristic for the Sum Coloring Problem. Applied Mathematical Sciences 5(63), 3121–3129 (2011)
18. Dorigo, M.: Optimization, learning and natural algorithms, Ph.D. Thesis, Politecnico di Milano, Italy (1992)
19. Dorigo, M., Maniezzo, V., Colorni, A.: Ant system: Optimization by a colony of cooperating agents. IEEE Transaction on Systems, Man, and Cybernetics - Part B 26(1), 29–41 (1996)
20. Douiri, S.M., Elbernoussi, S.: An Ant Algorithm for the Sum Coloring Problem. International Journal of Applied Mathematics and Statistics 27(3), 102–110 (2012)

21. Douiri, S.M., Elbernoussi, S.: A new ant colony optimization algorithm for the lower bound of sum coloring problem, Journal of Mathematical Modelling and Algorithms. Journal of Mathematical Modelling and Algorithms 11(2), 181–192 (2012)
22. Dorigo, M., Caro, G.D.: Ant colony optimization: A new meta-heuristic. In: Proceedings of the Congress on Evolutionary Computing (1999)
23. `http://mat.gsia.cmu.edu/COLOR03/`

A Genetic Algorithm with Multiple Mutation which Solves Orienteering Problem in Large Networks

Jolanta Koszelew and Krzysztof Ostrowski

Faculty of Computer Science, Bialystok University of Technology, Poland
http://www.wi.pb.edu.pl

Abstract. In this paper we present a genetic algorithm (GA) which solves the Orienteering Problem (OP). In the article, performance of the algorithm is analysed as a function of mutations number. In addition, GA results were compared with known local search methods: greedy randomized adaptive search procedure (GRASP) and guided local search method (GLS). The computer tests were conducted on large network of 908 cities in Poland. As a result, the GA performance was considerably better then local search methods in terms of both: results quality and execution time.

Keywords: genetic algorithm orienteering problem OP multiple mutation Poland large network.

1 Introduction

The Orienteering Problem (OP) can be modelled as a weighted and complete graph problem. Let G be a graph with n vertices, each vertex i has a nonnegative profit p_i. Each edge between vertices i and j has a nonnegative cost t_{ij} associated with it. The triangle inequality is satisfied for G. The starting point s and the ending point e are given. The objective is to find a path from s to e which maximizes the total profit. In addition, the total cost of the edges on this path should be limited by a given constraint t_{max} and each point on the path is visited only once (except the situation when $s = e$).

The problem name comes from the sport game of orienteering. The competitors start and end the game at a specific control point. Their purpose is to collect as much points as possible at given checkpoints and return to the control point within a given period of time. The OP alternative name is Selective Travelling Salesman Problem as it is a generalization of the well known problem.

The OP has many practical applications and tourism is one of the most common (practical systems implemented)[1][2]. When tourists plan a tour in some region, their purpose is to visit as many attractive places as possible. However, they usually are limited by time or money and cannot see all of them. Thus, the OP implementations can be adopted to help tourists plan their trips. The OP can also be applied in logistics[3]. Production scheduling problem is another example of OP application - the purpose is to maximize production profit without exceeding given time constraints.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 356–366, 2013.

There are many variants of the OP - one of them is OPTW (Orienteering Problem with Time Windows)[4] in which every point can be visited only in a given time interval. Another one is TOP (Team Orienteering Problem)[5] in which the goal is to maximize the sum of profits of m paths. The most complex variant is TPP (Tour Planning Problem) [21] in which a multi day-tour is generated with both time constraint and budget constraint. One of the most challenging variant is TDOP (Time Dependant Orienteering Problem)[6] in which edge costs vary in time - it can be adapted to trip planners which take into account public transport schedule.

The OP is an NP-hard problem[7][8] and exact solutions can be very time consuming. Thus, a lot of approximation algorithms (i.e. local search methods[5][9][10], genetic algorithms[11] and taboo search[12]) were proposed to deal with the problem. Benchmark instances, on which various methods are tested, are usually small- or medium-size networks (up to 500 vertices)[9][13][14]. In practice, solutions for larger graphs can also be needed i.e. while travelling around a whole country with many points of interest. Our method has been tested on large network of 908 cities in Poland[15] and it gives satisfactory results in reasonable time compared to other tested methods.

The paper is organised as follows. In section 2 there is literature review. Section 3 presents details of the genetic algorithm. Section 4 includes experimental results and comparison with GLS and GRASP methods. Conclusions are included in section 5.

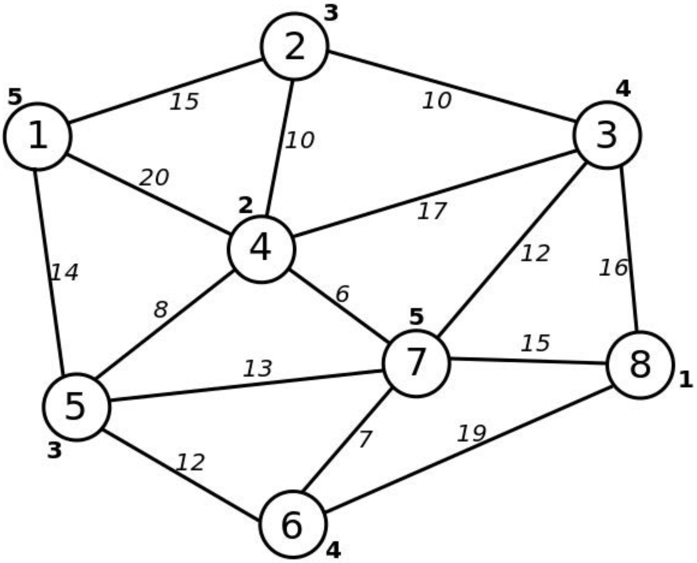


Fig. 1. A graph representing an exemplary transport network. Profits (in bold) are marked besides vertices and costs are marked besides edges.

2 Literature Review

A lot of different approaches has been applied to solve the OP. They include both exact solutions and approximation algorithms. Fischetti et al. found exact solution for instances up to 500 vertices using branch and cut method[12]. Gendreau[16] and Ramesh[17] also solved the OP for medium-size networks (up to 300 vertices). The computational hardness of the OP caused many researchers to develop approximation solutions.

Tsiligrides[13] introduced one of the first heuristic approaches to the OP. His algorithm is based on the Monte Carlo method - routes are generated randomly and the best one is chosen. The probability of vertex choose depends on both distance from the vertex and its profit. Another heuristic was presented by Golden. It cosists of route construction (using greedy algorithm and centre of gravity) and route improvement using 2 -op procedure. Golden et al. worked out the improved iterative heuristic, which links Tsiligrides S-algorithm concept, the center of gravity and learning capabilities.

Chao et al. [8] introduced a two-step iterative heuristic, consisting of initialization and improvement steps. The method only considers vertices lying in the ellipse using t_{max} as a length of the major axis. Several routes are generated with the help of the greedy algorithm and the one with the highest score is chosen. In the improvement phase a two-point exchange in the cheapest way is applied and next, improvement point is used in order to increase the total score. Finally, 2-op procedure is applied to decrease the total length of solution and when the t_{max} is exceeded. Tasgetiren et al. [10] worked out the first genetic algorithm for the OP. The tournament selection, the injection crossover as well as the mutation using elements of the local search method (add, omit, replace and swap operators) were created by them. In addition, they applied penalty function.

Vansteenwegen et al. [4] developed the guided local search method (GLS) for the OP. For each local search iteration, they used special penalties e.g. increase of the score of the non-included locations and decrease of the score of the included location. First, route generation is based on Chao method. Each searching iteration consists of 3 phases: TSP (2-op procedure which reduces the length of the route), Insert and Replace. Additionally, after reaching local optimum and applying penalty function, a procedure of diversification is executed - it removes a part of the tour in order to better explore solution space. The GLS method is applied in the Mobile Tourist Guide[2] and it gives satisfactory results for small and medium networks.

Souffriau et al. presented GRASP with Path relinking (GRASPwPR) method to solve the TOP[18]. Later it was adapted to solve the OP by Campos et al[10]. They obtained high quality results on benchmark instances - better than other methods. First, initial path is generated partially a in greedy way and partially in a random way (ratio between greediness and randomness applied). Afterwards, the procedures of exchange and insert are executed to decrease the tour length and increase its profit. A set of paths is generated in such a way and for each pair of them path relinking is performed: it gradually transforms one path into another by exchanging vertices.

Solutions mentioned in the literature were usually tested in small or medium networks (the number of vertices in Tsiligirides [14], Chao et al. [9] and Fischetti et al. [13] benchmark instances vary between 21 and 500). In practice there also could be need to obtain satisfactory solutions on larger networks and that is the goal of our research.

3 The GA Description

The method is an improved version of [19][20] with different, searching crossover operator and multiple mutation. Major changes include:

- Modified crossover procedure analyses all the possibilities of segment exchange between common points and chooses the one with highest child fitness whereas procedure from [19][20] randomly chooses a crossing point and exchanges chromosome fragments after the point.
- In removing mutation from [19][20] only duplicate genes are considered while in the modified version removal of all vertices is considered (for longest paths exceeding $0.9 \cdot t_{max}$).
- In this method procedure of mutation is performed multiply, which is opposite to mutation from [19][20].

The algorithm works on stardard OP version with triangle inequility satisfied. Path are encoded into chromosomes as a sequence of vertices. First, a population of P_{size} individuals is generated. Afterwards, operators of selection, crossover and mutation are applied repeatedly in each generation. The algorithm ends after a fixed number of generations or when it converges earlier. The GA result is a path with highest profit from the final population. The pseudocode:

```
Initial_Population_Creation;
Current_Generation=0;
while Current_Generation < Ng do
    Tournament_Selection;
    Crossover;
    Mutation;
    Current_Generation=Current_Generation+1;
    if No_Improvement_In_Last_100_Generations then
        break;
    end
end
return Best_Solution;
```

3.1 Evaluation

To estimate the quality of individuals we use fitness function F which is equal to $\frac{TotalProfit^3}{PathCost}$. It takes into account both path score and its potential of further growth. The total profit is sum of vertices profits. A given vertex can appear on the path twice but its profit is taken into account only once.

3.2 Initialization

First, the initial population of P_{size} solutions is generated. At the beginning a random vertex v, which is adjacent to vertex s (the start point) is chosen and path length is increased. Now we start at vertex v and choose a random vertex u adjacent to v. At every step we exclude the previously visited vertex from the set of candidate vertices and we consider only vertices which can be added without exceeding $0.5 \cdot t_{max}$. If there is no appropriate vertex to chose we return to vertex s the same way in reversed order. This way of generating the individuals of the initial populations means that they are symmetrical in respect to the middle vertex in the tour. However, these symmetries are removed by the algorithm. Initial population example is in tab. 1. All the examples base on the graph from fig. 1.

Table 1. An initial population example ($P_{size} = 5$, $t_{max} = 80$)

No	Individual	Profit	Cost	Fitness
1	(1, 2, 3, 7, 3, 2, 1)	17	74	66.4
2	(1, 5, 6, 7, 6, 5, 1)	17	66	74.4
3	(1, 4, 7, 6, 7, 4, 1)	16	66	62.1
4	(1, 5, 7, 4, 7, 5, 1)	15	66	51.1
5	(1, 5, 4, 7, 6, 7, 4, 5, 1)	19	77	89.1

3.3 Selection

We apply tournament selection. In each tournament we take t_{size} different, random individuals. The fittest of them is copied into the next generation. After P_{size} tournaments a new population is created. If we choose individuals no 1, no 2 and no 3 in a tournament ($t_{size} = 3$) from our population (tab. 1) the winner is individual no 2. It has the same profit as individual no 1 but lower cost. It also has the same cost as individual no 3 but higher profit.

3.4 Crossover

In the crossover process two random individuals are selected - they are parents. Afterwards, a searching procedure is performed - we determine all genes which are common in parents (the set of intersections). If there are no common genes, crossover cannot be done and no changes to the chosen chromosomes are applied. Next, we look for such pairs of successive points in the set of intersections for which two following conditions are satisfied: (1) the t_{max} limit after exchanging parts of individuals between considered points is preserved for at least one child, (2) one child has better fitness than its both parents and does not exceed t_{max}. From these pairs we choose the one which gives the highest fitness of the better child (after crossover). If one of the children does not preserve t_{max} constraint, the fitter replaces it in a new population. Let (1, 4, 7, 6, 7, 3, 4, 1) and (1, 5, 6, 8, 3, 2, 1) be two parents. They heve 4 common genes (1, 6, 3 and 1) so there are 3 options of crossover (presented in tab. 2). The best of them is exchanging fragments between 6 and 3 - the better child has fitness value of 197.5. The other child has cost which exceed t_{max} (105) and should be replaced by the fitter parent (1, 5, 6, 8, 3, 2, 1).

Table 2. An example of crossover between (1, 4, 7, 6, 7, 3, 4, 1) and (1, 5, 6, 8, 3, 2, 1) for $t_{max} = 90$. Chosen option and best child in bold.

	Individuals	Profit	Cost	Fitness
parents data	(1, 4, 7, 6, 7, 3, 4, 1)	20	89	89.9
	(1, 5, 6, 8, 3, 2, 1)	20	86	93.0
fragments exchange	(1, 5, 6, 7, 3, 4, 1)	23	82	148.4
between 1 and 6	(1, 4, 7, 6, 8, 3, 2, 1)	24	93	148.6
fragments exchange	(1, 4, 7, 6, 8, 3, 4, 1)	21	105	88.2
between 6 and 3	**(1, 5, 6, 7, 3, 2, 1)**	**24**	**70**	**197.5**
fragments exchange	(1, 4, 7, 6, 7, 3, 2, 1)	23	77	158.0
between 3 and 1	(1, 5, 6, 8, 3, 4, 1)	19	98	70.0

Table 3. An example of inserting mutation of (1, 5, 6, 8, 3, 2, 1) for $t_{max} = 100$. Chosen option in bold.

new vertex and insertion place	$\frac{NewVertexProfit^2}{PathCostIncrease}$	mutated individual	Total cost
4 (between 1 and 5)	0.286	(1, 4, 5, 6, 8, 3, 2, 1)	100
7 (between 5 and 6)	3.125	(1, 5, 7, 6, 8, 3, 2, 1)	94
7 (between 6 and 8)	**8.333**	**(1, 5, 6, 7, 8, 3, 2, 1)**	**89**
7 (between 8 and 3)	2.273	(1, 5, 6, 8, 7, 3, 2, 1)	97
4 (between 3 and 2)	0.235	(1, 5, 6, 8, 3, 4, 2, 1)	103
4 (between 2 and 1)	0.267	(1, 5, 6, 8, 3, 2, 4, 1)	101

3.5 Mutation

First, a random individual is selected to be mutated. Then a mutation procedure is performed N_m times on the chosen path. During a single mutation we perform gene insertion or gene removal (each with the probability of 0.5). The probability of 0.5 gives some balance between insertions and removals especially for very long paths (close to t_{max}) or for paths which have duplicate vertices.

When inserting a new gene (not present in the path) all the possibilities are considered. When choosing a new vertex and an insertion place the searching procedure maximizes ratio $\frac{NewVertexProfit^2}{PathCostIncrease}$. This step is time-consuming especially if the mutation of a given individual is performed multiple times. Thus, some optimizations were applied - for a limited number of insertion places (each defined by a pair of vertices) a ranking of potentially new vertices is memorized (sorted by $\frac{NewVertexProfit^2}{PathCostIncrease}$). Furthermore, if path cost is already close to t_{max} we consider only closest neighbours.

When removing a gene first we consider vertices which repeat in a path and choose one of them in order to reduce the path cost as much as possible. If there are no duplicates in a path and its length exceeds $0.9 \cdot t_{max}$ (path is almost full) then we remove a vertex in order to minimize ratio $\frac{RemovedVertexProfit^2}{PathCostDecrease}$. It gives a chance for further path improvement. An example of inserting mutation is presented in tab. 3. There are six options of inserting a new vertex and the chosen option has highest $\frac{NewVertexProfit^2}{PathCostIncrease}$ ratio. Last two options are impossible to apply - mutated individual exceeds t_{max} (100) constraint.

4 Experimental Results

In this section results of GA performance with various mutation numbers are presented as well as GA comparison to local search methods (GLS and GRASP) and previous algorithm version (tagged as pGA)[19]. The experiments have been carried out on the network, which contains 908 cities all over the territory of Poland. Each city is described by its longitude, latitude as well as its profit. The spherical distances between any c_1 and c_2 cities are calculated (in meters) as follows:

$$D = 6378137 \cdot arccos(arcsin(lat_1) \cdot sin(lat_2) + cos(lat_1) \cdot cos(lat_2) \cdot cos(long_1 - long_2))$$

where lat_1, lat_2, $long_1$, $long_2$ denote the latitude and the longitude of the cities (given in radians). Profit associated with a given city is proportional to its population. Profit of each city is equal to the number of inhabitants divided by 1000 (profit varies between 1 and 1720). The capital city of Poland, Warsaw, is established as the starting and the ending point. Costs limits (t_{max}) tested in the experiment are 500, 1000, 1500, 2000, 2500, 3000 kilometers.

Due to randomization the GA results can vary so the algorithm is executed 30 times and the best path is a result. In addition, this procedure was performed multiply and 95 percent confidence intervals (CI) for the best results were computed. Execution time relates to 30 algorithm runs altogether. GA parameters were $P_{size} = 300$, $t_{size} = 5$, $N_g = 1000$. Population size of 300 seems to be enough for the solution space of the experimental network - further increasing

Table 4. Comparison of best results from 30 runs (with confidence intervals) for different N_m values

	$N_m = 1$		$N_m = 3$		$N_m = 5$		$N_m = 10$		$N_m = 30$	
t_{max}	score	CI	score	CI	score	CI	score	CI	score	CI
500	3680	±6	3687	±6	3688	±5	3692	±5	3693	±5
1000	7690	±28	7750	±24	7744	±23	7768	±24	7780	±17
1500	9776	±52	10032	±33	10058	±24	10047	±20	10056	±27
2000	11275	±64	11576	±36	11616	±33	11619	±21	11621	±26
2500	12524	±47	12898	±38	12903	±37	12955	±38	12963	±29
3000	13497	±76	13922	±52	14018	±37	14013	±34	14027	±40
Avg.	**9740**	**±21**	**9977**	**±16**	**10005**	**±11**	**10016**	**±9**	**10023**	**±12**

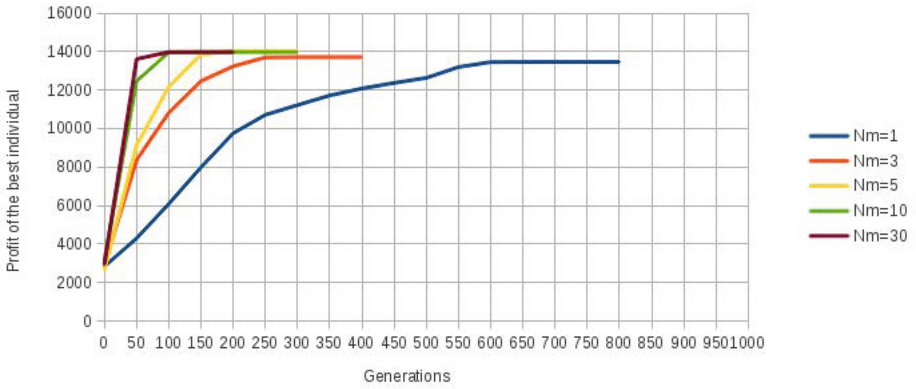


Fig. 2. GA convergence for different N_m values

of $P_{size} = 300$ gives practically no improvement. Parameter t_{size} was set to 5 but algorithm results were similar for different, tested t_{size} values (2, 3, 4 and 5). For such population size the algorithm almost always converges before the 1000th generation - N_g is an upper bound for convergence. Mutation repetition parameter (N_m) has values: 1, 3, 5, 10 and 30. In the comparison with other methods the GA is run with $N_m = 5$ as it has satisfactory results and relatively short execution time.

In the tests the GLS parameters are the same as in [4] with the exception of the replace phase. In this step, the maximum number of iterations is determined by tests and it is set to 700[19]. GRASP(wPR) implementation is based on the article [10]. Four constructive methods (alpha denoted as C1, C2, C3, C4), each of them with four values of randomness (alfa equal to 0.2, 0.4, 0.6 or 0.8) were tested in the GRASP. The method C2 (alpha=0.2), giving the best profit result in a relatively short time, was chosen to the comparison[20]. The experiments have been performed on the Intel Core 2 duo 2.8GHz CPU.

First, results of GA for different N_m values were compared (tab. 4). One can see that best results of 30 GA runs generally are better as number of muta-

Table 5. Comparison of execution times (s) of 30 runs for different N_m values

t_{max}	$N_m = 1$	$N_m = 3$	$N_m = 5$	$N_m = 10$	$N_m = 30$
500	3.1	3.2	3.2	3.6	5.3
1000	5.7	5.7	5.5	6	9.4
1500	7.8	7.7	7.5	7.9	11.5
2000	9.0	9.3	9.4	9.7	14.7
2500	12.1	11.7	11.8	13.1	17.8
3000	13.3	13.7	14.5	16	21.9
Avg.	**8.5**	**8.6**	**8.7**	**9**	**13.5**

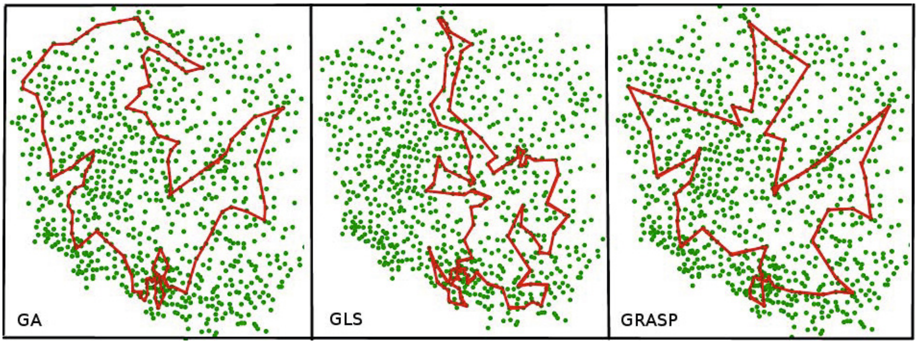


Fig. 3. Paths generated by the GA, GLS and GRASP ($t_{max} = 3000$)

tions grows - in this case individuals reach their local optimums faster. It is clearly seen especially for higher t_{max} values as paths have more potential to grow and improve. However, this effect is more pronounced as N_m grows from 1 to 5. Further grow gives little improvement. What is more, execution time grows significantly (tab. 5) for larger N_m - procedure of repetitive mutation starts to be dominant operation in the algorithm. However, this effect is partially compensated by faster algorithm convergence (fig. 2). N_m values in a range of 5-10 look to be reasonable for this large network in terms of both results quality and execution time.

Table 6. Comparison of GA ($N_m = 5$), pGA, GLS, GRASP (abbr. GR), GRASPwPR (abbr. GRPR) results and execution time

	GA		pGA		GLS		GR		GRPR		% gap GA with		
t_{max}	score	time	score	time	score	time	score	time	score	time	GLS	GR	GRPR
500	3688	3.2	3666	2.1	3456	4.0	3652	16.6	3631	27.1	6.7	1.0	1.6
1000	7744	5.5	7621	4.2	6659	14.7	7267	21.8	7472	31.1	16.3	6.6	3.6
1500	10058	7.5	9671	5.5	9750	19.1	8859	20.9	8862	37.2	3.2	13.5	13.5
2000	11616	9.4	10527	6.9	10236	17.8	10280	24.4	10256	45.0	13.5	13.0	13.3
2500	12903	11.8	12280	12.9	12027	22.5	11234	31.6	11400	54.1	7.3	14.9	13.2
3000	14018	14.5	13237	17.2	12464	21.5	12595	41.5	12472	67.4	12.5	11.3	12.4
Avg.	**10005**	**8.7**	**9500**	**8.1**	**9099**	**16.6**	**8981**	**26.1**	**9016**	**43.6**	**9.9**	**11.4**	**10.9**

The second part of the experiment was comparing GA with other methods (pGA, GLS, GRASP, GRASPwPR). One can see (tab. 6) that for large network the GA is considerably faster than local search methods and its results are also significantly better (see confidence intervals in table 4) especially for higher t_{max} values. One can see also a noticable improvement from the previous GA version. Its contributors are new, searching crossover and multiple mutation. Due to multiple mutation average execution time of GA increased (compared to pGA). However, this effect was minimized thanks to optimizations mentioned in section 3.5.

5 Conclusions and Further Work

The experiment shows that the OP solving genetic algorithm when operating on large network and long paths (chromosomes) can be further improved by intensifying the mutation phase. It also shows that GA algorithm can succesfully compete with local search methods on large networks (in terms of both execution time and results quality). There can still be room for GA improvement in both mutation and selection procedures. We plan to test it on other networks (including smaller benchmark instances) and to adopt the GA for Orienteering Problem with Time Windows and Tour Planning Problem. There is also a possibility of creation a new, hybrid algorithm which would base on both: genetic algorithm and local search method.

References

1. Souffriau, W., Vansteenwegen, P.: Tourist Trip Planning Functionalities: State–of–the–Art and Future. In: Daniel, F., Facca, F.M. (eds.) ICWE 2010. LNCS, vol. 6385, pp. 474–485. Springer, Heidelberg (2010)
2. Vansteenwegen, P., Souffriau, W., Vanden Berghe, G., Van Oudheusden, D.: The City Trip Planner: An expert system for tourists, vol. 38. Elsevier (2011)
3. Koszelew, J.: Logistics for globetrotters - innovative software component for etourism. Logistics 6, 54–56 (2010)
4. Vansteenwegen, P., Souffriau, W., Vanden Berghe, G., Van Oudheusden, D.: Iterated local search for the team orienteering problem with time windows. Computers O.R. 36, 3281–3290 (2009)
5. Vansteenwegen, P., Souffriau, W., Vanden Berghe, G., Van Oudheusden, D.: A guided local search metaheuristic for the team orienteering problem. European Journal of Operational Research 196, 118–127 (2009)
6. Garcia, A., Arbelaitz, O., Vansteenwegen, P., Souffriau, W., Linaza, M.T.: Hybrid approach for the public transportation time dependent orienteering problem with time windows. In: Corchado, E., Graña Romay, M., Manhaes Savio, A. (eds.) HAIS 2010, Part II. LNCS, vol. 6077, pp. 151–158. Springer, Heidelberg (2010)
7. Golden, B., Levy, L., Vohra, R.: The orienteering problem. Naval Research Logistics 34 (1987)
8. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman (1979)
9. Chao, I.M., Golden, B.L., Wasil, E.A.: A Fast and effective heuristic for the orienteering. European Journal of Operational Research 88, 475–489 (1996)
10. Campos, V., Marti, R.: Grasp with Path Relinking for the Orienteering Problem. Technical Raport, 1-16 (2011)
11. Tasgetiren, M.F., Smith, A.E.: A genetic algorithm for the orienteering problem. In: Proceedings of the 2000 Congress on Evolutionary Computation San Diego, vol. 2, pp. 1190–1195 (2000)
12. Tang, H., Miller-Hooks, E.: A tabu search heuristic for the team orienteering problem. Comput. Oper. Res. 32(6), 1379–1407 (2005)

13. Fischetti, M., Salazar, J.J., Toth, P.: Solving the Orienteering Problem through Branch-and-Cut. INFORMS Journal on Computing 10, 133–148 (1998)
14. Tsiligirides, T.: Heuristic methods applied to orienteering. Journal of the Operational Research Society 35(9), 797–809 (1984)
15. Network of 908 polish cities, `http://jolantakoszelew.pl/attachments/File/Poland_908_cities.txt`
16. Gendreau, M., Laporte, G., Semet, F.: A branch-and-cut algorithm for the undirected selective traveling salesman problem. Networks 32(4), 263–273 (1998)
17. Ramesh, R., Yoon, Y., Karwan, M.H.: An optimal algorithm for the orienteering tour problem. INFORMS Journal on Computing 4(2), 155–165 (1992)
18. Souffriau, W., Vansteenwegen, P., Vanden Berghe, G., Van Oudheusden, D.: A path relinking approach for the team orienteering problem. Computers & Operational Research 37, 1853–1859 (2010)
19. Karbowska-Chilinska, J., Koszelew, J., Ostrowski, K., Zabielski, P.: Genetic algorithm solving Orienteering Problem in large networks. Frontiers in Artificial Intelligence and Applications 243 (2012)
20. Karbowska-Chilińska, J., Zabielski, P.: A genetic algorithm vs. Local search methods for solving the orienteering problem in large networks. In: Graña, M., Toro, C., Howlett, R.J., Jain, L.C. (eds.) KES 2012. LNCS, vol. 7828, pp. 11–20. Springer, Heidelberg (2013)
21. Zhu, C., Hu, J.Q., Wang, F., Xu, Y., Cao, R.: On the tour planning problem. Annals of Operations Research 192, 67–86 (2012)

Parallel Genetic Algorithm for Creation of Sort Algorithms

Igor Trajkovski

Faculty of Computer Science and Engineering,
"Ss. Cyril and Methodius" University in Skopje,
Rugjer Boshkovikj 16, P.O. Box 393, 1000 Skopje, Macedonia
trajkovski@finki.ukim.mk
http://www.finki.ukim.mk/en/staff/igor-trajkovski

Abstract. In this paper we present parallel genetic algorithm that was used to the task of evolving imperative sort programs. A variety of interesting lessons were learned. With proper selection of the primitives, sorting programs were evolved that are both general and non-trivial. Unique aspect of our approach is that we represent the individual programs with simple assembler code, rather than usual tree like structure. We also report the effect of different parameters on quality of the programs and time needed for finding the solution.

Keywords: genetic algorithm, sort algorithm, parallel algorithm.

1 Introduction

Genetic Programming (GP) derives from Genetic Algorithms (GA), developed initially by John Holland [1] and extended since then by many others. Genetic Algorithms often work on fixed length, linear representations of genetic material, and operate on this material with fitness proportionate selection, reproduction, crossover, and mutation [2], [3].

John Koza has developed GP, using analogues of GA genetic operators to directly modify and evolve tree structured programs (typically LISP functions) [4], [5]. While genetic programming does not imitate nature as closely as do genetic algorithms, it does offer the opportunity to evolve programs of a high complexity, without having to define the size or the structure of the program in advance. Koza has shown in [5] that GP can be effectively applied to an unusual variety of problem areas.

Sorting is a procedure of rearranging the elements of a given input sequence into ascending or descending order. Sorting has a rich history in computer science and is a procedure that requires complex control structures. There are a lot of solutions to this problem domain - some are simple and intuitive, while others are complicated but of greater efficiency. Sorting algorithms fall into two classes of time complexity, specifically, $O(n^2)$ and $O(n \times \log(n))$. The algorithms with quadratic complexity are usually expressed iteratively, whereas algorithms with $O(n \times \log(n))$ complexity leads more naturally to recursive expressions.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 367–376, 2013.

The sorting problem has attracted a great deal of research, due to its ubiquity, and due to the challenge of solving it efficiently.

This paper presents work on evolving general iterative sorting algorithms represented as linear list of instructions, contrary to the traditional representation of programs in GP, as trees. Sorting is clearly a general problem for GP since the evolved hypothesis will have to sort the comparable elements of any input sequence of arbitrary length into ordered sequence. This is a challenging problem, and in this paper we will show that it can be solved with the proposed method.

The remainder of the paper is organized as follows: first we provide a literature review on the evolution of sorting programs. Then the fitness function used to guide the evolutionary search is described. We then proceed to describe the experimental context: instruction set used for construction of the programs, evolutionary algorithms control parameters, parallelization strategies followed by a section that details and discusses the experimental results, and the computational effort involved. Finally, conclusions are drawn and ideas for future work are presented.

2 Related Work on the Evolution of Sorting Algorithms

A literature survey on evolution of sorting programs evolution revealed several attempts in this problem domain.

Kinnear [**??**] evolved iterative sorting algorithms, mainly of bubble sort's simplicity. He investigated the relative probability of success and the difficulty resulting from varying the primitive terminal and non-terminal elements. The primitive alphabet contained elements that could result in an exchange-oriented sorting strategy. Primitive functions were defined for swapping sequence elements, comparing elements in specified sequence index values, incrementing and decrementing arithmetic variables. Control functions contained conditionals and a bounded iteration construct. It considered the addition of a linear parsimony function in order to discourage the individuals' increasing size as well as a disorder penalty, in the case where the remaining disorder was greater than the initial sequence disorder.

O'Reilly and Oppacher [7], [8] also investigated ways of evolving iterative sorting algorithms. Their first attempt [7] failed to produce a 100% correct individual. The representational constructs and fitness function used were different than those used in Kinnear's experiments. Specifically, primitive functions and control structures included decrementing a variable, accessing indexed sequence elements, swapping adjacent sequence elements, bounded looping, and conditional. The fitness function counted the number of out-of-place items. Their second attempt [8] yielded a successful outcome. It considered different primitive constructs and two fitness functions; the first fitness function was the same as in [7] whereas the second was based on permutation order [9], the count for each sequence element of the smaller elements that follow it.

The most recent attempt to the evolution of sorting algorithms is presented in the work of [10] with their PushGP system. They used primitives along the lines of earlier investigations: swapping and comparing indexed elements, getting the list length, accessing list elements. The Push3 programming language offers a variety of explicit iteration instructions but also allows for the evolution of novel control manipulation structures. They evolved an $O(n^2)$ general sorting algorithm and suggested an efficiency component addition to the fitness function as a precursor to the evolution of ingenious $O(n \times \log(n))$ algorithms, though they report no experiments towards that direction.

3 Sorting Fitness Function

When evolving a sorting algorithm, the following two problems arise:
How to quantitatively measure the algorithm's ability to sort?, and
How to Recognize Success?, or how can we be sure that we found the solution.

The fitness function is the driving force behind any evolutionary algorithm. The choice of fitness function can play a critical part in the success of a given run. In this work, we consider fitness functions based on different measures of change in the sequence disorder.

We first familiarize the reader with a concepts of sequence disorder, namely, *inversions.*

Let $a_1\ a_2\ \ldots\ a_n$ be a permutation of the set $1,\ 2,\ \ldots,\ n$. The pair $(a_i,\ a_j)$ is called an inversion of the permutation if and only if $i < j$ and $a_i > a_j$. Inversion pairs represent pairs of sequence elements that are out-of-order, so that a completely sorted sequence is a permutation with no inversions. Let $Inv(A)$ represent the number of inversions in sequence A. Using these notations we define the following fitness measure F of the program P on given input sequence A:

$$F(P) = Inv(A) - Inv(B)$$

where $B = P(A)$ is the output sequence of the program P executed on input sequence A. In some cases B is the modified, by program P, input sequence A.

We can see that programs which completely sort the input sequence will have maximum fitness. Off course we do not test the program on only one input sequence, but on several (A_i) test sequences. In this case the fitness measure F is calculated with the following formula:

$$F(P) = \sum_{i=1}^{N} Inv(A_i) - Inv(P(A_i))$$

where $N = |A|$.

When evolving a sorting program, a problem presents itself. Frequently, an evolved program which performs successfully all of the fitness tests used (a *completely fit individual*) will fall short of complete generality. Therefore, a particular run is not terminated with the appearance of a completely fit individual. Instead, additional testing is performed on every completely fit individual and the run is terminated as successful only when one individual passes these more tests.

The fitness tests used to drive evolution consist of K random tests, with a maximum length of 30, with the random tests changed in each evaluation. The additional tests for generality (used only to terminate the run) consist of 256 random tests of maximum length 32, and all 256 sequences of length 8 consisting only of 0 and 1.

While many of the completely fit individuals fail to pass the additional tests (usually only 60% to 80% pass), no algorithm that has passed all of the additional tests has ever been shown to fail on any sequence subsequently presented to it. Many of these evolved and potentially fully general algorithms have been tested with thousands of random sequences of lengths of up to 600 with not a single failure. Thus, while testing can never prove generality, the sorting problem as expressed with these primitives appears to have a certain threshold of testing that, once passed, ensures at least a very high likelihood of generality.

The last fitness measure F has a small shortage. It prefers moderate partial sorting of long sequences than perfect complete sorting of short sequences because the number of fixed inversion is bigger in the first case. Therefore, the following fitness measure is used that normalize the number of fixed inversions relative to the number of inversions sequence had in the first place:

$$F(P) = \sum_{i=1}^{N} \frac{Inv(A_i) - Inv(P(A_i))}{Inv(A_i)}$$

4 Program Representation and Genetic Operators

The basic approach to genetic programming is to generate a random population of individuals, evaluate their fitness, perform various genetic operations on them based in some way on their fitness, and then go back and evaluate the results again. The desired function of the individuals (programs) is to reorder a sequence of integers so as to leave it in "sorted" order, small to large. In our approach, the individuals of the population are pseudo assembler programs. Assembler instructions are defined to compare and swap the various values in the sequence, to increase, decrease and compare various registers, and control the flow of the execution by various conditional jump instructions.

To evaluate the fitness of an individual, it is presented with an "unsorted" sequence and then executed. The "disorder" of the sequence is measured before and after execution, and the fitness is based on the decrease in disorder. If the program access out-of-memory address, or it does not end in predefined maximum number of cycles, it gets lowest possible fitness value.

4.1 Instruction Set

In order to apply genetic programming to a problem, it must be cast into a form that can be evolved. A set of assembler instructions must be created that is sufficient to solve the problem. Arguments to these instructions will be registers and addresses (or values) of sequence elements.

A variety of instruction sets were tried in order to successfully evolve a general and yet non-trivial sorting algorithm. The instructions described here are the least complex, and therefore most realistic that would reliably allow evolution of general sorting algorithms within a reasonable time.

Our instruction set is composed of the following 12 classes of instructions:

```
 1. MOV  Ri, Rj(const)
 2. XCHG mem[Ri], mem[Rj]
 3. INC  Ri
 4. DEC  Ri
 5. JE   Ri, Rj(const), line
 6. JA   Ri, Rj(const), line
 7. JB   Ri, Rj(const), line
 8. JMP  line
 9. JA   mem[Ri], mem[Rj], line
10. JB   mem[Ri], mem[Rj], line
11. JE   mem[Ri], mem[Rj], line
12. END
```

$R_j(const)$ means R_j or *constant value*, which will get us to 16 different types of instructions.

Instruction 1 moves the value of register R_j to register R_i, or moves a constant value in register R_i. Instruction 2 swap the values of sequence elements at positions R_i and R_j. Instructions 3-4 increase/decrease the value of register R_i. Instructions 5-7 are conditional jumps, comparing register R_i with another register R_j or constant value, to instruction *line*. Instruction 8 is a unconditional jump to instruction *line*. Instructions 9-11 are conditional jumps, comparing the values of sequence elements at positions R_i and R_j, to instruction *line*. If instruction 12 is executed the program terminates.

One thing worth noting is that this instruction set will not allow the individuals in the population to modify the data in any way, they can only change the order.

4.2 Initial Population

Initial population was constructed by generating 1000 random programs with fixed length of L instructions. The effect of L was investigated in the experiments. Each instruction was randomly chosen from the 16 different types of instructions.

4.3 Mutation

Mutation works by selecting an individual by uniform random selection, selecting an instruction within the program and then one of the three types of mutations is done: mutating (changing) the instruction arguments, changing the type of the instruction or generating completely new instruction. When instruction argument is changed, if it is a constant, random value with power low distribution is added/subtracted, if it is a register it is substituted by another random register. When type of the instruction is changed, another type is chosen that has the same type of arguments. For example, if the type of INC instruction is mutated, it can become DEC, but not XCHG.

4.4 Crossover

Crossover is done in the following way: two individuals are selected with fitness proportional selection, than two continuous blocks of instructions, equal size, one from each individual were selected, and than they were exchanged. In this way we get two new individuals. The size of the blocks is a random number from 4 to 8 instructions. The position of the block was also random number from 0 to L - $block\_size$.

5 Experimental Setup and Parallelization

A population of 10.000 programs were randomly generated with a fixed length L, generated by uniform selection over the list of instructions. Evaluation of the programs was done by their execution and analyzing the disorder of the input sequence before and after the execution of the program in order to compute its fitness. The number of registers used by the programs was fixed to NR. Before the execution of the programs, it is assumed that the length of the input sequence N is stored in register R_0 and that numbers of the input sequence are located on the memory locations 0 to $N - 1$. The memory working area was set to be from $-3 * N$ to $3 * N$. If a program access a memory location outside this interval, the fitness of the program was set to minimum possible value. Also, if the program did not finish its execution after 5000 executed instructions it is asumed that it does not halt, and it is given the lowest possible fitness value. The initial value of all registers, except R_0, was set to 0. The initial value of all memory locations, except from 0 to $N - 1$, was also set to 0.

5.1 Parallelization Strategies

The basic idea behind most parallel programs is to divide a task into chunks and to solve the chunks simultaneously using several processors. This divide-and-conquer approach can be applied to GA in many different ways, and the literature contains many examples of successful parallel implementations. Some parallelization methods use a single population, while others divide the population into several relatively isolated subpopulations. Some methods can exploit massively parallel computer architectures, while others are better suited to multicomputers with fewer and more powerful processing elements.

There are three main types of parallel GA: (1) global single-population master-slave GA, (2) single-population fine-grained, and (3) multiple-population coarse-grained GA. In a master-slave GA there is a single population (just as in a simple GA), but the evaluation of fitness is distributed among several processors. Since in this type of parallel GA selection and crossover consider the entire population, it is also known as global parallel GA. Fine-grained parallel GA are suited for massively parallel computers and consist of one spatially structured population. Selection and mating are restricted to a small neighbourhood, but neighbourhoods overlap permitting some interaction among all the individuals. The ideal case is to have only one individual for every processing element available. Multiple-population (or multiple-deme) GA are more sophisticated, as they consist of several subpopulations which exchange individuals occasionally.

This exchange of individuals is called migration and it is controlled by several parameters. Multiple-deme GA are very popular, but they are also the class of parallel GA which is most difficult to understand, because the effects of migration are not fully understood. Multiple-deme parallel GA introduce fundamental changes in the operation of the GA and have a different behaviour than simple GA. Multiple-deme parallel GA are known with different names. Sometimes they are known as "distributed" GA, because they are usually implemented on distributed memory MIMD computers. Since the computation-to-communication ratio is usually high, they are occasionally called coarse-grained GA. Finally, multiple-deme GA resemble the "island model" in population genetics which considers relatively isolated demes, so the parallel GA are also known as "island" parallel GA.

Another method to parallelize GA combines multiple population with master-slave or fine-grained GA. We call this class of algorithms hierarchical parallel GA, because at a higher level they are multiple-deme algorithms with single-population parallel GA (either master-slave or fine-grained) at the lower level. A hierarchical parallel genetic algorithm (HPGA) combines the benefits of its components, and it promises better performance than any of them alone.

We implemented a HPGA. In our test computations, we ran the same sequential version on several processors, and after every 100 generations 3 randomly selected individuals from 10 best individuals of every population were sent to all other populations. The best known solutions are always found for all instances. According to the results, we observe that the algorithm is well fitted to a large number of instances, and that speedup is proportional to the number of the processors. We can mention that there is a case with superlinear speedup where the speedup is more than 8 when we run our GA on 8-processor machine. This is due to the fact that GA are non-deterministic algorithms.

6 Evolved Sort Programs

Let's look at some of the sort programs that have evolved using the presented approach, $L = 20$, $NR= 5$.

Example 1:

```
0  XCHG mem[R4], mem[R3]
1  INC R4
2  MOV R1, R4
3  DEC R0
4  JA mem[R0], mem[R1], 18
5  JA mem[R4], mem[R0], 17
6  XCHG mem[R0], mem[R1]
7  MOV R4, R3
8  JB mem[R1], mem[R4], 17
9  JNE R0, R2, 2
10 MOV R2, 2
11 END
12 INC R1
13 JB R0, -1, 1
14 DEC R0
15 DEC R0
16 DEC R4
17 XCHG mem[R0], mem[R4]
18 INC R0
19 JNE R4, R0, 0
```

Example 2:

```
0  DEC R0
1  MOV R4, R3
2  JB mem[R1], mem[R0], 5
3  JMP 16
4  XCHG mem[R4], mem[R3]
5  MOV R3, R2
6  JB R1, R1, 0
7  XCHG mem[R0], mem[R3]
8  JB mem[R0], mem[R1], 16
9  JE R0, R4, 0
10 XCHG mem[R4], mem[R3]
11 JNE R4, R2, 2
12 DEC R0
13 MOV R1, R1
14 MOV R3, -1
15 MOV R4, R4
16 INC R4
17 XCHG mem[R0], mem[R3]
18 JB R1, R0, 4
19 END
```

The first impression is that they look very messy, and it is hard to follow the flow of the algorithm. The second thing is that they have a lot of unnecessary instructions that do not have any effect on the functionality. But on the other hand it was expected in some way, because also the DNA of almost any species has these features. If we remove instructions that do not have impact on the semantic of the programs and rewrite them in more readable PASCAL-like language, the second program for example, will look like this:

Example 2 (PASCAL-like):

```
    r0 := n; r1 = r2 = r3 = r4 := 0;
0:  dec(r0);
1:  r4 := 0;
2:  if smaller(0, r0) then goto l5;
3:  goto l16;
4:  swap(r4, 0);
5:
6:
7:  swap(r0, 0);
8:  if smaller(r0, 0) then goto l16;
9:  if r0 = r4 then goto l0;
10: swap(r4, 0);
11: if r4 <>0 then goto l2;
12:
13:
14:
15:
16: inc(r4);
17: swap(r0, 0);
18: if 0 <r0 then goto l4;
19: goto end;
    end:
```

swap(X, Y) is a procedure for swapping the values on memory locations X and Y, and *smaller(X, Y)* is a boolean function that compares values on memory locations X and Y. Empty instructions represent instructions that do not have any effect on the semantic of the program, something like a NOP (no operation assembler instruction).

If we look closely we can see that this is INSERTION SORT algorithm. It is written in a very messy way, but if we follow the execution of the program it is doing exactly the same thing as insertion sort algorithm. Also the other program is strange implementation of insertion sort algorithm. Most of the discovered sorting programs were implementation of insertion sort algorithm, but there were also some versions of bubble sort. In any case all of them had at least $O(n^2)$ complexity to sort n elements of an array. If we want to discover $O(n \times log(n))$ sort algorithm we gonna need procedure calls necessary for implementing recursion.

7 Results

Results of the experiments are presented in Table 1. Experiments consist of sets of runs compared to each other. Each set of runs consists of 20 runs, where a run uses 1000 individuals and processes them for up to 1000 generations. The significant conditions of the experiments are to the left of the vertical double line, and the results are to the right.

The principal metric for each set of 20 runs is the number of runs that were completely successful in evolving at least one individual that correctly sorted the fitness tests presented to it (removing 100% of the disorder in the sequences), and this is recorded in the column under # SUCC RUNS.

Additional information is presented as to how many runs produced at least one individual which removed 90% of the disorder in the fitness tests and 75% of the disorder in the fitness tests. The 90% and 75% categories are cumulative, in that each includes the count of runs that did better as well as the count of runs that made it past the 90% or 75% boundary but didn't reach the next highest. The column # GEN RUNS indicates for how many of the 100% successful runs passed all of the postrun generality tests.

AVG INDS records the average number of individuals that had to be processed for the 100% successful runs (averaged only over those runs that reached 100%). Thus, the number of 100% SUCC RUNS indicates how effectively the parameters used by a particular set of runs evolved a sort. The # GEN RUNS give some indication of the generality of the resulting sorts. The AVG INDS indicates how quickly the 100% SUCC RUNS reached 100%.

As we can see, when having more registers it was harder (slower) to find the solutions, because the search space was much bigger. In experiments where programs were 10 instructions long, it was harder finding the solution because the programs were to short and there is no space for redundancy. Also the general conclusion is that number of tests should be bigger, even if that takes longer to evaluate the individuals, but on aggregate the search time was shorter, because with clever evaluation the search procedure earlier removes partial solutions.

Table 1. Experimental Results

Set	# registers	# instructions	# tests	75%	90%	# SUCC	# GEN	AVG
	NR	*L*				**RUNS**	**RUNS**	**INDS**
A	5	10	15	18	15	0	0	/
B	5	10	30	20	16	6	6	2943893
C	5	20	15	20	20	19	10	4149037
D	5	20	30	20	18	17	16	1157621
E	10	10	15	5	3	0	0	/
F	10	10	30	7	3	0	0	/
G	10	20	15	11	8	2	0	6551290
H	10	20	30	12	10	5	0	2922396

8 Conclusion and Future Work

The process of evolving general sorting algorithms provided an interesting testbed for assessing the power of genetic programming on solving a challenging problem. In this paper we showed that evolving imperative programs, instead of tree programs, is possible and effective. This investigation has been another step towards allowing the evolution of more complex algorithms. In the future we plan to apply this methodology for evolving programs that even when they are executed probabilistically their functionality is unaffected.

References

1. Holland, J.H.: Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor (1975)
2. De Jong, K.A.: On Using Genetic Algorithms to Search Program Spaces. In: Grefenstette, J. (ed.) Proceedings of the 2nd International Conference on Genetic Algorithms. Lawrence Erlbaum Associates, Hillsdale (1987)
3. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, MA (1989)
4. Koza, J.R.: Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems, Technical Report No. STAN-CS-90-1314, Computer Science Department, Stanford University (1990)
5. Koza, J.R.: Genetic Programming. MIT Press, Cambridge (1992)
6. Kinnear, K.E.: Evolving a sort: Lessons in genetic programming. In: Proceedings of the 1993 International Conference on Neural Networks, vol. 2. IEEE Press, San Francisco (1993)
7. O'Reilly, U.-M., Oppacher, F.: An experimental perspective on genetic programming. In: Parallel Problem Solving from Nature 2 (1992)
8. O'Reilly, U.-M.: A comparative analysis of Genetic Programming. In: Advances in Genetic Programming 2. MIT Press, Cambridge (1996)
9. Knuth, D.E.: The art of computer programming, volume 3, sorting and searching. Addison Wesley Longman Publishing Co., Redwood City (1998)
10. Spector, L., Klein, J., Keijzer, M.: The push3 execution stack and the evolution of control. In: GECCO 2005: Proceedings of the 2005 Conference on Genetic and Evolutionary Computation. ACM Press, New York (2005)

Comparison of Evolving Fuzzy Systems with an Ensemble Approach to Predict from a Data Stream

Zbigniew Telec[1], Bogdan Trawiński[1], Tadeusz Lasota[2], and Krzysztof Trawiński[3]

[1] Wrocław University of Technology, Institute of Informatics,
WybrzeżeWyspiańskiego 27, 50-370 Wrocław, Poland
[2] Wrocław University of Environmental and Life Sciences, Dept. of Spatial Management
ul. Norwida 25/27, 50-375 Wrocław, Poland
[3] European Centre for Soft Computing, Edificio Científico-Tecnológico, 3ª Planta,
C. Gonzalo Gutiérrez Quirós S/N, 33600 Mieres, Asturias, Spain
{zbigniew.telec,bogdan.trawinski}@pwr.wroc.pl,
tadeusz.lasota@up.wroc.pl, krzysztof.trawinski@softcomputing.es

Abstract. An approach to apply ensembles of regression models, built over the chunks of a data stream, to aid in residential premises valuation was proposed. The approach consists in incremental expanding an ensemble by systematically generated models in the course of time. The output of aged component models produced for current data is updated according to a trend function reflecting the changes of premises prices since the moment of individual model generation. The method employing general linear model, multiple layer perceptron, and radial basis function networks was empirically compared with evolving fuzzy systems designed for incremental learning from data streams.The results showed thatevolving fuzzy systems and general linear models outperformed the ensembles built using artificial neural networks.

Keywords: neural networks, general linear models, evolving fuzzy systems, data stream, sliding windows, ensembles, trend functions, property valuation.

1 Introduction

We have been investigating techniques for developing an intelligent model of real estate market to aid in premises valuation for a few years. The general schema of the system devoted to a cloud computing platform is presented in Fig. 1. Cadastral systems and public registers contaning both descriptive and geospatial data constitue complex data sources for the intelligent system of real estate market. The core of the system are appraisal models constructed according to the professional standards as well as data-driven models generated using machine learning algorithms. At present, a broad spectrum of users are interested in the premises values and can be supported by computer systems called Automated Valuation Models (AVM) and Computer Assisted Mass Appraisal (CAMA). These systems incorporate models built based on statistical multiple regression [1], neural networks [2], rough set theory [3], and hybrid approaches [4],and the combination of soft computing and spatial analysis

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 377–387, 2013.

methods using data provided by geographic information systems (GIS) [5]. However, the need to provide meaningful and interpretable models resulted in the proposals of fuzzy and neuro-fuzzy systems [6], [7] as alternative solutions. The application of incremental online approaches could be especially useful because each day a cadastral information centre in a big city registers dozens of real estate sales transactions which are the base of sales comparison valuation methods, the most commonly used by the professionals. If all the data would be ordered by the transaction date, they would constitute some form of a data stream which in turn could reflect the changes of real estate market in the course of time. This motivated us also to use evolving fuzzy models, which are able to process streams of data and to learn, update, expand their memory on-line on demand. Up to date the authors of the paper have already proposed and studied evolving fuzzy real estate appraisal models [8], [9], [10] which revealed a good predictive accuracy.

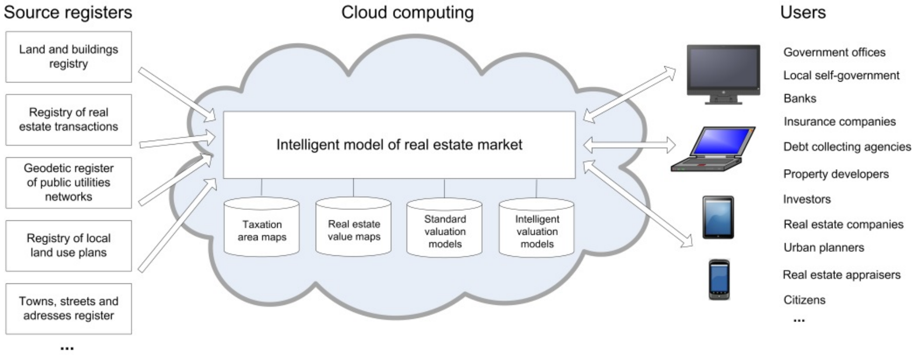

Source registers
Land and buildings registry
Registry of real estate transactions
Geodetic register of public utilities networks
Registry of local land use plans
Towns, streets and adresses register
...
Cloud computing
Intelligent model of real estate market
Taxation area maps
Real estate value maps
Standard valuation models
Intelligent valuation models
Users
Government offices
Local self-government
Banks
Insurance companies
Debt collecting agencies
Property developers
Investors
Real estate companies
Urban planners
Real estate appraisers
Citizens
...

Fig. 1. General schema of the intelligent system of real estate market

We have recently developed an approach to predict from a data stream of real estate sales transactions using ensembles of genetic fuzzy systems [11], [12], [13] and neural networks [14]. Working out the methods of processing data streams poses a considerable challenge because they require taking into account memory limitations, short processing times, and single scans of arriving data. Many strategies and techniques for mining data streams have been devised. Gaber in his recent overview paper categorizes them into four main groups: two-phase techniques, Hoeffding bound-based, symbolic approximation-based, and granularity-based ones [15]. Much effort is devoted to the issue of concept drift which occurs when data distributions and definitions of target classes change over time [16], [17], [18]. Among the instantly growing methods of handling concept drift in data streams Tsymbal distinguishes three basic approaches, namely instance selection, instance weighting, and ensemble learning [19]. The latter has been systematically overviewed in [20], [21]. In adaptive ensembles, component models are generated from sequential blocks of training instances. When a new block arrives, models are examined and then discarded or modified based on the results of the evaluation. Several methods have been proposed for that, e.g. accuracy weighted ensembles [22] and accuracy updated ensembles [23].

In [24], [25] Bifet et al. proposed two bagging methods to process concept drift in a data stream: ASHT Bagging using trees of different sizes, and ADWIN Bagging employing a change detector to decide when to discard underperforming ensemble members.

The goal of the paper is to compare in respect of predictive accuracy our method employing general linear model, multiple layer perceptron, and radial basis function networks with the *Flexfis* algorithm based onevolving fuzzy systems for incremental learning from data streams [26], [27]. The experiments were conducted in Matlab environment usingusing real-world data taken from a dynamically changing real estate market. We consider employing all above mentioned approaches in our prospect intelligent system of real estate market.

2 Ensemble Approach to Predict from a Data Stream

Our approach consists in the utilization of aged models to compose ensembles and correction of the output provided by component models by means of trend functions reflecting the changes of prices in the market over time. The outline of an approach to predict from a data stream using ensembles of regression models (*RM*) is depicted in Fig. 2. The data stream is partitioned into data chunks according to the periods of a constant length t_c. Each time interval determines the shift of a sliding time window which comprises training data to create *RM* models. The window is shifted step by step of a period t_s in the course of time. The length of the sliding window t_w is equal to the multiple of t_c so that $t_w=jt_c$, where j=1,2,3,... . The window determines the scope of training data to generate from scratch a property valuation model, in our case RM_i. It is assumed that the models generated over a given training dataset is valid for the next interval which specifies the scope for a test dataset. Similarly, the interval t_t which delineates a test dataset is equal to the multiple of t_c so that $t_t=kt_c$, where k=1,2,3,.. . The sliding window is shifted step by step of a period t_s in the course of time, and likewise, the interval t_s is equal to the multiple of t_c so that $t_s=lt_c$, where l=1,2,3,… .

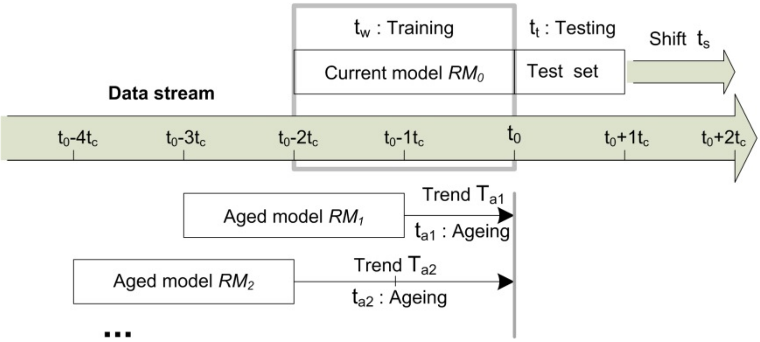


Fig. 2. Outline of an ensemble approach to predict from a data stream

We consider in Fig. 2 a point of time t_0 at which the current model RM_0 was built over data that came in between time t_0–$2t_c$ and t_0. The models created earlier, i.e. RM_1, RM_2, etc. have aged gradually and in consequence their accuracy has deteriorated. However, they are neither discarded nor restructured but utilized to compose an ensemble so that the current test dataset is applied to each component RM_i. However, in order to compensate ageing, their output produced for the current test dataset is updated using trend functions $T(t)$. As the functions to model the trends of price changes the polynomials of the degree from one to five were employed, denoted in the rest of the paper by *T1*, *T2*,..,*T5*, respectively. The trends were determined over two time periods: shorter and longer ones. The shorter periods encompassed the length of a sliding window plus model ageing intervals, i.e. t_w plus, t_{ai} for a given aged model RM_i. In turn, the longer periods took into account all data since the beginning of the stream. Hence, the shorter periods are denoted by *Age* and the longer periods are marked by *Beg* in the symbols of methods used in tables presenting the experimental results further on in the paper. In order to be concise, in remaining text of the paper we will call the former *Age Trends* and the latter *Beg Trends*.

Moreover, we proposed two different methods of updating the prices of premises according to the trends of the value changes over time. The first one based on the difference between a price and a trend value in a given time point and we called it the *Delta* method. In turn, the second technique utilized the ratio of the price to the trend value and it was named the *Ratio* method of price correction [12]. In the present paper we utilize only the former method.

3 Evolving Fuzzy System Method- Flexfis

The *Flexfis* method [27] incrementally evolves clusters (which are associated with rules) and performs a recursive adaptation of consequent parameters by using local learning approach. The *Flexfis*approach, short for FLEXible Fuzzy Inference Systems was first introduced in [26] and significantly extended version in [27], and designed for the purpose of incremental learning of Takagi-Sugeno fuzzy systems from data streams in a sample-wise single-pass manner. This means that always one sample can be loaded, sent into the *Flexfis*learning engine where the model is updated and immediately discarded, afterwards. In this sense, the method needs low resources 1.) with respect to computational complexity and 2.) with respect to virtual memory and hence is feasible for on-line modelling applications, where models should be kept-up-to-date as early as possible in order to account for new operating conditions, systems states etc. and to prevent extrapolation situations when performing predictions on new samples. The basic steps in *Flexfis*approach can be summarized as follows:

1. Rule evolution and updating of antecedent parameters in the cluster space with the help of an incremental evolving clustering variant.
2. Recursive adaptation of consequent parameters exploiting the local learning approach (parameters are updated for each rule separately).
3. Balancing out a non-optimal situation by adding a correction vector to the vector of consequent parameters.

4. In the extended version: Detecting redundant rules with the help of specific overlap measures (two variants: one-dimensional intersection points of fuzzy sets and inclusion metric) and performing on-line rule merging/pruning after each incremental update cycle.

4 Experimental Setup and Results

The experiments were conducted usingthe *Flexfis* application andour experimental system implemented in Matlab environment which was recently extended to include the functions of building ensembles over a data stream.The trends are modelled using the Matlab function *polyfit(x,y,n)*, which finds the coefficients of a polynomial *p(x)* of degree *n* that fits the *y* data by minimizing the sum of the squares of the deviations of the data from the model (least-squares fit).

Real-world dataset used in experiments was drawn from an unrefined dataset containing above 100 000 records referring to residential premises transactions accomplished in one Polish big city with the population of 640 000 within 14 years from 1998 to 2011. In this period the majority of transactions were made with non-market prices when the council was selling flats to their current tenants on preferential terms. First of all, transactional records referring to residential premises sold at market prices were selected. Then, the dataset was confined to sales transaction data of residential premises (apartments) where the land was leased on terms of perpetual usufruct. The other transactions of premises with the ownership of the land were omitted due to the conviction of professional appraisers stating that the land ownership and lease affect substantially the prices of apartments and therefore they should be used separately for sales comparison valuation methods. The final dataset counted 9795 samples. Due to the fact we possessed the exact date of each transaction we were able to order all instances in the dataset by time, so that it can be regarded as a data stream. Four following attributes were pointed out as main price drivers by professional appraisers: usable area of a flat (*Area*), age of a building construction (*Age*), number of storeys in the building (Storeys), the distance of the building from the city centre (*Centre*), in turn, price of premises (*Price*) was the output variable. In order to characterize quantitatively the data stream the sizes of one-year datasets are given in Table 1.

Table 1. Number of instances in one-year datasets

1998	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
446	646	554	626	573	790	774
2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
740	776	442	734	821	1296	577

Based on the results of our previous research [13] we were able to determine the following parameters of our experiments including two phases: *1. Generating single* *models* and 2. *Building ensembles*. As single models, general linear model (*Glm*), multilayer perceptron (*Mlp*) and radial basis function neural networks (*Rbf*) were built using *glm*, *mlp*, and *rbf*Matlab functions, respectively. In each function the number of inputs and outputs was set to four and one, respectively. In *glm*and *mlp*linear output unit activation functionsand in *rbf*a radially symmetric Gaussian function as hidden unit activation function were used. The functions*mlp*and *rbf*wererun with three neurons in a hidden layer. The number of epochs to learn each network was equal to 100. As the performance measure the root mean square error (*RMSE*) was used.

Phase 1. Generating single Glm, Mlp, and Rbf models

- Set the length of the sliding window to 12 months,$t_w = 12$.
- Set the starting point of the sliding window, i.e. its right edge, to 2000-01-01 and the terminating point to 2010-12-01.
- Set the shift of the sliding window to 1 month, $t_s = 1$.
- Move the window from starting point to terminating point with the step $t_s = 1$.
- At each stage generate a model from scratch over a training set delineated by the window. In total 108 models were built using each algorithm.

Phase 2. Building Glm, Mlp, and Rbfensembles

- Starting from the time point 2012-01-01 for each algorithm built ensembles composed of 24 ageing models each. The ensembles are created in the way described in Section 2 with the shift equal to one month, $\mathrm{t_s} =1$.
- Take test sets actual for each $\mathrm{t_0}$ over a period of 3 months, $\mathrm{t_t} =3$.
- Compute the output of individual models and update it using trend functions of degree from one to five determined for *Age Trends* and *Beg Trends*.
- As the aggregation function of ensembles use the arithmetic mean.
- Conduct statistical analysis of the results obtained.

In the case of *Flexfis* algorithm only single models were generated over chunks of data stream determined by a sliding window which was 12 months long. Analogously, test sets were determined over periods of three months.

The analysis of the results was performed using statistical methodology including nonparametric tests followed by post-hoc procedures designed especially for multiple *N*×*N* comparisons [28], [29], [30], [31]. The idea behind statistical methods applied to analyse the results of experiments was as follows. The commonly used paired tests i.e. parametric t-test and its nonparametric alternative Wilcoxon signed rank tests are not appropriate for multiple comparisons due to the so called family-wise error. The proposed routine starts with the nonparametric Friedman test, which detect the presence of differences among all algorithms compared. After the null-hypotheses have been rejected the post-hoc procedures should be applied in order to point out the particular pairs of algorithms which produce significant differences. For *N*×*N* comparisons nonparametric Nemenyi's, Holm's, Shaffer's, and Bergamnn-Hommel's procedures are recommended.

5 Statistical Analysis of Results

The goal of the statistical analysis was to compare the accuracy of the predictive models generated by *Flexfis* with ensembles composed of *Glm, Mlp,* and*Rbf* models which output was updated using trend functions *T1, T2, T3, T4,* and *T5* for*Age Trends* and *Beg Trends*. Additionally, the results not updated with any trend function were considered, they were denoted by *noT*. Nonparametric tests of statistical significance adequate for multiple comparisons were conducted for all years together. Within this period 108 values of *RMSE,* determined for each *RM* and trend function, *Flexfis*, and *noT* constituted the points of observation.

Table 2. Average rank positions of models produced by Friedman tests

	Glm		***Mlp***		***Rbf***	
No.	Trend	Rank	Trend	Rank	Trend	Rank
1.	AgeT2	4.98	Flexfis	1.30	Flexfis	1.30
2.	AgeT4	5.03	BegT4	5.39	BegT4	5.39
3.	AgeT5	5.09	BegT5	5.56	BegT5	5.55
4.	Flexfis	5.19	BegT3	5.69	BegT3	5.83
5.	AgeT3	5.23	AgeT4	6.44	AgeT4	6.41
6.	BegT4	5.51	AgeT5	6.44	AgeT5	6.42
7.	BegT5	5.60	AgeT2	6.46	AgeT2	6.45
8.	AgeT1	6.14	AgeT3	6.58	AgeT3	6.55
9.	BegT3	7.20	AgeT1	7.52	AgeT1	7.49
10.	BegT1	8.15	BegT2	8.00	BegT2	8.06
11.	BegT2	9.80	BegT1	8.16	BegT1	8.10
12.	noT	10.08	noT	10.46	noT	10.45

Table 3. Null hypotheses rejected by Schaffer's post-hoc procedure for *N*×*N* comparisons indicating statistically significant differences in performance between pairs of models.

	Glm	***Mlp***	***Rbf***
No.	Model vs Model	Model vs Model	Model vs Model
1.	FlexfisvsnoT	FlexfisvsnoT	FlexfisvsnoT
2.	Flexfisvs BegT3	Flexfisvs BegT5	Flexfisvs BegT5
3.	Flexfisvs BegT2	Flexfisvs BegT4	Flexfisvs BegT4
4.	Flexfisvs BegT1	Flexfisvs BegT3	Flexfisvs BegT3
5.	BegT5 vsnoT	Flexfisvs BegT2	Flexfisvs BegT2
6.	BegT5 vs BegT3	Flexfisvs BegT1	Flexfisvs BegT1
7.	BegT5 vs BegT2	Flexfisvs AgeT5	Flexfisvs AgeT5
8.	BegT5 vs BegT1	Flexfisvs AgeT4	Flexfisvs AgeT4
9.	BegT4 vsnoT	Flexfisvs AgeT3	Flexfisvs AgeT3
10.	BegT4 vs BegT3	Flexfisvs AgeT2	Flexfisvs AgeT2
11.	BegT4 vs BegT2	Flexfisvs AgeT1	Flexfisvs AgeT1
12.	BegT4 vs BegT1	BegT5 vsnoT	BegT5 vsnoT
13.	BegT3 vsnoT	BegT5 vs BegT2	BegT5 vs BegT2

Table 3. (*Continued*)

	Glm	***Mlp***	***Rbf***
No.	Model vs Model	Model vs Model	Model vs Model
14.	BegT3 vs BegT2	BegT5 vs BegT1	BegT5 vs BegT1
15.	BegT1 vsnoT	BegT5 vs AgeT1	BegT5 vs AgeT1
16.	BegT1 vs BegT2	BegT4 vsnoT	BegT4 vsnoT
17.	AgeT5 vsnoT	BegT4 vs BegT2	BegT4 vs BegT2
18.	AgeT5 vs BegT3	BegT4 vs BegT1	BegT4 vs BegT1
19.	AgeT5 vs BegT2	BegT4 vs AgeT1	BegT4 vs AgeT1
20.	AgeT5 vs BegT1	BegT3 vsnoT	BegT3 vsnoT
21.	AgeT4 vsnoT	BegT3 vs BegT2	BegT3 vs BegT2
22.	AgeT4 vs BegT3	BegT3 vs BegT1	BegT3 vs BegT1
23.	AgeT4 vs BegT2	BegT3 vs AgeT1	BegT3 vs AgeT1
24.	AgeT4 vs BegT1	BegT2 vsnoT	BegT2 vsnoT
25.	AgeT3 vsnoT	BegT1 vsnoT	BegT1 vsnoT
26.	AgeT3 vs BegT3	AgeT5 vsnoT	AgeT5 vsnoT
27.	AgeT3 vs BegT2	AgeT5 vs BegT2	AgeT5 vs BegT2
28.	AgeT3 vs BegT1	AgeT5 vs BegT1	AgeT5 vs BegT1
29.	AgeT2 vsnoT	AgeT4 vsnoT	AgeT4 vsnoT
30.	AgeT2 vs BegT3	AgeT4 vs BegT2	AgeT4 vs BegT2
31.	AgeT2 vs BegT2	AgeT4 vs BegT1	AgeT4 vs BegT1
32.	AgeT2 vs BegT1	AgeT3 vsnoT	AgeT3 vsnoT
33.	AgeT1 vsnoT	AgeT3 vs BegT1	AgeT3 vs BegT1
34.	AgeT1 vs BegT2	AgeT2 vsnoT	AgeT2 vsnoT
35.	AgeT1 vs BegT1	AgeT2 vs BegT1	AgeT2 vs BegT2
36.		AgeT1 vsnoT	AgeT2 vs BegT1
37.			AgeT1 vsnoT

The Friedman test performed in respect of *RMSE* values provided by the ensembles showed that there were significant differences among models. Average ranks of individual models for *Glm, Mlp,* and *Rbf* produced by the test are shown in Table 2, where the lower rank value the better model. However, please note, it does not mean that each model placed in the lower position significantly outperforms the ones ranked higher on the list. The post-hoc procedures for *N*×*N* comparisons should be applied to point out the pairs of models between which statistically significant differences occur.

The null hypotheses rejected by the most powerful Schaffer's post-hoc procedure for *N*×*N* comparisons are shown in Table 3. They indicate statistically significant differences in performance between pairs of models. The number of null hypotheses rejected by Shaffer's post-hoc procedure out of 66 is equal to 35, 36, and 37 for *Glm*, *Mlp*, and *Rbf*, respectively.

Following main observations can be done based on the results provided by the most powerful Shaffer's post-hoc procedure. *Flexfis* models outperform significantly all *Mlp* and *Rbf* based ensembles. Compared to *Gml* ensembles the *Flexfis* models hold position 4 determined by the Friedman tests. However, they are statistically equivalent to all preceding *Glm* ensembles. For all machine learning algorithms among *BegT4, BegT5, AgeT2, AgeT3, AgeT4, AgeT5* ensembles no significant difference is observed. In turn, *BegT1*, *BegT2*, and *noT* ensembles reveal significantly the worst performance. The performance of the *Mlp* and *Rbf* ensembles is similar and the rejected hypotheses are the same but one. Both methods differ from *Glm* with respect of their performance; the number of rejected hypotheses that are identical equals 20. Additional statistical analysis employing the Shaffer's post-hoc procedure showed that *Gml* ensembles surpass both *Mlp* and *Rbf* models in predictive accuracy.

6 Conclusions and Future Work

In the paper we reported our study of the method to predict from a data stream of real estate sales transactions based on ensembles of regression models. Our ensemble approach consists in incremental expanding an ensemble by models built from scratch over successive chunks of a data stream determined by a sliding window. The predicted prices of residential premises computed by aged component models for current data are updated according to trend functions which model the changes of the market. The method employing general linear model, multiple layer perceptron, and radial basis function networks was compared in terms of predictive accuracy with the *Flexfis* algorithm which is the implementation of evolving fuzzy systems.

The comparativeexperiments were carried out using real-world data taken from cadastral systems. They consisted in generating ensembles for 108 points of time and then comparing their predictive accuracy using nonparametric tests of statistical significance adequate for multiple comparisons. As the functions to model the trends of price changes, the polynomials of degree from one to five were employed. The trends were determined over two time periods: shorter (*AgeTrends*) and longer (*BegTrends*) ones. The method of correcting the output of component models was based on the difference between a predicted price and a trend value in a given time point.

Following main conclusions can be drawn from our study. *Flexfis* and *Glm* models outperform significantly all *Mlp* and *Rbf* based ensembles. There are not significant differences between *Flexfis* and the best *Glm*ensembles. The models with output corrected by linear trend functions and without correction revealed significantly worst performance. No clear differences among models corrected using *AgeTrends* and *BegTrends* could be observed.

The study opens the area for our further research into the selection of the best parameters of the proposed method. It includes the number of aged models encompassed by an ensemble as well as the selection of the degree of a trend function adequate for the dynamics of a given time period. Moreover, we plan to compare the outcome produced by proposed ensembles with human based predictions.

Acknowledgments. This paper was partially supported by the "MłodaKadra" funds of the Wrocław University of Technology and the Polish National Science Centre under grant no. N N516 483840.Many thanks also to Edwin Lughofer for granting us his FLEXFIS algorithm code, thereby allowing us to conduct the experiments.

References

1. Nguyen, N., Cripps, A.: Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. J. of Real Estate Research 22(3), 313 (2001)
2. Selim, H.: Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. Expert Systems with Applications 36, 2843–2852 (2009)
3. D'Amato, M.: Comparing Rough Set Theory with Multiple Regression Analysis as Automated Valuation Methodologies. Int. Real Estate Review 10(2), 42–65 (2007)
4. Kontrimas, V., Verikas, A.: The mass appraisal of the real estate by computational intelligence. Applied Soft Computing 11(1), 443–448 (2011)
5. García, N., Gámez, M., Alfaro, E.: ANN+GIS: An automated system for property valuation. Neurocomputing 71(4-6), 733–742 (2008)
6. González, M.A.S., Formoso, C.T.: Mass appraisal with genetic fuzzy rule-based systems. Property Management 24(1), 20–30 (2006)
7. Guan, J., Zurada, J., Levitan, A.S.: An Adaptive Neuro-Fuzzy Inference System Based Approach to Real Estate Property Assessment. J. of Real Estate Res. 30(4), 395–421 (2008)
8. Lughofer, E., Trawiński, B., Trawiński, K., Kempa, O., Lasota, T.: On Employing Fuzzy Modeling Algorithms for the Valuation of Residential Premises. Information Sciences 181, 5123–5142 (2011)
9. Lasota, T., Telec, Z., Trawiński, B., Trawiński, K.: Investigation of the eTS Evolving Fuzzy Systems Applied to Real Estate Appraisal. Journal of Multiple-Valued Logic and Soft Computing 17(2-3), 229–253 (2011)
10. Trawiński, B., Trawiński, K., Lughofer, E., Lasota, T.: Investigation of Evolving Fuzzy Systems Methods FLEXFIS and eTS on Predicting Residential Prices. In: Petrosino, A. (ed.) WILF 2011. LNCS (LNAI), vol. 6857, pp. 123–130. Springer, Heidelberg (2011)
11. Trawiński, B., Lasota, T., Smętek, M., Trawiński, G.: An Attempt to Employ Genetic Fuzzy Systems to Predict from a Data Stream of Premises Transactions. In: Hüllermeier, E., Link, S., Fober, T., Seeger, B. (eds.) SUM 2012. LNCS (LNAI), vol. 7520, pp. 127–140. Springer, Heidelberg (2012)
12. Trawiński, B.: Evolutionary fuzzy system ensemble approach to model real estate market based on data stream exploration. J. Univers. Comput. Sci. 19(4), 539–562 (2013)
13. Trawiński, B., Lasota, T., Smętek, M., Trawiński, G.: Weighting Component Models by Predicting from Data Streams Using Ensembles of Genetic Fuzzy Systems. Accepted for the Tenth International Conference on Flexible Query Answering Systems, FQAS 2013, Granada, Spain, September 18-20 (2013)
14. Telec, Z., Lasota, T., Trawiński, B., Trawiński, G.: An Analysis of Change Trends by Predicting from a Data Stream Using Neural Networks. Accepted for the Tenth International Conference on Flexible Query Answering Systems, FQAS 2013, Granada, Spain, September 18-20 (2013)
15. Gaber, M.M.: Advances in data stream mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(1), 79–85 (2012)
16. Elwell, R., Polikar, R.: Incremental Learning of Concept Drift in Nonstationary Environments. IEEE Transactions on Neural Networks 22(10), 1517–1531 (2011)

17. Maloof, M.A., Michalski, R.S.: Incremental learning with partial instance memory. Artificial Intelligence 154(1-2), 95–126 (2004)
18. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine Learning 23, 69–101 (1996)
19. Tsymbal, A.: The problem of concept drift: Definitions and related work. Technical Report. Department of Computer Science, Trinity College, Dublin (2004)
20. Kuncheva, L.I.: Classifier Ensembles for Changing Environments. In: Roli, F., Kittler, J., Windeatt, T. (eds.) MCS 2004. LNCS, vol. 3077, pp. 1–15. Springer, Heidelberg (2004)
21. Minku, L.L., White, A.P., Yao, X.: The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift. IEEE Transactions on Knowledge and Data Engineering 22(5), 730–742 (2010)
22. Wang, H., Fan, W., Yu, P.S., Han, J.: Mining concept-drifting data streams using ensemble classifiers. In: Getoor, L., et al. (eds.) KDD 2003, pp. 226–235. ACM Press, New York (2003)
23. Brzeziński, D., Stefanowski, J.: Accuracy Updated Ensemble for Data Streams with Concept Drift. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part II. LNCS (LNAI), vol. 6679, pp. 155–163. Springer, Heidelberg (2011)
24. Bifet, A., Holmes, G., Pfahringer, B., Gavaldà, R.: Improving Adaptive Bagging Methods for Evolving Data Streams. In: Zhou, Z.-H., Washio, T. (eds.) ACML 2009. LNCS (LNAI), vol. 5828, pp. 23–37. Springer, Heidelberg (2009)
25. Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., Gavalda, R.: New ensemble methods for evolving data streams. In: Elder IV, J.F., et al. (eds.) KDD 2009, pp. 139–148. ACM Press, New York (2009)
26. Lughofer, E., Klement, E.P.: FLEXFIS: A variant for incremental learning of Takagi-Sugeno fuzzy systems. In: Proc. of FUZZ-IEEE 2005, Reno, USA, pp. 915–920 (2005)
27. Lughofer, E.: FLEXFIS: A robust incremental learning approach for evolving TS fuzzy models. IEEE Transactions on Fuzzy Systems 16(6), 1393–1410 (2008)
28. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
29. García, S., Herrera, F.: An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)
30. Graczyk, M., Lasota, T., Telec, Z., Trawiński, B.: Nonparametric Statistical Analysis of Machine Learning Algorithms for Regression Problems. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part I. LNCS (LNAI), vol. 6276, pp. 111–120. Springer, Heidelberg (2010)
31. Trawiński, B., Smętek, M., Telec, Z., Lasota, T.: Nonparametric Statistical Analysis for Multiple Comparison of Machine Learning Regression Algorithms. International Journal of Applied Mathematics and Computer Science 22(4), 867–881 (2012)

Exploding Balls Principle as a Method of Global Optimization

Aleksander Nowodzinski, Leszek Koszalka,
Iwona Pozniak-Koszalka, and Andrzej Kasprzak

Department of Systems and Computer Networks
Wroclaw University of Technology
Wroclaw, Poland
aleksander@nowodzinski.pl, {leszek.koszalka,iwona.pozniak-koszalka,andrzej.kasprzak}@pwr.wroc.p

Abstract. This paper reflects on the alternative methodology for finding global minima of one-dimensional cost-functions by using Exploding Balls Principle. In this paper, the simulation scenario is presented which effectively compares the new procedure with existing methods such as Simulated Annealing and Genetic Algorithm. Finally, the results of the experiments are presented and summarized with necessary conclusions and new ideas for further research. Additionally, the software simulation environment (EbSim) is introduced with an emphasis on the features and the opportunities it provides.

Keywords*:* global optimization, algorithm, heuristics, exploding balls principle**.**

1 Introduction

Numerical optimization is a branch of the mathematics which focuses onsearching for optimal solutions and making correct decisions.Finding global minimumof a cost-function is a common task of the numerical optimization. It provides solutions for many real-life problems related to physics, mechanics, economics, electronics, communication and many others. Despite the fact it is a very common problem; no simple methodology has been found which would solve all the possible cases. It is mostly due to the diversity of functions which may be investigated. Usually the algorithms for finding global minimumare chosen and fine-tuned for a particular problem instance so there is no way to keep the performance level high if the cost-function (problem) changes.

The more complex is the cost-function, the harder is to solve the optimization problem it drives. There are three main approaches which can deal with this kind of problem [1], [2]. The first one is the brute-force (BF) approach. The problem is that this method is useless when the solution-space is wide as we usually expect the performance to be fair. Another method is the gradient-based approach which is good for regular cost-functions with distinct minimum. Gradient methods do not converge efficiently for irregular functions thought. Some basic knowledge such as gradient

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 388–398, 2013.

(or Hessian) of the function makes the whole process much easier and faster but never guarantees finding an optimal solution.

The third way of solving global minimum problem is called heuristic approach. In general, heuristic algorithms trade off accuracy for speed. They offer polynomial bounds of the computational time atthe expense of the quality of the solution [3]. Although there are many meta-heuristics capable of solving different optimization problems, researchers are still looking for new, robust procedures [4] which would be faster and smarter. The aim of this paper is to focus on the new Exploding Balls Principle algorithm which is to be compared with other commonly used algorithms such as Simulated Annealing [5], [6],and Genetic Algorithm[7].

The paper is organized as follows: Section 2contains the mathematical formulation of the considered optimization problem. In Section 3Exploding Balls Principle is presented and explained. Section 4 covers briefly both Simulated Annealing and Genetic Algorithmalgorithms. Section 5 is a short overview of the created experimentation system*EbSim*which is used during the simulation experiments,and describes the setup of experiments, including datasets and quality measures. Section 6 is devoted to experiments, results and observations. In Section 7 the basic conclusions regarding the experiments are drawn, and plans for further research are shortly presented.

2 Problem Statement

The problem of finding global minima requires some basic terms to be clearly defined in order to avoid ambiguities. The examined function*f*is assumed to be defined by (1).

$$\boldsymbol{I} \in \boldsymbol{R} \qquad \boldsymbol{f} : \boldsymbol{I} \rightarrow \boldsymbol{R}(1)$$

Moreover, the function *f*has to be everywhere continuousand differentiable (2).

$$\{x,a\} \in I \ \lim_{x \to a} f(x) = f(a) \ \forall_{a \in I} f'(a) = \lim_{h \to 0} \frac{f(a+h)-f(a)}{h} \tag{2}$$

Expression(1) restricts the one-dimensional function f to the set I which determines a specific range in which global minima are located. Set I is chosen arbitrarily so the process of finding extremes is finite and fully controlled. Restrictions (2) let us avoid pathologically irregular functions. It may happen that the global minimum is defined by the fractal-like process [1], which introduces many difficulties and it is obviously undesirable.If the assumptions above are satisfied, the global minimum x^* within *I,* and the approximated solution denoted by $\hat{x}$, are defined by (3).

$$[x_1^*, x_2^*, ..., x_n^*] = \arg\min_{x \in I}(f(x)) \ \hat{x} \to x_k^* \quad k \in (1,n) \quad n \in N \tag{3}$$

A simple problem instance is presented in Fig. 1. Global optimization can be performed both on continuous and discrete solution spaces.The discrete approach is however much simpler to implement and it solvessome problems automatically.Discretization eliminates the problem of discontinuity and differentiability. Discrete solution-spaces can be also easily optimized, provided the cost-function is real and finite. Moreover, in most cases the real-life data are collected in a form of discrete computer-friendly datasets, e.g., CSV and XML files that can be immediately analyzed.

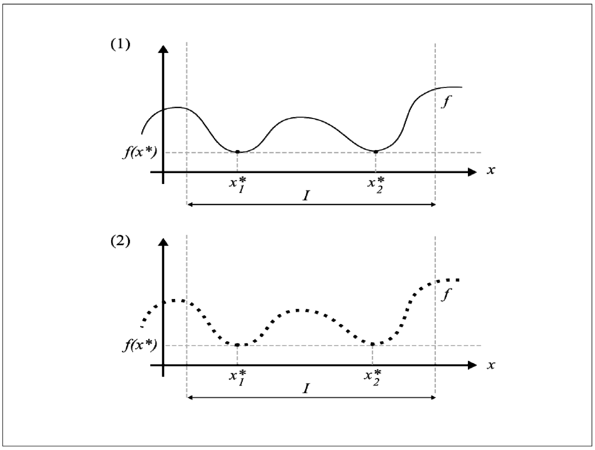


Fig. 1. The analysis of a sample solution space determined by the cost-function f(x) within subset I. Optimization of a continuous process – up (1).Optimization of a discrete process - down (2).

In this paper the discrete model of a solution space is default.The input-output model of the problem looks as follows:

INPUT - A finite set Y of the values (samples) of the discrete cost-function*f(x)* within the subset *I*is given by (4)

$$Y = [y_1, y_2, ... y_n] \quad f(x_i) = y_i \quad i, n \in N \; n = |I| \geq 1 \quad 1 \leq i \leq n \tag{4}$$

OUTPUT- A finite set of the global minima of the cost-function *f(x)*. The number of the minima cannot exceed the number of samples (5):

$$X^* = [x_1^*, x_2^*, ..., x_m^*] = \arg\min_{x \in I}(f(x)) \; m \in N \quad m \leq n \tag{5}$$

A global optimum of the function y^* is expressed by (6).

$$y^* = f(x_1^*) = f(x_2^*) = f(x_3^*) = ... = f(x_m^*) \tag{6}$$

3 Exploding Ball Principle

Exploding Balls Principle (EBP) is a hybrid, heuristic algorithm. Itcombines basic features of the population-oriented methods with the power of the gradient-based approach. The original idea of the procedure has been proposed by Daniel Davies [8].

The general idea of EBP is to explore the solution-space gradually (and randomly) with a number of balls that rolldown. The rolling procedure is similar to the approach used in gradient methods; however, theprocess of ball distribution draws from the genetic approach. The idea is thata finite number of balls can be distributed throughout the solution space. Eachball rolls down towards some local optimum according to the local gradientof the explored process. Once it reaches the local optimum, it explodes andcreates a new generation of balls which is dispersed in the vicinity. Process repeats iteratively until some stop-condition is satisfied. (Fig. 2).

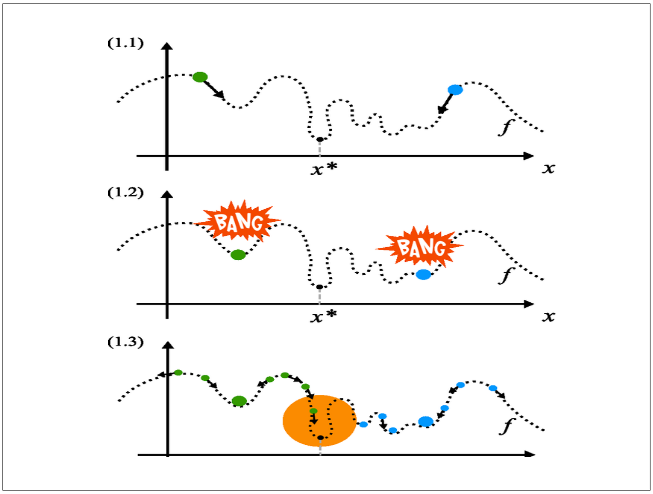


Fig. 2. The principle behind EBP. Two ballsexplore the solution space (1.1). They reach some local minima. Once theystop the explosion is initialized (1.2). The next generation explores the solution space and, hopefully, it reaches some better solution (1.3).

With every iteration (rolling step), each ball holds the minimal valueof the solution space that has been found so far. The final result of theprocedure is a value stored by the ball that reached the deepest ravines ofthe solution space.What makes the EBPinteresting is that thesubsequent generations of the balls explore the solution space around theirparents with a higher precision. This increases the probability of finding theoptimum, as each generation derives from its predecessor. Unlike SimulatedAnnealing, the algorithm does not tend to get stuck in local optimum sinceit explores the whole solution space.The pseudo-code of the algorithm - presented in Fig. 3 - corresponds with the implementation provided by *EbSim* environment (see: Section 5).The crucial part of the procedure is a moment of explosion. With eachexplosion, a number of children are dropped like pieces of a real explosive.Thanks to that randomness, the algorithm is able to explore some parts ofthe solution-space that cannot be easily accessed. This is a great advantage assome cost-functions may be really unpredictable in terms of the global optimum location.

Step 1. Initialize a first generation on balls according to the algorithm parameters.
Step 2. Let the balls roll down the function until explosion criterion is fulfilled. Each ball holds the value of the function. The best position is determined by the ball which is the lowest.
Step 3. Perform series of explosions in which a new generation of balls is created around the existing ones. Use specific distribution to control their position.
Step 4. Repeat steps 2 and 3 until the stop condition is fulfilled.

Fig. 3. EBP algorithm, procedure explained step-by-step

There are several parameters which determine the EBP algorithm allowing fine-tuning:

- **Initial ball density**(EBP$_{\mathrm{idens}}$) - A measure of how many balls are distributed throughout the solution space at the very beginning of the procedure run.
- **Initial ball radius**(EBP$_{\mathrm{irad}}$) - An initial radius of the balls (1stgeneration).
- **Ball radius reduction factor**(EBP$_{\mathrm{brf}}$— › *A* factor which controls howthe ball radius is reduced between the generations.
- **Maximal number of generations**(EBP$_{\mathrm{geners}}$) - A stop condition parameter.
- **Number of offspring**(EBP$_{\mathrm{noff}}$) - A factor that determines how manychild balls are produced during each explosion.
- **Ball distribution**(EBP$_{\mathrm{bdist}}$) - Two types of distribution are considered: uniform and normal. This parameter allows choosingwhich ball distribution is used in the moment of the explosion.
- **Explosion radius std. deviation factor** (EBP$_{\mathrm{sigma}}$) - A parameter that determines the standard deviation of the normal distribution $\mathrm{N}(\mu,\sigma^2)$.

4 Known Heuristics

Although this paper is mostly focused on EBP, two different meta-heuristics are also considered as a reference for further benchmarking and final comparison. We implemented Simulated Annealing and Genetic Algorithm. They are very powerful methods with a wide range of applications, e.g., [7].

Simulated Annealing(SA) is a meta-heuristic algorithm based on the analogy ofthe annealing in metallurgy. The algorithm directly refers to the coolingprocess of a metal which is eventually frozen into a crystalline structure ofthe lowest energy. The state of the lowest energy is, basically, considered theglobal optimum [4], [5], [6]. SA follows a particle during the annealing process, jumping from one energy state to another. The procedure expects the particle to fluctuate between differentenergy states but the fluctuations are damped with time. The algorithm is terminated if the temperature reaches some level or a fixednumber of iterations have passed. The optimum is the lowest energy state the particle went through.Our implementation of SAcontains parametersthat drive the algorithm's behavior:

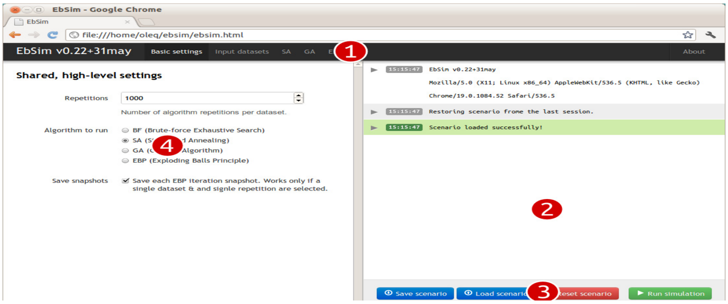
- **Improvement stop criterion**(SA_{impr}) -Determines the number ofiterations without any improvement.
- **Iteration stop condition**(SA_{iters}) - Determines the number of iterations that the algorithm is allowed to run.
- **Initial temperature** (SA_{itemp}) - Determines the initial temperature.
- **Temperature loss factor** (SA_{tloss}) - Controls the cool-down process. It determines the relationship between temperaturesin subsequent iterations.
- **Acceptance level** (SA_{pwrong}) - The probability of the worse solution tobe accepted.

Genetic Algorithm(GA) is based on evolutionary biology, there are many variations ofgenetic approach that share the same idea but differ in details [7]. GA assumes that populations of living organisms change across successive generations because ofeither the pressure of the environment or unpredictable fluctuations. Biological evolution is mostly driven by the natural selection, which is a slow,non-random process of eliminating weak individuals while the best adaptedsurvive.The procedure of an algorithm begins with a number of randomly created solutions. Each individual (solution) is encoded as a simple, numerical(mostly binary) pattern called a chromosome. Some of the solutions are, ofcourse, better than the others (in terms of the cost-function). They are more likely to reproduce. The nextstep, when the subset of good solutions is randomly split into pairs, is calledthe cross-over.Each pair produces a descendant that partially (but still randomly) inherits from each parent. Once a descendant is made, the mutation may occur. The stop condition isusually a number of generations or a number of successive iterations withoutany improvement. The global optimum is the individual (solution) which is the best adapted one. In this paper, there are implemented control parameters which drive the algorithm:

- **Target population** (GA_{pop}) - Determines the number of individuals tobe held in the population.
- **Iteration stop condition** (GA_{iters}) - Determines the number of iterations that the algorithm is allowed to run.
- **Max. Number of generations** (GA_{gener}) - Determines the numberof generations the algorithm is allowed to create.
- **Probability of mutation** (GA_{pmut}) - Determines the probability of arandom change in a chromosome.

5 Experimentation System

***EbSim* Environment.** To obtain reliable results, a powerful simulation environment,calledEbSim,has been designed and implemented. It allows for EBP algorithmtesting directly in a web browser (Fig. 4)providing cross-platform compatibility.The motivation for development of such software is the fact that mostof the simulators are Operating System-dependent, so they require a specific system setup.



Shared, high-level settings

Repetitions 1000

Number of algorithm repetitions per dataset.

Algorithm to run
○ BF (Brute-force Exhaustive Search)
● SA (S d Annealing)
○ GA (Algorithm)
○ EBP (Exploding Balls Principle)

Save snapshots ☑ Save each EBP iteration snapshot. Works only if a single dataset & and signle repetition are selected.

```
15:15:47 EbSim v0.22+31may
         Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/536.5 (KHTML, like Gecko)
         Chrome/19.0.1084.52 Safari/536.5
15:15:47 Restoring scenario frome the last session.
15:15:47 Scenario loaded successfully!
```

Fig. 4. EbSim user interface with Console window (2), Scenario controls (3),Toggle able settings area (4) -controlled by the main menu (1)

The created software is a robust, thread-based and research-oriented utility. There are a number of features available:(i) scenario support, (ii) session support, (iii) batch processing, (iv) dataset management, (v) different heuristic algorithms –EBF, SA, GA; and BF (Brute Force), (vi) user Interface. It works off-line with all modern web browsers. Everything is configured stating that *EbSim* is open-source software.

Experiment Design. The series of experiments should benchmark EBP and determine its pros and cons, while SA and GA have been used as a reference. The selected datasets differ in CV (coefficient of variation), length, and location of the minima not to be biased (Table 1).

Table 1. Datasets used in experiments

Name	CV	Length	Minimum
Add10 dataset: 1	0.57	9792	1
Add10 dataset: 2	0.57	9792	1
Add10 dataset: 3	0.57	9792	1
UCI adult: hours-per-week	0.30	48842	27
UCI adult: age	0.35	48842	595
UCI adult: fnlwgt	0.55	48842	1
Hwang dataset: 3	0.25	27200	1
Hwang dataset: 4	0.47	27200	1
UCI: Japanese Vowels	20.81	51288	1
UCI: El Nino s.s.temp.	0.78	160805	1
UCI: El Nino air temp	0.099	159835	1

Quality Measures. *EbSim* supports six different indices which have to be gathered for further analysis. However, to introduce any of them, a few terms need to be defined:

Y - A set of datasets to be explored.
y - A single dataset . $y \in Y$.
$|y|$ - A length of a single dataset.
$\bar{y}$ - A mean value of a single dataset.
s - A single simulation run *s*.
S - A set of repetitions to be performed.
$x_{s,y}$ - An argument for the optimum of the dataset *y* found during the run *s*.
x_y^* - An argument for the optimum of the dataset *y* (if many with the same value, then the closest to $x_{s,y}$ is taken).
$t_{s,y}$ - An execution time of a single run *s* withdataset *y*.

Considering the variables defined above, there are two key indices which have been extensively used in this paper due to their relevance:

Average Execution Time(AET in *ms*)defined by *(*7) as the average time needed to find the result. It isexpressed by the following formula:

$$AET = \frac{1}{|Y| \cdot |S|} \sum_{s \in S} \sum_{y \in Y} t_{s,y} \quad ms \tag{7}$$

Normalized average solution-optimum difference(NADO$_{\text{diff}}$ in %) defined by (8) as the difference between the real optimum and the approximated one. It is relative to the mean value of the dataset so it can be used across different datasets :

$$NASO_{diff} = \frac{1}{|Y| \cdot |S|} \sum_{s \in S} \sum_{y \in Y} \frac{|y(x_{s,y}) - y(x_y^*)|}{\bar{y}} \% \tag{8}$$

These indices are important as they clearly determine both the accuracy and the performance of the algorithm. Those two factors are crucial for drawing conclusion.

6 Investigation

Many different experiments have been performed during the research. Due to the space limitations only the most significant are included in this section.

Experiment #1. Initial Ball Density Parameter Study. EBP$_{\text{idens}}$ is responsible for the initial distribution of balls. It has a large impact on the accuracy of the entire procedure. The relationship is logarithmic and it indicates that low values of EBP$_{\text{idens}}$ ensure the best quality of the solution (Fig. 5).

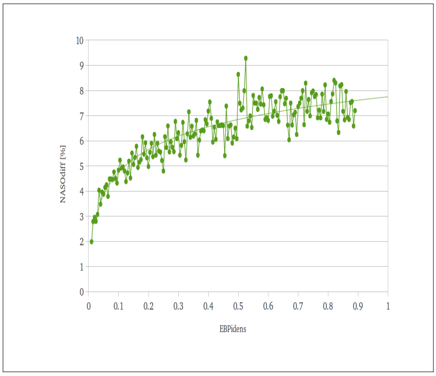


Fig. 5. NASOdif $_{f}$ with respect to EBP$_{\text{idens}}$

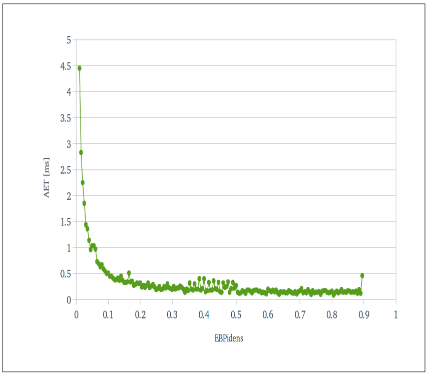


Fig. 6. AET with respect to EBP$_{\text{idens}}$

However, on the other side, low EBP$_{\text{idens}}$significantly degrade the performance of the algorithm (Fig. 6). Thus, the most beneficial rangeof values is ca. $EBP_{idens} \in [0.1,0.2]$.

Experiment #2. **Number of Offspring Control Parameter Study.** A greater number of offspring significantly increases AET (Fig. 7). It is mostly because EBP$_{\text{noff}}$ has a direct impact on the overall number of balls participating in the optimization. As each ball rolls separately, the more balls are involved, the lower the performance.On the other side the more offspring is created, the higher the quality of the solution NASO$_{\text{diff}}$ (Fig. 8).Thus, the value of EBP$_{\text{noff}}$ should be based on the golden mean between AET and NASO$_{\text{diff}}$ if we do not want to degrade the accuracy. Considering that fact, it may be observed, that $EBP_{noff} \in [5,8]$ is satisfactory.

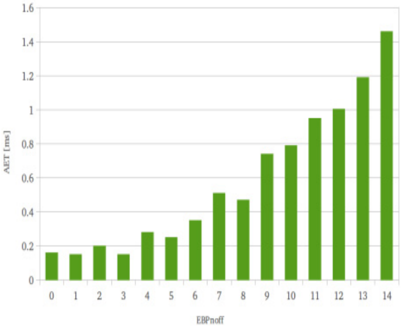


Fig. 7. AET with respect to EBP$_{\text{noff}}$

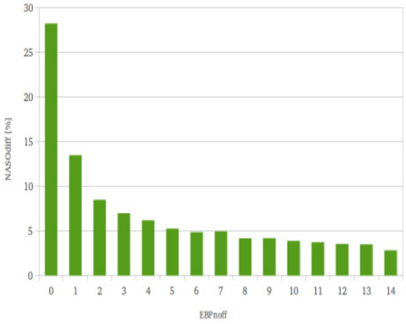


Fig. 8. NASOdif $_{f}$ with respect to EBP$_{\text{noff}}$

Experiment #3. Exploding Balls Principle vs. Other Algorithms. The goal was the comparison of different heuristics.The parameters of the algorithms were tuned according to manual tests and results of previous experiments. We observed the performance factors AET and NASO$_{\text{diff}}$ of each algorithm with respect to |Y|.Each time the |Y| changed, the dataset was randomly scrambled and sliced. The number of repetitions within a single value of |Y| was 150. The experiment was performed for |Y| = [1000,2000,…, 149000], which gave 22200 individual runs per algorithm.

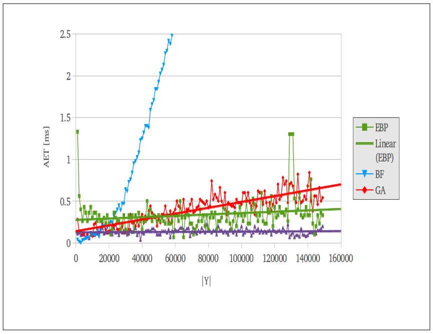


Fig. 9. Comparison of algorithmsin terms of AET. Brute-force (BF)is out of scope (grows too fast).

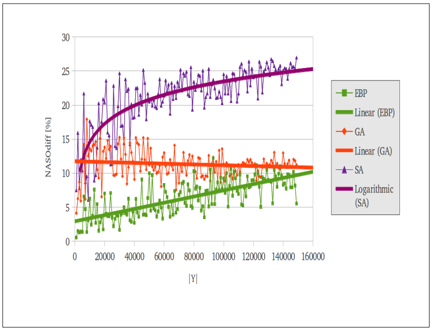


Fig. 10. Comparison of algorithmsin terms of AET. Brute-force (BF)is out of scope (grows too fast).

Fig. 9 shows that EBP is generally comparable to other algorithms in terms of AET. The relationship is linear for both EBP and SA which proves stability and reliability. Unlike GA and BF, there is no degradation of the performance of EBPand SA with the size of the dataset. The relationship between the size of the dataset and the quality of solution is illustrated in Fig. 10. The index NASO$_{\text{diff}}$ fits between 0.5% and 27% for all the algorithms except BF, whichalways finds the exact solution. According to the linear and logarithmic approximation models, EBP seems to be the most accurate algorithm for |Y| less than 170000 samples. However, the accuracy degrades linearly with the size of the solution-space. GA, on the other hand is the most stable algorithm with the lowest variance, and the mean value of NASO$_{\text{diff}}$ = 11.29%. The worst algorithm in this comparison is SA, of which accuracy degrades logarithmically with the highest variance amongthe competitors.

7 Conclusions

EBP seems to be a really promising algorithm which has been proved during the research. EBF provides the best accuracy in a narrow range of the solutionspace. However, more experiments should be made following idea of multistage designing experiments [9], and comparing results produced by EBF with other evolutionary algorithms [10].We also plan to develop the EBF implementing the following ideas: (*i*) *Adaptive EBP*.The parameters of EBF could be adaptively adjusted: by using asymmetric distributions when converging towards some "valley" and the uniform one when the "terrain is flat" or by changing the produced number of balls due to complexity of the solution-space (i.e. variation in vicinity),(*ii*) *Multi-layer heuristics.*That means creating heuristics (i.e. an appropriate GA) to estimate the control parameters of EBP and next fine-tuningby adjusting parametersof EBP which depend on each other, (*iii*) *Advanced physics.* There are physical factors that could be implemented, i.e., mass, momentum and flexibility - balls could bounce and collide like real particles. They also could speed up when going downhill towards some ravine, and slow down when rolling uphill.

Acknowledgement. This work was supported by the statutory funds of Dept. of Systems and Computer Networks, Faculty of Electronics, Wroclaw University of Technology, No S20010W4.

References

1. Mandelbrot, B.B.: The Fractal Geometry of Nature. W. H. Freeman and Co. (1982)
2. Mallat, S., Wavelet, A.: Tour of Signal Processing. Academic Press (1999)
3. Talbi, E.G.: Metaheuristics: From Design to Implementation. Wiley Series on Parallel and Distributed Computing, Wiley (2009)
4. Reeves, C.R.: Modern Heuristic Techniques for Combinatorial Problems. McGraw-Hill (1995)
5. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220(4598), 671–680 (1983)
6. Bertsimas, D., Nohadani, O.: Robust optimization with simulated annealing. Journal of Global Optimization 48(2), 323–334 (2010)
7. Wong, K.-C., Wu, C.-H., Mok, R.K.P., Peng, C., Zhang, Z.: Evolutionary multimodal optimization using the principle of locality. Information Sciences 194(1), 138–170 (2012)
8. Davies, D.: A methodology for finding global minima: an exploding balls principle. Research Proposal, Wroclaw University of Technology (October 2010)
9. Zydek, D., Selvaraj, H., Koszalka, L., Pozniak-Koszalka, I.: Evaluation scheme for NoC-based CMP with integrated processor management system. International Journal of Electronics and Telecommunications 56(2), 157–167 (2010)
10. Kakol, A., Pozniak-Koszalka, I., Koszalka, L., Kasprzak, A., Burnham, K.J.: An experimentation system for testing bee behavior based algorithm to solving a transportation problem. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS, vol. 6592, pp. 11–20. Springer, Heidelberg (2011)

Comparison of Reinforcement and Supervised Learning Methods in Farmer-Pest Problem with Delayed Rewards

Bartłomiej Śnieżyński

AGH University of Science and Technology
Faculty of Computer Science, Electronics and Telecommunications
Department of Computer Science
Al. Mickiewicza 30, 30-059 Kraków, Poland
bartlomiej.sniezynski@agh.edu.pl

Abstract. In this paper we propose a method based on the time-window idea which allows agents to generate their strategy using supervised learning algorithms in environments with delayed rewards. It is universal and can be used in various environments. Learning speed of the proposed method and reinforcement learning algorithm are compared in a Farmer-Pest problem with delayed rewards. Farmer-Pest problem is chosen for the comparison because it is designed especially for learning algorithms benchmarking. It has several dimensions which change environment characteristics and allows to test algorithms in various conditions. This paper presents results for one reinforcement learning method (SARSA) and three supervised learning algorithms (Naïve Bayes, C4.5 and Ripper). These algorithms are tested on configurations with various complexity.

Keywords: agent learning, supervised learning, reinforcement learning.

1 Introduction

The problem of learning naturally appears in multi-agent systems [11] which are efficient architectures for decentralized problem solving. In complex or changing environments it is very difficult, sometimes even impossible, to design all system details a priori. To overcome this problem one can apply a learning algorithm which allows to adapt the system to the environment. To apply learning in a multi-agent system, one should choose a method of learning, which fits well to the problem. There are many algorithms developed so far. However, in multi-agent systems most applications use reinforcement learning or evolutionary computations [7]. These methods are relatively slow and the knowledge learned is difficult to analyze by humans.

The goal of this research is to show that supervised learning algorithms may be applied by agents to generate their strategy in an environment with delayed rewards. We also compare learning speed of supervised and reinforcement learning algorithms. For this purpose we use a Farmer-Pest problem, a learning environment designed especially for benchmarking. Reinforcement learning

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 399–408, 2013.

is designed to work well in environments with delayed rewards. For supervised learning it is not straightforward. In this paper we propose a solution based on a time-window approach to take into account several time stamps. It makes the search space larger, but still supervised learning algorithms achieve good results faster than reinforcement learning. The other advantage is that the knowledge learned is readable. This paper is an continuation of [15], where comparison of reinforcement learning algorithm (SARSA) and three supervised learning algorithms (Naïve Bayes, C4.5 and Ripper) was performed in the environment without delays. The same learning algorithms are compared here.

In our research, we make the following contributions to the state of the art: we propose a method how supervised learning may be applied to generate agent strategy in environment with delayed rewards, we present a multi-dimensional domain allowing comparison of leaning algorithms; we show that methods other than reinforcement learning can be used for strategy generation; we compare learning algorithms in configurations with various complexity and show that supervised learning can improve performance of agents much faster that reinforcement learning.

In the following sections we overview learning in multiagent systems, present how supervised learning may be applied to problems with delayed rewords, and describe the proposed problem domain. Next, experimental results are described. Finally, conclusions and the further work are outlined.

2 Learning Agents

Good survey of learning in multi-agent systems working in various domains can be found in [7] and [11]. Experiments are performed in various environments, using many kinds of learning algorithms. Below we demonstrate variety of these works using three example domains.

Very popular in multi-agent systems is a soccer domain. The environment consist of a soccer field with two goals and a ball. Two teams of simulated or real robots are controlled by the agents. The performance is measured by the difference of scored goals. In [6] genetic programming is utilized to learn behavior-based team coordination. In [10] C4.5 algorithm is used to classify the current opponent into predefined adversary classes. Reinforcement learning can be also applied. Good example is [16], where a "keep away" version of the domain is used. Agents learn when to hold the ball an when to pass it. It appears that this domain is difficult for the learning algorithms. The state space is large, uncertainty and long delays of action effects should be taken into account.

One of environments, which is often used in research is Predator-Prey domain. It is a simple simulation with two types of agents: predators, and preys. The aim of a predator is to hunt for a prey. Prey is captured if predator (or several predators if cooperation is tested) is close enough. In [18] predator agents use reinforcement learning to learn a strategy minimizing time to catch a prey. Additionally, agents can cooperate by exchanging sensor data, strategies or episodes. Experimental results show that cooperation is beneficial. Other researchers working on this domain successfully apply genetic programming [5] and evolutionary computation [4]. Results of rule induction application in this domain can be found in [13,14].

Target observation is another interesting problem domain. Good example is [21], where rules are evolved to control large area surveillance from the air. In [8] Parker et al. present cooperative observation task to test autonomous generating of cooperative behaviors in robot teams. Agents cooperate to keep targets within specific distance. Lazy learning based on reinforcement learning is used to generate strategy better than random, but worse than a manually developed one. Results of application of reinforcement learning mixed with state space generalization can be found in [3].

3 Reinforcement Learning

Classical learning method in MAS is reinforcement learning. There are several algorithms using this strategy. In experiments we used SARSA algorithm [17] which is described below. Knowledge is stored in a function Q that estimates quality value of the action in a given state: $Q : Act \times X \rightarrow \mathcal{R}$, where Act is a set of actions, X is a set of possible states, and $\mathcal{R}$ is a set of real numbers. Reinforcement learning agent gets description of the current state $x_t \in X$, and using its current strategy chooses an appropriate action $a_t \in Act$. Usually, action with the highest Q value is chosen. Next, using reward r_t obtained from the environment, the next state description $x_{t+1} \in X$, and action $a_{t+1} \in Act$ that will be next executed, it updates its strategy:

$$\Delta := r_t + \gamma Q(a_{t+1}, x_{t+1}) - Q(a_t, x_t) \tag{1}$$

$$Q(a_t, x_t) := Q(a_t, x_t) + \beta\Delta \tag{2}$$

where $\gamma \in [0, 1]$ is a discount rate (importance of the future rewards), and $\beta \in (0, 1)$ is a learning rate. The reward characteristics depends on the problem. It represents quality of the action. It is positive in case of achiving a goal (like hiting the target), and negative in case of a failure (like hitting an obstacle).

To speed up the learning process various techniques are developed. One of them is temporary differences mechanism TD($\lambda > 0$) [20], which updates not only last state but also these visited recently. Parameter $\lambda \in [0, 1]$ is a recency factor. Values close to 0 mean that traces are very short.

In reinforcement learning various techniques are used to prevent from getting into a local optimum. The idea is to explore the solution space better by choosing not optimal actions from time to time (e.g. random or not performed in a given state yet). Boltzmann selection [17] is used in experiments for this purpose. Instead of selecting the action with highest value, the action a^* is selected in state x with probability $P(a^*, x)$ calculated according to the following formula:

$$P(a^*, x) = \frac{e^{Q(a^*,x)/\tau}}{\sum_a e^{Q(a,x)/\tau}}, \tag{3}$$

where $\tau > 0$ is a temperature parameter. High values of τ make probability of all actions almost the same, regardless their quality.

4 Supervised Learning

Generally, supervised learning allows to generate an approximation of a function $f : X \to C$ which assigns labels from the set C to objects from set X. To generate knowledge a supervised learning algorithm needs labeled examples which consist of pairs of f arguments and values. Let us assume that elements of X are described by set of attributes $A = (a_1, a_2, \ldots, a_n)$, where $a_i : X \to D_i$, and D_i is a domain of attribute a_i. Therefore $x^A = (a_1(x), a_2(x), \ldots, a_n(x))$ is used instead of x. If size of C is small, like in this research, the learning is called classification, C is set of classes, and the approximation h is called classifier.

In experiments we use three algorithms: Naïve Bayes (NB), C4.5 and Ripper. NB is a simple probabilistic classifier, in which every attribute node a_1, a_2, ..., a_n depends on the class attribute c. Learning is a process of calculation of a priori probabilities $P(c)$ and conditional probabilities $P(a_i|c)$. C4.5 is an decision tree learning algorithm developed by Ross Quinlan [9]. Ripper developed by William W. Cohen [2] generates decision rules instead of trees.

4.1 Agent Architecture for Supervised Learning

In [12] we propose architecture for agent applying supervised learning for strategy generation (see Fig. 1). The agent consists of four modules:

Processing Module is responsible for basic agent activities, storing training data, executing learning process, and using learned knowledge;

Learning Module is responsible for execution of learning algorithm and giving answers for problems with use of learned knowledge;

Training Data is a storage for examples used for learning;

Generated Knowledge is a storage for learned knowledge.

These components interact in the following way. *Processing Module* receives *Percepts* from the environment. It may process them and execute *Actions*. If during processing learned knowledge is needed, it formulates a *Problem* and sends it to

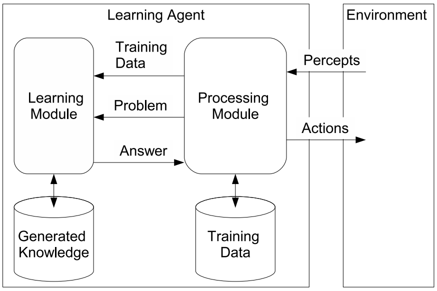


Fig. 1. Learning agent architecture

the *Learning Module*, which generates an *Answer* for the *Problem* using *Generated Knowledge*. *Processing Module* decides also what data should be stored in the *Training Data* storage. When necessary (e.g. periodically, or when *Training Data* contains many new examples) it calls the *Learning Module* to execute the learning algorithm to generate new knowledge from *Training Data*. The learned knowledge is stored in the *Generated Knowledge* base. Learning module may be defined as a four-tuple: (*Learning Algorithm*, *Training Data*, *Problem*, *Answer*). Characteristics of the training data, the problem and the answer depend on the learning strategy used in the learning module.

5 Supervised Learning for Delayed Rewards

Below we propose a method which allows supervised learning to work in environments with delayed rewards. It is based on a time window method.

During learning, instead of the single example x, we use a sequence of examples $\mathbf{X} \ni \boldsymbol{x_t} = (x_t, x_{t-1}, \ldots, x_{t-dt+1})$ to represent situation in time stamp t with window of size dt. Every x_i is described by attributes $A : a_1(x_i), a_2(x_i), \ldots,$ $a_n(x_i), a(x_i))$. The last attribute $a(x_i) \in Act$ represents an action executed in time step i. Two categories are used: $C = \{+, -\}$ representing success $(+)$ or failure $(-)$ of executing actions $a(x_i)$ in the sequence $\boldsymbol{x_t}$. In case of continuous rewards, success can be defined as a reward above certain threshold. Training data can be defined as a sequence $((\boldsymbol{x_1}, c_1), (\boldsymbol{x_2}, c_2), \ldots, (\boldsymbol{x_m}, c_m))$.

After learning we get the classifier $h : \mathbf{X} \to C$, or more precisely, $h : \mathbf{X}, C \to \mathcal{R}$ which for given $\boldsymbol{x_t}$ and c returns certainty of this classes. Our goal is to select the best action $a^* \in Act$ for the current time t using h. We decided to apply the following method:

1. Prepare $\boldsymbol{x_t}$ by setting all attributes describing current and previous states and leaving action attributes unknown. Classifier should be able to work with unknown attribute values. If $t < dt$ then attributes describing non-existing states are also set to unknown value.
2. Find $t^* \in \{t, t-1, \ldots, t-dt+1\}$ for which the dispersion of certainties of positive class for various actions is the highest, assuming that actions in other time stamps are unknown:
   (a) For $\boldsymbol{x_t}$ prepare a set of examples $\{\boldsymbol{x}_k^{act}\}$, for all $act \in Act$ and $k = t, t-1, \ldots, t-dt+1$ where $\boldsymbol{x}_t^{act}$ is created from $\boldsymbol{x_t}$ by setting $a(x_k) = act$.
   (b) Calculate $\delta_k = \max_{act} h(\boldsymbol{x}_k^{act}, +) - \min_{act} h(\boldsymbol{x}_k^{act}, +)$ for every k.
   (c) Choose k^* which maximizes δ_k.
   (d) Return action to execute: $a^* = \mathrm{argmax}_{act} h(\boldsymbol{x}_{k^*}^{act}, +)$.

We tried also entropy measure instead of the maximal dispersion but experimental results were worse. Experimental results showed that exploration improves results of supervised learning too. Therefore Boltzmann selection was also applied here.

6 The Farmer-Pest Problem

We tested the learning algorithm described above on a Farmer-Pest problem. This environment borrows the concept from the specific aspect of real world, in which farmers struggle to protect their fields and crops from pests. Each farmer (this is the only type of agent in the problem) can manage multiple fields. On each field, a multiple types of pests can appear. Each pest has a specific set of attributes e.g. number of legs, color. Values of these attributes depend on the pest type. To protect the field, the farmer can take the advantage of multiple means (e.g. pesticides) called *actions*. However, each pest type has different resistance to each farmer's action (hereinafter referred to as *resistance matrix*). Usually, the problem is time-limited to discrete number of turns. In every turn an agent can execute one action only. It makes simple strategies like applying all actions for every pest inefficient.

The key assumption here is that the farmer agent is not aware of the possible types of the pests nor the resistance matrix. What he can see are the pests' attributes. Based on them, he needs to learn how to recognize different pest types. To learn the resistance matrix, the agent needs to experiment with different actions and observe their effects (i.e. whether the pet dies or not). To make the problem more complicated, the effects are not always immediate and they depend on the resistance matrix. The resistance of the specific pest type to a specific action is valued by the time after which the pets dies (pest's immunity to the action is valued as infinite time). Pests can also have a maximum life-span after which they die regardless of the agent's actions. This maximum life span is called *alive time*.

The problem can be further extended by introduction of deviations to the observed values of the pests' attributes or limiting the number of attributes the farmer agent can see.

7 Experiments

Using the Farmer-Pest problem we are able to make comparison of reinforcement and supervised learning algorithms. Two dimensions are chosen to define various versions of the environment: attribute distributions and the delay of action results. The objective is to analyze how quickly agents may improve their performance and show that various conditions favor different learning algorithms.

7.1 Setup

Two experiments with various settings were performed: simple and complex. Four learning agents take part in every experiment. They use SARSA, Naïve Bayes, C4.5 or Ripper learning algorithms. Their results are compared to those obtained by a fifth agent that executes random actions. Pests are described by four attributes. There are eight actions $Act = \{act_1, act_2, \ldots act_8\}$, and every pest p^t can be killed by act_t only, after delay d. Three values of d were used in experiments: 2, 4 and 8. As a consequence, we used $dt = 8$.

Table 1. Pest's attribute probability distributions for Experiment 2

Pest	size		legsno		speed			jump			Pest	size		legsno		speed			jump		
type	40	10	40	30	40	20	10	40	20	10	type	40	10	40	30	40	20	10	40	20	10
1	x		x		x			x			5	x			x	x			x		
2	x		x		x				x		6	x			x	x				x	
3		x	x		x			x			7		x		x	x			x		
4		x	x		x				x		8		x		x	x				x	

Data was collected by running simulation software developed in Java (Weka's J48 and JRip implementations of C4.5 and Ripper were used). Every experiment consists of 100 simulations. Every simulation consists of 20 consecutive games. 80 pests appear in every game. Knowledge of agents is preserved from game to game. Although, it is cleared between simulations. Supervised learning is executed between games while reinforcement learning is performed after every turn. In the experiments we check how the performance of agents, measured by pests eliminated, changes during simulation. Figure 2 present means of numbers of pests eliminated by every agent in consecutive games.

Experiment 1 – Simple Environment We start with simple environment. Every pest p^t of type t has unique values of all attributes: $a_i(p^t) = 10t, \quad i = 1..4, t = 1..8$. It allows to recognize every type using single value of any attribute.

Experiment 2 – Complex Environment In this experiment four attributes with domains of size two or three are used. The attribute values are presented in Tab. 1. As we can see, distributions are chosen in such a way that no single attribute can be used to recognize pest type – tests on several attributes are necessary. Therefore, this domain is significantly more complex than the previous one, even though sizes of attribute domains are smaller.

Parameters were tuned using hill-climbing algorithm on the complex environment with $d = 4$. The following values were chosen for SARSA: $\lambda = 0.9, \gamma = 0.8, \beta = 0.3, \tau = 0.05$. For supervised learning algorithms only τ had to be tuned: for Naïve Bayes $\tau = 0.15$, for C4.5 $\tau = 0.15$ and for Ripper $\tau = 0.1$.

7.2 Discussion of the Results

Results for the simple case are presented in Fig. 2 (a-c). Random agent and NB agent have poor results. Explanation for NB is simple. This model is not able to take into account dependency between state attributes and action attributes. Remaining learning agents perform much better. If d is larger, learning is slower, but finally results are still good. The only exception is result for SARSA and $d = 8$. Always in the last game Ripper and C4.5 are better than random choice and NB and the difference is statistically significant (using t-test, at $p < 0.05$).

Results for the Complex Environment are presented in Fig. 2 (d-f). NB is not working again. The less d, the better results can be finally achieved. For $d = 2$ C4.5 and SARSA learn quickly and outperform other agents (result is statistically significant at $p < 0.05$). For $d = 8$ C4.5 is statistically better then others (at $p < 0.05$) and Ripper is better than SARSA.

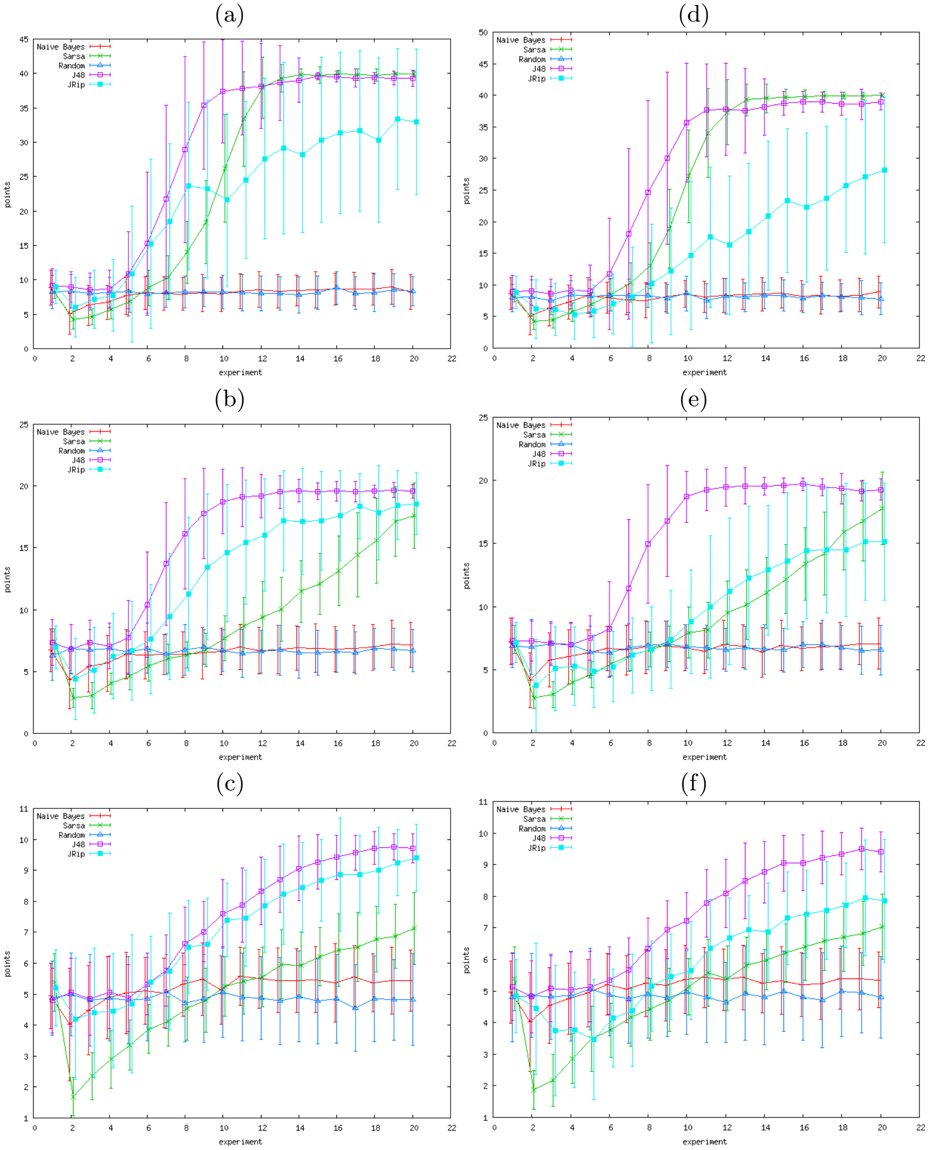

(a) (d) (b) (e) (c) (f)

Fig. 2. Means of numbers of pests eliminated by every agent in consecutive games in Experiment 1 for $d = 2$ (a), $d = 4$ (b) and $d = 8$ (c) and in Experiment 2 for $d = 2$ (e), $d = 4$ (f) and $d = 8$ (g); the following symbols are used: $\times$ for SARSA, + for Naïve Bayes, □ for C4.5, ■ for Ripper, and △ for random

Comparing results for both environments, we can observe, that change in the complexity affects Ripper algorithm more than SARSA and C4.5. For the same *d* SARSA have similar learning speed in both cases. It is coused by the tabular Q-function representation, which has the same performance for both cases. C4.5 is only slightly slower for complex environment. This classifier is able to separate classes well even in more complex environment. Ripper classifier has worse accuracy for the second one.

8 Conclusion and Further Research

In this paper we present a method of agent strategy generation using supervised learning in environments with delayed rewards. We compare performance of the reinforcement and supervised learning algorithms: SARSA, Naïve Bayes, C4.5 and Ripper. These algorithms were used by agents taking part in a simulation of Farmer Pest Problem which is a scalable multi-dimensional problem domain for testing agent learning algorithms. This environment provides a large number of configurable dimensions which enables preparation of different testing conditions.

Experimental results show that Naïve Bayes can not be used for this purpose because of the attribute independence assumptions. Performance of other algorithms depends on the environment configuration (e.g. *d* value). For simple environments, in which any attribute value allows to decide on action, every learning algorithm give fast improvements, unless the reward delay *d* is too large (it is a problem especially for SARSA). If environment is more difficult, when it happens that the agent should take into account values of several attributes to choose an appropriate action, and *d* is not small ($d \geq 4$) C4.5 and Ripper supervised learning algorithms perform better than SARSA. It appears that the most difficult for learning algorithm is uniform distribution of most of the attributes which generates a noise, especially dangerous for SARSA algorithm.

It should be noted that in cases when the knowledge learned by agents should be interpreted by humans, supervised learning algorithms with symbolic knowledge representation should be preferred (if it gives acceptable results). Knowledge stored during the learning process can have a form which makes it possible to interpret by human. Especially decision rules or trees are good choices.

Future work will be concentrated on the execution of experiments with more algorithms. Another aspect of work will be the extension of testing environment to cover cooperation between agents and delays in action results. Symbolic knowledge representation allows to take into account complex dependencies between environment attributes and decisions to be made. Next, we are planning other applications like resource allocation [1] and Focused Web Crawling [19].

Acknowledgments The research presented here has received founding from Polish National Science Centre under the grant no. UMO-2011/01/D/ST6/06146.

References

1. Cetnarowicz, K., Drezewski, R.: Maintaining functional integrity in multi-agent systems for resource allocation. Computing and Informatics 29(6), 947–973 (2010)

2. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning (ICML 1995), pp. 115–123 (1995)
3. Fernández, F., Borrajo, D., Parker, L.E.: A reinforcement learning algorithm in cooperative multirobot domains. Journal of Intelligent Robotics Systems (2005)
4. Giles, C.L., Jim, K.C.: Learning communication for multi-agent systems. In: WRAC, pp. 377–392 (2002)
5. Haynes, T., Sen, I.: Evolving behavioral strategies in predators and prey. In: Weiss, G., Sen, S. (eds.) IJCAI-WS 1995. LNCS, vol. 1042, pp. 113–126. Springer, Heidelberg (1996)
6. Luke, S., Hohn, C., Farris, J., Jackson, G., Hendler, J.: Co-evolving soccer softbot team coordination with genetic programming. In: Kitano, H. (ed.) RoboCup 1997. LNCS, vol. 1395, pp. 398–411. Springer, Heidelberg (1998)
7. Panait, L., Luke, S.: Cooperative multi-agent learning: The state of the art. Autonomous Agents and Multi-Agent Systems 11 (2005)
8. Parker, L.E., Touzet, C.: Multi-robot learning in a cooperative observation task. In: Distributed Autonomous Robotic Systems 4, pp. 391–401. Springer (2000)
9. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
10. Riley, P., Veloso, M.: On behavior classification in adversarial environments. In: Distributed Autonomous Robotic Systems 4. pp. 371–380. Springer (2000)
11. Sen, S., Weiss, G.: Learning in multiagent systems, pp. 259–298. MIT Press, Cambridge (1999)
12. Śnieżyński, B.: An architecture for learning agents. In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2008, Part III. LNCS, vol. 5103, pp. 722–730. Springer, Heidelberg (2008)
13. Śnieżyński, B.: Agent strategy generation by rule induction in predator-prey problem. In: Allen, G., Nabrzyski, J., Seidel, E., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2009, Part II. LNCS, vol. 5545, pp. 895–903. Springer, Heidelberg (2009)
14. Śnieżyński, B.: Agent strategy generation by rule induction. Computing and Informatics 32(5) (2013)
15. Śnieżyński, B., Dajda, J.: Comparison of strategy learning methods in farmer–pest problem for various complexity environments without delays. Journal of Computational Science 4(3), 144–151 (2013)
16. Stone, P., Sutton, R.S., Kuhlmann, G.: Reinforcement learning for robocup-soccer keepaway. Adaptive Behavior 13 (2005)
17. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning). The MIT Press (March 1998)
18. Tan, M.: Multi-agent reinforcement learning: Independent vs. cooperative agents. In: Proceedings of the Tenth International Conference on Machine Learning, pp. 330–337. Morgan Kaufmann (1993)
19. Turek, W., Opalinski, A., Kisiel-Dorohinicki, M.: Extensible web crawler – towards multimedia material analysis. In: Dziech, A., Czyżewski, A. (eds.) MCSS 2011. CCIS, vol. 149, pp. 183–190. Springer, Heidelberg (2011)
20. Watkins, C.J.C.H.: Learning from Delayed Rewards. Ph.D. thesis, King's College, Cambridge (1989)
21. Wu, A.S., Schultz, A.C., Agah, A.: Evolving control for distributed micro air vehicles. In: IEEE Conference on Computational Intelligence in Robotics and Automation, pp. 174–179 (1999)

Classification of Committees with Veto and Stability of Power indices

Jacek Mercik

Wroclaw University of Technology, Wroclaw, Poland
jacek.mercik@pwr.wroc.pl

Abstract. Decision making by a committee may be modelled by simple games. Some of committee's members are equipped with a veto, i.e. they may stop an action temporarily or permanently (via transforming a winning coalition into a losing coalition). Classification of such games and of power indices is presented in the paper. Special emphasis is given to particular characteristics of winning coalitions and consequently to a priori power indices and their stability.

Keywords: voting, veto, power index, classification.

1 Introduction

The analysis of a veto in committee decision making has wide spectrum of aspects. From a political science analysis (for example: Tsebelis (2002) and state-of-art paper of Ganghof (2003)), more toward theory of game analysis (see for example: Mercik (2011, 2012) till practical evaluation of a veto influence on a decision with special emphasis done on a power evaluation of a player (power indices). The last one may be seen for example in: Fragnelli, Chessa (2011, 2013) or Mercik (2009).

This article has three related goals. First, we want to classify observed cases of veto during decision making depending on number of vetoers and on type of veto, focusing especially on differences between types of vetoes. In doing this, we may do an analysis of conditional veto or unconditional veto. Second, we try to estimate number of winning coalitions with different types of veto and with different number of vetoers. And finally, third, we try to evaluate stability of power indices as direct result of changes in number of winning coalitions.

Almost all a priori power indices are constructed as a ratio of particular winning coalitions (with restrictions following from given assumptions about it) to all coalitions (sometimes to majority coalitions only). Therefore, known number of winning coalitions allows us to evaluate a given power index and, what probably is more useful, lets us to estimate a tendency of changes. Consequently, in practice, number of winning coalitions will measure a position of a given decision-maker acting under different regulations. This will help us to make a choice of proper voting system. Such analytical evaluation is not met in the literature up to now, except NP-hard permutation algorithms.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 409–416, 2013.

The article is set up as follows. The next section outlines the preliminaries of simple games with veto and power indices. The next one presents the classification of cases depending on veto's type, on number of vetoers and on yes-no or yes-no-abstain voting. The next section is devoted to problems of stability of power indices. Finally, there are some conclusions and suggestions for future research.

2 Preliminaries

Let N be a finite set of committee's members, q be a quota, w_j be a voting weight of member $j \in N$.

A game on N is given by a map $v : 2^N \to R$ with $v(\emptyset) = 0$. The space of all games on N is denoted by G. A coalition $T \in 2^N$ is called a carrier of v if $v(S) = v(S \cap T)$ for any $S \in 2^N$. The domain $SG \subset G$ of simple games on N consists of all $v \in G$ such that

(i) $v(S) \in \{0,1\}$ for all $S \in 2^N$;

(ii) $(N) = 1$;

(iii) v is monotonic, i.e. if $S \subset T$ $then$ $v(S) \leq v(T)$.

A coalition S is said to be winning in $v \in SG$ if $v(S) = 1$ and losing otherwise. Therefore, the voting upon a bill is equivalent to formation of a winning coalition consisting of voters. A simple game (N,v) is said to be proper, if and only if it is satisfied that for all$\mathrm{T} \subset \mathrm{N}$, if $v(T) = 1$ $then$ $v(N\backslash T) = 0$.

We analyse only simple and proper games where players may vote in yes-no or yes-no-abstains respectively.

By $(N, q, \mathrm{w}) = (\mathrm{N}, \mathrm{q}, \mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_\mathrm{n})$ we shall denote a committee (weighted voting body) with member set N, quota q and weights w_j, $j \in N$. We shall assume that w_j are nonnegative integers. Let $t = \sum_{j=1}^{n} w_j$ be the total weight of the committee.

Let V denotes a set of all committee's members equipped with veto. We assume that cardinality of the set V, $V \subset N$, is equal greater 1, i.e. $card(V) = c_v \geq 1$.

A power index is a mapping $\varphi\colon SG \to R^n$. For each $i \in N$ and $v \in SG$, the i^{th} coordinate of $\varphi(v) \in R^n, \varphi(v)(i)$, is interpreted as the voting power of player i in the game v. In the literature there are two dominating power indices: the Shapley-Shubik power index and the Banzhaf power index. Both are based on the Shapley value concept.

The Shapley-Shubik (Shapley, Shubik (1954)) power index for simple game is the value $\varphi\colon SG \to R^n$, $v \to (\varphi_1(v), \varphi_2(v), \ldots, \varphi_n(v))$ where for all $i \in N$, $card\,\{N\} = n; card\{S\} = s$

$$\varphi_i^{SS}(v) = \sum\nolimits_{S \subset N, i \notin S} \frac{s!(n-s-1)!}{n!}.$$

The Banzhaf (1965) power index for simple game is the value: $\varphi: SG \rightarrow R^n$, $v \rightarrow (\varphi_1(v), \varphi_2(v), \ldots, \varphi_n(v))$ where for all $i \in N$; $card\ \{N\} = n; card\{S\} = s$

$$\varphi_i^B(v) = \frac{1}{2^{n-1}} \sum_{S \subseteq N\{i\}} [v(S \cup \{i\}) - v(S)]$$

The above definitions of power indices are directly obtained from characteristic function games where a marginal value of power excess introduced into winning coalition is calculated. In the paper Turnovec et al. (2008) one can find introduction of power indices without theory of games but based on concept of permutations and of their probability. For calculations, both attempts base on winning coalitions. So, in the following chapter we will evaluate the number of such coalitions and its influence on power index.

If a given coalition's member by using a veto may transform a given coalition from winning into non-winning, then we call that veto the veto of the first degree.

If a veto can be overruled, we call that veto as the veto of the second degree (Mercik, 2011).

3 Classification

Analysing possible combinations of veto types, a cardinality of vetoes and voting systems (yes-no and yes-no-abstain) one may find the following list of all cases (tab.1)

Table 1. Classification of committee's decision situations with different number of vetoers and different types of vetoes

		card {V}=1	card {V}>1
Veto of the I degree (veto is definitive)	Yes-No	Case #1	Case #2
	Yes-No-Abstain	Case #3	
Veto of the II degree (veto can be overruled)	Yes-No	Case #4	?
	Yes-No-Abstain		Case #5

Case #1 represents committee where there is yes-no voting and one and only one member, $i_v \epsilon V$, $card(V) = 1$, is equipped with veto (which cannot be overruled). This member of committee has to be necessarily included into any winning coalition S. In fact, this member is a dictator. Hence, each winning coalition may be transformed by veto of that member into losing coalition. Consequently, a priori Shapley-Shubik and the standardized Banzhaf indices equal $\varphi_{i_v}^{SS} = \varphi_{i_v}^{B} = 1$ and they equal 0 for the rest of the members according to the axioms of power indices (Mercik, 2012). Any a priori power index should give the same results.

Case #2 is a modification of case #1 where $card(V) > 1$, i.e. there is more than one member with veto attribute. The direct impacts of that fact are:

- Every winning coalition must include all veto players,

- Number of winning coalitions S is a function of number of regular players belonging to $N\backslash V$ and a new quota $q - c_v$,
- An a priori power index equals $\frac{1}{c_v}$ for veto players and 0 for the rest of the players,
- If veto is of I-type then $G = \{N, q, w_j = 1,\} = \{N - c_v, q - c_v, w_j = 1\}$ for $j = 1,2, \dots, N$.
- Let W_G denotes set of all winning coalitions in regular SG, i.e. SG where there is no veto player and W_{c_v} denotes set of all winning coalitions in SG where players c_v are equipped with a veto. For yes-no voting if $w_j = 1$ for $j = 1,2, \dots, n$ then $card\ W_G > card\ W_{c_v}$.

  Proof: For $r = q, q + 1, \dots, N,\quad card\ W_G = \sum_r \binom{N}{r}$ and for $r = q - c_v, \dots, N - c_v\quad card\ W_{c_v} = \sum_r \binom{N - c_v}{r}$. The conclusion is obtained directly from the binomial coefficient definition.

Case #3 is similar to the case #1 with one exception: a player equipped with veto may vote yes, no and she/he may abstain. The veto, if any, may not be overruled. The real existing situation is for example a presidential veto where a president may accept a bill, may veto that bill or may ignore the bill. To ignore in this case means in fact to vote for in a game where a winning coalition needs quota q. In consequence:

- If $w_j = 1$ for $j = 1,2, \dots, n$ then $card\ W_{\#3} > card\ W_{\#1}$, where $W_{\#i}$ denotes the set of winning coalitions in case *#i*. The same may be said about $card\ W_{\#3} > card\ W_{\#2}$.
- For a priori power indices: $\sum_1^{c_v} \varphi_i \leq 1 \Longrightarrow \varphi_j \geq 0$ for the player j being non-vetoers and the player i being vetoers respectively, i.e. not only vetoers may have some a priori power. The exact result depends on ratio between a quota q and number of vetoers c_v,
- For $c_v > 1$ typical example is the UN Security Council where among 15 members 5 of them (permanent members) are equipped with veto[1].

Case #4 starts cases with veto of the second degree, when a veto may be overruled. In case #4 there is one member with a veto, and voting is leaded according to yes-no or yes-no-abstain technique. The very typical example of such case is eventual veto of the president of Poland (Mercik, 2009). The following is true:

- If $w_j = 1$ for $j = 1,2, \dots, n$ then $card\ W_{\#4} > card\ W_{\#3}$, where $W_{\#i}$ denotes the set of winning coalitions in case *#i*,
- For a priori power indices: $\sum_1^{c_v} \varphi_i \leq 1 \Longrightarrow \varphi_j \geq 0$ for the player j being non-vetoers and the player i being vetoers respectively, i.e. not only vetoers may have some a priori power. The exact result depends on the ratio between the quota q and the number of vetoersc_v.

[1] Each Council member has one vote. Decisions on procedural matters are made by an affirmative vote of at least 9 of the 15 members. Decisions on substantive matters require nine votes, including the concurring votes of all five permanent members. This is the rule of “great Power unanimity”, often referred to as the veto power.

Case #5 is in fact duplication of case #4 because abstain by a veto player means no veto at all and the game is a simple game *SG* played outside of the veto player.

The case #? is the most intriguing case. A little bit more complicated combinatorial calculation when calculating a priori power index but unclear when analysing relations between veto players. For example, we shall know whether it is possible to overrule one veto player's by another veto player. If it is a case then hierarchy among vetoers should be introduced. In fact, there is no evidence of such veto schema in real life.

4 Stability of Power

Analysis a stability of power means an analysis of a priori power indices due to changes of chosen parameter. Such parameter could be:

- A number of players (in all cases),
- The quota (in all cases),
- The weights (in all cases),
- A number of vetoers, c_v (for type II vetoes),
- Type of a voting (yes-no or yes-no-abstain).

The analysis of influence of changes in number of players shows that the ratio between number of non-veto players and veto players is crucial but in no one case increasing number of non-veto players may not evaporate the a priori power of vetoers. For the veto of the first degree vetoers remain dictators as their veto may not be overruled. For the veto of the second degree also vetoers' a priori power remains unchangeable – for normalized power indices it means that part of the a priori power belonging to all non-veto players is the same but it is distributed among increasing number of non-veto players. In a consequence, a priori power of the non-veto player may decrease only.

Let's analyse sensitivity of cases to changes of quota. By $W(N,q,\mathbf{w}) = \{ S \subseteq N : \sum_{j\in S} w_j \geq q \}$ we shall denote the set of all winning configurations for a committee [*N*, *q*, ***w***].

By weight allocation space we shall call a set $\Omega_N^t = [\mathbf{w} : \mathbf{w} \in R_n, \sum_{j\in N} w_j = t,\ w_j \geq 0]$ of all weight allocations in committees with the same member set *N* and total weight *t*.

We shall say that two committees [*N*, *q*, ***u***] and [*N*, *q*, ***v***] are strategically equivalent if *W(N, q, **u**) = W(N, q, **v**)*. A partition $P_1, P_2, ..., P_m$ of weight allocation space Ω_N^t such that for any two allocations $\boldsymbol{u}, \boldsymbol{v} \in P_i$ it holds that $W(N,q,\mathbf{u}) = W(N,q,\mathbf{v})$ and for any $\mathbf{u} \in P_r$ and $\mathbf{v} \in P_k$ $(r \neq k)$ $W(N,q,\mathbf{u}) \neq W(N,q,\mathbf{v})$ we shall call a constant power partition. All committees from the same set P_i of constant power partition are strategically equivalent. It is easy to show that any well-defined measure of power of the same member in all of these committees is the same.[2]

Analysing each of the P_i, one by one, we are finding subsets of the players with the same "power status" from the point of view of ability of the committee members to form winning coalitions. In all committees corresponding to the same subset each particular member of the committee is the member of the same winning coalitions (and the same losing coalitions). Moving from one subset to another the structure of winning and losing coalitions is changing. So we can use this partition as a starting point for some general considerations about distribution of power in a committee system including that one's having veto players among them.

Having a power index π of a committee $[N, \gamma, \omega]$ with a total weight t, quota γ_0 and an allocation ω^0, let us consider a response of the measure of power to changes of quota[3]. By $y_S(\gamma, \boldsymbol{\omega}) = \sum_{i \in S} \omega_i - \gamma$ we shall denote the surplus of the total weight of a coalition $S \in N$ over the quota γ with respect to the allocation $\boldsymbol{\omega}$. Let us suppose that the quota is changing: $\gamma(\delta) = \gamma^0 + \delta$, where γ^0 denotes initial quota, $\delta > 0$, both γ^0 and δ are integer. We shall say that an allocation $\boldsymbol{\omega}^0$ is stable with respect to a change δ of quota if the structure of winning coalitions remains the same. In this case any power index does not change.

Let's now analyse changing of weights in simple weighted game with veto player(s). We fixed all other parameters.

If we increase weight of veto player than we increase in fact number of c_v (there is assumption on weights being integer number). Therefore: if initial value of c_v is 1 than we switch from case #1 to case #2 or from case #4 to case #5 respectively. Case #3 could remain the same.

If we change weights of non-veto players, i.e. we propose another distribution of weights among them fixing sum total of weights (by assumption c_v is fixed too) than we do not change the classification of cases. The only observed changes may be values of power indices of non-veto players due to new weights.

Changing type of voting is the most sophisticated part of our analysis. Usually, yes-no or yes-no-abstain voting is observed. Generally, the way of aggregation of individual preferences via voting generates the set of winning coalitions and therefore may have direct impact on values of power indices (at least on a priori ones). And this impact is observed only on that. Being veto player is so strong characteristic that for veto of the first type (cases #1, #2 and #3) it should not change the position of a vetoer. For veto of the second type (the rest of cases) changes of aggregation (voting) may cause the value not only non-veto players but veto-player too, but the future research is necessary here.

[2] It follows from the fact that in characteristic function representation of the weighted voting committee (simple game) the characteristic function for all committees is the same, so power indices defined in terms of characteristic function (Shapley-Shubik, Penrose-Banzhaf, Holler-Packel, Johnston, Deegan-Packel) will generated the same values.

[3] For simplicity we shall assume integer weights and quotas.

Changing number of vetoers produces the same as above observations. In the process of aggregation the very crucial is position of veto-players and their number c_v . For the cases connected with veto of the first type the number of non-veto players does not influence on power of veto-players. For other cases (the second type of veto) one may expect some changes for power of veto of the veto-player(s) too.

5 Conclusions

There are still open problems in estimation of stabilisation intervals for a priori power indices being a function of winning coalitions for voting with veto-players. Such analytical evaluation is not met in the literature up to now, except NP-hard permutation algorithms. Presented attempt lets to enlisting of those coalitions depending on the number of veto-players, the type of veto, the number of committee's members and the quota.

Known number of winning coalitions allows us to evaluate a given power index and, what probably is more useful, lets us to estimate a tendency of changes. Consequently, in practice, the number of winning coalitions will measure a position of a given decision-maker acting under different regulations. This will help us to make a choice of proper voting system.

Future research should be focused on proposition of measures of stability of committee's members, including synthetic measure (index) of stability. For the cases connected with veto of the first type the number of non-veto players does not influence on power of veto-players. For other cases (the second type of veto) one may expect some changes for power of veto of the veto-player(s) too. It is quite possible that applications may reduce some of the cases as an unrealistic, however as it is now, case #? should be omitted only.

References

1. Banzhaf III, J.F.: Weighted voting doesn't work: a mathematical analysis. Rutgers Law Review 19, 317–343 (1965)
2. Fragnelli, V., Chessa, M.: Quantitative evaluation of veto power. Operations Research and Decisions 21(3-4), 5–19 (2011)
3. Fragnelli, V., Chessa, M.:: Competition among parties and power: An empirical analysis. Annals of Operations Research (to be published, 2013)
4. Ganghof, S.: Promises and Pitfalls of Veto Player Analysis. Swiss Political Science Review 9(2), 1–25 (2003)
5. Mercik, J.W.: A priori veto power of the president of Poland. Operations Research and Decisions 19(4), 61–75 (2009)
6. Mercik, J.: On a priori evaluation of power of veto. In: Herrera-Viedma, E., García-Lapresta, J.L., Kacprzyk, J., Fedrizzi, M., Nurmi, H., Zadrożny, S., et al. (eds.) Consensual Processes. STUDFUZZ, vol. 267, pp. 145–156. Springer, Heidelberg (2011)
7. Mercik, J.: On axiomatization of power index of veto. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part II. LNCS, vol. 7654, pp. 192–200. Springer, Heidelberg (2012)

8. Penrose, L.S.: The Elementary Statistics of Majority Voting. Journal of the Royal Statistical Society 109, 53–57 (1946)
9. Shapley, L.S.: A Value for n-person Games. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games, vol. II. Annals of Mathematical Studies, vol. 28, pp. 307–317 (1953)
10. Shapley, L.S., Shubik, M.: A method of evaluating the distribution of power in a committee system. American Political Science Review 48(3), 787–792 (1954)
11. Tsebelis, G.: Veto Players. How Political Institutions Work. Princeton University Press, Princeton (2002)
12. Turnovec, F., Mercik, J., Mazurkiewicz, M.: Power Indices Methodology: Decisiveness, Pivots, and Swings. In: Braham, M., Steffen, F. (eds.) Power, Freedom, and Voting, pp. 23–37. Springer, Heidelberg (2008)

Investigation of Property Valuation Models Based on Decision Tree Ensembles Built over Noised Data

Tadeusz Lasota[1], Tomasz Łuczak[2], Michał Niemczyk[2],
Michał Olszewski[2], and Bogdan Trawiński[2]

[1] Wrocław University of Environmental and Life Sciences, Dept. of Spatial Management
ul. Norwida 25/27, 50-375 Wrocław, Poland
[2] Wrocław University of Technology, Institute of Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
tadeusz.lasota@up.wroc.pl,
{tomasz.luczak,bogdan.trawinski}@pwr.wroc.pl,
{michal.niemczyk,michal.olszewski}@student.pwr.wroc.pl

Abstract. The ensemble machine learning methods incorporating bagging, random subspace, random forest, and rotation forest employing decision trees, i.e. *Pruned Model Trees*, as base learning algorithms were developed in WEKA environment. The methods were applied to the real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions. The accuracy of ensembles generated by the methods was compared for several levels of noise injected into an attribute, output, and both attribute and output. Ensembles built using rotation forest outperformed other models. In turn, random subspace method resulted in the models that were the most resistant to noised data.

Keywords: pruned model trees, bagging, random subspaces, random forest, rotation forest, cross-validation, property valuation, noised data.

1 Introduction

The issue of dealing with noisy data is one of key aspects in supervised machine learning to create reliable data-driven models. Noisy data may strongly affect the accuracy of resulting data models and can result in decreasing system performance in terms of predictive accuracy, processing efficiency and the size of the learner. Several works on the impact of noise, mainly in the context of classification problems and class noise, have been published. In [1] increasing the size of training set by adding noise to the training objects was explored for different amount and directions of noise injection. It was shown theoretically and empirically that the k-nearest neighbors directed noise injection was preferable over the Gaussian spherical noise injection when using multilayer perceptrons. In [2] noise was injected to both input attributes and output classes. The results varied depending on the noise type and the specific data set being processed. Naïve Bayes turned out to be the most robust algorithm, and SMO (support vector machine) the least. In [3] it was observed that the attribute noise

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 417–426, 2013.

was less harmful in comparison with class noise. Moreover, the higher the correlation between an attribute and the class, the more negative impact the attribute noise may have. The authors recommend to handle noisy instances before a learner is generated. In [4] two different class noise types were applied to training sets. Fuzzy Rule Based Classification Systems revealed a good tolerance to class noise in comparison to the C4.5 crisp algorithm which is considered resistant to noise. In [5] the performance of several ensemble models learned from imbalanced and noisy binary-class data was compared. As the result clear preference of bagging over boosting was shown. We have studied recently the impact of noised data on the performance of ensemble models for regression problem [6]. We injected noise to output values and showed that random subspace and random forest techniques, where the diversity of component models is achieved by manipulating of features, were more resistant to noise than classic resampling techniques such as bagging, repeated holdout, and repeated cross-validation.

For a few years we have been investigating techniques for developing an intelligent system to assist with of real estate appraisal devoted to a broad spectrum of users interested in the premises management. The outline of the system to be exploited on a cloud computing platform is presented in Fig. 1. Public registers and cadastral systems create a complex data source for the intelligent system of real estate market. The core of the system are valuation models including models constructed according to the professional standards as well as data-driven models generated using machine learning algorithms. So far, we have investigated several methods to construct ensembles of regression models to be incorporated into the system including various resampling techniques, random subspaces, random forests, and rotation forests. As base learning algorithms weak learners as evolutionary fuzzy systems, neural networks, and decision trees were employed [7], [8], [9], [10], [11], [12], [13].

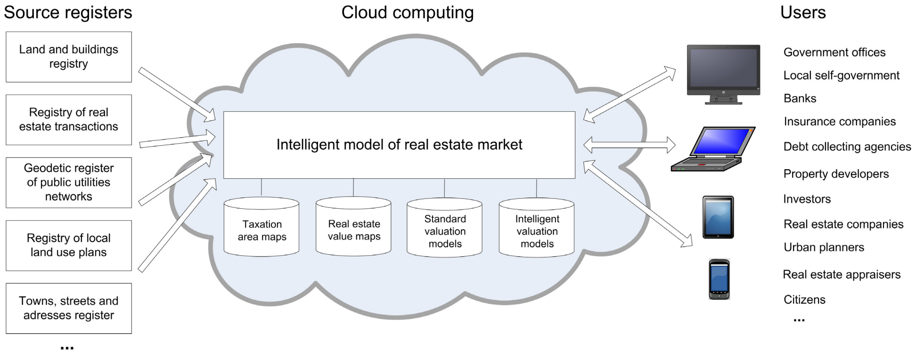

Source registers
Land and buildings registry
Registry of real estate transactions
Geodetic register of public utilities networks
Registry of local land use plans
Towns, streets and adresses register
...
Cloud computing
Intelligent model of real estate market
Taxation area maps
Real estate value maps
Standard valuation models
Intelligent valuation models
Users
Government offices
Local self-government
Banks
Insurance companies
Debt collecting agencies
Property developers
Investors
Real estate companies
Urban planners
Real estate appraisers
Citizens
...

Fig. 1. Outline of the intelligent system of real estate market

The first goal of the investigation presented in this paper is to compare empirically ensemble machine learning methods incorporating bagging, random subspace, random forest, and rotation forest employing decision trees as base learners. Bagging, which stands for bootstrap aggregating, devised by Breiman [14] is one of the most intuitive and simplest ensemble algorithms providing good performance. Another approach to ensemble learning is called the random subspaces, also known as attribute bagging. This approach seeks learners diversity in feature space subsampling [15]. The method called random forest merges these two approaches was worked out by Breiman [16]. Random forest uses bootstrap selection for supplying individual learner with training data and limits feature space by random selection. Rodríguez et al. [17] proposed in 2006 a new classifier ensemble method, called rotation forest, applying Principal Component Analysis (PCA) to rotate the original feature axes in order to obtain different training sets for learning base classifiers.

The second goal is to examine the performance of the ensemble methods dealing with noisy data. The noise was artificially injected into an attribute, output, and both attribute and output. The susceptibility to noised data can be an important criterion for the selection of appropriate machine learning methods to our automated valuation system. We do not konw the purpose of property valuation. For example, the prices estimated to secure loans may differ substantially from the prices appraised to calculate taxes. We do not what sort of properties and their locations were in vogue at the moment of the sale. Moreover, the market instability and uncertainty cause the investors to take irrational sales/purchase decisons. Hence, we may assume that the historical data, we use to create real estate valuation models, contain much noise.

2 Methods Used and Experimental Setup

We conducted a series of experiments to compare bagging (*Bag*), random subspace (*RaS*), random forest (*RaF*), and rotation forest (*RtF*) models with and single models (*Sgl*) in respect of its predictive accuracy using cadastral data on sales/purchase transactions of residential premises. All tests were accomplished using *WEKA (Waikato Environment for Knowledge Analysis),* a non-commercial and open source data mining system [18]. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. The decision tree WEKA algorithm, very often used for building and exploring ensemble models, namely *Pruned Model Tree* (*M5P*), was employed to carry out the experiments. *M5P* implements routines for generating M5 model trees The algorithm is based on decision trees, however, instead of having values at tree's nodes, it contains a multivariate linear regression model at each node. The input space is divided into cells using training data and their outcomes, then a regression model is built in each cell as a leaf of the tree.

Real-world dataset used in experiments was drawn from an unrefined dataset containing above 100 000 records referring to residential premises transactions accomplished in one Polish big city with the population of 640 000 within 14 years from 1998 to 2011. The final dataset counted the 9795 samples. Four following attributes were pointed out as main price drivers by professional appraisers: usable area of a flat (*Area*), age of a building construction (*Age*), number of storeys in the building (Storeys), the distance of the building from the city centre (*Centre*), in turn, price of premises (*Price*) was the output variable. For random subspace, random forest, and rotation forest approaches four more features were employed: number of rooms in the flat including a kitchen (*Rooms*), geodetic coordinates of a building (*Xc* and *Yc)*, and its distance from the nearest shopping center (*Shopping*).

Due to the fact that the prices of premises change substantially in the course of time, the whole 14-year dataset cannot be used to create data-driven models using machine learning. Therefore it was split into subsets covering individual years, and we might assume that within one year the prices of premises with similar attributes were roughly comparable. Starting from the beginning of 1998 the prices were updated for the last day of subsequent years using the trends modelled by polynomials of degree four. We might assume that one-year datasets differed from each-other and might constitute different observation points to compare the accuracy of ensemble models in our study and carry out statistical tests. The sizes of 14 one-year datasets are given in Table 1.

Table 1. Number of instances in one-year datasets

1998	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
446	646	554	626	573	790	774
2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
740	776	442	734	821	1296	577

Following methods were applied in the experiments:

Sgl – M5P algorithm with the number of features equal to 4. In this case single models were built, therefore there was only one iteration of the algorithm.

Bag – Bagging with M5P algorithm, size of each bag was set to 100% of the training set, number of bagging iterations was set to 50.

RaS – Random subspace with M5P algorithm, size of each subspace was set to 75% of all attributes, number of random subspace iterations was set to 50.

RaF – Random forest – Bagging with M5P as the Filtered Classifier and Random Subset as a filter. Size of each bag was set to 100% of the training set, number of bagging iterations was set to 50. Number of attributes in random subset was set to 75% of all attributes.

RtF – Rotation forest with M5P algorithm set as a classifier. Maximum and minimum number of groups was set to 4. The percentage of instances to be removed was set to 20%. As a projection filter the Principial Components with default parameters were used. Number of rotation forest iterations was set to 50.

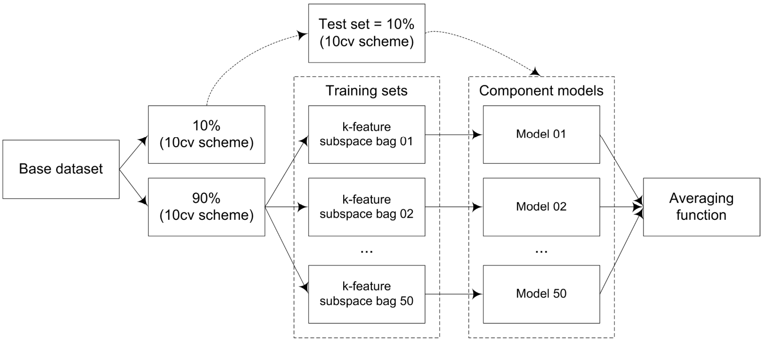


Fig. 2. Outline of experiment with random forest method within 10cv frame. The procedure is repeated 10 times according to 10cv schema.

For each method 10-fold cross validation repeated ten times was used as a result generator. Schema of an experiment using RaF within WEKA 10cv frame is shown in Fig. 2. As a performance function the root mean square error (RMSE) was used, and as aggregation functions of ensembles arithmetic mean was employed.

During our research we analyzed the impact of data noise on the performance of the described ensemble methods. During the first run of experiment no value was changed. Next, we replaced 1%, 5%, 10%, 20%, 30%, 40%, 50% randomly selected input values (Area) in training and in testing set with noised values. Then, we did the same processing with the output value (Price). Finally, we replaced in the same way both input values (Area) and output values (Price) simultaneously. The noised values were generated randomly from range [Q1- 1.5 x IQR, Q3+1.5 x IQR], where Q1 and Q3 denote value of first and third quartile, and IQR stands for the interquartile range. This assured that the numbers replacing the original values were not outliers. The schemata illustrating three modes of noise injection are given in Figures 3, 4, and 5.

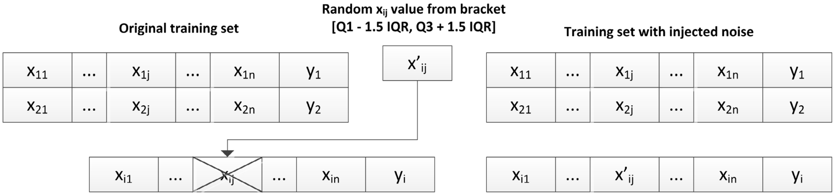

Random x_{ij} value from bracket
[Q1 - 1.5 IQR, Q3 + 1.5 IQR]
Original training set
Training set with injected noise
x'_{ij}
x_{11} ... x_{1j} ... x_{1n} y_1
x_{21} ... x_{2j} ... x_{2n} y_2
x_{i1} ... x_{ij} ... x_{in} y_i
x_{i1} ... x'_{ij} ... x_{in} y_i

Fig. 3. Schema illustrating injection of noise into input variable (*A*)

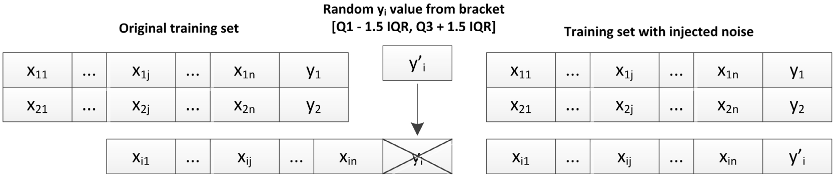

Random y_i value from bracket
[Q1 - 1.5 IQR, Q3 + 1.5 IQR]
Original training set
Training set with injected noise
y'_i
x_{11} ... x_{1j} ... x_{1n} y_1
x_{21} ... x_{2j} ... x_{2n} y_2
x_{i1} ... x_{ij} ... x_{in} y_i
x_{i1} ... x_{ij} ... x_{in} y'_i

Fig. 4. Schema illustrating injection of noise into output variable (*O*)

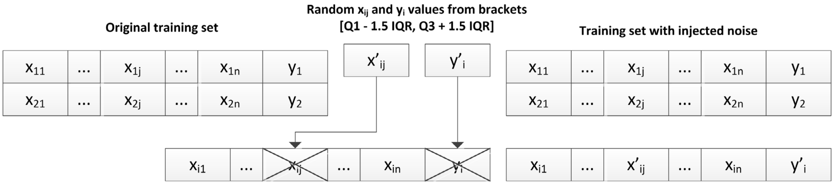

Random x_{ij} and y_i values from brackets
[Q1 - 1.5 IQR, Q3 + 1.5 IQR]
Original training set
Training set with injected noise
x'_{ij}
y'_i
x_{11} ... x_{1j} ... x_{1n} y_1
x_{21} ... x_{2j} ... x_{2n} y_2
x_{i1} ... x_{ij} ... x_{in} y_i
x_{i1} ... x'_{ij} ... x_{in} y'_i

Fig. 5. Schema illustrating injection of noise into both an input and output variables (*AO*)

3 Results of Experiments

The accuracy of *Sgl, Bag, RaS, RaF,* and *RtF* models created using *M5P* for non-noised data, data with 10% injected noise into the attribute *Area* (*A*), the output

Price (*O*), and both attribute and output (*AO*) is shown in Figures 6-9, respectively. In the charts it is clearly seen that *RtF* ensembles reveal the best performance, whereas the biggest values of RMSE provide the *Sgl* and *RaF* models. Moreover, noise injected into the output results in higher error rate than noise introduced into the attribute.

The Friedman tests performed in respect of *RMSE* values of all models built over 14 one-year datasets showed that there are significant differences among models for each noise injection mode considered. Average rank positions, determined by Friedman test, of single and ensemble models for different levels of injected noise into the attribute (*A*), the output (*O*), and both attribute and output (*AO*) are shown in Tables 2, 3, and 4, respectively. In all tables the lower rank value the better model. In each table the *RtF* models are in the first place and *Sgl* and *RaF* ones occupy the last positions. The further Wilcoxon paired tests indicated that there were no statistically significant differences between *RtF* and *Bag* models for (*O*), and between *RtF* and *RaS* models for (*AO*), as well as between *RaF* and *Sgl* models for all noise injection modes.

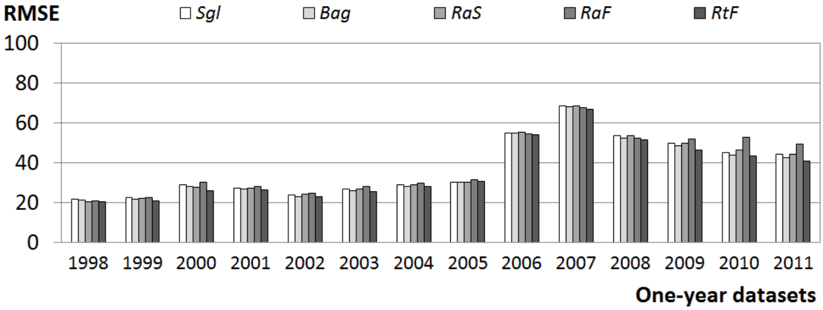


Fig. 6. Performance of single and ensemble models for *non-noised* data

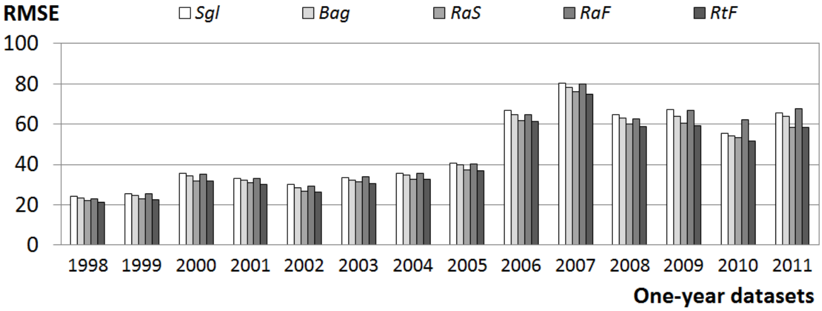


Fig. 7. Performance of single and ensemble models for *10% noise* injected into attribute (A)

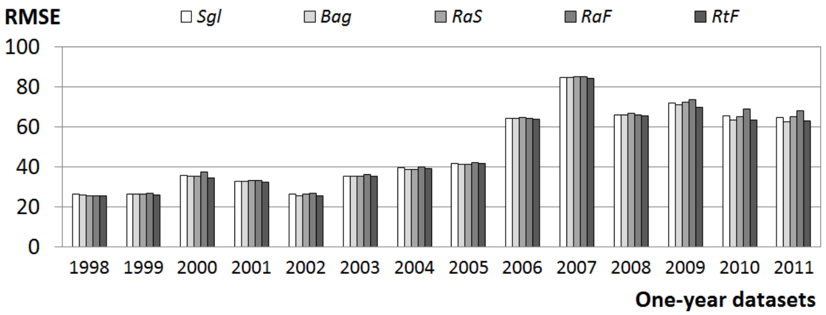


Fig. 8. Performance of single and ensemble models for *10% noise* injected into output (O)

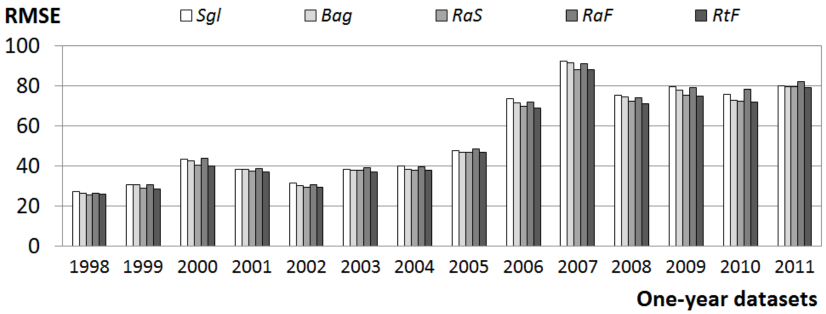


Fig. 9. Performance of models for *10% noise* injected into both attribute and output (AO)

Table 2. Average rank positions of single and ensemble models for different levels of injected noise into attribute (A) determined during Friedman test

Noise/Rank	1st	2nd	3rd	4th	5th
0%	RtF (1.29)	Bag (2.36)	RaS (3.36)	Sgl (3.79)	RaF (4.21)
..5%	RtF (1.14)	RaS (2.07)	Bag (3.00)	RaF (4.36)	Sgl (4.43)
10%	RtF (1.07)	RaS (1.93)	Bag (3.21)	RaF (4.14)	Sgl (4.64)
20%	RtF (1.21)	RaS (1.79)	Bag (3.29)	RaF (4.00)	Sgl (4.71)
30%	RtF (1.14)	RaS (1.86)	Bag (3.36)	RaF (3.79)	Sgl (4.86)
40%	RtF (1.14)	RaS (1.86)	Bag (3.29)	RaF (3.86)	Sgl (4.86)
50%	RtF (1.00)	RaS (2.00)	Bag (3.29)	RaF (3.93)	Sgl (4.79)

Table 3. Average rank positions of single and ensemble models for different levels of injected noise into output (O) determined during Friedman test

Noise/Rank	1st	2nd	3rd	4th	5th
0%	RtF (1.29)	Bag (2.36)	RaS (3.36)	Sgl (3.79)	RaF (4.21)
..5%	RtF (1.43)	Bag (2.21)	RaS (3.43)	Sgl (3.79)	RaF (4.14)
10%	RtF (1.43)	Bag (2.00)	RaS (3.50)	Sgl (3.50)	RaF (4.57)
20%	RtF (1.57)	Bag (2.43)	Sgl (3.36)	RaS (3.43)	RaF (4.21)
30%	RtF (1.79)	Bag (2.36)	Sgl (3.21)	RaS (3.50)	RaF (4.14)
40%	RtF (2.00)	Bag (2.07)	Sgl (3.14)	RaS (3.50)	RaF (4.29)
50%	RtF (2.00)	Bag (2.14)	Sgl (3.07)	RaS (3.64)	RaF (4.14)

Table 4. Average rank positions of single and ensemble models for different levels of injected noise into both attribute and output (AO) determined during Friedman test

Noise/Rank	1st	2nd	3rd	4th	5th
0%	RtF (1.29)	Bag (2.36)	RaS (3.36)	Sgl (3.79)	RaF (4.21)
..5%	RtF (1.07)	RaS (2.50)	Bag (2.86)	RaF (4.14)	Sgl (4.43)
10%	RtF (1.29)	RaS (1.86)	Bag (3.07)	RaF (4.29)	Sgl (4.50)
20%	RtF (1.36)	RaS (1.64)	Bag (3.57)	RaF (3.86)	Sgl (4.57)
30%	RtF (1.14)	RaS (1.86)	Bag (3.43)	RaF (4.07)	Sgl (4.50)
40%	RtF (1.36)	RaS (1.64)	Bag (3.43)	RaF (3.79)	Sgl (4.79)
50%	RtF (1.14)	RaS (1.86)	Bag (3.21)	RaF (3.86)	Sgl (4.93)

Table 5. Median of percentage loss of performance for data with noise vs non-noised data for different levels of injected noise into attribute (A)

Noise	Sgl	Bag	*RaS*	RaF	RtF
1%	4.7%	4.7%	*3.6%*	3.9%	*4.3%*
5%	13.8%	14.2%	*8.3%*	10.3%	*10.9%*
10%	22.4%	22.9%	*12.7%*	18.1%	*16.0%*
20%	35.8%	37.0%	*17.7%*	30.5%	*22.1%*
30%	44.0%	44.8%	*22.6%*	39.1%	*27.3%*
40%	54.4%	52.4%	*27.1%*	46.8%	*29.0%*
50%	54.8%	57.9%	*28.4%*	49.5%	*33.5%*

Table 6. Median of percentage loss of performance for data with noise vs non-noised data for different levels of injected noise into output (O)

Noise	Sgl	Bag	*RaS*	*RaF*	RtF
1%	2.6%	2.7%	*2.8%*	*2.8%*	2.9%
5%	16.4%	15.6%	*14.8%*	*13.8%*	16.4%
10%	23.6%	25.4%	*28.6%*	*26.4%*	32.5%
20%	45.3%	45.7%	*41.4%*	*41.8%*	46.7%
30%	61.8%	64.2%	*61.4%*	*58.3%*	67.1%
40%	72.9%	75.7%	*72.1%*	*68.3%*	80.3%
50%	84.6%	86.2%	*82.8%*	*79.4%*	88.9%

Table 7. Median of percentage loss of performance for data with noise vs non-noised data for different levels of injected noise into both attribute and output (AO)

Noise	Sgl	Bag	*RaS*	RaF	RtF
1%	7.2%	6.0%	*5.9%*	5.9%	6.4%
5%	22.9%	22.7%	*21.3%*	21.6%	23.4%
10%	41.1%	41.9%	*37.4%*	39.8%	40.1%
20%	64.7%	68.1%	*55.9%*	60.4%	62.1%
30%	79.6%	80.1%	*68.1%*	71.9%	75.6%
40%	90.6%	91.5%	*80.1%*	84.5%	86.2%
50%	93.7%	90.7%	*82.5%*	89.9%	90.5%

As for the susceptibility to noise of individual ensemble methods the general outcome is as follows. Injecting subsequent levels of noise results in worse and worse accuracy. Percentage loss of performance for data with 1%, 5%, 10%, 20%, 30%, 40%, and 50% noise versus non-noised data was computed for each one-year dataset. The aggregate results in terms of median over all datasets are presented in Tables 5, 6, and 7. The amount of loss is different for individual datasets and it increases with the growth of percentage of noise. The most important observation is that in each case the average loss of accuracy for *RaS* is lower than for the other models. We obtained similar results in our previous research into susceptibility to noise of ensemble models built with genetic fuzzy systems as basic learning methods [6].

4 Conclusions and Future Work

A series of experiments aimed to compare ensemble machine learning methods encompassing bagging, random subspace, random forest, and rotation forest was conducted. The ensemble models were created using decision tree algorithm over real-world data taken from a cadastral system. Moreover, the susceptibility to noise of these ensemble methods was examined. The noise was injected into an attribute, output, and both attribute and output by replacing the original values by the numbers randomly drawn from the range of values excluding outliers.

The overall results of our investigation were as follows. Ensembles built using rotation forest outperform any other models. On the other hand, single models and ensembles created with random forests revealed the worst performance. In turn, random subspace method resulted in the models the most resistant to noised data.

We intend to continue our research into resilience to noise of regression algorithms employing other machine learning techniques such as neural networks and support vector regression. We also plan to noise data using different probability distributions.

Acknowledgments. This paper was partially supported by the Polish National Science Centre under grant no. N N516 483840.

References

1. Skurichina, M., Raudys, S., Duin, R.P.W.: K-Nearest Neighbors Directed Noise Injection in Multilayer Perceptron Training. IEEE Transactions on Neural Networks 11(2), 504–511 (2000)
2. Nettleton, D.F., Orriols-Puig, A., Fornells, A.: A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial Intelligence Review 33(4), 275–306 (2010)
3. Zhu, X., Wu, X.: Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts. Artificial Intelligence Review 22, 177–210 (2004)
4. Sáez, J.A., Luengo, J., Herrera, F.: Fuzzy Rule Based Classification Systems versus Crisp Robust Learners Trained in Presence of Class Noise's Effects: A Case of Study. In: 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011), Córdoba, Spain, pp. 1229–1234 (2011)
5. Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data With Noisy and Imbalanced Data. IEEE Transactions on System, Man, and Cybernetics–Part A: Systems and Humans 41(3), 552–568 (2011)

6. Lasota, T., Telec, Z., Trawiński, B., Trawiński, G.: Investigation of Random Subspace and Random Forest Regression Models Using Data with Injected Noise. In: Graña, M., Toro, C., Howlett, R.J., Jain, L.C. (eds.) KES 2012. LNCS (LNAI), vol. 7828, pp. 1–10. Springer, Heidelberg (2013)
7. Graczyk, M., Lasota, T., Trawiński, B., Trawiński, K.: Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010, Part II. LNCS, vol. 5991, pp. 340–350. Springer, Heidelberg (2010)
8. Kempa, O., Lasota, T., Telec, Z., Trawiński, B.: Investigation of bagging ensembles of genetic neural networks and fuzzy systems for real estate appraisal. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS (LNAI), vol. 6592, pp. 323–332. Springer, Heidelberg (2011)
9. Lasota, T., Telec, Z., Trawiński, G., Trawiński, B.: Empirical Comparison of Resampling Methods Using Genetic Fuzzy Systems for a Regression Problem. In: Yin, H., Wang, W., Rayward-Smith, V. (eds.) IDEAL 2011. LNCS, vol. 6936, pp. 17–24. Springer, Heidelberg (2011)
10. Lasota, T., Telec, Z., Trawiński, G., Trawiński, B.: Empirical Comparison of Resampling Methods Using Genetic Neural Networks for a Regression Problem. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part II. LNCS (LNAI), vol. 6679, pp. 213–220. Springer, Heidelberg (2011)
11. Lasota, T., Łuczak, T., Trawiński, B.: Investigation of Random Subspace and Random Forest Methods Applied to Property Valuation Data. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 142–151. Springer, Heidelberg (2011)
12. Lasota, T., Telec, Z., Trawiński, B., Trawiński, G.: Investigation of Rotation Forest Ensemble Method Using Genetic Fuzzy Systems for a Regression Problem. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part I. LNCS, vol. 7196, pp. 393–402. Springer, Heidelberg (2012)
13. Lasota, T., Łuczak, T., Trawiński, B.: Investigation of Rotation Forest Method Applied to Property Price Prediction. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2012, Part I. LNCS, vol. 7267, pp. 403–411. Springer, Heidelberg (2012)
14. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
15. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), 832–844 (1998)
16. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
17. Rodrígeuz, J.J., Kuncheva, I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(10), 1619–1630 (2006)
18. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, San Francisco (2011)

Online Classifiers Based on Fuzzy C-means Clustering

Joanna Jędrzejowicz[1] and Piotr Jędrzejowicz[2]

[1] Institute of Informatics, Gdańsk University,
Wita Stwosza 57, 80-952 Gdańsk, Poland
jj@inf.ug.edu.pl

[2] Department of Information Systems, Gdynia Maritime University,
Morska 83, 81-225 Gdynia, Poland
pj@am.gdynia.pl

Abstract. In the online approach a classifier is, as usual, induced from the available training set. However, in addition, there is also some adaptation mechanism providing for a classifier evolution after the classification task has been initiated and started. In this paper two algorithms for online learning and classification are considered. These algorithms work in rounds, where at each round a new instance is given and the algorithm makes a prediction. After the true class of the instance is revealed, the learning algorithm updates its internal hypothesis. Both algorithms are based on fuzzy C-means clustering followed by calculation of distances between cluster centroids and the incoming instance for which the class label is to be predicted. The proposed approach is validated experimentally. Experiment results show that both proposed classifiers can be considered as a useful extension of the existing range of online classifiers.

Keywords: online learning, fuzzy C-means clustering.

1 Introduction

Data mining is an important part of the Knowledge Discovery in Databases (KDD) process, which, according to [6], aims at 'identifying valid, novel, potentially useful, and ultimately understandable patterns in data'. One of the data mining basic tasks is the classification that is identification of some unknown object or phenomenon as a member of a known class of objects or phenomena. In machine learning and statistics, classification is usually understood as the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing instances (observations) whose category membership is known. The idea in the machine learning is to produce a so called classifier, which can be viewed as the function induced by a classification algorithm that maps input data to a category.

There are two basic approaches to solving classification tasks through applying machine learning: batch and online (incremental). In the batch approach the

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 427–436, 2013.

available training set, known before any classification decisions are taken, is used by a classification algorithm to induce a classifier, which can be used in a later stage to predict class labels of incoming instances. The approach does not provide for any adaptation or modification of the classifier once it has been induced. In the online approach a classifier is also induced from the available training set. However, in addition, there is also some adaptation mechanism providing for a classifier evolution after the classification task has been initiated and started. In each round a class label of the incoming instance is predicted and afterwards information as to whether the prediction was correct or not, becomes available. Based on this information adaptation mechanism may decide to leave a classifier unchanged, or modify it, or induce a new one. Online learning is considered to be of increasing importance to deal with never ending and usually massive stream of received data such as sensor data, traffic information, economic indexes, video streams, etc. Online approach is, as a rule, required when the amount of data collected over time is increasing rapidly. This is especially true in the data stream model where the data arrive at high speed so that the algorithms used for mining the data streams must process them in very strict constraints of space and time [14]. A data stream can roughly be thought of as an ordered sequence of data items, where the input arrives more or less continuously as time progresses (see for example [9]). Reviews of algorithms and approaches to data stream mining can be found in [7], [8] and [14].

Online approach can be also useful in situations where the available training dataset does not allow inducing classifier assuring the required generalization level. This can happen because of the inadequate training dataset size or because of the concept drift property. Concept drift is understood as a phenomenon where the statistical properties of the target variable, which the model is predicting, change over time in unpredictable manner. Unfortunately, designing an ideal data mining system using only a limited amount of memory with short and constant processing time of a single record, producing in an incremental way classifier equivalent to the one that would have been obtained through common batch learning (on all instances received so far) and reacting properly to concept drift always reflecting the current concept, is not a simple task. In an attempt to meeting the above requirements formulated in [4], numerous approaches and online data mining algorithms have been proposed and still are pursued. These are, for example described in [13], [2] and [3]. Some interesting, recently proposed, online classification algorithms include [16] and [15].

Two algorithms for online learning and classification are considered. These algorithms work in rounds, where at each round a new instance is given and the algorithm makes a prediction. After the true class of the instance is revealed, the learning algorithm updates its internal hypothesis. Both algorithms are based on fuzzy C-means clustering followed by calculation of distances between cluster centroids and the incoming instance for which the class label is to be predicted. A variant of majority voting is used to decide on the prediction result. In the first algorithm training dataset is extended in each round by adding a new example after the incoming instance class has been revealed. In the second algorithm the size of the training dataset is kept constant and new examples replace some of the old ones in each round. The paper is organized as follows. Section 1 contains introduction. Section 2 gives details of the proposed algorithms and their computational complexity. Section 3 provides information on the computational experiment carried-out to validate the approach. In Section 4 conclusions and directions for future research are presented.

2 Online Classification Algorithm

The general data classification problem is formulated as follows. Let C be the set of categorical classes which are denoted $1, \dots, |C|$. The learning algorithm is provided with the learning instances $\hat{LD} = \{< d, c > \mid d \in D, c \in C\} \subset D \times C$, where D is the space of attribute vectors $d = (w_1^d, \dots, w_n^d)$ with w_i^d being a numeric value, n - the number of attributes. The algorithm is used to find the best possible approximation $\hat{f}$ of the unknown function f such that $f(d) = c$. Then $\hat{f}$ can be used to find the class $c = \hat{f}(d)$ for any d such that $(d, c) \notin \hat{LD}$, that is the algorithm will allow to classify instances not seen in the process of learning.

The considered algorithms make use of fuzzy C-means clustering (see [5]), that is an iterative method which allows one row of data to belong to two or more clusters. The method is based on minimization of the objective function

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{noCl} u_{ij}^m \cdot d(x_i, c_j)$$

where m is a fixed number greater than 1 (in the experiments the value was fixed and equal 2), N is the number of data rows, $noCl$ is the number of clusters, c_j is the center of the j-th cluster, u_{ij} is the degree of membership of the i-th data row x_i in cluster j and d is a fixed metric to calculate the distance from the vector x_i to cluster centroid c_j. Fuzzy C-means clustering is an iterative process. In each iteration step the membership factors u_{ij} and cluster centers c_j are updated.

The following metrics were used (and compared) in the experiments. For s, t being attribute vectors of dimension n we have:

– Euclidean metrics: $d_e(s,t) = \sqrt{(s_1 - t_1)^2 + \dots + (s_n - t_n)^2}$
– Manhattan metrics: $d_m(s,t) = \sum_{k=1}^{n} \mid s_k - t_k \mid$
– discrete metrics: $d_d(s,t) = \sum_{k=1}^{n} g_k$, where

$$g_k = \begin{cases} 0 & s_k = t_k \\ 1 & \text{otherwise} \end{cases}$$

The preliminary step of both online algorithms uses fuzzy C-means clustering to partition the training set into clusters. Namely, for a given set of training data, a given lower limit of elements in each cluster and lower and upper limit of clusters, **Algorithm 1** ensures the partition of learning data into clusters satisfying the limitations. Note that the partition of the training set TD is

Algorithm 1. Fuzzy C-means clustering to fix the number of clusters and partition of training data

```
Require: training data $TD = \bigcup_{c \in C} TD^c$, $noClMin$, $noClMax$ - respectively minimal
    and maximal number of clusters, $noEl$ - minimal number of elements in each cluster
Ensure: number of clusters $noCl$ satisfying $noClMin \leq noCl$, $noCl \leq noClMax$,
    partition of each $TD^c$ into $noCl$ clusters each containing at least $noEl$ elements
initialize $noCl \leftarrow noClMin$
$mayIncrease \leftarrow true$
while $mayIncrease$ do
    for all $c \in C$ do
        use fuzzy C-means clustering to find $noCl$ centroids $ct_1, \ldots, ct_{noCL}$ for $TD^c$
        for all $l \in TD^c$ do
            find the nearest centroid $ct^l$
            $ct^l = arg \min_{1 \leq j \leq noCl} dist(l, ct_j)$
        end for
        $rightClustering^c = \bigwedge_{j \leq noCl}(|T_j^c| \geq noEl)$
    end for
    $mayIncrease = \bigwedge_{c \in C} rightClustering^c$
    if $mayIncrease \wedge (noCl < noClMax)$ then
        $noCl \leftarrow noCl + 1$
    end if
end while
if $mayIncrease = false$ then
    $noCl \leftarrow noCl - 1$
end if
```

performed in rounds, starting with the minimal number of clusters $noClMin$ for each class, and increasing the number in each round, if possible. Since at some step (before reaching $noClMax$) the procedure may fail, it is necessary to remember not only the current partition but also the previous one. It is not annotated in Algorithm 1.

Given the partition of the training data $TD = \bigcup_{c \in C} TD^c$, where each TD^c contains exactly the same number $noCl$ of clusters, for the next step of classification a parameter x fixing the number of neighbours is specified. Let $dist(*, *)$ stand for a fixed metric and assume that r stands for a data row to be classified. For each class $c \in C$ and each cluster $T_j^c \subset TD^c$, for $j = 1, \ldots, noCl$, the distance from r to any element from cluster T_j^c is calculated. Then the distances are sorted in a non-decreasing order so that l_1^{cj} points to the element nearest to r, l_2^{cj} - to the second nearest etc. The coefficient S_{cj}^x measures the distance from x neighbours:

$$S_{cj}^x = \sum_{i=1}^{x} dist(r, l_i^{cj}) \tag{1}$$

For each class $c \in C$ and each cluster $T_j^c \subset TD^c$ the coefficient S_{cj}^x is calculated. Finally, the row r is classified as class c, for which the value (1) is minimal:

$$class(r) = arg \min_{1\leq c\leq |C|} S^x_{cj} \tag{2}$$

The details of this step are given in **Algorithm 2**.

Algorithm 2. Classification

```
Require: training data $TD = \sum_{c\in C} TD^c$, data row $r$, $x$ -number of neighbours
Ensure: class for row $r$
apply Algorithm 1 to partition each $TD^c$ into $noCl$ clusters
for all $c \in C$ do
    for $j = 1, \ldots, noCl$ do
        calculate the coefficient $S^x_{cj}$ as in (1)
        define the class for $r$ as in (2)
    end for
end for
```

Algorithm 3. Online 1

```
Require: training data $LD$, containing $k$ data rows, initial training set $IN$, $noClMin$,
    $noClMax$ - respectively minimal and maximal number of clusters,
Ensure: $qc$ - quality of classification
initialize training set $TD \leftarrow IN$, $testSize \leftarrow 0$, $correctClsf \leftarrow 0$
while $testSize < k$ do
    $testSize \leftarrow testSize + 1$
    let $(r, c)$ randomly chosen row from $LD$
    apply Algorithm 2 to find class $\hat{c}$ for row $r$ using training set $TD$
    if $c = \hat{c}$ then
        $correctClsf \leftarrow correctClfs + 1$
    end if
    delete (r,c) from $LD$
    $TD \leftarrow TD \cup (r, c)$
end while
$qc \leftarrow \frac{correctClfs}{testSize}$
```

Both online algorithms work in rounds, where at each round a new instance is given and the algorithm, using **Algorithm 2** makes a prediction. After the true class of the instance is revealed, the learning algorithm updates its internal hypothesis. In the first algorithm *(Online 1)* training dataset is extended in each round by adding a new example after the incoming instance class has been revealed. In the second algorithm *(Online 2)* the size of the training dataset is kept constant and new examples replace some of the old ones in each round. For each of the algorithms the quality of classification is calculated.

Algorithm 4. Online 2

```
Require: training data $LD$, containing $k$ data rows, $m < k$ - size of training set in each round, $noClMin$, $noClMax$ - respectively minimal and maximal number of clusters,
Ensure: $qc$ - quality of classification
initialize training set $TD$ drawing randomly $k$ distinct rows from $LD$,
$testSize \leftarrow 0$, $correctClsf \leftarrow 0$
for $i = 1 \rightarrow m - k$ do
    $testSize \leftarrow testSize + 1$
    draw random row $(r, c)$ from $LD$ not used before
    apply Algorithm 2 to find class $\hat{c}$ for row $r$ using training set $TD$
    if $c = \hat{c}$ then
        $correctClsf \leftarrow correctClfs + 1$
    end if
    replace one randomly chosen row from $TD$ by $(r, c)$
end for
$qc \leftarrow \frac{correctClfs}{testSize}$
```

2.1 Computational Complexity of the Algorithms

To estimate the computational complexity of the algorithms note that fuzzy C-means clustering has the complexity $O(t \cdot N \cdot m)$, where t is the number of iterations of the C-means algorithm, N is the number of data rows and m is the number of clusters. Thus the complexity of Algorithm 1 is $O(M^2 \cdot t \cdot N \cdot |C|)$, where $|C|$ is the number of classes and M is the maximal number of considered clusters. Since Algorithm 1 is the first step of Algorithm 2 and also Algorithm 2 makes use of sorting, its complexity is $O(M^2 \cdot t \cdot N \cdot |C|) + O(M^2 \cdot |C|) \approx O(t \cdot N \cdot M^2 \cdot |C|)$. Finally, online algorithms perform classification for each data row which gives the complexity $O(t \cdot N^2 \cdot M^2 \cdot |C|)$.

3 Computational Experiment Results

To test the performance of the proposed *Online 1* and *Online 2* algorithms we run them on several publicly available benchmark datasets from UCI and IDA repositories [1] and [10], as shown in Table 1.

In case of the *Online 1* algorithm experiment plan has been based on the 10-cross- validation scheme. The original dataset is partitioned into 10 subsets. In each run one of these subsets is used as the initial training set and the remaining 90% of instances arrive, one in each round. Immediately after an instance arrival *Online 1* predicts its class label. Then the true label is revealed and cumulative error of the algorithm is calculated. Finally, arrived instance extends the current training set and the next round begins. The whole procedure is repeated with each of 10 subsets serving as the initial training set. Final evaluation is based on averages from 10 repetitions of thus defined 10-cross-validation scheme.

Table 1. Benchmark datasets used in the experiment

Dataset	Instances	Attributes	Dataset	Instances	Attributes
ACredit	690	16	Ionosphere	351	35
Banana	5300	3	Magic	19020	11
Bank Mark.	4522	17	Sonar	208	61
Breast Cancer	263	10	Spam	4601	58
CMC	1473	10	Thyroid	215	6
Diabetes	768	9	Twonorm	7400	21
GCredit	999	21	Waveform	5000	41
Heart	303	14	WBC	630	11
Hepatitis	155	20	Winered	1599	12
Image	2086	19	Winewhite	4898	12

In case of the *Online 2* algorithm the number of instances in the initial training set has been set arbitrarily for each considered dataset as shown in Table 2. In each run a new instance arrives and the algorithm predicts its class label. Then the true label is revealed and cumulative error of the algorithm is calculated. Finally, arrived instance replaces another instance from the current training

Table 2. Performance of the Online 1 and Online 2

Dataset	Online 1				Online 2				
	x	Accuracy	St. dev.	Inst sec.	Init. Size	x	Accuracy	St. dev.	Inst sec.
Acredit	3	0.869	0.031	101	50	5	0.854	0.025	320
Banana	2	0.876	0.007	4	200	2	0.859	0.033	12
Bank Marketing	20	0.922	0.021	1	200	2	0.922	0.015	3
Breast Cancer	2	0.753	0.019	320	20	4	0.747	0.013	1500
CMC	2	0.548	0.038	3	100	2	0.510	0.023	10
Diabetes	3	0.763	0.016	5	50	2	0.753	0.016	240
Gcredit	2	0.754	0.014	32	50	2	0.753	0.029	480
Heart	5	0.842	0.012	232	20	2	0.792	0.021	720
Hepatitis	2	0.774	0.031	161	20	2	0.768	0.023	500
Image	2	0.968	0.022	22	100	2	0.924	0.019	32
Ionosphere	2	0.888	0.017	30	20	2	0.884	0.012	150
Magic	2	0.761	0.023	1	500	2	0.727	0.019	4
Sonar	2	0.786	0.003	142	20	2	0.786	0.017	340
Spam	2	0.803	0.015	7	200	2	0.789	0.022	30
Thyroid	1	0.944	0.011	93	20	2	0.926	0.020	1150
Twonorm	2	0.977	0.009	2	500	2	0.974	0.013	8
Waveform	5	0.884	0.016	2	500	2	0.873	0.022	30
WBC	15	0.975	0.020	9	50	2	0.975	0.016	256
Winered	2	0.611	0.031	7	200	2	0.601	0.036	30
Winewhite	2	0.569	0.023	3	500	2	0.564	0.032	28

Table 3. Accuracy of example online algorithms

Problem	Online 1	Online 2	FPA	incSVM	PA
Banana	0.876	0.859	0.887	**0.892**	0.874
Breast Cancer	**0.753**	0.747	0.719	0.722	0.686
Diabetes	0.763	0.753	0.754	**0.774**	0.725
Heart	**0.842**	0.792	0.826	0.838	0.806
Thyroid	0.944	0.926	0.954	**0.958**	0.954
Twonorm	**0.977**	0.974	0.976	0.974	0.973
Waveform	0.884	0.873	**0.904**	0.897	0.900

set and the next round begins. Final evaluation is based on averages from 20 repetitions of the above described scheme.

For both considered online algorithms the above described experiment plan has been executed 3 times, once for each of the considered distance metrics. For each dataset the value of the parameter x denoting the number of considered neighbors has been set arbitrarily. Performance measures included average cumulative classification error, its standard deviation and number of classified instances per second using notebook equipped with Intel Core i7 processor.

In Table 2 experiment results produced by the *Online 1* and *Online 2* algorithm using the Manhattan metrics are shown.

It is worth noting that using Euclidean metrics instead of the Manhattan one produces statistically worse performance in case of both algorithms. For example, in case of the *Online 2* algorithm, accuracy of classification is, on average, 9.5% better using Manhattan metrics then using Euclidean one as calculated over all considered datasets. Using the discrete metrics leads even to statistically much worse accuracy then using Euclidean one.

In Table 3 performance of the proposed algorithms is compared with performance of the state-of-the-art online algorithms as shown in [16].

FPA in Table 3 stands for the Fuzzy Passive-Aggressive algorithm and IncSVM for the Incremental SVM, both based on the RBF kernel. PA stands for a

Table 4. Online versus batch classifier accuracy comparison

Problem	Online 1	Online 2	GEPC-ad	RF-GEP	SVM	C4.5
Acredit	0.869	0.854	0.893	0.905	0.856	0.887
Diabetes	0.763	0.753	0.826	0.812	0.720	0.751
Gcredit	0.754	0.753	0.877	0.803	0.671	0.639
Heart	0.842	0.792	0.843	0.842	0.837	0.744
Hepatitis	0.774	0.768	0.798	0.821	0.756	0.708
Ionosphere	0.888	0.884	0.969	0.962	0.853	0.892
Sonar	0.786	0.786	0.845	0.896	0.758	0.743
WBC	0.975	0.975	0.978	0.983	0.968	0.955

standard Passive Aggressive algorithm proposed in [3]. In Table 4 comparison with the classification accuracy of several high quality batch classifiers is shown.

Results for cellular gene expression programming with Adaboost (GEPC-ad) were reported in [11]. Results for gene expression programming with rotation forest (RF-GEP) were reported in [12]. Results for SVM and C4.5 classifiers have been calculated using WEKA environment.

4 Conclusions

The paper proposes two simple online classifiers based on fuzzy C-means clustering and comparison of distance between cluster centroids and incoming instances, for which class label is to be predicted. The algorithms *Online 1* and *Online 2* differ in their approach to adapting to concept drift that may occur within an incoming stream of data. While *Online 1* updates its current training set extending it with each incoming instance, *Online 2* uses the mechanism similar to a time window concept.

The reported computational experiment has been carried out to validate the proposed classifiers. From the experiment results one can draw the following conclusions:

- *Online 1* performance in terms of classification accuracy is quite good and comparable to the performance of standard batch classifiers.
- *Online 2* classification accuracy is statistically slightly worse than the performance of the *Online 1*. However in terms of the number of instances classified per second *Online 2* significantly outperforms *Online 1*.
- Both proposed classifiers can be considered as a useful extension of the existing range of online classifiers. Their performance in terms of classification accuracy is comparable with other online classifiers known from the literature.
- In case of both proposed algorithms best performance is assured using the Manhattan metrics as a distance measure.

Future research will be focused on refining the adaptation mechanism of the online algorithms with a view to improve their performance in both dimensions: classification accuracy and speed.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California. School of Information and Computer Science (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Bouchachia, A., Mittermeir, R.: Towards incremental fuzzy classifiers. Soft Computing 11, 193–207 (2007)
3. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive–aggressive algorithms. Journal of Machine Learning Research 7, 551–585 (2006)

4. Domingos, P., Hulten, G.: A General Framework for Mining Massive Data Streams. Journal of Computational and Graphical Statistics 12, 1–6 (2003)
5. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32–57 (1973)
6. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 1–30. AAAI/MIT Press (1996)
7. Gaber, M.M., Zaslavsky, A., Krishnaswamy, S.: Mining data streams: a review. ACM SIGMOD Record 34(1), 18–26 (2005)
8. Gaber, M.M., Zaslavsky, A., Krishnaswamy, S.: Data Stream Mining. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, Part 6, pp. 759–787 (2010)
9. Gama, J., Gaber, M.M.: Learning from Data Streams. Springer, Berlin (2007)
10. IDA Benchmark Repository, `http://mldata.org/repository/tags/data/IDA_Benchmark_Repository/` (January 12, 2013)
11. Jędrzejowicz, J., Jędrzejowicz, P.: Cellular GEP-Induced Classifiers. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part I. LNCS (LNAI), vol. 6421, pp. 343–352. Springer, Heidelberg (2010)
12. Jędrzejowicz, J., Jędrzejowicz, P.: Rotation Forest with GEP-Induced Expression Trees. In: O'Shea, J., Nguyen, N.T., Crockett, K., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2011. LNCS (LNAI), vol. 6682, pp. 495–503. Springer, Heidelberg (2011)
13. Laskov, P., Gehl, C., Kruger, S., Muller, K.R.: Incremental support vector learning: analysis, implementations and applications. Machine Learning 7, 1909–1936 (2006)
14. Pramod, S., Vyas, O.P.: Data Stream Mining: A Review on Windowing Approach. Global Journal of Computer Science and Technology Software & Data Engineering 12(11), 26–30 (2012)
15. Shaker, A., Senge, R., Hllermeier, E.: Evolving fuzzy pattern trees for binary classification on data streams. Information Sciences 220, 34–45 (2013)
16. Wang, L., Ji, H.-B., Jin, Y.: Fuzzy Passive-Aggressive classification: A robust and efficient algorithm for online classification problems. Information Sciences 220, 46–63 (2013)

Agent-Based Natural Domain Modeling for Cooperative Continuous Optimization

Tom Jorquera, Jean-Pierre Georgé, Marie-Pierre Gleizes, and Christine Régis

IRIT (Institut de Recherche en Informatique de Toulouse)
Paul Sabatier University, Toulouse, France
{jorquera,george,gleizes,regis}@irit.fr

Abstract. While multi-agent systems have been successfully applied to combinatorial optimization, very few works concern their applicability to continuous optimization problems. In this article we propose a framework for modeling a continuous optimization problems as multi-agent system, which we call NDMO, by representing the problem as an agent graph, and complemented with optimization solving behaviors. Some of the results we obtained with our implementation on several continuous optimization problems are presented.

Keywords: Multi-Agent System, Continuous Optimization, Problem Modeling.

1 Introduction

One of the major areas in which Multi-Agent Systems (MAS) have been successfully applied is the domain of combinatorial optimization. Using DCOP (Distributed Constraint Optimization Problem), numerous agent-based algorithms have been proposed in the field. However, the applicability of MAS to another optimization domain, continuous optimization, is still mostly unexplored, despite the existence of whole categories of complex optimization problems. Continuous optimization presents interesting challenges concerning the scalability of optimization methods. The various topologies of optimization problems couple has lead to highly specialized methods targeting specific problems types. However, the industrial development created a need for solving highly complex optimization problems involving heterogeneous models and complex interdependencies. Some methods adapted to handle such types of problems have been proposed, but are often unwieldy and lack flexibility. Arguably, the scalability and adaptability properties of MAS indicate their strong potential for the field.

We present in this article a way of modeling a continuous optimization problem as an agent graph, akin to how DCOP is used to model combinatorial optimization problems, named Natural Domain Modeling for Optimization (NDMO). We propose to complement this graph representation with specific agent behaviors tailored for continuous optimization.

In the next part (section 2), we begin by introducing complex continuous optimization methods and the existing works on MAS for optimization. We then

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 437–446, 2013.

present in section 3 a new generic agent-based modeling for continuous optimization problems. To complement this modeling, we present in section 4 specific agent behaviors for continuous optimization as well as some results, and finish by perspectives about future improvements based on the current work.

2 Context

2.1 Complex Continuous Optimization Methods

Continuous optimization problems are optimization problems where the shape of the search space is defined by one (or several) continuous function. Contrary to combinatorial optimization, where the difficulty lies in the combinatorial explosion of possible states, continuous optimization is more concerned with the efficient exploration of search spaces of heterogeneous (and sometimes exotic) topology. However, a common point between these two fields is the fact that the problems encountered can vary from very simple ones to problems of tremendous complexity.

A good example of complex continuous optimization problems are multidisciplinary optimization (MDO) problems. Sobieszczansky-Sobieski and Haftka proposed to define MDO *as methodology for the design of systems in which strong interaction between disciplines motivates designers to simultaneously manipulate variables in several disciplines* [1]. Designers have to simultaneously consider different disciplines (such as, for example, aerodynamics, geometrics and acoustics for an aircraft engine) which are often not only complex by themselves but also strongly interdependent, causing the classical optimization approaches to struggle handling them.

Currently, MDO problems require specific strategies to be solved, and a major part of the research in the field has been focusing on providing these strategies. For example Multi-Disciplinary Feasible Design, considered to be one of the simplest methods [2], consists only in a central optimizer taking charge of all the variables and constraints *sequentially*, but gives poor results when the complexity of the problem increases [3]. Other approaches, such as Collaborative Optimization [4] or Bi-Level Integrated System Synthesis [5], are said bi-level. They introduce different levels of optimization [6], usually a local level where each discipline is optimized separately and a global level where the optimizer tries to reduce discrepancies among the disciplines. However these methods can be difficult to apply since they often require to heavily reformulate the problem [7],and can have large computation time [3].

One of the major shortcomings of these methods is that they require a lot of work and expertise from the engineer to be put in practice. To actually perform the optimization process, one must have a deep understanding of the models involved as well as of the chosen method itself. This is mandatory to be able to correctly reformulate the models according to the formalism the method requires, as well as to work out what is the most efficient way to organize the models in regard to the method. Since by definition MDO involves disciplines of different natures, it is often impossible for one person to possess all the required knowledge, needing the involvement of a whole team in the process. Moreover, answering all these requirements implies a lot of work *before* even starting the optimization process.

It should be clear from these shortcomings that MDO methods are not universal optimization methods but specialized tools to be applied in specific contexts. However the nature of the problems they aim to solve is in this regard not different from the one of "smaller" continuous optimization problems. Their usefulness lies in the fact that classical optimization techniques are not able to handle the increased complexity incurred by the combinatorial explosion of the interactions between the disciplines and their subproblems. This statement brings to light the fact that one of the fundamental difficulties of continuous optimization problems concerns the scalability of the approaches, which in itself is a strong indication of the potential of MAS techniques for this domain.

2.2 Multi-Agent Systems for Optimization

While multi-agent systems have already been used to solve optimization problems, the existing works concern their application to *Combinatorial* Optimization, mainly in the context of the DCOP framework [8].

In DCOP, the agents try to minimize a global cost function (or alternatively, maximize a global satisfaction) which depends on the states of a set of design variables. Each design variable of the optimization problem is associated to an agent. The agent controls the value which is assigned to the variable. The global cost function is divided into a set of local cost functions, representing the cost associated with the conjoint state of two specific variables. An agent is only aware of the cost functions which involve the variable it is responsible for.

While some works successfully used DCOP in the context of continuous optimization [9], it is not adequate to handle the type of problems we propose to solve here. DCOP problems are supposed to be easily decomposable into several cost functions, a property which is not true for continuous optimization problem in general. Moreover the cost values associated to the variables states are supposed to be known. This major assumption does not stand for MDO problem, where the complexity of the models and their interdependencies cause this information to be unavailable in most cases.

DCOP is a very successful framework since it provides a common ground for researchers to propose and to compare new agent-based algorithms for combinatorial optimization. No equivalent has been proposed for continuous optimization using MAS, an obvious impediment to the development of the field.

3 Problem Modeling with NDMO

In answer to the previous shortcomings, we propose a generic approach called Natural Domain Modeling for Optimization (NDMO) that relies on a natural or intrinsic description of the problem (*i.e.* close to the reality being described).

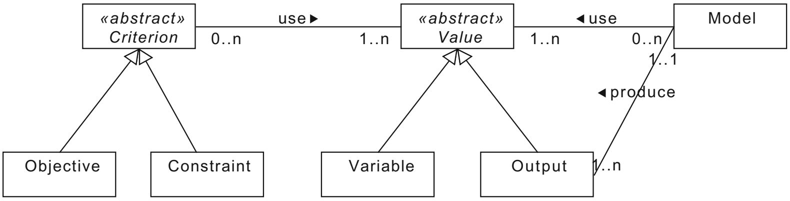


Fig. 1. Class diagram of continuous optimization problems

In order to identify the elements of a generic continuous optimization model, we worked with experts from several related fields: numerical optimization, mechanics as well as aeronautics and engine engineers. As a result, we identified five classes of interacting entities: *models*, *design variables*, *output variables*, *constraints* and *objectives*. These entities and their relations are represented by the diagram in Fig. 1.

To illustrate how an optimization problem is modeled, we use a simplified Turbofan optimization problem, provided by one of our industrial partners. In Fig. 2, the analytic expression of this optimization problem is given,with the corresponding relations graph. The design variables of this problem are *pi_c* and *bpr*, which indicate respectively the compressor pressure ratio and the bypass ratio of the engine. The turbofan model produces three outputs: $Tdm0$, s and fr, representing respectively the thrust, fuel consumption and thrust ratio of the engine. In this problem we try to maximize the thrust and minimizing the fuel consumption while satisfying some feasibility constraints.

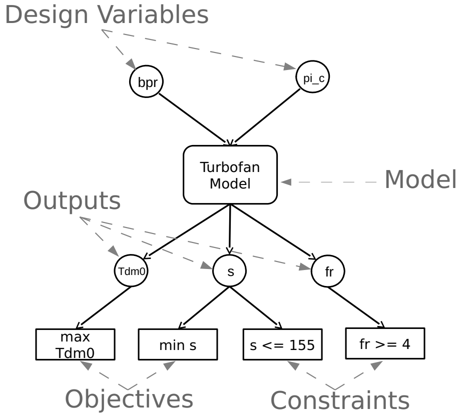
Let's now see in more details the roles of each of these fives entities: *model*, *variable*, *output*, *constraint* and *objective*.

Models. In the most general case, a *model* can be seen as a black box which takes input values (which can be *design variables* or *output variables*) and produces output values. A *model* represents a technical knowledge of the relations between different parts of a problem and can be as simple as a linear function or a much more complex algorithm requiring several hours of calculation. Often some properties are known (or can be deduced) about a model and specialized optimization techniques can exploit this information. In our Turbofan example, a *model* entity is the $Turbofan$ function which calculate the three outputs using the values of *bpr* and *pi_c*.

Design Variables. These are the inputs of the problem and can be adjusted freely (within their defining boundaries). The goal is to find the set(s) of values for these variables that maximize the objectives while satisfying the constraints. *Design variables* are used by *models* to calculate their outputs and by constraints and objectives to calculate their current value. A *design variable* can be shared

$$\begin{gathered}(Tdm0, s, fr) = Turbofan(pi\_c, bp \\ max\ Tdm0 \\ min\ s \\ subject\ to \\ s \leq 155 \\ fr \geq 4\end{gathered}$$

(a) mathematical formulation.



(b) corresponding entities graph.

Fig. 2. Turbofan problem

among several *models*, objectives and constraints. Keeping with our example, *bpr* and *pi_c* are the two *design variables* of our optimization problem.

Output Variables. These values are produced by a *model*, and consequently cannot be changed freely. As for the *design variables*, the *output variables* are used by *models* to calculate their outputs and by constraints and objectives to calculate their current value. In our example, $Tdm0$, s and fr are *output variables* produced by the $Turbofan$ model.

Constraints. These are strict restrictions on some parts of the problem, represented as functional constraints defined by equalities and/or inequalities. These can be the expression of a physical constraint, or a requirement concerning the problem. Regarding the Turbofan, the two *constraints* are $s <= 155$ and $fr >= 4$.

Objectives. The goals to be optimized. In the general case, different objectives are often contradictory. The two *objectives* of the Turbofan problems are to maximize $Tdm0$ and to minimize s.

An interesting and important point is that models, constraints as well as objectives involve computation. Often the most heavyweight calculus is encapsulated inside a model and the calculi concerning criteria tend to be simple equations, but this is neither an absolute requirement nor a discriminating characteristic.

The NDMO modeling aims to provide the most complete and natural representation of the problem. This modeling preserves the relations between the domain entities and is completely independent of the solving process. Since we now have a way to model optimization problems as graphs of entities, we now present the multi-agent algorithm proposed to solve them.

4 Agent Behavior and Experiments

In complement to this modeling of the problem, we propose for NDMO a multi-agent system and associated solving behaviors where each domain entity is associated with an agent. Thus, the multi-agent system is the representation of the problem to be solved with the links and communication between agents reflecting its natural structure. It is worth underlining the fact that this transformation (*i.e.* the agentification) can be completely automatic as it is fully derived from the analytical expression of the problem.

The solving process relies on two continuous simultaneous flow of information: downward (from design variables to criteria) with new values computed by models, and upward (from criteria to design variables) with change-value requests that drive the movements of the design variable in the search space. Intuitively, by emitting requests, criteria agents are "pulling" the different design variables, through the intermediary agents, in multiple direction in order to be satisfied. The system thus converges to an equilibrium between all these "forces", especially in the case of multiple contradicting criteria, which corresponds to the optimum to be found.

A summary of the basic principles of each agent type is given in Algorithm 1.

The functioning of the system can be divided into two main tasks: problem simulation and collective solving.

Problem simulation can be seen as the equivalent of the analysis of classical MDO method. The agents behavioral rules related to problem simulation concern the propagation of the values of design variables to the models and criteria based on the value. For this part, the agents will exchange *inform* messages which contains calculated values. The "message flow" is top-down: the initial inform messages will be emitted by the variable agents and will be propagated down to the criteria agents.Collective solving concerns the optimization of the problem. The agent behavioral rules related to collective solving are about satisfying the constraints while improving the objectives. For this part, the agents will exchange *request* messages which contains desired variations of values. The "message flow" is bottom-up: the initial request messages will be emitted by the criteria agents and propagated up to variable agents.

These basic mechanisms are in themselves not sufficient to handle some of the specificities of complex continuous optimization problems such as MDO. We introduced several specific mechanisms used in conjunction with the previously presented behaviors. The mechanisms have been designed to handle specific challenges related to complex continuous optimization, such as conflicting objectives, cycle handling, hidden dependencies *etc.* The exact working of these mechanisms is of little interest here and will not be detailed. The interested reader can refers to [10] for more detailed explanations.

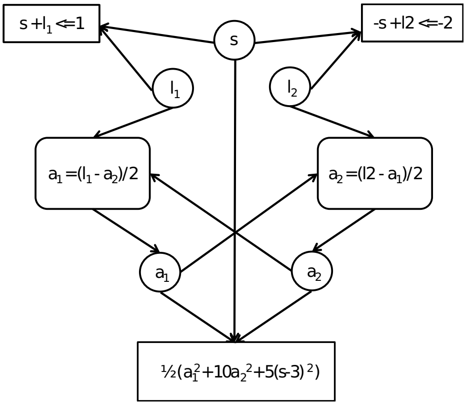
Algorithm 1. Agents Behaviors

```
procedure Model Agent Behavior
    loop
        analyze received messages
        if received new information messages then
            recalculate outputs
            inform depending agents
        end if
        if received new requests then
            use optimizer to find adequate inputs
            propagate requests to input agents
        end if
    end loop
end procedure

procedure Variable Agent Behavior
    loop
        analyze received messages
        if received new requests then
            select most important
            adjust value
            inform depending agents
        end if
    end loop
end procedure

procedure Output Agent Behavior
    loop
        analyze received messages
        if received new information messages then
            update its value
            inform depending agents
        end if
        if received new requests then
            select most important
            transmit selected request to model agent
        end if
    end loop
end procedure

procedure Constraint/ Objective Agent Behavior
    loop
        analyze received messages
        if received new information messages then
            update its value
            use optimizer to find adequate inputs
            send new requests to variable/output agents
        end if
    end loop
end procedure
```

In order to validate our prototype, we experimented on several test cases. We present here synthetic results on three of them: Alexandrov, Turbofan and Viennet1 test cases.

$$
\begin{gathered}
a_1 = (l_1 - a_2)/2 \\
a_2 = (l_2 - a_1)/2 \\
min\ \tfrac{1}{2}(a_1^2 + 10a_2^2 + 5(s-3)^2) \\
subject\ to \\
s + l_1 \leq 1 \\
-s + l_2 \leq -2
\end{gathered}
$$

(a) mathematical formulation.



(b) corresponding agent graph.

Fig. 3. Alexandrov problem

The Alexandrov test case is inspired from an academic example taken in literature by Alexandrov and al [6]. This simple example presents some of the commons characteristics of MDO problems, such as interdependent disciplines and multiple criteria. In the original article, the example was used to illustrate some properties of Collaborative Optimization, which we presented earlier, in terms of reformulation. While the paper only gave the structure of the problem, we adapted it with meaningful values and equations.

The mathematical formulation of the problem and the corresponding agent graph can be seen in Fig. 3. Interestingly, the NDMO representation is quite similar to the one adopted by the original authors of the problem.

The Viennet1 test case is part of a series of problems proposed in [11] to evaluate multi-criteria optimization techniques. This problem involves three objectives.

In each test case, the MAS is executed 100 times with random starting points for each *design variable* and consistently converges towards the best (or one of the best) solution. As the performances of the system are not central to the topic at hand, we will only present some summarized results illustrating the convergence of the system. Once more, the interested reader can refer to [10] for a more detailed analysis of the performances.

These results are presented on Table 1. The first group of values represents the number of evaluations which was needed for respectively 10%, 50% and 90% of the instances to find the best solution. The second group represent the average distance to the best solution (trucated at 10^{-3}) among all instances at different times (0% being the start 100% being the end of the solving in the worst case).

Table 1. Summary of experiments results for the tests cases

	nb. evaluations to best			average distance to best			
	10%	50%	90%	0% (start)	30%	60%	100% (end)
Alexandrov	29	52	79	13109.169	803.126	5.685	0.059
Turbofan_o1	16	38	50	67.654	14.971	0.743	0.313
Turbofan_o2	10	23	35	23.876	1.853	0.143	0.101
viennet_o1	4	17	31	8.514	0.300	0.025	0.021
viennet_o2	4	15	30	9.412	0.320	0.02	0.02
viennet_o3	5	14	27	10.622	0.063	4.40E-004	1.68E-004

5 Conclusion

We have presented a model of numerical optimization problem and an agent-based optimization algorithm. While classical methods often have difficulties to handle complex continuous problems and require the use of specific methodologies, we distribute the problem among the agents in order to keep a low local complexity.

One of our concerns has been to facilitate the work of the engineer and allow him to express his problem in a way which is the most natural to him, instead of restricting him to a specific formulation. By analyzing the different concepts involved in the expression of a continuous problem, we extracted several atomic roles upon which we based the relations between the entities of our system. With these low-level entities, we are able to propose a way to represent continuous optimization problems as agents graphs ,which we name NDMO. This representation can reconstruct a great variety of problems while mirroring their original formulation. We applied NDMO by proposing a MAS capable of solving continuous optimization problem.

In the same way DCOP is a framework used by several MAS techniques to translate combinatorial optimization problems, this agent graph representation is not restricted to the specific MAS we presented but can be used as a base for multiple different techniques. With this work we achieved a proof-of-concept for the mostly unexplored field of continuous optimization using MAS.

We continue to work on the validation of our approach, as well as of the performances of our MAS, and already obtained interesting preliminary results concerning the extension of this modeling for uncertainty propagation, as well as optimization of scaled-up industrial test cases.

Acknowledgements. This work has been supported by French National Research Agency (ANR) through COSINUS program with ANR-09-COSI-005 reference.

References

1. Sobieszczanski-Sobieski, J., Haftka, R.T.: Multidisciplinary aerospace design optimization: Survey of recent developments. Structural Optimization 14, 1–23 (1996)

2. Cramer, E., Dennis Jr, J., Frank, P., Lewis, R., Shubin, G.: Problem formulation for multidisciplinary optimization. SIAM Journal on Optimization 4(4), 754–776 (1994)
3. Yi, S.I., Shin, J.K., Park, G.J.: Comparison of mdo methods with mathematical examples. Structural and Multidisciplinary Optimization 35(5), 391–402 (2008)
4. Kroo, I.M., Altus, S., Braun, R.D., Gage, P.J., Sobieski, I.P.: Multidisciplinary optimization methods for aircraft preliminary design. In: AIAA 5th Symposium on Multidisciplinary Analysis and Optimization (September 1994) AIAA 1994-4325
5. Sobieszczanski-Sobieski, J., Agte, J., Sandusky, R.: Bi-Level Integrated System Synthesis. NASA Langley Technical Report Server (1998)
6. Alexandrov, N., Lewis, R.: Analytical and computational aspects of collaborative optimization for multidisciplinary design. AIAA Journal 40(2), 301–309 (2002)
7. Perez, R., Liu, H., Behdinan, K.: Evaluation of multidisciplinary optimization approaches for aircraft conceptual design. In: AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY (2004)
8. Modi, P.J., Shen, W.M., Tambe, M., Yokoo, M.: An asynchronous complete method for distributed constraint optimization. In: International Conference on Autonomous Agents: Proceedings of the Second international Joint Conference on Autonomous Agents and Multiagent Systems, vol. 14, pp. 161–168 (2003)
9. Stranders, R., Farinelli, A., Rogers, A., Jennings, N.R.: Decentralised coordination of continuously valued control parameters using the max-sum algorithm. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems. International Foundation for Autonomous Agents and Multiagent Systems, vol. 1, pp. 601–608 (2009)
10. Jorquera, T., Georgé, J.P., Gleizes, M.P., Couellan, N., Noel, V., Régis, C.: A natural formalism and a multi-agent algorithm for integrative multidisciplinary design optimization. In: AAMAS 2013 Workshop: International Workshop on Optimisation in Multi-Agent Systems (to be published, May 2013)
11. Viennet, R., Fonteix, C., Marc, I.: Multicriteria optimization using a genetic algorithm for determining a pareto set. International Journal of Systems Science 27(2), 255–260 (1996)

Agent-Based Data Reduction Using Ensemble Technique

Ireneusz Czarnowski and Piotr Jędrzejowicz

Department of Information Systems, Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
{irek,pj}@am.gdynia.pl

Abstract. The problem addressed in this paper concerns data reduction. In the paper the agent-based data reduction algorithm is extended by adding mechanism of integration of the multiple learning models into a single multiple classification system called ensemble model. The paper includes the overview of the proposed approach and discusses the computational experiment results.

Keywords: learning from data, data reduction, ensemble, population learning algorithm, multi-agent system.

1 Introduction

The key objective of the machine learning is to design algorithms that are able to improve performance at some task through experience [20]. Such algorithms are called learners. The learner is learning using an example solution or a set of solutions to the particular problem type. Learning from examples is one of most popular paradigms of the machine learning. It deals with the problem of identifying regularities between a number of independent variables (attributes) and a target or dependent categorical variable observing and analyzing some given dataset [26]. Learning from examples is understood as the process of finding a model (or function) that describes and distinguishes data classes. The model produced under the machine learning process is called the classifier.

One of the current focuses of research in the field of machine learning are methods of selecting relevant information. Selection of the relevant information is the process often referred to as a *data reduction* [18], [23]. Data reduction techniques aim at decreasing the quantity of information required to learn a high quality classifiers. Data reduction can be achieved by selection of instances, by selection of attributes or by simultaneous reduction in both dimensions [6]. In practice, it means that data reduction approaches are concerned with selecting informative instances and, finally, with producing a minimal set of instances or prototypes to represent a training set and presenting the reduced dataset to a machine learning algorithm [28]. The data reduction through removing attributes which are irrelevant for the classification results is equally important. Many of attributes are not only irrelevant from the point of view of classification results but may also have a negative influence on the accuracy of machine classification and on the time required for classifier learning. Thus, data reduction carried-out without losing extractable information is considered as an important approach to increasing the effectiveness of the learning process.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 447–456, 2013.

In recent years, multiple model approaches are also rapidly growing and enjoying a lot of attention from the machine learning community due to their potential to greatly increase accuracy of the machine classification. Multiple model approaches aim at integration of the multiple learning models into one multiple classification system which is called the combined classification system or ensemble model. Classification systems using the multiple learning models are practical and effective solutions to different classification problems, when the single algorithms are rather specialized to solve some learning problems and cannot perform best in all situations [24]. As it was mentioned in [14], the aim of the integration of the multiple learning models is to achieve best possible accuracy. In [14] strategies for classifier integration are also referred to as the coverage optimization methods.

The paper deals with the learning from examples with data reduction. To solve the learning problem with data reduction an agent-based population learning algorithm is proposed. The basic assumption behind the population-based multi-agent approach (originally proposed in [12]) to learning from data is that the data are reduced through the instance selection. The approach solves the problem of deriving the classification model by integrating data reduction and learning stages using the advantages of the A-Team capabilities. It is generally accepted that the improvement of the quality of machine learning tools can be achieved through the integration of the data reduction with the learning process stage [1], [6], [8]. This has been also confirmed in one of the previous work of the authors [11].

In this paper the extended version of the agent-based data-reduction algorithm is proposed. The new feature of the proposed version of the algorithm is constructing the final classifier model through integration of multiple learning models stored in the common memory and produced by the so-called optimizing agents during the evolutionary search.

The goal of the paper is to show through computational experiment that the extended version of the agent-based data-reduction algorithm brings about an improvement to the classifier performance. In addition it is claimed that the current version can be competitive to its earlier version presented in [12], as well as to other combined classification systems known from the literature.

The paper is organized as follows. Section 2 contains some remarks on the scope of the ensemble learning and a review of techniques using the multi model approach. Main idea of the agent-based population learning algorithm is presented in Section 3. The discussed Section also provides details of the agent-based approach for learning classifiers from the reduced data. Section 4 gives description of the suggested extension involving the ensemble technique. Section 5 provides details on the computational experiment setup and discusses its results. Finally, the last section contains conclusions and suggestions for future research.

2 Ensemble Learning

Integration of the multiple learning models can be achieved in many ways. The so-called weak classifiers can be constructed by the same - homogenous, or different - heterogeneous, learning algorithms. One multiple classification system can use different learning algorithms on the same dataset. For example, based on such an approach, the stacking method [27] and meta-learning [9] have been proposed. Others combine multiple, homogenous learners, induced using the same dataset, but with different distributions, as it happens in the cases of bagging, boosting or random subspace methods [5]. Another approach to heterogeneous ensemble members is based on manipulation of attribute space. For example, the random subspace method constructs ensemble of classifiers by random selection of attributes for randomly formed dataset [15]. Heterogeneous learners are also produced in case of the random forests algorithm. This algorithm merges bagging strategy with random selection of attributes for different bootstrapped subset of instances [7]. In [19] the problem of integration of the multiple learning algorithms is discussed in the context of hybrid systems integrating some properties of both - heterogeneous and homogenous ensemble models.

The base models can be aggregated into a single multiple learning model also in several ways. Most ensemble models are constructed basing on the following architectures:

- Parallel – base models are independent but their results are aggregated.
- Sequential – an intermediate result has an impact on the next step result,
- Hybrid – combines features of the above architectures.

In the multiple classification system the base models can be linked in many ways depending on the adopted aggregation rule. It must be emphasized though, that in reality what is linked are the results of the predictions of these models. Having at one's disposal a fixed set of classifiers, the aim is to find an optimal combination of their decisions. More often strategies for optimal combination are referred to as decision optimization methods or decision combination methods [14]. In [29] decision combination methods are divided with regard to the method of generating the classification result into the following groups:

- Abstract-based methods - the final model predicts a class the observation belongs to. Among such approaches are the following: majority voting, weighted majority voting, plurality voting and Behaviour-Knowledge Space.
- Rank-based methods - the final model generates the set of classes arranged by their posterior probabilities.
- Measurement-based methods - the final model generates the vector of posterior probabilities. Typical aggregation rules of this group of methods are: generalized mean, weighted average, arithmetic average, probabilistic product, fuzzy integral and others.

In this paper the ensemble approaches is proposed for the purpose of improving quality of the agent-based data-reduction algorithm, originally shown in [12]. In this paper the ensemble technique is applied to combine results (base models) stored in the common memory. Each of base models is induced on the independent set of instances (set of prototypes) produced in the data reduction process. In particular, the discussed ensemble implementation bases on the parallel architecture. In next sections details of the proposed approach are described.

3 Agent-Based Algorithm for Learning Classifiers from the Reduced Data

The paper deals with the problem of learning from data with data reduction through applying the agent-based population learning algorithm. The main goal is to find the optimal learning model based on joint reduction of the original dataset and maximization of the classification quality criterion or criteria.

3.1 Main Features of the Agent-Based Population Learning Algorithm

It is well known that data reduction as well as classifier learning are computationally difficult combinatorial optimization problems (see, for example [10]). To overcome some of the difficulties posed by computational complexity of the data reduction problem it is proposed to apply the population-based approach with optimization procedures implemented as an asynchronous team of agents (A-Team).

The A-Team concept was originally introduced in [25]. The design of the A-Team architecture was motivated by other architectures used for optimization including blackboard systems and genetic algorithms. Within the A-Team multiple agents achieve an implicit cooperation by sharing a population of solutions, also called individuals, to the problem to be solved. An A-Team can be also defined as a set of agents and a set of memories, forming a network in which every agent remains in a closed loop. All the agents can work asynchronously and in parallel. Agents cooperate to construct, find and improve solutions which are read from the shared, common memory.

Main functionality of the agent-based population learning approach includes organizing and conducting the process of search for the best solution. It involves a sequence of the following steps:

- Generation of the initial population of solutions to be stored in the common memory.
- Activation of optimizing agents which execute some solution improvement algorithms applied to solutions drawn from the common memory and, subsequently, store them back after the attempted improvement in accordance with a user defined replacement strategy.
- Continuation of the reading-improving-replacing cycle until a stopping criterion is met. Such a criterion can be defined either or both as a predefined number of iterations or a limiting time period during which optimizing agents do not manage to improve the current best solution. After computation has been stopped the best solution achieved so far is accepted as the final one.

To implement the agent-based population learning algorithm one has to set and define the following:

- Solution representation format,
- Initial population of individuals,
- Fitness function,
- Improvement procedures,
- Replacement strategy implemented for managing the population of individuals.

More information on the population learning algorithm with optimization procedures implemented as agents within an asynchronous team of agents (A-Team) can be found in [3]. In [3] also several A-Team implementations are described.

3.2 Agent-Based Algorithm for Data Reduction

In our case the shared memory is used to store a population of solutions to the data reduction problem. The initial population is generated randomly. Each solution is represented by the set of prototypes i.e. by the compact representations of the original datasets. A feasible solution to the data reduction problem is encoded as a string consisting of numbers of selected reference instances. Such string is constructed at the initial population generation phase. When the initial population is generated at first, for each instance from the original set, the value of its similarity coefficient, proposed in [12], is calculated. Then instances with identical values of this coefficient are grouped into clusters. Further on, selection of the representation of instances through population-based search carried out by the team of agents for each cluster and removal of the remaining instances constitute basic steps of the first stage of the proposed procedure.

Each solution from the population is evaluated and the value of its fitness is calculated. The evaluation is carried out by estimating classification accuracy of the classifier, which is constructed taking into account the instances as indicated by the solution.

To solve the data reduction problem, the following two types of optimizing agents carrying-out different improvement procedures have been implemented: local search with the tabu list for instance selection and a simple local search for instance selection.

The first procedure - local search with tabu list for instance selection, modifies a solution by replacing a randomly selected reference instance with some other randomly chosen reference instance thus far not included within the improved solution. The modification takes place providing the replacement move is not on the tabu list. After the modification, the move is placed on the tabu list and remains there for a given number of iterations.

The second procedure - simple local search for instance selection, modifies the current solution either by removing the randomly selected reference instance and by adding some other randomly selected reference instance thus far not included within the improved solution.

In each of the above cases solutions that are forwarded to the optimizing agents for the improvement are selected from the population by the so-called working strategy.

4 Learning Model Integration for Data Reduction

In the proposed approach agents cooperate through selecting and modifying solutions according to the working strategy. The modifications of the solution carried out by agents can be constructive or destructive, that means, that an attempt to improve a solution can be successful or unsuccessful.

The working strategy is an implementation of a set of rules for agent communication, selection of solution to be improved and management of the population of solutions which are kept in the common memory. Different working strategies with respect to selecting solutions to be improved by the A-Team members and replacing the solutions stored in the common memory by the improved ones can be applied. However, as it was shown in [4], the choice of the working strategy influences the quality of solutions and working strategies are not statistically equally effective. Thus, the quality of solutions produced by the A-Team depends also, among some other factors, on the working strategy used.

It is also possible to use in parallel a number of different working strategies defined within the A-Team, each of them specifying:

- How the initial population of solutions is created.
- How to choose solutions which are forwarded to the optimizing agents for improvement.
- How to merge the improved solutions returned by the optimizing agents with the whole population (for example they may be added, or may replace random or worse solutions).
- When to stop searching for better solutions (for example after a given time, or after no better solution has been found within a predefined period of time).

In [10] two working strategies have been evaluated with respect to the quality of data reduction solutions obtained by the A-Team algorithm. These included:

- RM-RR – selection of individuals for improvement is a purely random move i.e. each optimizing agent receives a solution drawn at random from the population of solutions. A returning individual replaces randomly selected individual in the common memory.
- RM-RW - selection of individuals for improvement is a purely random move i.e. each optimizing agent receives a solution drawn at random from the population of solutions. A returning individual replaces first worst individual. Replacement does not take place if there is no worse individual in the common memory.

The experiment reported in [10], shows that in a majority of cases working strategy based on selection with random moves and random replacement assures the best level of performance of the agent-based data reduction algorithm.

In this paper an extended version of the agent-based approach to learning from the reduced data is proposed. The new version is based on the belief that the quality of the system can be increased by including all solutions from the common memory in the final classification model. As it has been observed by [17] "Empirically, ensembles tend to yield better results when there is significant diversity among the models, for that reason, many ensemble methods seek to promote diversity among the models they combine".

In case of the agent-based algorithm the population, in each step, is updated according to the implemented working strategy. If the working strategy is seeking best solutions among the possible, storing potential candidates in the common memory, it seems reasonable to use such a population of solutions after they have become mature, to form an ensemble solution. Based on this assumption, we propose to use all thus evolved solution as base models immediately after the stopping criterion has been reached by the respective A-Team. Next the base models are aggregated into a single multiple learning model.

For classification problems the final model predicts the class the observation belongs to. This prediction is obtained through application of the majority voting scheme.

5 Computational Experiment

This section contains the results of several computational experiments carried out with a view to evaluate the performance of the proposed approach measured in terms of the classification accuracy. In particular the reported experiment aimed at answering the question whether the new extended version of the agent-based data-reduction called *ABDRE* (Agent-Based Data-Reduction with Ensemble) performs better than its earlier version – *ABIS*, introduced in [12]. The experiment allowed also to study and evaluate how the choice of the working strategy may influence the quality of the classification system. The *ABDRE* has been also compared with other well known ensemble classifiers i.e. AdaBoost, Bagging and Random Subspace Method.

Generalization accuracy has been used as the performance criterion. The learning tool was the C4.5 algorithm [21]. The C4.5 algorithm has been also applied to induce all of the base models for all ensemble classifiers.

The size of bags has been set to 50% of the original training set for Bagging and Random Subspace Method, respectively. The number of base models for the compared ensemble classifiers has been set to 40. Population size for each investigated A-Team, following earlier experiment results presented in [10], was set to 40.

The process of searching for the best solution of each A-Team has been stopped either after 100 iterations or after there has been no improvement of the current best solution for one minute of computation. Values of these parameters have been set arbitrarily.

In all cases the algorithms have been applied to solve respective problems using several benchmark datasets obtained from the UCI Machine Learning Repository [2]. Basic characteristics of these datasets are shown in Table 1.

Each benchmark problem has been solved 50 times, and the experiment plan involved 10 repetitions of the 10-cross-validation scheme. The reported values of the quality measure have been averaged over all runs. The quality measure was the correct classification ratio – accuracy.

Table 2 shows mean values of the classification accuracy of the classifiers obtained using the *ABDRE* approach (i.e. using the set of prototypes found by selecting instances and using different implemented working strategies). Table 2 also contains results obtained by other ensemble classifiers and some example non-ensemble classifiers.

Table 1. Datasets used in the reported experiment

Dataset	Number of instances	Number of attributes	Number of classes	Best reported results classification accuracy
Heart	303	13	2	90.0% [8]
Diabetes	768	8	2	77.34%[16]
WBC	699	9	2	97.5% [2]
Australian credit (ACredit)	690	15	2	86.9% [2]
German credit (GCredit)	1000	20	2	77.47%[16]
Sonar	208	60	2	97.1% [2]

Table 2. Accuracy of the classification results (%)

Algorithm	Heart	Diabetes	WBC	ACredit	GCredit	Sonar
ABDRE with RM-RR	**92.84**	**80.4**	96.4	90.8	**78.2**	83.4
ABDRE with RM-RW	90.84	78.07	**97.6**	89.45	76.28	81.75
ABIS (earlier version)	91.21	76.54	97.44	90.72	77.7	83.65
AdaBoost	82.23	73.55	63.09	**91.05**	73.01	86.09
Bagging	79.69	76.37	95.77	85.87	74.19	76.2
Random Subspace Method	84.44	74.81	71.08	82.14	75.4	85.18
C 4.5 [13]	77.8	73	94.7	84.5	70.5	76.09
SVM [13]	81.5	77	97.2	84.813	72.5	**90.413**
DROP 4 [28]	80.90	72.4	96.28	84.78	**-**	82.81

When the versions of the *ABDRE* algorithm are compared as shown in Table 2, it can be observed that the best results have been obtained by the *ABDRE* algorithm with the RM-RR strategy. The RM-RR strategy outperforms the RM-RW strategy and such observation confirms earlier findings presented in [10]. It should be also noted that the experiment supported hypotheses that the working strategy influences the quality of the agent-based data reduction classification system with ensemble technique.

The *ABDRE* algorithm assures the required size of the reduced dataset and is competitive to the it earlier version. The proposed algorithm is also competitive, in some cases, to the other ensemble classifiers, i.e. AdaBoost or Bagging, and extends set of the ensemble approaches.

6 Conclusions

Main result of the reported research is proposing and evaluating an extended version of the agent-based data reduction algorithm incorporating the mechanism of integration of the multiple learning models into a single multiple classification system called ensemble model. From the reported experiment it can be concluded that the proposed approach is a useful tool allowing to obtain high quality ensemble models for machine learning. The experiment, due to its limited scope, allows only for a preliminary validation of the approach. Further research should aim at investigating different techniques for producing the ensemble model as well as carrying-out more extensive experiments.

Acknowledgments. This research has been supported by the Polish Ministry of Science and Higher Education with grant no. N N519 576438 for years 2010-2013.

References

1. Aksela, M.: Adaptive Combinations of Classifiers with Application to On-line Handwritten Character Recognition. Department of Computer Science and Engineering. Helsinki University of Technology, Helsinki (2007)
2. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
3. Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: e-JABAT - An Implementation of the Web-based A-Team. In: Nguyen, N.T., Jain, L.C. (eds.) Intelligence Agents in the Evolution of Web and Applications. SCI, vol. 167, pp. 57–86. Springer, Heidelberg (2009)
4. Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: Influence of the Working Strategy on A-Team Performance. In: Szczerbicki, E., Nguyen, N.T. (eds.) Smart Information and Knowledge Management. SCI, vol. 260, pp. 83–102. Springer, Heidelberg (2010)
5. Bauer, E., Kohavi, R.: An Empirical Comparison of Voting Classification Algorithhms: Bagging, Boosting and Variants. Machine Learning 36(1-2), 691–707 (1994)
6. Bhanu, B., Peng, J.: Adaptive Integration Image Segmentation and Object Recognition. IEEE Trans. on Systems, Man and Cybernetics 30(4), 427–444 (2000)
7. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
8. Bull, L.: Learning Classifier Systems: A Brief Introduction. In: Bull, L. (ed.) Applications of Learning Classifier Systems. STUDFUZZ, vol. 150, pp. 1–12. Springer, Heidelberg (2004)
9. Chan, P.K., Stolfo, S.J.: Experiments on Multistrategy Learning by Meta-Learning. In: Second International Conference on Information and Knowledge Management, pp. 31–45 (1993)
10. Czarnowski, I., Jędrzejowicz, P.: Experimental Evaluation of the Agent-Based Population Learning Algorithm for the Cluster-Based Instance Selection. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS (LNAI), vol. 6923, pp. 301–310. Springer, Heidelberg (2011)

11. Czarnowski, I., Jędrzejowicz, P.: An Approach to Data Reduction and Integrated Machine Classification. New Generation Computing 28, 21–40 (2010)
12. Czarnowski, I.: Distributed Learning with Data Reduction. In: Nguyen, N.T. (ed.) TCCI IV 2011. LNCS, vol. 6660, pp. 3–121. Springer, Heidelberg (2011)
13. Datasets used for classification: comparison of results. In. directory of data sets, `http://www.is.umk.pl/projects/datasets.html` (accessed September 1, 2009)
14. Ho, T.K.: Data Complexity Analysis for Classifier Combination. In: Kittler, J., Roli, F. (eds.) MCS 2001. LNCS, vol. 2096, pp. 53–67. Springer, Heidelberg (2001)
15. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transaction on PAMI 19(8), 832–844 (1998)
16. Jędrzejowicz, J., Jędrzejowicz, P.: Cellular GEP-Induced Classifiers. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part I. LNCS, vol. 6421, pp. 343–352. Springer, Heidelberg (2010)
17. Kuncheva, L., Whitaker: Measures of diversity in classifier ensembles. Machine Learning 51, 181–207 (2003)
18. Liu, H., Lu, H., Yao, J.: Identifying Relevant Databases for Multidatabase Mining. In: Proceeding of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 210–221 (1998)
19. Michalski, R.S., Tecuci, G.: Machine Learning. A Multistrategy Approach, vol. IV. Morgan Kaufmann (1994)
20. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
21. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, SanMateo (1993)
22. Rozsypal, A., Kubat, M.: Selecting Representative Examples and Attributes by a Genetic Algorithm. Intelligent Data Analysis 7(4), 291–304 (2003)
23. Silva, J., Giannella, C., Bhargava, R., Kargupta, H., Klusch, M.: Distributed Data Mining and Agents. Engineering Applications of Artificial Intelligence Journal 18, 791–807 (2005)
24. Stefanowski, J.: Multiple and Hybrid Classifiers. In: Polkowski, L. (ed.) Formal Methods and Intelligent Techniques in Control, Decision Making. Multimedia and Robotics, Warszawa, pp. 174–188 (2001)
25. Talukdar, S., Baerentzen, L., Gove, A., de Souza, P.: Asynchronous Teams: Co-operation Schemes for Autonomous, Computer-Based Agents, Technical Report EDRC 18-59-96, Carnegie Mellon University, Pittsburgh (1996)
26. Tsoumakas, G., Angelis, L., Vlahavas, I.: Clustering Classifiers for Knowledge Discovery from Physically Distributed Databases. Data & Knowledge Engineering 49, 223–242 (2004)
27. Wei, Y., Li, T., Ge, Z.: Combining Distributed Classifies by Stacking. In: Proceedings of the Third International Conference on Genetic and Evolutionary Computing, pp. 418–421 (2009)
28. Wilson, D.R., Martinez, T.R.: Reduction Techniques for Instance-based Learning Algorithm. Machine Learning 33(3), 257–286 (2000)
29. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of Combining Multiple Classifiers and their Application to Handwriting Recognition. IEEE Transaction on Systems, Man and Cybernetics 22, 418–435 (1992)

Reinforcement Learning Strategy for A-Team Solving the Resource-Constrained Project Scheduling Problem

Piotr Jędrzejowicz and Ewa Ratajczak-Ropel

Department of Information Systems
Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
{pj,ewra}@am.gdynia.pl

Abstract. In this paper the strategy for the A-Team with Reinforcement Learning (RL) for solving the Resource Constrained Project Scheduling Problem (RCPSP) is proposed and experimentally validated. The RCPSP belongs to the NP-hard problem class. To solve this problem a team of asynchronous agents (A-Team) has been implemented using JABAT multiagent system. An A-Team is the set of objects including multiple agents and the common memory which through interactions produce solutions of optimization problems. These interactions are usually managed by the static strategy. In this paper the dynamic learning strategy is suggested. The proposed strategy based on reinforcement learning supervises interactions between optimization agents and the common memory. To validate the approach computational experiment has been carried out.

1 Introduction

Resource Constrained Project Scheduling Problem (RCPSP) have attracted a lot of attention and many exact and heuristic algorithms have been proposed for solving it [1, 12, 15]. The current approaches to solve this problem produce either approximate solutions or can only be applied for solving instances of the limited size. Hence, searching for more effective algorithms and solutions to the RCPSP problem is still a lively field of research. One of the promising directions of such research is to take advantage of the parallel and distributed computation solutions, which are the common feature of the contemporary multiagent systems [24].

Modern multiagent system architectures are an important and intensively expanding area of research and development. There is a number of multiple-agent approaches proposed to solve different types of optimization problems. One of them is the concept of an A-Team, originally introduced by Talukdar et al. [21]. The idea of A-Team was used to develop the environment for solving a variety of computationally hard optimization problems called JABAT [2, 16]. JADE based A-Team (JABAT) system supports the construction of the dedicated A-Team

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 457–466, 2013.

architectures. Agents used in JABAT assure decentralization of computation across multiple hardware platforms. Parallel processing results in more effective use of the available resources and ultimately, a reduction of the computation time.

Reinforcement Learning (RL) [6, 20, 17] belongs to the category of unsupervised machine learning algorithms. It can be described as learning which action to take in a given situation (state) to achieve one or more goal(s). The learning process takes place through interaction with an environment. RL is usually used in solving combinatorial optimization problems at three levels [23]:

The direct level - RL is directly applied to the problem.
The metaheuristic level - RL is used as a component of the respective metaheuristic.
The hyperheuristic level - RL is used as a component of the respective hyperheuristic.

The other field where RL is commonly used is Multi-Agent Reinforcement Learning (MARL) where multiple reinforcement learning agents act together in a common environment [10, 22]. In this paper the RL is used to support strategy of searching for optimal solution by the team of agents.

An A-Team is a system composed of the set of objects including multiple agents and the common memory which through interactions produce solutions of optimization problems. Several strategies controlling the interactions between agents and memories have been recently proposed and experimentally validated. The influence of such strategy on A-Team performance was investigated by Barbucha at al. [3]. In [5] the reinforcement learning based strategy for synchronous team of agents has been considered. The similar topics were also considered by other authors for different multi-agent systems, e.g. [11, 19].

In this paper the RL based on utility values to learn interaction strategy for the A-Team solving the RCPSP problem is proposed and experimentally validated. The new concept is using the RL to control the strategy parameters and individuals selection instead of metaheuristics parameters and selection. It is expected that introducing the proposed RL will result in obtaining high quality solutions in an efficient manner.

Optimization agents used to produce solutions to the RCPSP instances in the considered A-Team represent four metaheuristic algorithms: local search, tabu search, crossing and path relinking. JABAT system implementation adopted for solving RCPSP problem proposed in [13] has been used.

The paper is constructed as follows: Section 2 contains the RCPSP problem formulation. Section 3 gives some information on JABAT environment. Section 4 provides details of the proposed RL based JABAT implementation. Section 5 describes settings of the computational experiment carried-out with a view to validate the proposed approach and contains a discussion of the computational experiment results. Finally, Section 6 contains conclusions and suggestions for future research.

2 Problem Formulation

A single-mode resource-constrained project scheduling problem consists of the set of n activities, where each activity has to be processed without interruption to complete the project. The dummy activities 1 and n represent the beginning and the end of the project. The duration of an activity j, $j = 1, \ldots, n$ is denoted by d_j where $d_1 = d_n = 0$. There are r renewable resource types. The availability of each resource type k in each time period is r_k units, $k = 1, \ldots, r$. Each activity j requires r_{jk} units of resource k during each period of its duration, where $r_{1k} = r_{nk} = 0$, $k = 1, ..., r$. All parameters are non-negative integers. There are precedence relations of the finish-start type with a zero parameter value (i.e. $FS = 0$) defined between the activities. In other words activity i precedes activity j if j cannot start until i has been completed. The structure of a project can be represented by an activity-on-node network $G = (SV, SA)$, where SV is the set of activities and SA is the set of precedence relationships. SS_j (SP_j) is the set of successors (predecessors) of activity j, $j = 1, \ldots, n$. It is further assumed that $1 \in SP_j$, $j = 2, \ldots, n$, and $n \in SS_j$, $j = 1, \ldots, n-1$. The objective is to find a schedule S of activities starting times $[s_1, \ldots, s_n]$, where $s_1 = 0$ and resource constraints are satisfied, such that the schedule duration $T(S) = s_n$ is minimized.

The above formulated problem as a generalization of the classical job shop scheduling problem belongs to the class of NP-hard optimization problems [8]. The considered problem class is denoted as PS$|prec|C_{max}$ [9].

The objective is to find a minimal schedule in respect of the makespan that meets the constraints imposed by the precedence relations and the limited resource availabilities.

3 The JABAT Environment

JABAT is the environment facilitating the design and implementation of the A-Team architecture for solving various combinatorial optimization problems. The problem-solving paradigm on which the proposed system is based can be best defined as the population-based approach.

JABAT produces solutions to combinatorial optimization problems using a set of optimization agents, each representing an improvement algorithm. Each improvement (optimization) algorithm when supplied with a potential solution to the problem at hand, tries to improve this solution. The initial population of solutions (individuals) is generated or constructed. Individuals forming the initial population are, at the following computation stages, improved by independently acting optimization agents. The main functionality of the proposed environment includes organizing and conducting the process of search for the best solution.

The behavior of an A-Team is controlled by the, so called, interaction strategy defined by the user. An A-Team uses a population of individuals (solutions) and a number of optimization agents. All optimization agents within the A-Team work together to improve individuals from its population in accordance with the interaction strategy.

To implement the proposed architecture, the following classes of agents need to be designed and implemented:

SolutionManager - represents and manages the A-Team e.g. the set of optimization agents and the population of solutions stored in the common memory.

OptiAgent - represents a single improving algorithm (e.g. local search, simulated annealing, genetic algorithm etc.).

Other important classes in JABAT include TaskManager and PlatformManager which are used to initialize the agents and maintain the system. Objects of these classes also act as agents. To describe the problem Task class representing an instance of the problem and Solution class representing the solution is used. Additionally, an interaction strategy is used to define a set of rules applicable to managing and maintaining a population of current solutions in the common memory.

JABAT has been designed and implemented using JADE (Java Agent Development Framework), which is a software framework proposed by TILAB [7] supporting the implementation of the multiagent systems. More detailed information about the JABAT environment and its implementations can be found in [2, 4, 16].

4 A-Team with Interaction Strategy Controlled by RL

JABAT was successfully used by the authors for solving the RCPSP, MRCPSP and RCPSP/max problems ([13, 14, 4]. To extend JABAT to solving RCPSP problem the set of class and agents were implemented as described in [13, 15].

Classes describing the problem are responsible for reading and preprocessing of the data and generating random instances of the problem. The discussed set includes the following classes: RCPSPTask, RCPSPSolution, Activity, Mode, Resource.

The second set includes classes describing the optimization agents. Each of them includes the implementation of an optimization algorithms used to solve the RCPSP problem. All of them are inheriting from OptiAgent class. These implement specialist algorithms: LSA, TSA, CA and PRA described below. The prefix Opti is assigned to each agent with its embedded algorithm:

OptiLSA - implementing the Local Search Algorithm (LSA),
OptiTSA - implementing the Tabu Search Algorithm (TSA),
OptiCA - implementing the Crossover Algorithm (CA),
OptiPRA - implementing the Path Relinking Algorithm (PRA).

In our earlier approaches the static interaction strategies were used, including:

Basic - where random solution from the population is read and sent to the optimization agent. Next solution sent by an agent replaces random solution from the population.

Blocking1 - where additionally the solutions send to optimization agents are blocked. If the number of non blocked solutions is less then the number of solutions needed by optimization agent the blocked solutions are released.

Blocking2 - where a random worse solution from the population is replaced by the one sent by optimization agent and additionally one new solution is generated randomly every fixed number of iteration.

Other important parameters that describe the interaction strategies are Population size Stopping criterion.

Initially a solution returned by an optimization agent replaces its original version that is a version before the attempted improvement. RL is than used at the following levels:

RL1 - in which RL controls the replacement of one individual from the population by other randomly generated one.

RL2 - in which, additionally, the method of choosing the individual for replacement is controlled by RL.

RL3 - in which individuals in the population are grouped according to certain features, and next the procedure of choosing individuals to be forwarded to optimization agents from respective groups is controlled by RL.

For each level of learning the environment state is remembered and used in the learning procedure. This state includes: the best individual, the population average diversity and the features of each individual in the population. The state is calculated every fixed number of iteration

$$itNS = \lfloor PopulationSize/AgentsNumber \rfloor .$$

The proposed RL is based on utility values reinforcement proposed in [18, 5].

In case of RL1 the probability of randomly generating a new solution P_{rg} is calculated directly. Initially it is set to 0 to intensify the exploitation within an initial population. The P_{rg} is increased in two cases: where the average diversification in population decreases and where it decreases deeply. The P_{rg} is decreased in three environment states: where average diversity in the population increases, where the best solution is found and where the new solution is randomly generated.

To reduce the computation time, average diversity of the population is calculated by comparison with the best solution only. Diversity of two solutions for RCPSP problem is calculated as the sum of differences between activities starting times in the projects.

In the RL2, additionally, the probability of choosing the method of adding the new randomly generated individual to the population is considered. Three methods are possible: the new solution replaces the random one in the population, the new solution replaces the random worse one or the new solution replaces the worst solution in the population. Experiments show that replacing the worse and worst solution is good to intensify exploitation while replacing the random one intensifies exploration. The weight w_m for each method is calculated, where $m \in M$, $M = \{random, worse, worst\}$. The w_{random} is increased where the population average diversity decreases and it is decreased in the opposite case. The w_{worse} and w_{worst} is decreased where the population average diversity decreases and they are increased in the opposite case. The probability of choosing the method m is calculated as

$$P_m = \frac{w_m}{\sum_{i \in M} w_i} .$$

In case of the RL3 the similar RL utility values are introduced for groups of solutions and kinds of algorithms. The RL is used during solving one instance of the problem. For each solution the average number of unused resources in each period of time $\overline{f_{URP}}$ and average lateness of activities $\overline{f_{LA}}$ are remembered. These two features allow to group solutions from the population and allocate the optimization algorithms to such solutions for which they are expected to achieve the best results. For each algorithm the matrix of weights allocated to each solution feature is remembered $W^{Alg}_{\overline{f_{URP}},\overline{f_{LA}}} = \left[w^{Alg}_{i,j}\right]$, where $i \in F_{URP}$ and $j \in F_{LA}$. The appropriate weight is increased where the optimization agent implementing given kind of algorithm returns better or best solution (positive reinforcement). The appropriate weight is decreased where the optimization agent returns non better solution (negative reinforcement). The probability of choosing the appropriate algorithm is calculated as

$$P^{Alg}_{\overline{f_{URP}},\overline{f_{LA}}} = \frac{w^{Alg}_{\overline{f_{URP}},\overline{f_{LA}}}}{\sum_{i \in F_{URP}, j \in F_{LA}} w^{Alg}_{ij}} ,$$

where F_{URP} and F_{LA} denote the sets of average unused resources in each period of time and average activity lateness, respectively.

5 Computational Experiment

5.1 Settings

To evaluate the effectiveness of the proposed approach the computational experiment has been carried out using benchmark instances of RCPSP from PSPLIB[1] - test sets: sm30 (single mode, 30 activities), sm60, sm90, sm120. Each of the first three sets includes 480 problem instances while set sm120 includes 600. The experiment involved computation with a fixed number of optimization agents, fixed population size, and the stopping criteria indicated by the number of no improvement iterations where the environment state has been calculated ($itNSnoImp$).

In the experiment the following parameters have been used:

- $PopulationSize = 30$,
- $itNSnoImp = PopulationSize = 30$,
- $AgentsNumber = 4$ (The proposed TA-Teams includes four kinds of optimization agents representing the LSA, TSA, CA and PRA algorithms described in Section 4),
- $itNS = \lfloor PopualtionSize/AgentsNumber \rfloor = 7$,
- initial values for probabilities and weights:

[1] See PSPLIB at `http://129.187.106.231/psplib`

- P_{rg}=0,
- w_m: $w_{random} = 30$, $w_{worse} = 60$, $w_{worst} = 10$,
- $w^{Alg}_{f_{URP}, f_{LA}} = 1$.

In case of the positive reinforcement the additive adaptation for the weights is used: $w = w+1$, and in case of the negative reinforcement the additive $w = w-1$ or root adaptation $w = \sqrt{w}$ is used. If the utility value falls below 1, it is reset to 1; if the utility value exceeds a certain max_w, it is reset to max_w if the utility value is assigned a non-integer value, it is rounded down. These update schemes proved effective in [18, 5].

The experiment has been carried out using nodes of the cluster Holk of the Tricity Academic Computer Network built of 256 Intel Itanium 2 Dual Core 1.4 GHz with 12 MB L3 cache processors and with Mellanox InfiniBand interconnections with 10Gb/s bandwidth. During the computation one node per eight optimization agents was used.

5.2 Results

During the experiment the following characteristics of the computational results have been calculated and recorded: Mean Relative Error (MRE) calculated as the deviation from the optimal solution for sm30 set or from the best known results and the Critical Path Lower Bound (CPLB) for sm60, sm90 and sm120 sets, Mean Computation Time (MCT) required to find the best solution and Mean Total Computation Time (MTCT). Each instance has been solved five times and the results have been averaged over these solutions.

The computational experiment results are presented in Tables 1-2. In each case all solutions were feasible. The results obtained for the proposed approach are good and very promissing. The mean relative error below 1% in the case of 30, 60 and 90 activities and below 3% in the case of 120 activities have been obtained. The maximal relative error is below 7% and 9% respectively. It should be noted that introducing each level of learnig to the RL strategy improves the results and simultaneously insignificantly influences the computation time. Hence additional or other lesarning parameters or methods could be considerd.

The presented results are comparable with the results reported in the literature. In Table 3 the results obtained by the heuristic algorithms compared in [1, 12] are presented. However in case of the agent based approach it is difficult to compare computation times as well as the number of schedules, which is another widely used measure of the algorithm efficiency. In the proposed agent-based approach computation times as well as number of schedules differ between nodes and optimization agent algorithms working in parallel. The results obtained be a single agent may or may not influence the results obtained by the other agents. Additionally the computation time includes the time used by agents to prepare, send and receive messages.

The experiment results show that the proposed implementation is effective and using reinforcement learning to control different elements of the interaction strategy in A-Teams architecture is beneficial.

Table 1. Results for benchmark test set sm30 (RE from the optimal solution)

Strategy	MRE	MCT [s]	MTCT [s]
Blocking2	0.028%	6.43	72.62
RL1	0.031%	2.23	35.79
RL2	0.025%	2.29	36.02
RL3	0.018%	2.18	34.96

Table 2. Results for benchmark test sets sm60, sm90 and sm120 (RE from the best known solution and CPLB)

Set	Strategy	MRE from the best known solution	MRE from the CPLB	MCT [s]	MTCT [s]
sm60	Blocking2	0.64%	11.44%	32.70	75.56
	RL1	0.56%	11.23%	11.72	56.49
	RL2	0.51%	11.15%	12.84	64.41
	RL3	0.51%	11.16%	12.50	62.30
sm90	Blocking2	1.19%	11.38%	36.05	74.51
	RL1	1.08%	11.05%	19.66	56.20
	RL2	1.04%	11.00%	19.70	56.61
	RL3	0.98%	10.91%	24.23	77.29
sm120	Blocking2	3.17%	34.43%	78.38	137.10
	RL1	3.10%	34.25%	87.89	170.02
	RL2	2.98%	34.02%	90.21	183.65
	RL3	2.91%	33.27%	88.11	202.41

Table 3. Literature reported results [1, 12]

Set	Algorithm	Authors	MRE	MCT [s]	Computer
sm30	Decompos. & local opt	Palpant et al.	0.00	10.26	2.3 GHz
	VNS–activity list	Fleszar, Hindi	0.01	5.9	1.0 GHz
	Local search–critical	Valls et al.	0.06	1.61	400 MHz
sm60	PSO	Tchomte et al.	9.01	–	–
	Decompos. & local opt	Palpant et al.	10.81	38.8	2.3 GHz
	Population–based	Valls et al.	10.89	3.7	400 MHz
	Local search–critical	Valls et al.	11.45	2.8	400 MHz
sm90	Filter and fan	Ranjbar	10.11	–	–
	Decomposition based GA	Debels, Vanhoucke	10.35	–	–
	GA–hybrid, FBI	Valls at al.	10.46	–	–
sm120	Filter and fan	Ranjbar	31.42	–	–
	Population-based	Valls et al.	31.58	59.4	400 MHz
	Decompos. & local opt.	Palpant et al.	32.41	207.9	2.3 GHz
	Local search–critical	Valls et al.	34.53	17.0	400 MHz

6 Conclusions

The computational experiment results show that the proposed dedicated A-Team architecture supported by Reinforcement Learning to control the strategy parameters is an effective and competitive tool for solving instances of the RCPSP problem. Presented results are comparable with solutions known from the literature and in some cases outperform them. It can be also noted that they have been obtained in a comparable time. However, in this case time comparison may be misleading since the proposed TA-Teams have been run using different numbers and kinds of processors. In case of the agent-based environments the significant part of the time is used for agent communication which has an influence on both - computation time and quality of the results.

The presented experiment should be extended to examine the other RL schemes, especially other positive and negative reinforcement adaptations. The other policy iteration methods as for example Learning Automata should also be investigated. On the other hand different parameters of environment state and solutions could be considered.

Acknowledgments. The research has been supported by the Ministry of Science and Higher Education grant no. N N519 576438 for years 2010 – 2013. Calculations have been performed in the Academic Computer Centre TASK in Gdańsk.

References

[1] Agarwal, A., Colak, S., Erenguc, S.: A Neurogenetic Approach for the Resource–Constrained Project Scheduling Problem. Computers & Operations Research 38, 44–50 (2011)

[2] Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: E-JABAT - An Implementation of the Web-Based A-Team. In: Nguyen, N.T., Jain, L.C. (eds.) Intelligent Agents in the Evolution of Web and Applications. SCI, vol. 167, pp. 57–86. Springer, Heidelberg (2009)

[3] Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: Influence of the Working Strategy on A-Team Performance. In: Szczerbicki, E., Nguyen, N.T. (eds.) Smart Information and Knowledge Management. SCI, vol. 260, pp. 83–102. Springer, Heidelberg (2010)

[4] Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: Parallel Cooperating A-Teams. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS (LNAI), vol. 6923, pp. 322–331. Springer, Heidelberg (2011)

[5] Barbucha, D.: Search Modes for the Cooperative Multi-agent System Solving the Vehicle Routing Problem, Intelligent and Autonomous Systems. Neurocomputing 88, 13–23 (2012)

[6] Barto, A.G., Sutton, R.S., Anderson, C.W.: Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, 835–846 (1983)

[7] Bellifemine, F., Caire, G., Poggi, A., Rimassa, G.: JADE. A White Paper, Exp. 3(3), 6–20 (2003)

[8] Błażewicz, J., Lenstra, J., Rinnooy, A.: Scheduling subject to resource constraints: Classification and complexity. Discrete Applied Mathematics 5, 11–24 (1983)

[9] Brucker, P., Drexl, A., Möhring, R., Neumann, K., Pesch, E.: Resource-Constrained Project Scheduling: Notation, Classification, Models, and Methods. European Journal of Operational Research 112, 3–41 (1999)

[10] Busoniu, L., Babuska, R., De Schutter, B.: A Comprehensive Survey of Multiagent Reinforcement Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C:Applications and Reviews 38(2), 156–172 (2008)

[11] Cadenas, J.M., Garrido, M.C., Muñoz, E.: Using machine learning in a cooperative hybrid parallel strategy of metaheuristics. Information Sciences 179(19), 3255–3267 (2009)

[12] Hartmann, S., Kölisch, R.: Experimental Investigation of Heuristics for Resource-Constrained Project Scheduling: An Update. European Journal of Operational Research 174, 23–37 (2006)

[13] Jędrzejowicz, P., Ratajczak-Ropel, E.: New Generation A-Team for Solving the Resource Constrained Project Scheduling. In: Proc. the Eleventh International Workshop on Project Management and Scheduling, Istanbul, pp. 156–159 (2008)

[14] Jędrzejowicz, P., Ratajczak-Ropel, E.: Solving the RCPSP/max Problem by the Team of Agents. In: Håkansson, A., Nguyen, N.T., Hartung, R.L., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2009. LNCS, vol. 5559, pp. 734–743. Springer, Heidelberg (2009)

[15] Jędrzejowicz, P., Ratajczak-Ropel, E.: Team of A-Teams for Solving the Resource-Constrained Project Scheduling Problem. In: Grana, M., et al. (eds.) Advances in Knowledge-Based and Intelligent Information and Engineering Systems, pp. 1201–1210. IOS Press (2012)

[16] Jędrzejowicz, P., Wierzbowska, I.: JADE-Based A-Team Environment. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 719–726. Springer, Heidelberg (2006)

[17] Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal of Artificial Intelligence Research 4, 237–285 (1996)

[18] Nareyek, A.: Choosing search heuristics by non-stationary reinforcement learning. Metaheuristics: Computer Decision-Making, 523–544 (2001)

[19] Pelta, D., Cruz, D., Sancho-Royo, C., Verdegay, A., Using, J.L.: Using memory and fuzzy rules in a cooperative multi-thread strategy for optimization. Information Sciences 176(13), 1849–1868 (2006)

[20] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)

[21] Talukdar, S., Baerentzen, L., Gove, A.: P. de Souza: Asynchronous Teams: Cooperation Schemes for Autonomous, Computer-Based Agents. Technical Report EDRC 18-59-96. Carnegie Mellon University, Pittsburgh (1996)

[22] Tuyls, K., Weiss, G.: Multiagent learning: Basics, challenges, prospects. AI Magazine 33(3), 41–53 (2012)

[23] Wauters, T.: Reinforcement learning enhanced heuristic search for combinatorial optimization. Doctoral thesis, Department of Computer Science, KU Leuven (2012)

[24] Wooldridge, M.: An Introduction to MultiAgent Systems, 2nd edn. John Wiley & Sons (2009)

A Multi-agent Approach to the Dynamic Vehicle Routing Problem with Time Windows

Dariusz Barbucha

Department of Information Systems
Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
d.barbucha@wpit.am.gdynia.pl

Abstract. The term Dynamic Vehicle Routing refers to a wide range of transportation problems where the required information is not given a priori to the decision maker but is revealed concurrently with the decision-making process, where the goal of such process is to provide the required transportation and at the same time minimize service cost subject to various constraints including vehicle and fleet capacities. The most common source of dynamism in vehicle routing problem is the online arrival of customer during the operations, which increases the complexity of decisions and introduces new challenges while finding the optimal route plan. The paper proposes a new agent-based approach to the Dynamic Vehicle Routing Problem with Time Windows, in which two different dynamic order dispatching strategies are considered. Their influence on the results are investigated and identified in the computational experiment.

Keywords: Dynamic Vehicle Routing Problem with Time Windows, Multi-agent Systems, Cooperative Problem Solving.

1 Introduction

One of the most important problem in the contemporary transport companies is the Vehicle Routing Problem where a set of vehicles have to deliver (or pickup) goods or persons to (from) locations situated in a given area. While customer requests can either be known in advance or appear dynamically during the day, vehicles have to be dispatched and routed in the real time, possibly, by taking into account changing traffic conditions, uncertain demands, or varying service times [9].

Recent advances in information and communication technologies can help such companies to manage vehicle fleets in the real-time. When jointly used, technologies like Geographic Information Systems (GIS), Global Positioning Systems (GPS), traffic flow sensors and cellular telephones are able to provide relevant real-time data, such as current vehicle locations, and periodic estimates of road travel times [7].

There exist several important dynamic routing problems that are being solved in the real-time. Most representative examples, including transport of goods,

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 467–476, 2013.

arise in dynamic fleet management and courier services. Others, focused on personal transportation services, include dial-a-ride problems, emergency and taxi cab services [7].

The paper aims at proposing a new multi-agent approach for solving the Dynamic Vehicle Routing Problem with Time Windows (DVRPTW), where the set of customer requests has to be served by the set of available vehicles in order to minimize the vehicle fleet size and the total distance needed to pass by these vehicles, and satisfying several customers and vehicles constraints. According to Pillac [9] and Ghiani [7] classifications, the considered problem belongs to *dynamic* and *deterministic* routing class, where all data are known in advance, and the input orders are revealed dynamically and unpredictably during the execution of orders.

In contrast to its static counterpart, dynamic VRPTW involves new features that increase the complexity of decisions and introduces new challenges while finding the optimal route plan. Because of the fact, that new customer orders can continuously arrive over time, at any moment of time, there may exist customers already under servicing and new customers which need to be serviced. As a consequence, each newly arriving dynamic request, needs to be incorporated into the existing vehicles tours and the current solution may need to be reconfigured to minimize the goal functions.

The rest of the paper is organized as follows. Section 2 includes formulation of the VRPTW. An overview and the main elements of the proposed agent-based approach for VRPTW are presented in Section 3. Section 4 presents results of computational experiment, and finally, Section 5 concludes the paper and suggests the directions of future research.

2 Formulation of the Vehicle Routing Problem with Time Windows

The *static* Vehicle Routing Problem with Time Windows (VRPTW) can be formulated as the problem of determining optimal routes passing through a given set of locations (customers) and defined on the undirected graph $G = (V, E)$, where $V = \{0, 1, \ldots, N\}$ is the set of nodes and $E = \{(i, j)|i, j \in V\}$ is the set of edges. Node 0 is a central depot with NV identical vehicles of capacity W. Each node $i \in V \setminus \{0\}$ denotes a customer characterized by a non-negative demand d_i, and a service time s_i. Moreover, with each customer $i \in V$, a time window $[e_i, l_i]$ wherein the customer has to be supplied, is associated. Here e_i is the earliest possible departure (ready time), and l_i - the latest time the customer request has to be started to be served. The time window at the depot ($[e_0, l_0]$) is called the scheduling horizon.

Each link $(i, j) \in E$ denotes the shortest path from customer i to j and is described by the cost c_{ij} of travel from i to j by shortest path ($i, j \in V$). It is assumed that $c_{ij} = c_{ji}(i, j \in V)$. It is also often assumed that c_{ij} is equal to travel time t_{ij}.

The goal is to minimize the vehicle fleet size and the total distance needed to pass by vehicles in order to supply all customers in their required time windows (minimization of the fleet size is often considered to be the primary objective of the VRPTW), satisfying the following constraints:

- each route starts and ends at the depot,
- each customer $i \in V \setminus \{0\}$ is serviced exactly once by a single vehicle,
- the total load on any vehicle associated with a given route does not exceed vehicle capacity,
- each customer $i \in V$ has to be supplied whithin the time window $[e_i, l_i]$ associated with it (the vehicle arriving before the lower limit of the time window causes additional waiting time on the route),
- each route must start and end within the time window associated with the depot.

The *dynamic* VRPTW considered in the paper can be seen as an extension of the static VRPTW where all customer requests are dynamic (they arrive while the system is already running) and hence the optimisation process has to take place in the real-time. Let the planning horizon starts at time 0 and ends at time T. Let $t_i \in [0, T]$ $(i = 1, \ldots, N)$ denotes the time when the $i - th$ customer request is submitted.

Whereas during the recent years there have been many important advances in the field of static versions of different variants of VRP (the reader can find a review of different methods proposed for solving VRPTW for example in two papers of Braysy and Gendreau [4,5]), definitely much less has been done with respect to solving their dynamic versions. Dynamic vehicle routing problems, their variants, methods of solving them and examples of practical application can be found in one of the last paper of Pillac [9].

In recent years only few approaches based on using intelligent agents for solving some transportation problems have been proposed. The survey of existing research on agent-based approaches to transportation management presented by Davidson et al. [6] shows that this field is still promising and worth further exploration.

3 Multi-agent Approach for VRPTW

3.1 Overview

Technically, the proposed approach extends the platform for solving dynamic VRP proposed in [1,2]. Architecture of the approach is based on JADE (Java Agent Development Framework), a software framework for the development and run-time execution of peer-to-peer applications [3].

Within the proposed platform several types of autonomous agents are used. They are `GlobalManager`, `OrderGenerator`, `OrderManager`, and `Vehicle`. Each agent encapsulates a particular abilities and during the process of solving the problem agents play their roles cooperating in order to achieve the common goal. The list of agents and their characteristics (description, attributes and communication with other agents) are included in Tab. 1.

Table 1. Agents and their characteristics

`GlobalManager`	An agent which runs first and initializes all others agents.
Atributes:	

Communication:	`GlobalManager` $\longrightarrow$ `OrderGenerator`
	`GlobalManager` $\longrightarrow$ `OrderManager`
	`GlobalManager` $\longrightarrow$ `Vehicle`
	It initializes all agents.

`OrderGenerator`	An agent which generates (or reads) new orders.
Atributes:	l_R - list of all dynamic requests, l_dR - list of actual dynamic requests.
Communication:	`OrderGenerator` $\longrightarrow$ `OrderManager` It sends new orders to the `OrderManager` agent.
`OrderManager`	An agent which manages the list of requests.
Atributes:	$R = \{R_1, R_2, \ldots, R_{NV}\}$ - list of all routes,

Communication:	`OrderManager` $\longleftarrow$ `OrderManager`
	`OrderManager` $\longrightarrow$ `Vehicle`
	After receiving the new request from `OrderGenerator`, `OrderManager` inserts it to the list of dynamic orders and allocates it to the available `Vehicle` agents.

`Vehicle`	An agent that represents a vehicle and serves the customers' orders.
Atributes:	$(x0, y0)$ - depot's coordinates on the plane, W - the capacity of the vehicle, Ri - actual route assigned to this vehicle, $cost(Ri)$ - actual cost of the route, Wr - actual available space, v - speed, sV - vehicle state (*waiting*, *driving*, *stopped*), tsV - total time spend by vehicle in the system, twV - vehicle's total waiting time.
Communication:	`OrderManager` $\longleftrightarrow$ `Vehicle` Most of its lifetime, a vehicle spends serving requests. It starts after receiving and accepting the first request. After reaching the nearest customer it proceeds to the next customer belonging to the route or waits for next order. If the vehicle reaches the last customer on the current route, it waits in the location until a new request arrives. When all requests are dispatched among the available vehicles, the waiting vehicle returns back to the depot. Periodically `Vehicle` receives customer's request from the `OrderManager` one at a time, and tries to assign it to the existing route in order to perform it.

Additionally, all the above agents operate on customer order represented by `Customer` class in the system. Each order is described by the set of attributes described above, and, additionally, by: ts_i - time in which an order is served, sC_i - state of the order (*available*, *blocked*, *finished*, *canceled*), and twC_i - customer's waiting time for servicing.

3.2 Process of Solving DVRPTW

The process of solving DVRPTW (performed in the loop until end of request is reached) is divided into general steps, presented as Algorithm 1.

Algorithm 1. MAS-DVRPTW

```
Initialize all agents: GlobalManager, OrderGenerator, OrderManager, and Vehicle
while (end_of_orders has not been reached) do
    OrderGenerator agent generates (or reads) new order and sends it to the OrderManager,
    OrderManager incorporates it into the list of orders to be served and allocates it to the available Vehicle agents, using predefined dispatching strategy.
end while
```

GlobalManager agent runs first and next it initializes all other agents. After receiving messages from all agents about their readiness to act, the system is waiting for events. Three kinds of events are genertaed in the proposed approach. The first group includes *system events* (among them the most important is the end_of_orders event). The second group (*order events*) includes main event observed in the system - new_order event. The last group of events (*vehicle events*) are mainly sent by Vehicle agents which reports their states during serving a set of requests assigned to them (vehicle_drive, vehicle_stop, vehicle_wait, vehicle_reached_location(p), etc.). The above algorithm focuses on main steps of solving the problem without deeper insight into the simulation part, so details of the flow of messages are omitted here.

What is important, the above process of solving the problem is performed in a dynamically changing environment. It means that at the arrival of the new order, all vehicles are serving the customers already assigned to their routes, and introduction of a new request to the routes requires re-optimization procedure. In the proposed approach it has been decided to implement two strategies of dispatching dynamic requests to the available vehicles. They reflect the level at which the list of customer orders is maintained and are called *Decentralized* and *Centralized Dispatching Strategy*, respectively. The process of (re-)optimization is organized in different way in both of them.

3.3 Dispatching Strategies

Decentralized Dispatching Strategy. Using Decentralized Dispatching Strategy (DDS) the process of assigning a new customer order to the available vehicles is performed using the *Contract Net Protocol* (CNP) [11], which is composed of a sequence of five main steps, where agents, acting as *Initiator* or *Participant*, negotiate the contract. After receiving a new order the following steps are performed.

1. `OrderManager` initializes a session using the CNP and starts communication between `OrderManager` and `Vehicle` agents. As *Initiator* it announces the request to each `Vehicle` agent (moving and waiting vehicles) sending around the call for proposal (`cfp`) message. `Vehicle` (as *Participants* or *Contractors*) are viewed as potential contractors.
2. Each `Vehicle` agent after receiving the request (with customer data) from the `OrderManager`, calculates the cost of inserting a new customer into the existing route (using the Solomon's *I1* constructive heuristic for VRPTW [13]). If the insertion of a new customer into the existing route does not violate the problem constraints, the calculated cost of insertion is sent back (as the `propose` message) to the `OrderManager`. Otherwise, the `Vehicle` sends back the rejection (`reject`) message.
3. `OrderManager` after receiving proposals from all `Vehicle` agents, chooses the one with the lowest cost of insertion. Next, it sends the `accept-proposal` message to the `Vehicle` which is awarded and the `reject-proposal` to the others.
4. `Vehicle` which receives the `accept-proposal` message, inserts the customer into its current route and sends the `inform-done` message if the operation is performed successfully and `failure` message, otherwise.
5. If all `Vehicle` agents rejected the proposal, then new order is assigned to new `Vehicle` agent located at the depot.

Using DDS, each `Vehicle` agent is autonomous and maintains its own list of orders. Let $v(i)$ be a `Vehicle` agent and let $R^i = [r^i_1, r^i_2, \ldots, r^i_k, \ldots, r^i_{length(R^i)}]$ be a current route assigned to the vehicle $v(i)$ $(i = 1, \ldots, NV)$. Assume that r^i_k is a customer (location), the vehicle $v(i)$ is currently driving to. Thus, the part of route $[r^i_1, r^i_2, \ldots, r^i_k]$ is fixed, and the process of assigning a new order to the existing route is possible only on positions $k+1, \ldots, length(R^i)$ of route $v(i)$. It is easy to see that if a particular request is arrising close to the end of planning horizon (highly dynamic order), possibility of reoptimization of the route assigned to vehicle $v(i)$ is limited.

After reaching a location of the current customer, `Vehicle` sends a message to `OrderManager` informing about it and continues serving next requests according to its autonomous routing plan.

Centralized Dispatching Strategy. In Centralized Dispatching Strategy (CDS) the list of all orders is maintained by `OrderManager` agent, and `OrderManager` is responsible for planning each vehicle route. On the other hand, `Vehicle` agent maintains short-term list of requests including only the location of the next order.

Similarly, as in DDS, let $v(i)$ be a `Vehicle` agent and let R^i be a current route assigned to the vehicle $v(i)$ $(i = 1, \ldots, NV)$. Let $R = [R^1, R^2, \ldots, R^{NV}]$ be a list of all routes assigned to each `Vehicle` agent $v(i)$ $(i = 1, \ldots, NV)$. In fact, it can be viewed as a solution of the routing problem.

Opposite to DDS, here, after receiving a new request, the `OrderManager` does not announce it immediatelly to the `Vehicle` agents but it buffers the order and tries to insert to one of the active vehicles routes on positions $k+1, \ldots, length(R^i)$ of each route $v(i)$ (r^i_k is a customer (location), the vehicle $v(i)$ is driving to).

After reaching a location of the current customer r^i_k, each Vehicle agent $v(i)$ sends the message to OrderManager informing it about readiness for serving next requests. According to the current global routes plan, OrderManager sends the next customer's order for serving to the Vehicle agent $v(i)$.

In order to improve the solution currently maintained by OrderManager, parallel to the dispatching new orders to the available vehicles, OrderManager performs a set of operations trying to improve the current solution taking into account orders which have not been yet assigned to the vehicles yet. The process of improvement is performed by the set of four local search heuristics based on the following moves:

- a single customer from randomly selected position of individual is moved to another, randomly selected position,
- two customers from randomly selected positions of individual are exchanged,
- two randomly selected routes are disconnected and the remaining parts from different routes are reconnected again,
- two edges from two randomly selected routes are exchanged.

Centralized dispatching strategy can be viewed as a kind of *bufferring strategy*, where new requests are not allocated immediately to the vehicles but assignment of some requests to vehicles is delayed (see for example [10]).

4 Computational Experiment

In order to validate the proposed approach, a computational experiment has been carried out. It aimed at evaluating the influence of the kind of dispatching strategies on the performance of the system measured by the number of vehicles needed to serve all requests in the predefined time and the total distance needed to pass by these vehicles.

The proposed agent-based approach was tested on the classical VRPTW benchmark datasets of Solomon [13] transformed into the dynamic version through revealing all requests dynamically. The experiment involved six datasets of instances (R1, R2, C1, C2, RC1, RC2) [12] including 100 customers, each. Best known solutions identified by different heuristics for solving static VRPTW averaged over all solutions belonging to each group are presented in Tab. 2.

It is assumed that all requests are dynamic, and they may arrive with various frequencies. In the experiment arrivals of the dynamic requests have been generated using the Poisson distribution with λ parameter denoting the mean number of requests occurring in the unit of time (1 hour in the experiment). For the purpose of experiment λ was set to 5, 10, 15, and 30. It is also assumed that all requests have to be served. Additionally, it has been assumed that the vehicle speed was set at 60 km/h.

Table 2. Best known solutions identified by heuristics [12] averaged for each group of instances

Instance	Vehicles	Distance	Instance	Vehicles	Distance
R1	11.92	1209.89	R2	2.73	951.91
C1	10.00	828.38	C2	3.00	589.86
RC1	11.50	1384.16	RC2	3.25	1119.35

Each instance was repeatedly solved five times and mean results from these runs were recorded. All simulations have been carried out on PC Intel Core i5-2540M CPU 2.60 GHz with 8 GB RAM running under MS Windows 7.

The experiment results for both dispatching strategies: decentralized (DDS) and centralized (CDS), are presented in Tab. 3 and 4, respectively. Beside the name of the instance set and the value of λ parameter, next columns include the average number of vehicles used and the average distance travelled by all vehicles.

Table 3. Results obtained by the proposed approach (DDS strategy)

Instance	λ	#Vehicles	Distance	Instance	λ	#Vehicles	Distance
R1	5	13.72	2231.99	R2	5	4.11	1808.62
	10	13.12	1954.31		10	3.69	1370.79
	15	13.28	1694.18		15	3.89	1234.29
	30	13.20	1288.40		30	3.64	1024.23
C1	5	11.59	1499.37	C2	5	3.38	1114.83
	10	11.25	1100.71		10	3.00	732.04
	15	10.00	949.38		15	3.00	728.85
	30	10.00	916.49		30	3.00	619.84
RC1	5	13.54	2394.60	RC2	5	4.48	2160.35
	10	13.76	1884.80		10	4.60	1977.57
	15	13.61	1585.51		15	4.11	1328.20
	30	13.00	1412.81		30	4.15	1136.27

Analysis of the results presented in both tables allows one to conclude that dynamization of the VRPTW almost always results in deterioration of the results in comparison with its static counterpart. For the datasets used in the experiment, it is especially observed for instances with tight time horizon (C1, R1, RC1). It is easy to observe, that such deterioration also often depends on level of dynamism of the instance. For instances with a low ratio of request arrivals ($\lambda = 5 - 10$), the results are worse in comparison with the cases where customer request are known in advance or they arrive at early stage of computation ($\lambda = 30$).

An interesting conclusions can be provided by focusing observation on dispatching strategies proposed in the paper. The influence of the kind of strategies on the performance of the system is observed for almost all tested cases. The centralized strategy (CDS) outperforms decentralized (DDS) one, taking into account both observed factors: the number of vehicles needed to serve requests and the total distance needed to pass by these vehicles while serving them, but not with the same strength.

Table 4. Results obtained by the proposed approach (CDS strategy)

Instance	λ	#Vehicles	Distance	Instance	λ	#Vehicles	Distance
R1	5	13.58	2114.09	R2	5	3.87	1827.66
	10	13.54	1736.71		10	3.50	1191.66
	15	13.32	1469.36		15	3.48	1171.94
	30	13.09	1261.63		30	3.81	982.37
C1	5	10.97	1532.50	C2	5	3.00	1103.04
	10	10.55	1272.35		10	3.00	899.72
	15	10.00	870.94		15	3.00	718.83
	30	10.00	853.50		30	3.00	627.72
RC1	5	12.42	2336.13	RC2	5	4.21	2137.96
	10	13.40	1810.14		10	4.21	1800.48
	15	13.00	1516.53		15	4.04	1276.20
	30	13.00	1489.81		30	4.24	1149.08

5 Conclusions

Although there exists several methods for solving the Vehicle Routing Problem with Time Windows, the majority of them are focused on the static case where all input data are known in advance. This paper proposes a multi-agent approach to solving the Dynamic Vehicle Routing Problem with Time Windows, which allows to observe how different scenario of dispatching dynamic orders to the fleet of vehicles can influence the total cost (and other parameters) of servicing customer orders. Computational experiment showed that centralized dispatching strategy outperforms decentralized one in both dimensions: number of vehicles and the total distance.

Future research may aim at observation and comparison the values of different factors allowing a decision maker to measure performance of the proposed approach. These include: the time a customer must wait before its request is completed, the time spent by each vehicle in the system, the total waiting time of each vehicle, etc.

Another direction of future research is extending the proposed approach to other problems like for example Dynamic Pickup and Delivery Routing Problem with Time Windows (DPDPTW) and incorporating waiting [8] and buffering [10] strategies into the system. The former strategy consists in deciding whether a vehicle should wait after servicing a request, before heading toward the next customer or planning a waiting period on a strategic location. The later one consists in delaying the assignment of some requests to vehicles in a priority buffer, so that more urgent requests can be handled first. Their positive impact on the results has been confirmed by the authors of them.

Acknowledgments. The research has been supported by the Ministry of Science and Higher Education grant no. N N519 576438 (2010-2013).

References

1. Barbucha, D., Jędrzejowicz, P.: Multi-agent platform for solving the dynamic vehicle routing problem. In: Proc. of the 11th IEEE International Conference on Intelligent Transportation Systems (ITSC 2008), pp. 517–522. IEEE Press (2008)
2. Barbucha, D., Jędrzejowicz, P.: Agent-based approach to the dynamic vehicle routing problem. In: Demazeau, Y., Pavón, J., Corchado, J.M., Bajo, J. (eds.) 7th International Conference on PAAMS 2009. AISC, vol. 55, pp. 169–178. Springer, Heidelberg (2009)
3. Bellifemine, F., Caire, G., Greenwood, D.: Developing Multi-Agent Systems with JADE. John Wiley & Sons, Chichester (2007)
4. Braysy, O., Gendreau, M.: Vehicle routing problem with time windows, part i: Route construction and local search algorithms. Transportation Science 39, 104–118 (2005)
5. Braysy, O., Gendreau, M.: Vehicle routing problem with time windows, part ii: Metaheuristics. Transportation Science 39, 119–139 (2005)
6. Davidson, P., Henesey, L., Ramstedt, L., Tornquist, J., Wernstedt, F.: An analysis of agent-based approaches to transport logistics. Transportation Research Part C 13, 255–271 (2005)
7. Ghiani, G., Guerriero, F., Laporte, G., Musmanno, R.: Real-time vehicle routing: Solution concepts, algorithms and parallel computing strategies. European Journal of Operational Research 151, 1–11 (2003)
8. Mitrovic-Minic, S., Laporte, G.: Waiting strategies for the dynamic pickup and delivery problem with time windows. Transportation Research Part B: Methodological 38(7), 635–655 (2004)
9. Pillac, V., Gendreau, M., Guéret, C., Medaglia, A.: A review of dynamic vehicle routing problems. European Journal of Operational Research 225, 1–11 (2013)
10. Pureza, V., Laporte, G.: Waiting and buffering strategies for the dynamic pickup and delivery problem with time windows. INFOR 46(3), 165–175 (2008)
11. Smith, R.: The contract net protocol: High level communication and control in a distributed problem solver. IEEE Transactions on Computers 29(12), 1104–1113 (1980)
12. Solomon, M.: Vrptw benchmark problems, `http://w.cba.neu.edu/~msolomon/problems.htm`
13. Solomon, M.: Algorithms for the vehicle routing and scheduling problems with time window constraints. Operations Research 35, 254–265 (1987)

Operation of Cluster-Based Web System Guaranteeing Web Page Response Time

Krzysztof Zatwarnicki

Department of Electrical, Control and Computer Engineering
Opole University of Technology, Opole, Poland
k.zatwarnicki@gmail.com

Abstract. Guaranteeing Quality of Web Services (QoWS) is now very important for development of internet services. This paper describes a MLF (Mostly Loaded First) method of HTTP requests scheduling, and distribution in a cluster-based Web system. Its application enables keeping the quality of the Web service on the required level and makes the Web service behave in a predictable way, unlike to other Web services offering unpredictable and unreliable services of the best-effort type. The proposed system keeps the page response time within established boundaries, in such a way that at a heavy workload, the page response times for both small and complex pages, would not exceed the imposed time limit. We show thought simulation experiments the operation of the MLF system and the changes of distribution strategy while the load of the system increases.

Keywords: Quality of Web Services, Request Scheduling, Request Distribution.

1 Introduction

Quality of Web service is now one of the key elements which has a significant impact on the profitability of conducted Internet enterprises. Whereas the Internet users will not pay attention to the technical aspect of the Web service when it is of high quality, whereas a low level of quality may cause them to give up on the service.

Quality-based Web systems can be categorized as best-effort, and predictable or guaranteed [10]. Best-effort services does not provide any control over the quality of Web service, it is unpredictable how these services will satisfy the user requirements. Guaranteed and predictable services are reliable and enable the management of quality of service.

There are many different ways of improving QoWS [6, 12, 13]. The encountered solutions include the application of a more efficient server in the service, scheduling and admission control in the Web server [18], the application of a locally [7] and globally [4] distributed cluster-based Web system.

We have already developed systems minimizing HTTP request response times in a locally distributed Web cluster system [3], globally distributed Web cluster systems with a broker [4] and globally distributed Web cluster system without brokers [17]. Our later work concentrates on guaranteeing quality of service in Web systems with

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 477–486, 2013.

one Web server [18], locally distributed cluster of servers [16], and a globally distributed Web cluster system with brokers [19].

In this paper we present MLF systems applying in the construction locally distributed cluster of Web servers. The main element controlling operation of the system is a Web switch. Proposed Web switch schedule HTTP requests and distributes them among Web servers. Both scheduling and distribution algorithms operate together in this way as to not let the imposed maximal page response time exceed, or, if the time is exceeded, to minimize the time. The system operates so that at a heavy workload, the page response times both for small and complex pages, do not exceed the imposed time limit. The MLF system was already described in [16]. In this paper we will show, through simulation experiments, the way of operation and distribution strategy changes in a load function of the system.

The problem of building Web systems providing guaranteed services has been the topic of many research papers. Most of them, however, concern keeping the quality of the service for individual HTTP requests [1, 3, 4, 8, 9]. Few papers have been dedicated to the problem of guaranteeing web page response times to limited group of users [15]. Most works focus on providing differentiated services to different classes of clients using priority-based scheduling [1, 2, 9, 14, 15]. According to our knowledge there are no dedicated solutions guaranteeing page response time in cluster-based Web systems. In our solution the same level of quality of service is provided to all of the users.

The paper is divided into four sections. Section 2 presents design of the MLF Web switch. Section 3 describes the testbed used in the experiments and the research results. Finally, Section 4 contains concluding remarks.

2 Cluster-Based Web System Guaranteeing Page Response Time

The MLF system is constructed in this way to be able to guarantee the page response time and to distribute requests among Web servers. The main element responsible for controlling the system is a Web switch located between clients sending HTTP requests and Web servers servicing requests. Fig. 1a presents overall view of MLF system. The Web switch is composed of two separate sections cooperating with each other. The first, scheduling section, schedules incoming requests in such a way that the page response time is not longer than the imposed value $t_{\max}$. The second, switching section distributes requests among Web servers in this way to not overload any of the servers, and to service the request within time determined in the scheduling section. Fig 1.c show the structure of the MLF Web switch.

2.1 Scheduling Section

The web switch uses the scheduling section only when the number of requests serviced by the web servers is greater then a value $ar_{\max}$, otherwise the HTTP request x_i is passed directly to switching section. The value $ar_{\max}$ is the lowest number of request serviced by web servers in the cluster, for which the service reach the maximal throughput.

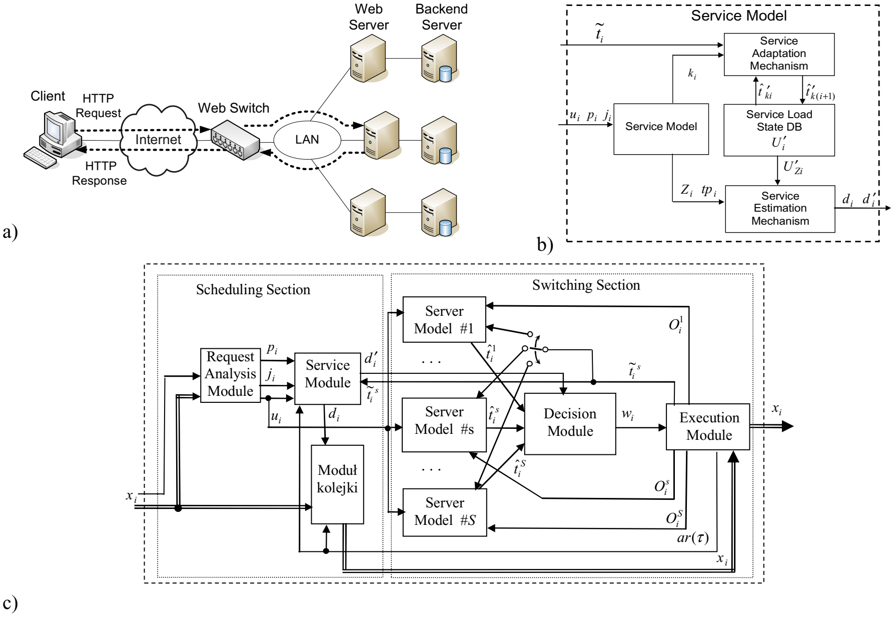


Fig. 1. MLF system: a) Overall view, b) Service module, c) Web switch

The scheduling section is composed of request analysis, service, and queue modules. The request analysis module analyzes the incoming request x_i (where i is index of request) and extracts from it address u_i of requested object, information of user identifier j_i and the web page identifier p_i to which the requested object belongs. Information j_i and p_i are placed in the cookie field and are set by the web service during the first visit.

Service module determine a deadline d_i and a term d'_i. The deadline d_i is a time when the service of the request should begin. The term d'_i is time when the service of the request should be ended. After determining terms d_i and d'_i the request x_i is placed in queue module containing queue Q_i. The politicy of the queue is EDF (Earliest Deadline First) and the requests are organized according to d_i deadlines. The request x_i can leave the Q_i queue if the request is first in the queue and the number $ar\left(\tau_i^{(1)}\right)$ of request been serviced by web servers (where $\tau_i^{(1)}$ is a moment of arrival x_i request) is not greater than $ar_{\max}$.

The aforementioned service module is the most important element of the scheduling section. It consists of service model, service estimation mechanism, service adaptation mechanism and service load state DB module. The service model stores information about HTTP objects available in the service. The inputs for the service module are: u_i, p_i and j_i. The output of the service module is a class k_i of the requested object ($k_i = 1,2,...,K$), a vector $Z_i = \left[k_i^1,...,k_i^l,...,k_i^L\right]$ of the object classes belonging to the page not requested for by clients yet within the page ($k_i^1 = k_i$), and information on the time tp_i measured from the moment of the arrival of the first request concerning page p_i to the moment of arrival of request x_i. The class of the object is determined on the basis of the object's size, in case of static objects. Objects of similar size belong to the same class whereas every dynamic object has its own individual class. Objects belonging to the same classes have similar request response times.

The service load state DB module stores information about service times of the objects belonging to individual classes. The load state DB contains information $U_i' = \left[\hat{t}_{1i}',...,\hat{t}_{ki}',...,\hat{t}_{Ki}'\right]$, where $\hat{t}_{ki}'$ is the estimated service time of the request belonging to the k-th class. The load state forwards service times $U_{Zi}' = \left[\hat{t}_{k^1i},...,\hat{t}_{k^li},...,\hat{t}_{k^Li}\right]$ of objects indicated in Z_i to the service estimation mechanism.

The service estimation mechanism determine the deadline d_i and the term d_i'. Deadline d_i is calculated as follows: $d_i = \tau_i^{(1)} + \Delta d_i - \hat{t}_{k^0i}'$, where $\Delta d_i = \hat{t}_{k^1i}'(t_{\max} - p_i) \Big/ (\lambda \sum_{l=1}^{L} \hat{t}_{k^li}')$ is a time which can be spent in the queue module. The λ is a concurrency factor, it depends on the number of objects being concurrently requested by the same client within the same web page, its value for modern web browsers is about 0.267 [20]. Term d_i' can be calculated according to following formula $d_i' = \tau_i^{(1)} + \Delta d_i$.

The service adaptation mechanism modifies information $U_i' = \left[\hat{t}_{1i}',...,\hat{t}_{ki}',...,\hat{t}_{Ki}'\right]$ for each k_ith class of serviced request x_i according to the formula $\hat{t}_{k\ (i+1)}' = \hat{t}_{k\ i}' + \hat{\eta}(\tilde{t}_i - \hat{t}_{k\ i}')$, where $\tilde{t}_i$ is a measured response time.

2.2 Switching Section

The x_i request is passed to the switching section then it leaves the queue Q_i or if it was not placed in the queue at all. The switching section distributes HTTP requests among Web servers in the cluster. This section contains the following modules: server model, decision and execution. The number of server model modules is equal to number of web serves in the cluster. Each server model estimates request response times $\hat{t}_i^1,...,\hat{t}_i^s,...,\hat{t}_i^S$ having information of assigned web server loads $O_i^s = \left[a_i^s, b_i^s\right]$, where s is the index of the server, $s = 1,...,S$, a_i^s is the number of static requests, and b_i^s is the number of dynamic requests concurrently serviced by sth server. Construction of server model module will be presented further.

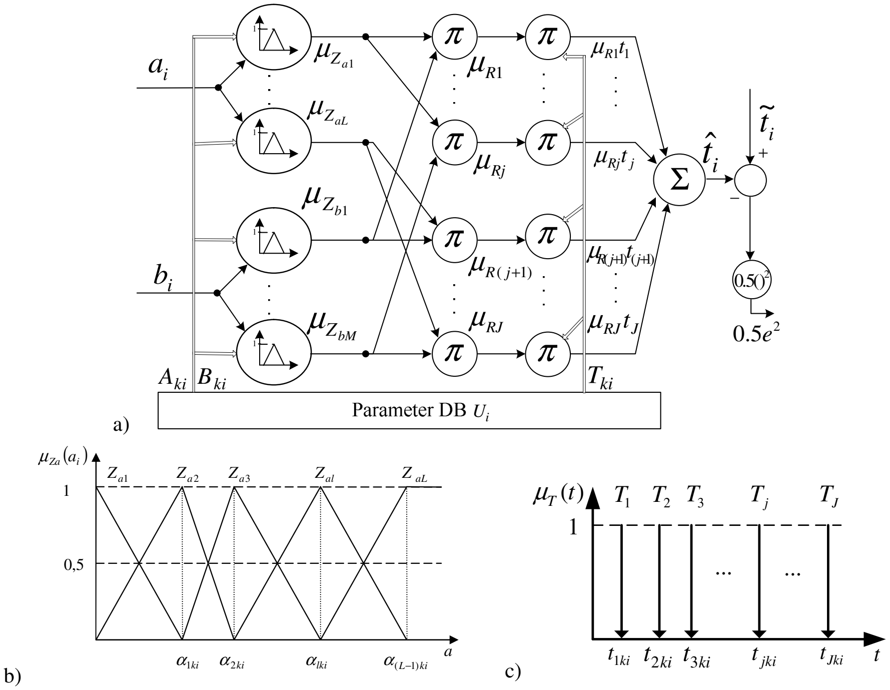


Fig. 2. Server model: a) neuro-fuzzy model, b) input fuzzy sets functions, c) output fuzzy sets functions

The decision module determines the Web server to service the request x_i. The decision is calculated in following way:

$$w_i = \begin{cases} \min\left\{s : s \in \{1,...,S\} \wedge \Delta d_i' \le \hat{t}_i^s\right\} \text{ if } ar\left(\tau_i^{(1)}\right) = ar_{\max} \text{ and } \underset{s \in \{1,...,S\}}{\exists} \hat{t}_i^s \ge \Delta d_i' \\ s_{\min} : \hat{t}_i^{s_{\min}} = \min\left\{\hat{t}_i^s : s \in \{1,...,S\}\right\} \text{ in other case} \end{cases} \quad (1)$$

where $\Delta d_i' = \tau_i^{(2)} - d_i'$, $\tau_i^{(2)}$ is the moment of the request x_i arrival to the execution module. According to the formula requests are assigned to servers with the lowest indexes (in most cases the most loaded ones) which are able to service the requests within time shorter then $\Delta d_i'$. If there is no server offering time shorter than $\Delta d_i'$ then the server with the shortest service time is chosen.

The aforementioned server model module is a neuro-fuzzy model which can adapt to changing behavior of Web server. Construction of adopted model is presented on Fig. 2a. The neuro-fuzzy model has different parameters $U_i = [U_{1i},...,U_{ki},...,U_{Ki}]$ for each class k_i of requested object, where $U_{ki} = [A_{ki}, B_{ki}, T_{ki}]$,

$A_{ki} = \left[\alpha_{1\,ki},...,\alpha_{l\,ki},...,\alpha_{(L-1)\,ki}\right]$, $B_{ki} = \left[\beta_{1\,ki},...,\beta_{m\,ki},...,\beta_{(M-1)\,ki}\right]$ are parameters of input fuzzy set functions, and $T_{ki} = \left[t_{1\,ki},...,t_{j\,ki},...,t_{J\,ki}\right]$ are parameters of output fuzzy set functions (Fig. 2c). Input fuzzy set functions are triangular (Fig. 2b) and are denoted as $\mu_{Z_{al}}(a_i)$, $\mu_{Z_{bm}}(b_i)$, where $l = 1,...,L$, $m = 1,...,M$, $\underset{a_i \in \langle 0,\infty)}{\forall} \sum_{l=1}^{L} \mu_{Z_{al}}(a_i) = 1$ and $\underset{b_i \in \langle 0,\infty)}{\forall} \sum_{m=1}^{M} \mu_{Z_{bm}}(b_i) = 1$.

The request response time is calculated as follows: $\hat{t}_i = \sum_{j=1}^{J} t_{jki} \mu_{R_j}(a_i, b_i)$, where $\mu_{R_j}(a_i, b_i) = \mu_{Z_{al}}(a_i) \cdot \mu_{Z_{bm}}(b_i)$. New values of parameters for the output membership functions are calculated with use of Back Propagation Method according to the formula $t_{j\,k(i+1)} = t_{j\,ki} + \eta_y (\tilde{t}_i - \hat{t}_i) \mu_{R_j i}$ whereas parameters for the input membership functions are calculated in the following way $\alpha_{\phi\,k(i+1)} = \alpha_{\phi ki} + \eta_a (\tilde{t}_i - \hat{t}_i) \sum_{m=1}^{M} \left(\mu_{Z_{bm}}(b_i) \sum_{l=1}^{L} \left(t_{((m-1) \cdot L + l)ki}\ \partial \mu_{Z_{al}}(a_i) / \partial \alpha_{\phi ki} \right) \right)$ and $\beta_{\gamma k(i+1)} = \beta_{\gamma k\,i} + \eta_b (\tilde{t}_i - \hat{t}_i) \sum_{l=1}^{L} \left(\mu_{Z_{al}}(a_i) \sum_{m=1}^{M} \left(t_{((l-1) \cdot M + m)ki}\ \partial \mu_{Z_{bm}}(b_i) / \partial \beta_{\gamma ki} \right) \right)$, where η_t , η_a , η_b are adaptation ratios, $\phi = 1,...,L-1$, $\gamma = 1,...,M-1$.

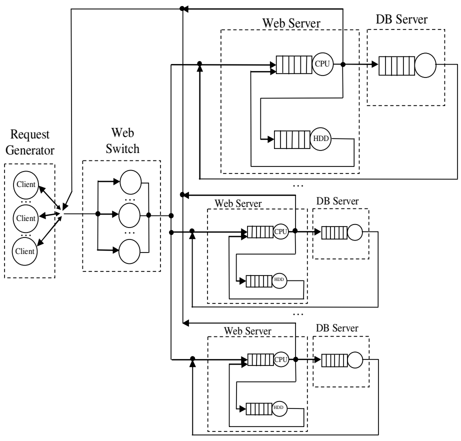
3 Simulation Model and Experiment Results

In order to evaluate the operations of Web cluster working under control of the MLF method simulation experiments were conducted. The main aim of the experiments was to observe behavior of the cluster and the load of Web server working under increasing number of incoming HTTP requests.

The simulation program was written with use of CSIM 19 package [11] and consists of the following modules: HTTP request generator, Web switch, Web and database servers (Fig. 3a).

During simulation, the Web switch could work under control of four different methods: MLF, LFNRD one of the best methods minimizing request response times, LARD distribution algorithm taking into account localization of previously requested object, CAP algorithm uniformly distributing HTTP requests of different types, RR algorithm distributing uniformly all incoming requests.

The service times were established on the base of experiments for the Apache Web server operating on a computer with Intel Pentium 4, a 2 GHz processor and a Seagate ST340810A IDE hard drive. The service of static and dynamic requests was provided. The dynamic requests, serviced by the Web and database server, were divided into three classes [5]: high intensive, medium intensive and low intensive. The service times of the dynamic requests were modeled according to hyperexponential distribution with parameters presented in Fig. 3b. The request generator module was working in the way that the generated request traffic complied with the traffic observed on the Internet, which is characterized by bursts and self-similarity [6] (Fig. 3c).



a)

Type	Mean service time	Frequency
High intensive	20 ms	0.85
Medium intensive	10 ms	0.14
Low intensive	5 ms	0.01

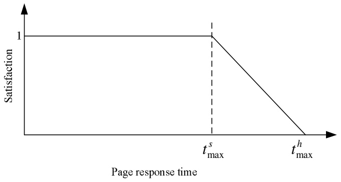
b)

Category	Distribution	Parameters
Requests per session	Inverse Gaussian	μ=3.86, λ=9.46
User think time	Pareto	α=1.4, k=1
Objects per request	Pareto	α=1.33, k=2

HTML object size	Lognormal
	Pareto
		μ=7.630, σ=1.001
		α=1, k=10240

Embedded object size	Lognormal	μ=8.215, σ=1.46

c)



d)

Fig. 3. a) Simulation model, b) Workload model parameters, c) Workload model parameters of dynamic objects, d) Satisfaction function

The new approach in the scope of request distribution requires the indication of an adequate quality factor. Therefore, the mean value of satisfaction was chosen as the quality factor. It is often used to evaluate real-time soft systems. The satisfaction is equal to 1 when the page response time is shorter then $t^s_{\max}$, and decreases to 0 then the time is longer then $t^h_{\max}$ (Fig. 3d). In all of the experiments it has been adopted that $t^s_{\max} = t_{\max}$ and $t^h_{\max} = 2t^s_{\max}$.

The experiments were conducted for two clusters of servers. The first cluster consisted of three identical sets of Web and database servers (denoted as Hom3s), the second one was built of two identical sets of servers and one in which service times were increased of 33% - this server had index $s = 1$ in the experiments (cluster denoted as Het1s/2s).

Fig. 4. presents results of conducted experiments. Diagrams 4 a, b, c show the mean value of satisfaction in load function. As it can be noticed the highest value of satisfaction was obtained for the MLF method for both of the clusters and for different adopted $t_{\max}$ values [16].

Fig. 4 d, e, f presents participation of individual web servers in request service in load function. The Web switch distribute request among servers according to formula 1 taking in to account the estimated for individual Web servers request service times. As it was mentioned before the distribution algorithm chooses Web server with the lowest index offering service time not longer than determined in scheduling section time. As one can notice, when the load of the cluster is low most of the request are serviced by the first server even in case of heterogeneous cluster where the first server is the least efficient (Fig. 4f). As the load increase the participation of request service for the first servers decrease. When the load is very high (close to the maximal load

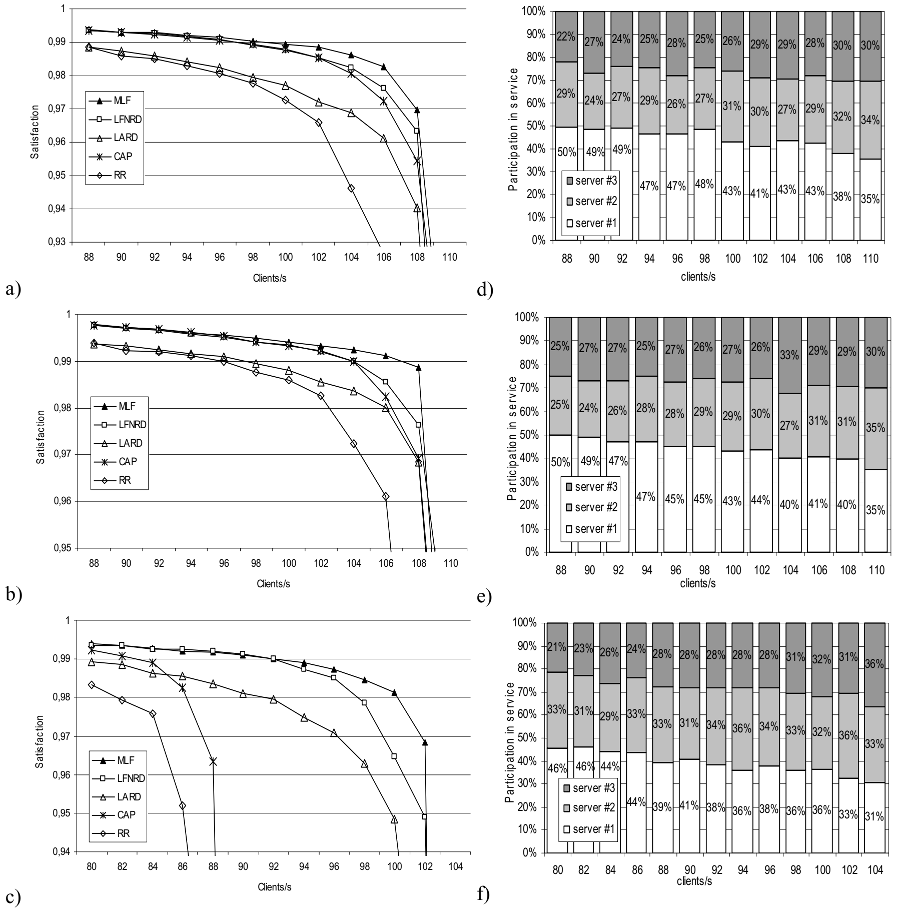


Fig. 4. Experiment results: Satisfaction vs. load a) $t_{\max} = 300ms$, Hom3s, b) $t_{\max} = 500ms$, Hom3s, c) $t_{\max} = 300ms$, Het1s/2s; Participation in service vs. load d) $t_{\max} = 300ms$, Hom3s, e) $t_{\max} = 500ms$, Hom3s, f) $t_{\max} = 300ms$, Het1s/2s

for given Web cluster) distribution algorithm chooses, in most cases, Web server offering the shortest request service times. It can be observed on the diagrams that in the case of the highest load the load on the servers even out, and in case of heterogeneous cluster the first (weakest) server receives the lowest number of request.

Results of experiment show that the distribution algorithm gradually changes the distribution strategy while the load increases. Thanks to this the MLF Web switch can gain higher satisfaction than the LFNRD Web switch minimizing the request service time optimally.

4 Summary

In this paper the HTTP request scheduling and distribution method enabling guaranteeing quality of the cluster-based Web service was presented. The proposed MLF method applies adaptive algorithms. According to the method, requests are scheduled at the front of the web switch and distributed among Web servers in the cluster. MLF web system operates in this way that the page response time does not exceed a demanded time under the assumptions that the number of requests simultaneously serviced by the system is lower then the maximal capacity of the system. Thanks to the new method, the clients using loaded service will obtain similar Web page request response times for both small simple pages as well as complex ones. The experiments showed that the MLF system gradually change its distribution strategy while the load increases. Thanks to this the, system is effective in cases of both low and high loads.

References

1. Abdelzaher, T.F., Shin, K.G., Bhatti, N.: Performance Guarantees for Web Server End-Systems: A Control-Theoretical Approach. IEEE Trans. Parallel and Distributed Systems 13(1), 80–96 (2002)
2. Blanque, R.J.M., Batchelli, A., Schauser, K., Wolski, R.: Quorum: Flexible Quality of Service for Internet Services. In: Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation, NSDI 2005, Berkeley, CA, USA, vol. 2, pp. 43–56 (2005)
3. Borzemski, L., Zatwarnicki, K.: Performance Evaluation of Fuzzy-Neural HTTP Request Distribution for Web Clusters. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 192–201. Springer, Heidelberg (2006)
4. Borzemski, L., Zatwarnicka, A., Zatwarnicki, K.: Global Adaptive Request Distribution with Broker. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 271–278. Springer, Heidelberg (2007)
5. Cardellini, V., Casalicchio, E., Colajanni, M., Mambelli, M.: Web Switch Support for Differentiated Services. ACM Perf. Eval. Rev. 29(2), 14–19 (2001)
6. Cardellini, V., Casalicchio, E., Colajanni, M., Yu, P.S.: The state of the art in locally distributed Web-server systems. ACM Computing Surveys 34(2), 263–311 (2002)
7. Gilly, K., Juiz, C., Puigjaner, R.: An up-to-date survey in web load balancing. World Wide Web, 10.1007/s11280-010-0101-5 (2010)
8. Harchol-Balter, M., Schroeder, B., Bansal, N., Agrawal, M.: Size-based scheduling to improve web performance. ACM Trans. Comput. Syst. 21(2), 207–233 (2003)
9. Kamra, A., Misra, V., Nahum, E.: Yaksha: A Self Tubing Controller for Managing the Performance of 3-Tiered Websites. In: Proc. Int'l Workshop Quality of Service, pp. 47–56 (2004)
10. McCabe, D.: Network analysis, architecture, and design. Morgan Kaufmann, Boston (2007)
11. Mesquite Software Inc. CSIM User's Guide. Austin, TX (2012), `http://www.mesquite.com`

12. Olejnik, R.: A Floor Description Language as a Tool in the Process of Wireless Network Design. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2009. CCIS, vol. 39, pp. 135–142. Springer, Heidelberg (2009)
13. Rzońca, D., Stec, A., Trybus, B.: Data Acquisition Server for Mini Distributed Control System. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2011. CCIS, vol. 160, pp. 398–406. Springer, Heidelberg (2011)
14. Schroeder, B., Harchol-Balter, M.: Web servers under overload: How scheduling can help. In: 18th International Teletraffic Congress, Berlin, Germany, pp. 20–52 (2003)
15. Wie, J., Xue, C.Z.: QoS: Provisioning of client-perceived end-to-end QoS guarantees in Web servers. IEEE Trans. on Computers 55(12), 1543–1556 (2006)
16. Zatwarnicki, K.: A cluster-based Web system providing guaranteed service. System Science 35(4), 68–80 (2009)
17. Zatwarnicki, K.: Neuro-Fuzzy Models in Global HTTP Request Distribution. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part I. LNCS, vol. 6421, pp. 1–10. Springer, Heidelberg (2010)
18. Zatwarnicki, K.: Providing Web Service of Established Quality with the Use of HTTP Requests Scheduling Methods. In: Jędrzejowicz, P., Nguyen, N.T., Howlet, R.J., Jain, L.C. (eds.) KES-AMSTA 2010, Part I. LNCS, vol. 6070, pp. 142–151. Springer, Heidelberg (2010)
19. Zatwarnicki, K.: Guaranteeing Quality of Service in Globally Distributed Web System with Brokers. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS, vol. 6923, pp. 374–384. Springer, Heidelberg (2011)
20. Zatwarnicki, K., Zatwarnicka, A.: Estimation of web page download time. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2012. CCIS, vol. 291, pp. 144–152. Springer, Heidelberg (2012)

Temporal Aggregation of Video Shots in TV Sports News for Detection and Categorization of Player Scenes

Kazimierz Choroś

Institute of Informatics, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
kazimierz.choros@pwr.wroc.pl

Abstract. A large amount of digital video data stored in Internet video collections, TV shows archives, video-on-demand systems, personal video archives offered by Internet services, etc. leads to the development of new methods and technologies of video indexing and retrieval. Content-based indexing of TV sports news is based on the automatic segmentation, then recognition and classification of scenes reporting the sports events. The automatic identification of sports disciplines in TV sports news will be less time consuming if the analysed video material is limited to player scenes. The shots detected can be grouped in scenes using a new proposed temporal aggregation method based on the length of the shot as a sufficient alone criterion. The tests have shown its good performance in detecting player scenes in TV sports news.

Keywords: content-based video indexing, digital video segmentation, scene detection, player scenes, temporal aggregation, AVI Indexer, TV sports news.

1 Introduction

The automatic recognition and classification of scenes reporting the sports events are often examined in research and experiments. The main purpose is to recognize sports disciplines presented in sports news. Due to the automatic categorization of sports events videos can be automatically indexed for content-based retrieval. So, the retrieval of news presenting the best or actual games, tournaments, matches, contests, races, cups, etc. in a desirable sports disciplines becomes possible. The recognition of sports disciplines is usually performed for all frames of videos. The processing time could be significantly reduced if we were able to detect sport scenes (player scenes) and to limit the analyses only to these scenes, ignoring other parts of TV sports news such as studio discussions, commentaries, interviews, charts, tables, announcements of future games, discussions of decisions of sports associations, etc. The player scenes presenting the best highlights of sports events are the most adequate for automatic categorization of sports events.

Digital video is hierarchically structured. Video is composed of acts, episodes (or sequences, stories, events), scenes, shots, and finally of single frames [1-3]. The length of shots in TV sports news is not accidental. Player scenes are usually very dynamic. Such differences of different categories of movies [3] can be applied to

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 487–497, 2013.

optimize temporal segmentation [4]. The method of scene detection presented in the paper based on spatiotemporal information will be called shot temporal aggregation. The temporal aggregation of shots leads to the detection not only of player scenes but in some cases also to the detection of the sequences of scenes, thus, of sports events.

The paper is organized as follows. The next section reminds some related works in the area of automatic shot grouping and clustering, scene detection, and finally scene selection in sports videos as well as content analysis of sports videos. The third section presents the scheme of temporal aggregation of video shots leading to the important reduction of analysed video material and to the detection of scenes with players or playing fields. In the forth section the experimental results for the temporal aggregation of shots for automatic filtering of scenes in the tested TV sports news videos are reported. The final conclusions are discussed in the last fifth section.

2 Related Works

A scene is defined as a group of consecutive shots sharing similar visual properties and having a semantic correlation – following the rule of unity of time, place, and action. Semantic grouping of shots in a video is the first step towards scene detection. Shot grouping can be based on a dominant colour measure of the semantic correlation of consecutive shots [5], on dominant shots [6], on a similarity measure of colour and motion information [7], or on clustering algorithms using keyframes and sampled frames from shots [8], as well as on shot similarity based on visual features [9]. Video shot clustering approaches can use different mathematical tools like tensor algebra with multilinear operators over a set of vector spaces [10] or rough sets [11].

Grouping or clustering of shots is also useful to detect scenes and actions reflecting in the best way the main events. In [12] a survey can be found of methods applied to different video domains, considering edited videos (i.e. videos that have been created from a collection of different video materials presenting also different sports events) and unedited videos (i.e. continuous video recording of a game).

The automatic recognition of a content of a video clip is one of the crucial problem that should be solved to retrieve efficiently videos stored in more and more huge multimedia databases. Many of the methods have been tested on TV news and TV sports videos. Many promising approaches, methods, algorithms, frameworks have been proposed for the most popular sports such as baseball, basketball, or tennis and of course for soccer [13-19]. A unified framework for semantic shot classification in sports videos has been defined in [13]. The proposed scheme makes use of domain knowledge of specific sport to perform a top-down video shot classification, including identification of video shots classes for each sport. The method has been tested over three types of sports videos: tennis, basketball, and soccer. Another approach and another kind of analyses of sports news were implemented in the system presented in [14] that performs automatic annotation of soccer videos. That approach has resulted in detecting principal highlights, and in recognizing identity of players based on face detection, and on the analysis of contextual information such as jersey's numbers and superimposed text captions. The procedure presented and tested in [20] has shown that it is possible to retrieve shots of a given sports discipline in TV sports news.

Other experiments have been carried out for example with baseball videos [15], with tennis videos [17], as well as with other sports.

There are also many promising experiments in which the specific features of sport playfields such as lines automatically extracted from a court view [21, 22] are used in content analysis of sports videos.

3 Temporal Aggregation of Shots

TV sports news video has a specific structure due to standard video editing style [23]. The analyses of TV sports news broadcasted in the first national Polish TV channel show that the video has its individual standard structure. It is composed of several highlights that are introduced and commented by anchorperson or numerical results presented in tables. These shots are called studio shots. Shots that belong to the scene are visually similar and they are located closely along the time axis. A player scene is formed by all successive shots reporting the same event between two studio shots.

Scenes presenting sports events differ from studio scenes. Player scenes are composed of a few very short shots, each of only a few seconds. Whereas a studio scene is usually a single shot but much more longer than any sport highlight shot. It was already noticed in our earlier experiments [4] with segmentation methods. The length of player shots varies form 2 to 6 seconds, whereas the shots of commentaries, interviews, tables, announcements of future games, discussions of decision of sports associations, etc. are always much longer, they take usually tens of seconds.

Temporal aggregation of shots should respect the following rules:

- single frame detected as a shot can be linked to the next shot,
- very short shots should be aggregated till their length attain a certain value,
- all long consecutive shots should be aggregated because these shots seem to be useless in further content analyses and categorization of sports events,
- all shots of the length between two a priori defined maximum and minimum values should remain unchanged – these shots are very probably the most informative shots for further content-based analyses.

A temporal aggregation is specified by three values: minimum shot length as well as lower and upper limits representing the length range for the most informative shots. The values of these parameters will be determined in the further experiments basing on the exhaustive analyses of TV sports news during several weeks preceding the main tests.

4 Tests of Temporal Aggregation

The tests have been performed first for five TV sports news from 1 to 5 February, 2012, and then repeated some months later for 13 TV sports news: five from 1 to 5 September, 2012, and eight sports news randomly recorded in October, 2012. All sports news were broadcasted in the first national Polish TV channel (TVP1) after the main evening TV news. The reason was to confirm that the structure of sports news in a given TV channel is unchanged and that the values of aggregation parameters do not need to be set again because they are specific for a given video structure of sports news. So, the tests of these February videos are treated as a training process leading to the determination of the values of the temporal aggregation parameters.

The characteristics of all 18 broadcasts are presented in Table 1.

Table 1. Characteristics of tested broadcasts of TV sports news from February, September, and October 2012 and of sports scenes in the tested news

	Video	Duration [min:sec: frames]	Number of frames	Number of sports scenes	Number of shots in sports scenes	Number of frames in sports scenes	Share of sports scenes in news
1.	Feb. 01, 2012	10:25:13	15638	14	50	4972	32 %
2.	Feb. 02, 2012	8:14:18	12368	9	47	3750	30 %
3.	Feb. 03, 2012	8:12:00	12300	13	39	3550	29 %
4.	Feb. 04, 2012	8:17:00	12425	13	70	5639	45 %
5.	Feb. 05, 2012	6:26:15	9666	14	53	4203	43 %
	Averages:	08:19:05	12480	12.60	51.80	4423	35.44 %
6.	Sep. 01, 2012	03:48:00	5700	6	19	1585	28 %
7.	Sep. 02, 2012	05:30:00	8250	15	61	4747	58 %
8.	Sep. 03, 2012	04:30:00	6750	13	47	3076	46 %
9.	Sep. 04, 2012	02:40:15	4015	5	14	1449	36 %
10.	Sep. 05, 2012	02:32:00	3800	7	12	1097	29 %
	Averages:	03:48:03	5703	9.20	30.60	2391	41.92 %
11.	Oct. 01, 2012	06:44:00	10100	14	46	3880	38%
12.	Oct. 02, 2012	03:19:18	4993	11	28	2026	41%
13.	Oct. 04, 2012	10:08:10	15210	25	76	5946	39%
14.	Oct. 05, 2012	04:34:19	6869	13	42	3924	57%
15.	Oct. 24, 2012	10:42:15	16065	24	70	5472	34%
16.	Oct. 25, 2012	09:11:10	13785	23	77	6316	46%
17.	Oct. 29, 2012	04:07:20	6195	6	20	1546	25%
18.	Oct. 31, 2012	03:50:00	5750	12	49	3126	54%
	Averages:	06:34:21	9871	16.00	51.00	4154	42.09 %
	Total averages:	06:17:13	9438	13.17	45.56	3684	39.44 %

The first observation is that the length of broadcast does not always depend on the number of sports scenes reported upon TV sports news. For example the TV sports news of February 1 has reported the same number of sports scenes as the TV sports news of February 5, although its duration was almost twice longer. The length of news depends much more on the duration of studio scenes, that is of commentaries, interviews, tables, announcements of future games, discussions of decision of sports associations, etc. Player scenes represent only about third of the TV sports news (Tab. 1, last column). It would be useful to reject these two thirds of TV sports news with non-player scenes before starting content analyses of shots to reduce computing time and to do these analyses more effective.

The process of temporal segmentation has been performed in the AVI Indexer [24]. Four methods of temporal segmentation have been implemented and applied to segment the tested TV sports news. Three methods for cut detection are:

- pixel pair differences,
- histogram differences of frames,
- likelihood ratio method based on block sampling of video frames,

and one method for dissolve detection:

- twin threshold comparison.

The cut and dissolve detection methods have been described and their effectiveness for different categories of videos have been discussed in [3].

Table 2 presents the results of temporal segmentation for all 18 tested TV sports news videos. The method 4 based on the twin threshold comparison is oriented to detect mainly cross dissolves. As a result the method gives a series of cuts in consecutive frames. For our analyses we treat this all cuts as individual cuts aggregated in a further procedure. The total number of detected shots without duplications is specified in the last column of Table 2.

Table 2. Results of temporal segmentation in the AVI Indexer

	Video	Method 1 (pixel pair differences)	Method 2 (histogram differences of frames)	Method 3 (likelihood ratio method)	Method 4 (twin threshold comparison)	Total number of shots detected
1.	Feb. 01	138	163	19	235	337
2.	Feb. 02	140	115	12	172	279
3.	Feb. 03	148	116	23	172	292
4.	Feb. 04	210	159	22	182	373
5.	Feb. 05	132	123	32	80	204
	Averages:	153.60	135.20	21.60	168.20	297.00
6.	Sep. 01	41	35	10	161	89
7.	Sep. 02	199	143	18	494	330
8.	Sep. 03	242	80	29	405	273
9.	Sep. 04	40	35	9	141	83
10.	Sep. 05	44	39	11	160	85
	Averages:	113.20	66.40	15.40	272.20	172.00
11.	Oct. 01	148	131	54	418	218
12.	Oct. 02	69	97	24	242	142
13.	Oct. 04	157	184	28	586	322
14.	Oct. 05	77	100	25	260	139
15.	Oct. 24	307	228	76	108	379
16.	Oct. 25	262	197	33	213	438
17.	Oct. 29	108	89	22	97	165
18.	Oct. 31	116	75	46	47	127
	Averages:	152.60	148.00	41.40	322.80	240.00
Total averages:		143.22	117.17	27.39	231.83	237.50

Table 2 shows that the numbers of shots detected in sports news are rather great and it would be very reasonable to aggregate mainly very short shots and to reduce the potential number of shots taking into account for content analysis. In most approaches of automatic scene detection and recognition, including the sports scene categorization, calculations and content analyses are performed for all frames or for their representatives called keyframes, i.e. the most representative frames, one for every given shot or scene. The segmentation of video should be precise mainly for player scenes whereas studio scenes that is commentaries, interviews, tables, announcements of future games, discussions of decision of sports associations, etc. usually not useful for sport categorization should be grouped and neglected in further content-based analyses. The segmentation of studio scenes is useless.

Therefore then, the share of short length shots in shots detected has been analysed before and after short shot aggregation. The next table (Tab. 3) presents the number of shots after grouping the most short shots. Generally very short shots of one or several frames are detected in case of dissolve effects or they are simply wrong detections. The causes are different. Most frequently it is due to very dynamic movements of players or of a camera during the game, as well as due to light flashes during the interviews. In our procedure these extremely short shots detected are joined with the next shot in a video. This step of temporal aggregation of shots leads to the significant reduction of wrong cuts incorrectly detected during temporal segmentation.

Table 3. Number of shots after aggregating very short consecutive shots

	Video	Total number of shots before aggregation	Aggregation of shots shorter than 5 frames	Aggregation of shots shorter than 8 frames	Aggregation of shots shorter than 25 frames
1.	Feb. 01	337	175	160	136
2.	Feb. 02	279	126	111	97
3.	Feb. 03	292	140	127	101
4.	Feb. 04	373	144	129	109
5.	Feb. 05	204	101	94	80
	Averages:	297.00	137.20	124.20	104.60
6.	Sep. 01	89	46	40	33
7.	Sep. 02	330	109	88	62
8.	Sep. 03	273	83	72	56
9.	Sep. 04	83	38	35	28
10.	Sep. 05	85	42	36	27
	Averages:	172.00	63.60	54.20	41.20
11.	Oct. 01	218	108	98	83
12.	Oct. 02	142	63	52	41
13.	Oct. 04	322	162	143	125
14.	Oct. 05	139	70	60	50
15.	Oct. 24	379	176	162	140
16.	Oct. 25	438	175	156	125
17.	Oct. 29	165	84	71	55
18.	Oct. 31	127	75	69	54
	Averages:	241.25	114.13	101.38	84.13
Total averages:		237.50 100 %	106.50 44.84 %	94.61 39.84 %	77.89 32.80 %

The aggregation of short shots as well as of long shots leads to the selection of shots of the length the most adequate for player scenes. Tables 4 and 5 present the most important results. The tests have been performed for two sets of aggregation parameters.

In the both tests the sets of shots typed as most probable to be shots belonging to the player scene have been then verified and standard two measures have been applied: recall and precision. The shot is treated as correctly recognized if it belongs to a player scene. The scene is correctly detected if at least one of the shot of a scene is detected. So, it means that no exact match of real player scene and detected scene timecodes are required. Such an evaluation has been applied because it is sufficient for further content analyses for sports discipline identification.

In the first test (Tab. 4) all shots shorter than 1 second (25 frames) have been jointed to the next shot, then all consecutive shots shorter than 70 frames and similarly all shots longer than 130 frames have been aggregated. Only shots of the length of 70 to 130 frames have been remained unchanged probably belonging to player scenes. These shots are of the length of 3 to 5 second ± 5 frames of tolerance.

Table 4. Number of selected player pseudo-scenes after aggregating very short consecutive shots as well as very long shots. In the first series of tests all shots shorter than 70 frames have been accepted as short shots and as long shots all shots longer than 130 frames.

	Video	Number of all aggregated pseudo-scenes	Number of selected player scenes	Recall of correct detections of player scenes [in %]	Number of frames in selected aggregated scenes	Share of selected aggregated scenes in total video [in %]	Precision of detection of scenes with player scenes [in %]
1.	Feb. 01	60	42	**85.71**	4264	**13**	**52.38**
2.	Feb. 02	53	36	**100.00**	3572	**29**	**86.11**
3.	Feb. 03	47	27	**53.85**	2604	**21**	**48.15**
4.	Feb. 04	45	29	**84.62**	2818	**23**	**72.41**
5.	Feb. 05	36	22	**71.43**	1875	**19**	**72.73**
	Averages:	48.20	31.20	**79.12**	3027	**21.00**	**66.36**
1.	Sep. 01	16	10	**83.33**	905	**16**	**100.00**
2.	Sep. 02	31	21	**66.67**	2095	**25**	**80.95**
3.	Sep. 03	28	18	**61.54**	1820	**27**	**72.22**
4.	Sep. 04	16	11	**60.00**	999	**25**	**63.64**
5.	Sep. 05	15	10	**85.77**	1052	**28**	**60.00**
6.	Oct. 01	47	29	**64.20**	2714	**27**	**68.97**
7.	Oct. 02	23	15	**63.64**	1436	**29**	**66.67**
8.	Oct. 04	58	36	**56.00**	3361	**23**	**63.89**
9.	Oct. 05	25	16	**76.92**	1484	**22**	**93.75**
10.	Oct. 24	74	46	**70.83**	4847	**31**	**71.74**
11.	Oct. 25	56	33	**69.57**	3166	**24**	**75.76**
12.	Oct. 29	29	16	**83.33**	1487	**25**	**37.50**
13.	Oct. 31	34	25	**83.33**	2427	**43**	**88.00**
Total averages for 13 videos:		**34.77**	**22.00**	**71.16**	**2138**	**26.54**	**72.55**

The aggregation parameters used in the first test enable us to limit news to only one quarter of the video material, but in this case about three quarters of sports scenes would be indexed.

The results in the second test are much more promising. For the 6 videos recall has attain the maximum value – 100 %. In other cases recall has been very height, close to the maximum value. The analysis of video content has shown that it was mainly due to the fact that in some news soccer game scenes have been edited of very similar shots and no method used in temporal segmentation process detected a cut between these shots and in consequence their total length was greater than the value of maximum duration of shot (160 frames) used in the test.

In the second test (Tab. 5) all shots shorter than only 15 frames have been jointed to the next shot, then all consecutive shots shorter than 40 frames and similarly all shots longer than 160 frames have been aggregated. So, shots of the length of 40 to 160 frames have been remained unchanged as the most probably belonging to player scenes. These shots are of the length of 2 to 6 second ± 10 frames of tolerance.

Table 5. Number of selected player pseudo-scenes after aggregating very short consecutive shots as well as very long shots. In the second series of tests all shots shorter than 40 frames have been accepted as short shots and as long shots all shots longer than 160 frames.

	Video	Number of all aggregated pseudo-scenes	Number of selected player scenes	Recall of correct detections of player scenes [in %]	Number of frames in selected aggregated scenes	Share of selected aggregated scenes in total video [in %]	Precision of detection of scenes with player scenes [in %]
1.	Feb. 01	115	94	**100.00**	7570	**48**	**45.74**
2.	Feb. 02	75	59	**100.00**	5415	**44**	**67.80**
3.	Feb. 03	89	67	**100.00**	5426	**44**	**52.24**
4.	Feb. 04	94	81	**100.00**	7188	**58**	**70.37**
5.	Feb. 05	71	56	**92.86**	4249	**44**	**66.07**
	Averages:	88.80	71.40	**98.57**	5970	**47.60**	**60.44**
1.	Sep. 01	30	23	**100.00**	2014	**35**	**73.91**
2.	Sep. 02	56	41	**86.67**	3496	**42**	**85.37**
3.	Sep. 03	49	42	**100.00**	3576	**53**	**83.33**
4.	Sep. 04	23	19	**80.00**	1485	**37**	**63.13**
5.	Sep. 05	22	17	**85.77**	1408	**38**	**58.82**
6.	Oct. 01	66	50	**92.86**	4113	**41**	**66.00**
7.	Oct. 02	33	26	**90.91**	2149	**44**	**84.62**
8.	Oct. 04	101	79	**88.00**	6969	**46**	**60.76**
9.	Oct. 05	45	38	**100.00**	3402	**50**	**81.58**
10.	Oct. 24	121	99	**100.00**	7890	**50**	**63.64**
11.	Oct. 25	101	81	**95.65**	7521	**56**	**75.31**
12.	Oct. 29	45	35	**100.00**	3146	**51**	**40.00**
13.	Oct. 31	45	38	**100.00**	3265	**58**	**86.84**
Total averages for 13 videos:		**56.69**	**45.23**	**93.84**	**3880**	**46.23**	**71.02**

In all cases the recall is optimal or very close to optimal. The precision is also on an acceptable level more than 71%. The total number of frames suggested for further content-based indexing is only 46.23%, so less than 50%. It means that it would be possible to reject more than half of video but despite this almost all sports scenes reported in TV sports news would be indexed.

This is the main advantage of applying the temporal aggregation method. Less than half of video source material can be included in the analyses but even then practically all sports scenes will be detected and analysed in sports categorization process.

5 Final Conclusions

A new method called temporal aggregation has been proposed for a significant reduction of video material analysed in content-based indexing of TV sports news. The method detects and aggregates two kinds of shots: sequences of long studio shots unsuitable for content-based indexing and player scene shots adequate for sports categorization. The results of tests performed have shown its high usefulness. The aggregation parameters used in the first test enable us to limit news to only one quarter of the video material, but in this case only about three quarters (71.16%) of sports scenes would be indexed with the precision of 72.55%. The effectiveness of the detection of player scenes using temporal aggregation measured by the recall of player scenes and the precision of selected shots was more satisfactory for the second set of parameters. Scene recall was almost optimal and attained the level of almost 94%, whereas shot precision was also acceptable – more than 71%. But in this second case it was possible to reject about half of video and despite this almost all sports scenes reported in TV sports news would be indexed. So, as in many computer science solutions we have two opportunities: high recall of the detection of player scenes and significant decrease of time processing due to the reduction of automatic analyses to the half of the video material or not perfect recall of the detection of player scenes but much more important decrease of time processing due to the reduction of automatic analyses down to only the quarter of the video material.

Then, in next tests the influence of the application of both temporal aggregation of shots leading to the reduction of video material to only player scenes as well as of the recognition of video editing patterns on the video retrieval will be examined. An important question is also if sport news broadcasted on other TV channels have so consistent structure that the temporal aggregation method can be applied. The first analyses suggest that all news broadcasts have specific structures and that these structures are relatively similar. It seems that it is a common standard worldwide.

In further research we will define a method of automatic clustering of player scenes of the same sports event. It should be also tested if the grouping of the player scenes of the same sports category but possibly of different sports events would be acceptable in the content-based video indexing process.

References

1. Film Terms Glossary, `http://www.filmsite.org/filmterms.html` (accessed November 15, 2012)
2. Zhang, Y.J., Lu, H.B.: A hierarchical organization scheme for video data. Pattern Recognition 35, 2381–2387 (2002)
3. Choroś, K., Gonet, M.: Effectiveness of video segmentation techniques for different categories of videos. In: New Trends in Multimedia and Network Information Systems, pp. 34–45. IOS Press, Amsterdam (2008)
4. Choroś, K.: Reduction of faulty detected shot cuts and cross dissolve effects in video segmentation process of different categories of digital videos. In: Nguyen, N.T. (ed.) Transactions on CCI V. LNCS, vol. 6910, pp. 124–139. Springer, Heidelberg (2011)
5. Lin, T., Zhang, H.-J.: Automatic video scene extraction by shot grouping. In: Proc. of the 15th International Conference on Pattern Recognition, vol. 4, pp. 39–42 (2000)
6. Jun Ye, J., Li, J.-L., Mak, C.M.: Video scenes clustering based on representative shots. World Journal of Modelling and Simulation 1(2), 111–116 (2005)
7. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. IEEE Trans. on Multimedia 7(6), 1097–1105 (2005)
8. Mohanta, P.P., Saha, S.K.: Semantic grouping of shots in a video using modified k-means clustering. In: Proc. of the 7th International Conference on Advances in Pattern Recognition, pp. 125–128 (2009)
9. Chasanis, V.T., Likas, A.C., Galatsanos, N.P.: Scene detection in videos using shot clustering and sequence alignment. IEEE Trans. on Multimedia 11(1), 89–100 (2009)
10. Liu, Y., Wu, F.: Multi-modality video shot clustering with tensor representation. Multimedia Tools and Applications 41(1), 93–109 (2009)
11. Zhe, W., Zhan-Ming, L., Yan-Fang, Q., Li-Dong, Z.: A novel rough sets based video shot clustering algorithm. Information Technology J. 10(5), 1056–1060 (2011)
12. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Event detection and recognition for semantic annotation of video. Multimedia Tools and Applications 51, 279–302 (2011)
13. Ling-Yu, D., Min, X., Qi, T., Chang-Sheng, X., Jin, J.S.: A unified framework for semantic shot classification in sports video. IEEE Trans. on Multimedia 7(6), 1066–1083 (2005)
14. Bertini, M., Del Bimbo, A., Nunziati, W.: Automatic annotation of sport video content. In: Sanfeliu, A., Cortés, M.L. (eds.) CIARP 2005. LNCS, vol. 3773, pp. 1066–1078. Springer, Heidelberg (2005)
15. Lien, C.-C., Chiang, C.-L., Lee, C.-H.: Scene-based event detection for baseball videos. J. of Visual Communication and Image Representation 18(1), 1–14 (2007)
16. Chena, L.-H., Laib, Y.-C., Liaoc, H.-Y.M.: Movie scene segmentation using background information. Pattern Recognition, 1056–1065 (2008)
17. Huang, Y., Choiu, C., Sandnes, F.E.: An intelligent strategy for the automatic detection of highlights in tennis video recordings. Expert Systems with Applications 36(6), 9907–9918 (2009)
18. Lin, C., Su, C.-H.: Using color strings comparison for video frames retrieval. In: Proc. of the International Conference on Information and Multimedia Technology, pp. 211–215 (2009)
19. Tapu, R., Zaharia, T.: High level video temporal segmentation. In: Bebis, G., et al. (eds.) ISVC 2011, Part I. LNCS, vol. 6938, pp. 224–235. Springer, Heidelberg (2011)

20. Choroś, K., Pawlaczyk, P.: Content-based scene detection and analysis method for automatic classification of TV sports news. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS (LNAI), vol. 6086, pp. 120–129. Springer, Heidelberg (2010)
21. Choroś, K.: Detection of tennis court lines for sport video categorization. In: Nguyen, N.-T., Hoang, K., Je̜drzejowicz, P. (eds.) ICCCI 2012, Part II. LNCS (LNAI), vol. 7654, pp. 304–314. Springer, Heidelberg (2012)
22. Chi-Kao, C., Min-Yuan, F., Chung-Ming, K., Nai-Chung, Y.: Event detection for broadcast tennis videos based on trajectory analysis. In: Proc. of the 2nd International Conference on Communications and Networks (CECNet), pp. 1800–1803 (2012)
23. Choroś, K.: Video structure analysis for content-based indexing and categorisation of TV sports news. Int. J. on Intelligent Information and Database Systems 6(5), 451–465 (2012)
24. Choroś, K.: Video structure analysis and content-based indexing in the Automatic Video Indexer AVI. In: Nguyen, N.T., Zgrzywa, A., Czyżewski, A. (eds.) Advances in Multimedia and Network Information System Technologies. AISC, vol. 80, pp. 79–90. Springer, Heidelberg (2010)

A Graph-Cut-Based Smooth Quantization Approach for Image Compression

Maria Trocan[1] and Beatrice Pesquet[2]

[1] Institut Superieur d'Electronique de Paris,
28 rue Notre Dame des Champs, Paris, France
maria.trocan@isep.fr
[2] Telecom ParisTech,
46 rue Barrault, Paris, France

Abstract. Quantization represents an important aspect in image acquisition and coding. However, the classical quantization algorithms lack of spatial smoothness, especially when dealing with low bitrate constraints. In this paper, we propose a graph-cut-based smooth quantization approach for image compression that can alleviate the artefacts driven by classical quantization algorithms. The best representation for an image using a finite number of levels is obtained by convex optimization, realized by graph-cut techniques and which considers the spatial correlation in the minimization process in addition to the classical distortion approach. We show that even when using a small number of reconstruction levels, our approach can yield better quality results, in terms of PSNR, than JPEG2000.

1 Introduction

Quantization plays a fundamental role in digital image acquisition, representation, processing and transmission. It represents the basis for all the families of lossy encoders. Moreover, it has a close similarity to other image processing tasks, as denoising, segmentation, and data classification. For example, the quantization of an image using Q reconstruction levels can be viewed as a classification or segmentation of the image in Q areas following an intensity homogeneity criterion.

Lloyd-Max algorithm [7] is the classical solution for optimally assigning the quantization levels to an image. However, the main drawback of this quantization method is the lack of spatial regularity of the quantized / reconstructed image. Spatially smooth properties may be useful especially in low-rate compression and can lead to good compression results when a smooth quantizer is implemented jointly with advanced entropy coding algorithms (e.g. JPEG2000 [9,10] image coding standard).

Generally, computer vision problems can be naturally expressed in terms of energy minimization. The methods in [1,3] consist in modelling a graph for an energy type, such that the minimum cut minimizes globally or locally an energy functional. Even with a simple, grid-like design, the graph presents an easy way for representing local segmentation decisions and has powerful computational mechanisms for extracting global segmentation decisions from these simple local (pairwise) pixel similarities. Good energy-optimization results based on graph cuts were obtained in image restoration [2], as well as in motion segmentation [4], texture synthesis in image and video [5], etc.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 498–506, 2013.

In this paper, we propose a graph-cut-based quantization method that enforces global spatial smoothness on the reconstructed image. The best representation for an image using a finite number of quantization levels is obtained by convex optimization, realized by graph-cut techniques and which considers the spatial correlation in the minimization process, in addition to the classical distortion approach. The presented method outperforms the standard JPEG2000, for both natural and noisy images, as it will be shown by the experimental results. The graph-cuts have been succesfully used in previous works [11,12] for subband-based rate-distortion optimization. In this work, the quantization is realized at pixel level and the quantization initialization is performed using the Lloyd-Max algorithm [7], ensuring thus faster convergence.Moreover, due to pixel-based optimization, smoother reconstructions are obatined for the quantized images.

The paper is organized as follows: Section 2 describes the graph-cut based solution for smooth image quantization. Some experimental results obtained with the proposed method for both natural and noisy images are presented in Section 3. Finally, conclusions and future work directions are drawn in Section 4.

2 Adaptive Quantization Using Graph-Cuts

2.1 Graph-Cut Based Minimization

Consider the graph $G = (V, E, W)$ with positive edge weights W, which have not only two, but a set of terminal nodes, V, and connected by a set of edges, E. In [2], Y. Boykov *et al.* find the minimal multiway cut of G by succesively finding the min-cut between a given terminal and the other terminals. This approximation guarantees a local minimization of the energy function that is close to the optimal solution for both concave and convex energy functionals. However, if the function to be minimized is convex, the global minimum can be reached.

In the sequel, we consider the quantization of a real valued image $I^{M\times N}$, $M\times N$ denoting the image resolution. For our optimization problem, the graph is designed as a grid, following the pixel positions within the image; therefore, the position (n,m) represents a vertex within the vertices set V. Having a grid design, the edges are defined as vertical (i.e., $((n-1,m)\rightarrow(n,m))\in E$) and horizontal (i.e., $((n,m-1)\rightarrow(n,m))\in E$) links in the graph edges set E.

Our adaptive quantizer is defined as follows: Q is a positive integer denoting the number of quantization values; for each $q\in Q$, a reconstruction value r_q is defined. The quantization problem can be stated as follows: find the optimal association of the Q-recontruction levels, such that the reconstructed (quantized) image has the maximal spatial coherence. A partition P of V is defined as the label image $(q_P(n,m))\in\{1,...,Q\}$, which translates into: for every pixel position $(n,m)\in V$ and quantizer number $q\in\{1,...,Q\}$, $q_P(n,m)=q$. The partition P denotes thus a particular assignement of the quantizers to the pixel nodes; in Fig. 1 such a partion is shown, using the recontruction values associated to four quantizers (or labels).

A quantized image over Q reconstruction values r and associated with a quantizer/label partition P is denoted in the sequel by $\tilde{I}_{q_P,r}$. Therefore, an optimal quantization $\tilde{I}_{q_{\overline{P}},\overline{r}}$ of the image I is usually given by the solution to the following minimization problem:

$$minimize_{(q_P,r)}\mathbf{f}(\tilde{I}_{l_P,r}, I) \tag{1}$$

where **f** represents some measure of the quantization error.

In our adaptive quantization framework, we have considered **f** as the l_1 and l_2 norm measure; therefore, the distortion associated to the recontruction $\tilde{I}_{q_P,r}$ is given by:

$$\mathbf{f}(\tilde{I}_{q_P,r}) = \sum_{n=1}^{N}\sum_{m=1}^{M} \left\| I(m,n) - \tilde{I}_{q_P,r}(m,n) \right\|^p, \tag{2}$$

where $p \in \{1,2\}$.

However, considering only the distortion induced by the quantization in the minimization process does not guarantee any spatial homogeneity of the resulting quantized image, $\tilde{I}$. To alleviate this problem, we propose to add a smoothness term to the above presented minimization, i.e.:

$$minimize_{(q_P,r)}\mathbf{f}(\tilde{I}_{q_P,r}) + \rho(q_P) \tag{3}$$

where ρ is some penalty function used to promote the spatial regularity of among the quantizers assignement q_P image.

This type of minimization can be optimally solved by graph-cuts [14,2]. In [2] are presented two graph-cut-based algorithms able to reach a minimum for an energy function of the form:

$$E(l) = E_{data}(l) + E_{smoothness}(l) \tag{4}$$

where $E_{smoothness}$ is a smoothness constraint, while E_{data} measures the distortion introduced by an l-partitioning with respect to the original data. In our adaptive quantization optimization, the data energy term is given by the quantization error, $\mathbf{f}(\tilde{I}_{q_P,r})$, while $\rho(q_P)$ represents the smoothness term.

Typical choices for ρ in equation (3) can be expressed in term of isotropic variation functions:

$$\rho(q_P) = \mu \sum_{n=1}^{N}\sum_{m=1}^{M} \mathbf{g}(\|\nabla q_P(n,m)\|), \quad \mu \geq 0 \tag{5}$$

where μ is a regularization constant,

$$\nabla q_P(n,m) = (q_P(n+1,m) - q_P(n,m), q_P(n,m+1) - q_P(n,m)) \tag{6}$$

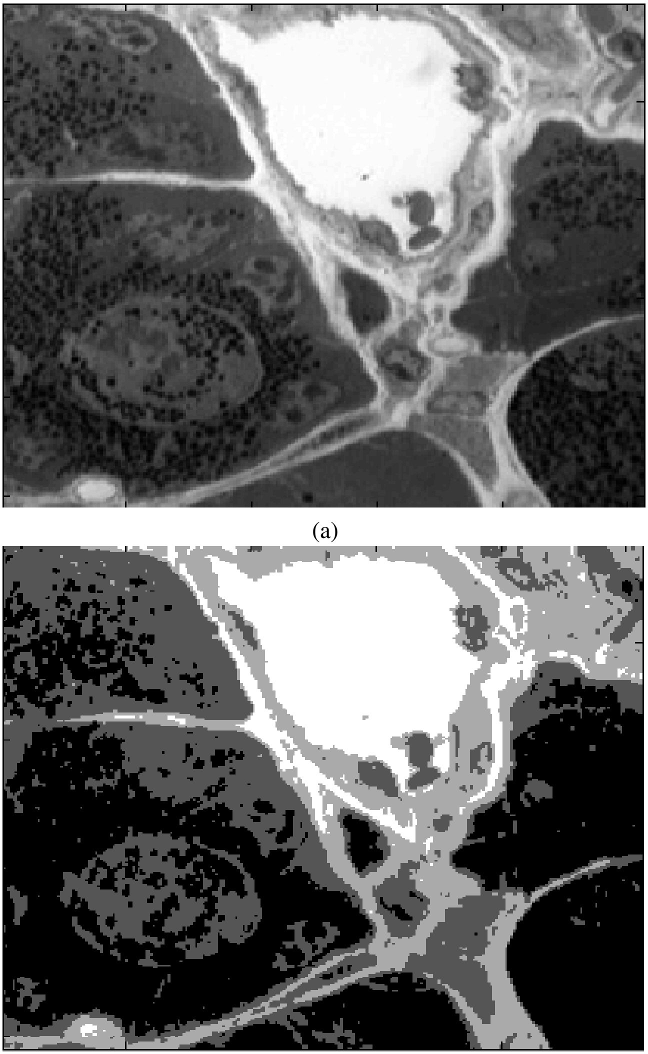
is the discrete gradient of q_P partitioning at location (n,m) and **g** is some linear function. For example, when **g** is the identity function, i.e.:

$$\forall x, \quad \mathbf{g}(x) = x, \tag{7}$$

ρ is then the classical isotropic total variation.

A more flexible form for **g** is given by the truncated linear function defined as:

$$\forall x, \quad \mathbf{g} = \begin{cases} x, & if \quad x < \zeta \\ \zeta, & otherwise \end{cases} \tag{8}$$



(a)

(b)

Fig. 1. Example of a 4-label graph-cut partitioning of the ”Pancreas” image: (a) original image, (b) 4-label partition

where $\zeta > 0$ is a limiting constant. Another interesting choice of **g** is the binary cost function:

$$\forall x, \quad \mathbf{g}(x) = \begin{cases} 0, & if \quad x = 0 \\ 1, & otherwise. \end{cases} \tag{9}$$

Note that several combinations of **f** and **g** could be considered for the minimization in equation(3).

2.2 Adaptive Quantization Algorithm

The proposed quantization method has two stages: in a first stage, we choose the number of quantization levels Q and obtain the first reconstruction values $r^{(0)}$, i.e. at iteration 0, using the Lloyed-Max algorithm [7] (I). Then, in a second stage, we pass to the minimization of (3) using the $\alpha-\beta$-swap algorithm as described in [2]; this is denoted by (II-1) in the algorithm description.

Algorithm pseudo-code:

```
Given $Q$,
(I) $r^{(0)} = lloyd\_max(I,Q)$
for i = 1, ..., max_iterations
    (II-1) $q_{P=i} = min_{q_P}\mathbf{f}(\tilde{I}_{q_P,r^{i-1}}) + \rho(q_P)$
    (II-2) $r^{i+1} = min_r\mathbf{f}(\tilde{I}_{q_{P=i},r})$.
```

The data energy term (E_{data} in (4)) is given by the quantization error, $\mathbf{f}(\tilde{I}_{q_P,r})$. The smoothness term ($E_{smoothness}$ in (4)) is given by $\rho(q_P)$ and is dynamically computed at each step of the $\alpha-\beta$-swap algorithm using one of the possible implementations, as in equations(7)-(9).

Using the partitioning in (II-1), $q_{P=1}$, obtained by graph-cut minimization, the reconstruction values set r is updated in a second step (II-2) (and denoted by r^1 using the combinatorial optimization in [13], and a new a graph-cut optimization of (3) is done using the new r^1. The algorithm stops after a finite number of iterations -user specified- or when the quality difference, in terms of PSNR, between two consecutive iterations is lower than a given threshold τ.

The proposed algorithm converges quickly—typically $3 \leq max\_iterations \leq 10$ are sufficient for PSNR convergence to the second decimal. The optimal solution of (3) is found then as $(q_{P=i}, \tilde{I}_{q_{P=i}}, r^{i+1})$.

3 Experimental Results

For our simulations, we have considered three representative test images: Peppers (256×256 pixels), Pancreas (256×256 pixels) and Cameraman (256×256 pixels), which have been selected for their different texture characteristics. We consider greyscale image quantization where the number of reconstruction levels ranges in $Q \in \{8, 13, 16, 23, 28, 32, 38\}$.

We have considered 4 implementations for the minimization in (3): the first one, that we refer in the followings as $GC1$, implements the function $\mathbf{f}$ as the l_1 norm and the function $\mathbf{g}$ as in (9). The second one, refered as $GC2$, implements $\mathbf{f}$ as the l_2 norm and $\mathbf{g}$ as the binary cost in (9). $GC3$ is given by the l_1 norm for $\mathbf{f}$ and $\mathbf{g}$ is defined as in (7); finally, $GC4$ is obtained using the l_2 norm for $\mathbf{f}$ and (7) is used for $\mathbf{g}$ definition. In order to prove the efficiency of the proposed quantization approach, the results obtained using the above setup have been compared to the ones obtained using JPEG2000 [6], in a spatial, transformless setup (realized using the parameter $-Flev0$). The bitrate estimation for our quantization method has been obtained by passing the label matrix l_P for the

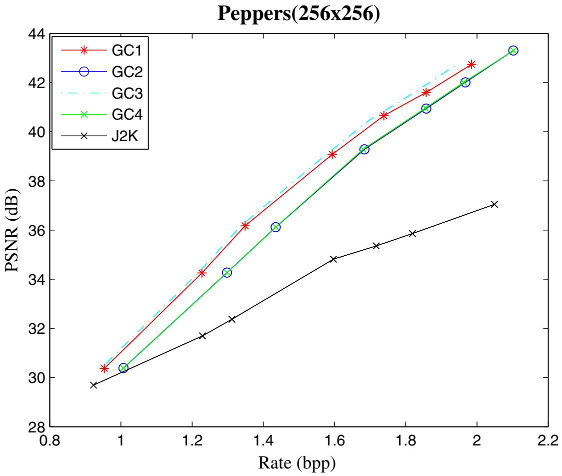


Fig. 2. Performance of the proposed quantization algorithm (dBs) for ”Peppers” test image, as a function of the rate (bpp)

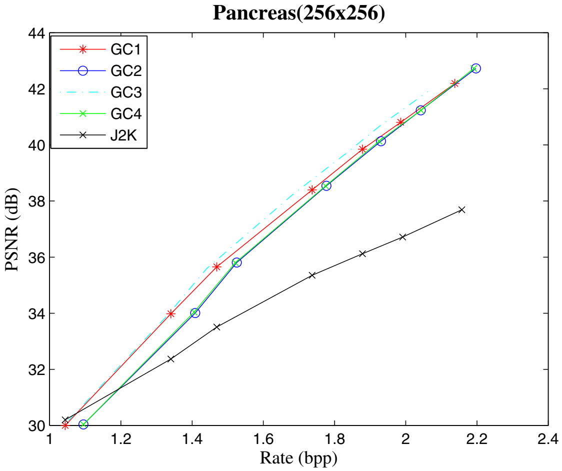


Fig. 3. Performance of the proposed quantization algorithm (dBs) for ”Pancreas” test image, as a function of the rate (bpp)

different Q-s to a lossless JPEG2000 setup (realized using the option $-Frev$). This way, by using the same entropy coder, a fair comparison of the quantization methods is realized. The proposed algorithm stops whenever the quality (PSNR) difference between two consecutive reconstructions is less than a threshold $\tau = 0.003$.

As it can be seen in Figs. 2 - 4, the proposed quantization method leads to higher-quality results, having an average gain of 0.5 dBs at low rates and more than 3 dBs at high bitrates w.r.t. JPEG2000 quantization (denoted by $J2K$). Moreover, it can be

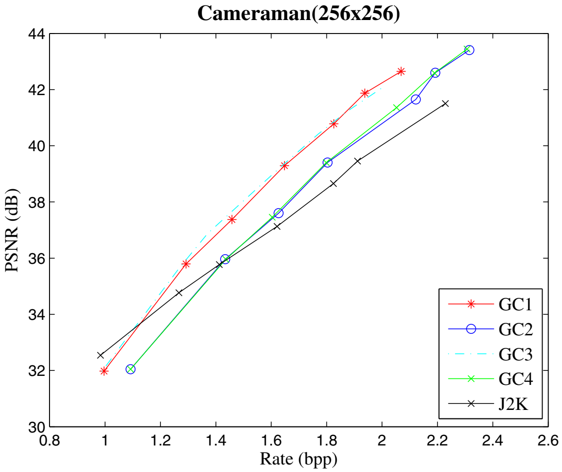


Fig. 4. Performance of the proposed quantization algorithm (dBs) for "Cameraman" test image, as a function of the rate (bpp)

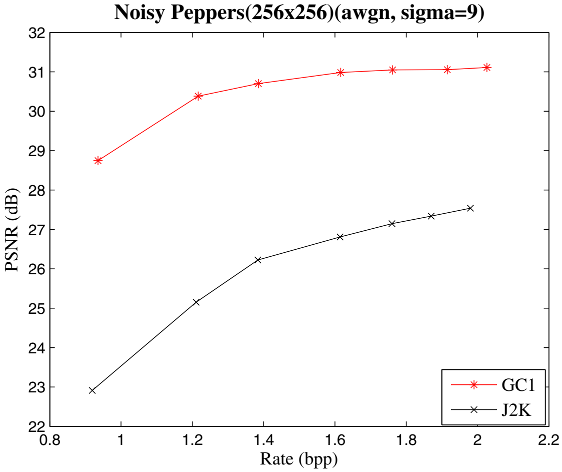


Fig. 5. Performance of the proposed quantization algorithm (dBs) for "Peppers" test image altered with zero-mean white gaussion noise ($\sigma = 9$), as a function of the rate (bpp).

seen that $GC1$ and $GC3$ setups have an average gain of ≈ 1 dBs w.r.t $GC2$ and $GC4$, therefore we can conclude that l_1 norm is not only a good choice from the computational point of view, having a reduced complexity w.r.t l_2 norm, but it also represents the best implementation of the Φ function. It should be noted that for the results in figures 2 - 4 the regularization constant μ has been empirically set to 1.

The proposed quantization method has been also tested in the presence of noise. We have altered the test images by adding zero-mean white gaussian noise of variance $\sigma = 9$.

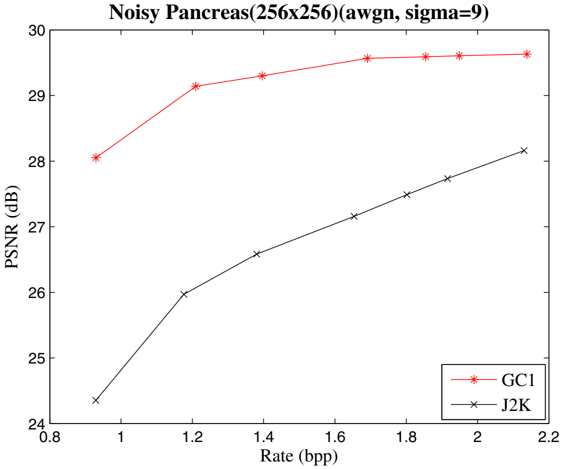


Fig. 6. Performance of the proposed quantization algorithm (dBs) for ”Pancreas” test image altered with zero-mean white gaussian noise ($\sigma = 9$), as a function of the rate (bpp)

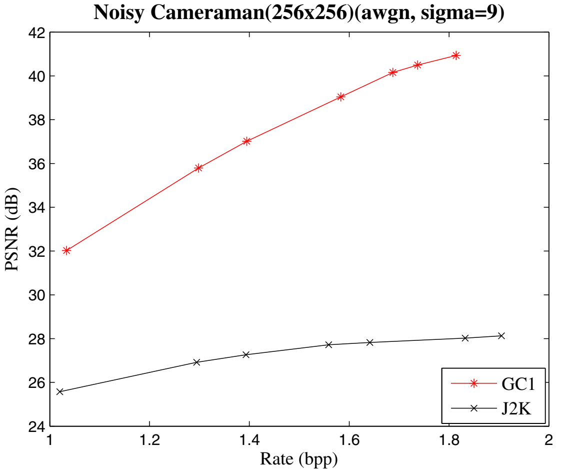


Fig. 7. Performance of the proposed quantization algorithm (dBs) for ”Cameraman” test image altered with zero-mean white gaussian noise ($\sigma = 9$), as a function of the rate (bpp)

As illustrated in figures 5 - 7, our approach outperforms JPEG2000 quantization with an average gain of ≈ 5.5 dBs (only $GC1$ setup has been considered, given the efficiency proved on the original images). In order to enforce the smoothness in the noisy environment, the regularization parameter μ has been experimentally chosen equal to 10.

4 Conclusion

In this paper, we have presented a new quantization method based on convex optimization implemented using graph-cuts and which minimizes not only a distortion measure but also imposes a spatial smoothness constraint. Unlike classical methods, the proposed approach allows to enforce spatial regularity in the quantized image. As shown by our simulation results, the presented method leads to high-quality results, for both natural and noisy images. In particular, in presence of noise our quantization approach has an average gain of ≈ 5.5dBs in comparison with JPEG2000.

References

1. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 1124–1137 (2004)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1222–1239 (2001)
3. Boykov, Y., Veksler, O.: Graph cuts in vision and graphics: Theories and applications. In: Handbook of Mathematical Models in Computer Vision, pp. 79–76. Springer (2006)
4. Schoenemann, T., Cremers, D.: Near Real-Time Motion Segmentation Using Graph Cuts (2006)
5. Zhang, Y., Ji, X., Zhao, D., Gao, W.: Video coding by texture analysis and synthesis using graph cut. In: Proceedings of Pacific-Rim Conference on Multimedia, China (2006)
6. Marcellin, M.W., Lepley, M.A., Bilgin, A., Flohr, T.J., Chinen, T.T., Kasner, J.H.: An overview of quantization in JPEG 2000. Signal Processing: Image Communication 17(1), 73–84 (2002)
7. Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28, 129–137 (1982)
8. Wu, X.: On convergence of lloyd's method. IEEE Transactions on Information Theory 38, 171–174 (1992)
9. JBIG-JPEG, JPEG2000 Part I Final Committee Draft 1.0, ISO/IEC JTC 1/SC 29/WG 1 ITU-T SG8, ISO (March 2000)
10. Taubman, D.S., Marcellin, M.W.: JPEG2000: Image Compression Fundamentals, Standards and Practice. Kluwer, Boston (2002)
11. Trocan, M., Pesquet-Popescu, B.: Graph-cut rate distortion optimization for subband image compression. In: Proc. of the IEEE European Signal Processing Conference, EUSIPCO 2007, Poznan, Poland (September 2007)
12. Trocan, M., Pesquet-Popescu, B., Fowler, J.E.: Graph-cut rate distortion algorithm for contourlet based image compression. In: Proc. of the IEEE Int. Conf. on Image Processing, San Antonio, Texas, U.S.A. (September 2007)
13. Chaux, C., Jezierska, A., Pesquet, J.-C., Talbot, H.: A spatial regularization approach for vector quantization. Springer Journal of Mathematical Imaging and Vision 41(1-2), 23–38 (2011)
14. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 147–159 (2004)
15. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient N-D image segmentation. International Journal of Computer Vision 70, 109–131 (2006)

Indexed Bloom Filters for Web Caches Summaries

Sylvain Lefebvre

LISITE - ISEP
28, rue Notre-Dame des Champs,75006 PARIS, France

Abstract. The popularity of web applications has shown rapid growth on Web Services and Cloud Computing. One of the major challenge in Web applications is that it needs to store lots of assets in cache memory to garantee responsive feeling for the user. User satisfaction is an important constraint imposing Quality of Service requirements. Therefore the use of distributed caches becomes necessary for numerous applications, fostering the need of efficient load distribution strategy. However some services (for example, on demand document searching) also require computations in addition to caching. This article presents a new load balancing policy called Block Based Workload And Cache Aware load balancing (BB-WACA) policy aimed at reducing the information retrieval time of the processor using Bloom Filter Based cache summaries. The BB-WACA is a centralized load balancing algorithm which routes the incoming requests to the machine where the underlying data will most probably be in its cache without overloading any nodes in the cluster.

1 Introduction

Large scale web caches represent a large part of todays performance hotspots of web based applications. It is therefore important to provide efficient algorithms to route requests towards the assets stored in these systems. Bloom filters or similar structures are often used to express caches summaries, in order to achieve distributed caching of web assets or for enhancing data processing services [1,2,3,4]. These strategies can be used when a certain amount of computation is needed in addition of the access to cached data, as in [1]. However, the question remains open on how to find which filter relates to which machine in a fast and efficient way. To achieve this particular goal we propose a new cache aware load balancing policy called Block Based Workload And Cache Aware algorithm (BB-WACA) which is an improvement of the WACA policy described in [4]. This new policy takes advantage of the sliced bloom filters data structure [5,3] to create an index associating requests hashes and the filters containing this hash.

This article will show that using a small and efficient index for retrieving bloom filters can provide performance enhancement compared to naive exhaustive search in summaries. Besides, most of similar algorithms do not take load

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 507–516, 2013.

of the machines into account, which can in turn affect performance in the case where requests follow a skewed distribution.

To demonstrate these improvements, background on Bloom filters is introduced in section 2. The Block Based Workload And Cache aware algorithm is then described in section 3 and related works in section 4. The algorithm is compared to previous versions with a discrete event simulator in section 5. Finally, conclusion and improvement tracks will be discussed in section 6.

2 Background on Bloom Filters

The WACA algorithm is based on the generation of compact summaries of machines cache content. In order to be able to generate these compact summaries, we need to use an efficient data structure. Such a data structure already used in this context is the Bloom Filter [6,2,3].

2.1 Bloom Filters

A Bloom Filter [6] is a compact probabilistic data structure, allowing fast membership query of an element in a set. For a Set $\{S\}$ of data elements of size N, a $m > N$ bit array will be created, with each bit initialized to *false* (0). Each Value (v) already present in $\{S\}$ is hashed with k discrete hash functions $h_0(v), ..., h_k(v) \mapsto [0, m]$ and each bucket (bit position) indicated by the hash functions is then set to *true* (1). It is therefore possible to know if a value is already present in $\{S\}$ or not by hashing it against all hash functions and for each hashed value, looking at the corresponding bucket in the bit array will return either $true(1)$ or $false(0)$. If one of these values is $false$, then we know with certainty that the element is not present in $\{S\}$. When all values are *true*, then we can assume that the value is present in $\{S\}$ with an error probability called the *false positive probability.* A *false positive* is the situation where the query to the bloom filter returns that the element belongs to the set, while it actually doesn't [6]. This probability depends on the ratio between the number of elements n inserted in the filter and the filter size m. As noted by Fan et al. in [2], the false positive probability $P(f)$ can be calculated from the following formula:

$$P(f) \approx \left(1 - e^{kn/m}\right)^k \tag{1}$$

From this equation, it is clear that the false positive probability depends on the filter filling ratio, eg. the number of bits set to 1 divided by the filter size m, and on the number of hash functions k. Therefore the size of the filter must be carefully set in order to minimize this false positive rate and minimize the number of hash functions to use. To use Bloom filters as cache summaries, it is necessary to calculate their optimal size efficiently given the expected number of stored objects in the summarized cache. Several works such as [5,3], show that dividing the filters into multiple "slices" reduces the number of false positive by avoiding collisions between different hash functions.

2.2 Block Bloom Filters

A block or sliced Bloom filter is a bloom filter divided into k slices of equal sizes. Each slice is associated to one of the k hash functions. Each hash function will set or query only 1 bit in its respective slice. The work from Almeida et al. in [3], shows that the false positive probability in a bloom filter rises with the number of bits set to one in the filter. This probability starts degrading when the filter is half full, therefore, it is possible to compute the size m of the slices with the following equation, where c is the expected number of objetcs to store in the filter:

$$m \approx \frac{c}{-\ln(1/2)} \tag{2}$$

It is also possible to compute the optimal number of hash functions, and hence the number of slices, from the wanted false positive rate f according to [2,3]:

$$k \approx \log 2(\frac{1}{f}) \tag{3}$$

3 Block-Based Workload and Cache Aware Algorithm

In the case of cache aware load load balancing, it is of importance to efficiently find the machine containing the targetted data, among all avilable machines. The BB-WACA load balancing policy has been designed regarding three conditions:

- Keep track of cache information of each machine,
- Maintain a balanced load among the machines,
- The algorithm should be fast.

The principle of the block based waca algorithm is similar to cache summaries based algorithms that can be found in [1,2,3,4] in that each machine in the system is associated to a Bloom Filter based summary of the content of its cache. However, most of these works will select the most appropriate node by iterating over the whole list of machines in the system. This slow selection process is generally compensated by the gain obtained from the fact that data for the request is in the cache and therefore the request will execute faster. However this gain can sometimes be mitigated by other factors such as network latency. Therefore it is important to reduce the selection time to minimum in order to achieve low latency and high throughput.

The Block Based Waca algorithm takes advantage of the sliced structure of the Sliced Bloom Filters to index the content of the first block of each machine's filter in the system. This indexing allows to reduce the number of possible candidates for each request thus reducing node selection time.

3.1 Algorithm Description

Block Based Waca uses efficient data structures to keep track of machines load and cache summaries information. Machines are sorted according to their current load with a Fibonacci Heap [7]. For each machine in the system, a Block Based Filter is set up to keep track of the requests that are sent to this specific server. The filters are dimensioned according to the formulas in section 2.2, based on the user configured false positive rate, and the memory capacity of each machine. The third main data structure is called the T index. This T index works as hash table of limited capacity, where the number of buckets is equal to the size of the largest block in the available block bloom filters. Each Bucket contains an unbounded linked list of block based filters. Every time a new entry is added to a block bloom filter, the first hash h_0 is used to determine a position $T[h_0]$ in the index, and the filter will be added to the linked list if it is not already present in it. The main use for this index is to enable a reduction of the possible candidate filters for a given hash.

The BB-WACA algorithm is written in listing 1. It runs every time a new request arrives for one of the machines in the system. On this event, the algorithm computes k_{max} hash values from the parameters of the request (line 4). The value k_{max} is the highest number of hash functions among all filters. The first hash (h_0) is used to query the corresponding position in the T index (line 5,6). If the queried position is not empty, the algorithm iterates over the filter list (line 8). On each filter, the algorithm checks wether the associated machine is overloaded or not and if the hashes are present in the filter (lines 9,10). If both these conditions are met the machine is selected as target and directly returned. Otherwise, if the hashes are not found in any of the filters, the algorithm selects the least loaded node in the current list (lines 12 to 18). If the load on this least loaded candidate is not too high, then this target machine is returned. If the load is too high on this machine, then the least loaded among all the system nodes is selected (line 19). In both cases, the k hashes are inserted in the corresponding filter before returning (line 23), this action in turn updates the T index to add the filter in the corresponding bucket.

The *least loaded* node is the node which has the lowest request count among the available nodes at that instant of time. The request count is nothing but the number of requests that the particular node is currently processing. Even though several approaches exist, for finding the least loaded node with the help of a request counter as shown in the Join Shortest Queue policy showed some good results over time [8]. A node is considered overloaded when its request count exceeds two times the number of logical cores available on the machine.

3.2 Hash Function Choice

In order to implement a fast hashing scheme for the requests parameters the MurmurHash function [9] is used. This hash function presents a good tradeoff between execution time and collision rate which makes it a good choice for use in hash tables and other randomized data structures.

Algorithm 1. Block Based WACA

```
function BBWACALOADBALANCE($nodesList$, $request$)
    $target \leftarrow$ Nil
    $bf \leftarrow$ Nil
    $H \leftarrow$ hash($request$)
    $i \leftarrow H[0]$
    if $T[i]$ not empty then
        $bf \leftarrow T[i]$.head
        while $bf not$ Nil do
            $load \leftarrow$ getNodeLoad($bf$)
            if $bf$.lookup($H$) AND $load =$ true then
                return $bf$.node
            else if $load =$ true then
                $target \leftarrow$ getLeastLoaded($target$, $bf$.node)
            end if
            $bf \leftarrow bf$.next
        end while
    end if
    if $target =$ Nil then
        $target \leftarrow$ getLeastLoaded($nodesList$)
        $bf \leftarrow target$.getFilter()
    end if
    if not $bf$.lookup($H$) then
        $bf$.insert($H$)
    end if
    return $target$
end function
```

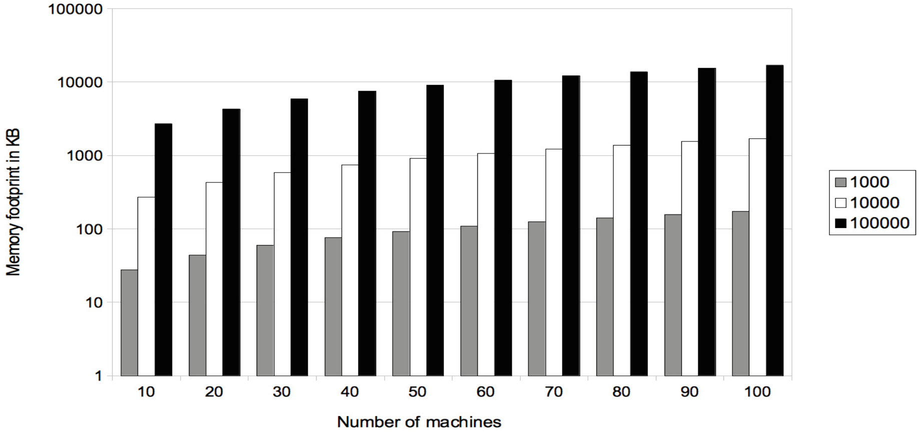


Fig. 1. Total filter size vs. number of machines

3.3 Algorithm Analysis

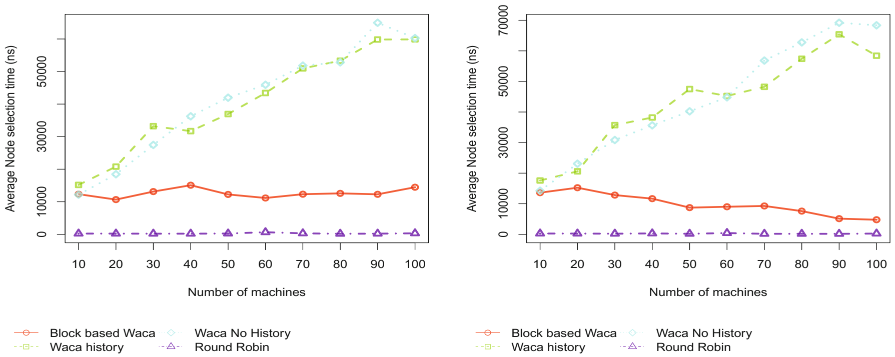
Memory Footprint. For each machine in the system, a block based bloom filter is created. The size of the bloom filter is set according to equations 3 and 2. Each filter depends on the cache capacity of the node it represents and on the configured false positive rate the user wishes. Hence, the memory footprint of the algorithm is given by the sum of all filters respective sizes, and grows linearly with the number of nodes. However, since the filters are implemented as bit arrays, the overall size does not take a large amount of virtual memory. As an illustration, figure 1 shows the total memory size consumed by filters setup for $1,000$, $10,000$ and $100,000$ objects, and 0.1% false positive rate. The filter size axis is in logarithmic scale to improve readability. In the maximum setting, which is 100 nodes, with each of them storing 100 000 objects, for a theorical total of 10 million objects, the size of the overall filters would be around 17Mb, which is still a bearable size for modern machines main memory.

Time Complexity. To evaluate execution time complexity of the WACA algorithm, we have to distinguish the possible cases: best, worst and average execution time. In the best case, the target machine will be selected in constant time by the algorithm. This case happens when the current request is not added to any filter and we choose the least loaded node as the target machine. Since nodes are stored in a sorted list, this selection is done in constant time. The worst possible case is bounded by the total number of nodes in the system. Each bucket in the index is only bounded in size by the total number of machines in the system. Average and worst case complexity therefore depend on the average number of elements in the T index. The frequency at which a new node is added to a given bucket i is difficult to evaluate because it depends on the requests distribution law: the more frequent the request, the more chances there is to add a node to the corresponding bucket. Figures 2a and 2b, compare the WACA algorithm average machine selection time with the ones of three other load balancing policies: a Round Robin strategy and two previous versions of the WACA algorithm described in [4]. This timing is compared to the number of machines in simulated system. As expected from the analysis, machine selection time for Block based WACA does not vary much with the number of machines, but appears to be higher when request distribution follows a Zipf Law (figure 2a), which is frequent in the case of web workloads according to [10]. In the case of the a Gaussian request distirbution, as shown in figure 2b, as the number of available nodes goes up, the selection time seems to be reduced. This is due to the fact that since request are more venly spread over the possible node lists (the buckets), the size of each list is reduced, and the number of iterations to find a node able to answer a given request is lower.

In a more realistic setting however, the number of possible machines for single type of request will be bounded by the number of machines having the capacity to answer the request.

4 Related Works

Content-Aware load distribution strategies generally aim to optimize requests allocation according to the information found in web requests body. In [11] Gou *et al.* use a Bloom Filter to achieve load balancing in a Network Intrusion Detection System. In order to maintain balanced load among processors of the NIDS, a weighted random load balancing algorithm is used. In this strategy, each processor is selected at random, with a probability affected by a weight. These weights



(a) Zipf law request distribution

(b) Gaussian Law request distribution

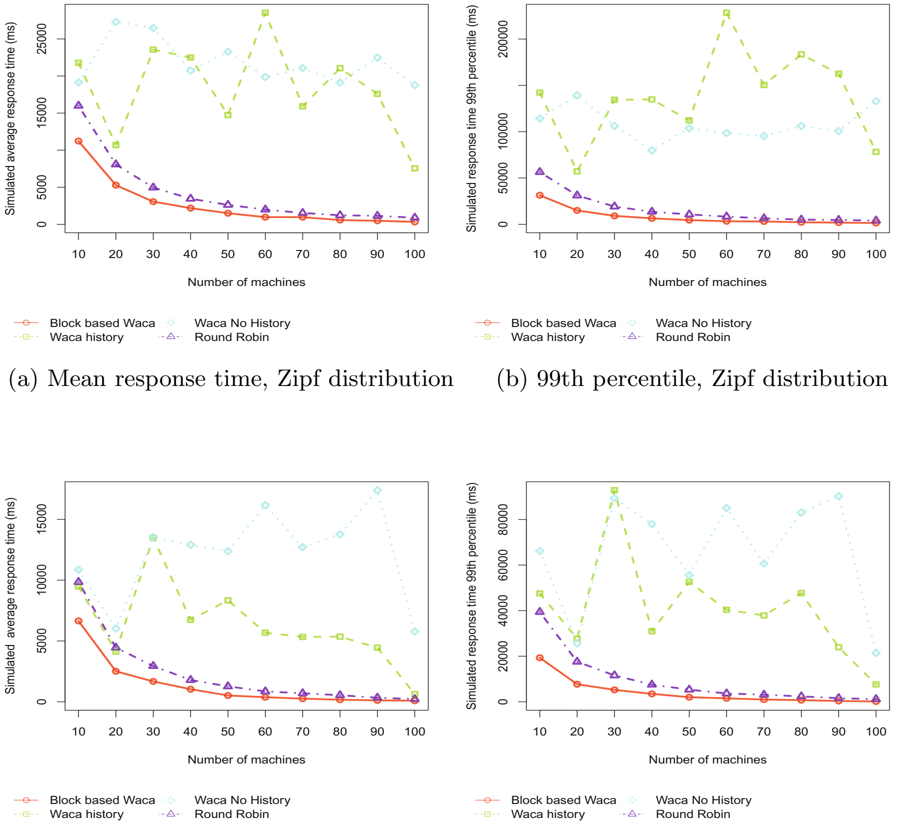
Fig. 2. Machine Selection timings

for each processors are periodically adapted according to preceding analysis of the network communication flows. In order to incrementally calculate a new set of weights, the algorithm builds a Bloom Filter identifying network flows used to calculate the current weights, and another bloom filter identifying network flows that were never seen before and are used to calculate the next set of weights.

In the domain of web services, request content analysis is commonly based on the Universal Request Locator string. For example in Pai *et al.*[12], the load balancing scheme is to forward requests concerning the same file to same node. The goal in these types of policies is to optimize disk accesses time.These type of strategies are also known as *cache aware* load balancing.

In [13] Rohm *et al.* focus on optimizing response time of black-box database components deployed on a cluster by generating signatures from received queries, and store them for each node. This gives a good approximation of the nodes cache content. In [14] O'Gorman *et al.* worked on synchronizing queries through a dedicated middleware to benefit from system cache access. This technique showed important speed-up on a well known benchmark. Dominguez *et al.* showed in [15] that cooperative caching, which is a common cache pool for all nodes in the cluster, can be efficiently used with cache-aware load balancing because this technique reduces the amount of network traffic for retrieving data.

In cache-aware load balancing solutions, the request is sent to the replica where the needed data will most probably be in the cache. The load balancer has some knowledge of the cache state of all the replicas stored in global Cache-Table (CT) local to the load balancer. The CT contains some knowledge of the request semantics and, sometimes, the requests' history. In WEB caches, CT can sometimes be an approximation of the cache state through a summary technique [2] based on Bloom Filters. Thanks to this knowledge, the proxy will be able to dispatch the requests to the nodes where the needed data is *most probably* within the disk cache. However, this technique does not address load-balancing among the machines in the case request distribution is skewed.



(a) Mean response time, Zipf distribution (b) 99th percentile, Zipf distribution

(c) Mean response time, Gaussian distribution (d) 99th percentile, Gaussian distribution

Fig. 3. Simulated response times

5 Simulation

In order to test and compare the Block Based Waca algorithm, we used a load balancing simulator called Simizer [16]. This tool is a library providing a way to programmatically evaluate and measure the behavior of different load balancing policies regarding different metrics such as response time, latency, and machine selection time. It is an event-based simulator. To generate requests, the user can specify different random laws to define the parameters distributions, requests frequencies, and number of available machines. BB-WACA was compared to previous versions of the Waca Algorithm described in [4] and to the Round Robin policy. The simulated workload consists in a one-minute long generative process with an average rate of 1000 requests per second. Requests are chosen among 1200 different parameters. These parameters determine which objects are accessed by the request. Each machine stores its accessed objects in a cache following the Least Recently Used object policy. If the requested object is not in the cache, it is retrieved from the simulated disk and put in the cache, causing the request to last longer, due to disk read latencies. Two types of parameter distributions were used: the first is a Zipf law with a skew parameter of 0.8, which is a reasonable assumption to model web requests distributions according to [10]. On the other hand, a Gaussian request distribution law was chosen, in order to establish the sensitivity of the algorithm to request distribution.

This simulation workload was ran against four different load balancing policies, namely the Block Based Waca policy, the Round Robin policy, and the WACA policies described in [4]. Figures 3a and 3c plot the mean of response times against the number of machines, respectively for Zipfian and a Gaussian parameters distribution. Figures 3b and 3d show the 99th percentile of the response times in the same fashion. It appears that old versions of the WACA policies are very sensitive to the number of objects and to the number of nodes in the system. At small scale (¡20 machines) the Waca History strategy shows good performance under a Gaussian request distribution, but as the number of machines grow, performance starts to degrade. After 20 machines, only small improvements to the mean response time and 99th percentile can be seen. Overall the mean response time and 99th percentile curve show similar trends for Block Based Waca and Round Robin policies under both request distributions. However a slightly better performance is shown when the Block Based Waca policy is used. As the number of machines grow, the Round Robin policy catches up with the Block Based Waca approach, as more ressources are available to answer to the requests. These results show that the Block Based Waca strategy can guarantee a better response time than a Round Robin based strategy with the same amount of ressources.

6 Conclusion and Discussion

We have shown that so far it is possible to significantly improve search time in multiple bloom filters by indexing the first block of the list of filters. This strategy was used to accomplish locality aware load balancing in a distiributed web cache cluster setting. Simulations of the algorithms shows significant improvements against previous versions of the algorithm and naive Round Robin distribution that do not use the indexing scheme. The same simulations also show little correlation between the number of machines in the system and machine selection time. This result is important because it shows that the algorithm can be used at large scale without reducing throughput of the system.

However the BB-WACA algorithm will have to be tested more extensively through different settings such as a larger number of objects and larger caches for the machines and compared to other existing locality based algorithms. In this regard, simple Bloom Filters can be replaced by Counting Bloom Filters like in [2] or a list of Counting Filters as in [1] in order to take cache eviction schemes into account. This improvement would have to be carefully implemented as frequent updates on the filters list could reduce the indexing table effectiveness.

References

1. Dominguez-Sal, D., Aguilar-Saborit, J., Surdeanu, M., Larriba-Pey, J.L.: Using Evolutive Summary Counters for Efficient Cooperative Caching in Search Engines. IEEE Transactions on Parallel and Distributed Systems 23(4), 776–784 (2012)
2. Fan, L., Cao, P., Almeida, J., Broder, A.Z.: Summary cache: A scalable wide-area web cache sharing protocol. Technical report (1998)
3. Almeida, P.S., Baquero, C., Preguiça, N., Hutchison, D.: Scalable bloom filters. Inf. Process. Lett. 101(6), 255–261 (2007)
4. Lefebvre, S., Raja Chiky, S.P.K.: Waca: Workload and cache aware load balancing policy for web services. In: 1st International Conference on Systems and Computer Science (2012)
5. Chang, F., Chang Feng, W., Li, K.: Approximate caches for packet classification. In: INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 4, pp. 2196–2207 (March 2004)
6. Bloom, B.H.: Space/time trade-offs in hash coding with allowable errors. Commun. ACM 13(7), 422–426 (1970)
7. Fredman, M.L., Tarjan, R.E.: Fibonacci heaps and their uses in improved network optimization algorithms. J. ACM 34(3), 596–615 (1987)
8. Bonomi, F.: On job assignment for a parallel system of processor sharing queues. IEEE Transactions on Computers 39(7), 858–869 (1990)
9. Appleby, A.: Murmurhash (2008)
10. Breslau, L., Cao, P., Fan, L., Phillips, G., Shenker, S.: Web caching and zipf-like distributions: Evidence and implications. In: Proceedings of the Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 1999, pp. 126–134. IEEE (1999)
11. Gou, C., Zhao, R., Diao, J.: A load-balancing scheme based on bloom filters. In: Second International Conference on Future Networks, ICFN 2010, pp. 404–407 (2010)
12. Pai, V.S., Aron, M., Banga, G., Svendsen, M., Druschel, P., Zwaenepoel, W., Nahum, E.: Locality-aware request distribution in cluster-based network servers. SIGPLAN Not. 33(11), 205–216 (1998)
13. Rohm, U., Bohm, K., Schek, H.J.: Cache-aware query routing in a cluster of databases. In: Proceedings of the 17th International Conference on Data Engineering, pp. 641–650 (2001)
14. O'Gorman, K., Agrawal, D., El Abbadi, A.: Multiple query optimization by cache-aware middleware using query teamwork. In: Proceedings of the 18th International Conference on Data Engineering, p. 274 (2002)
15. Dominguez-Sal, D., Perez-Casany, M., Larriba-Pey, J.: Cache-aware load balancing vs. cooperative caching for distributed search engines. In: 11th IEEE International Conference on High Performance Computing and Communications, HPCC 2009, pp. 415–423 (2009)
16. Lefebvre, S., Sathiya Prabhu, K., Chiky, R.: Simizer: A cloud simulation tool (2013), https://forge.isep.fr/projects/simizer/

The MapReduce Approach to Web Service Retrieval

Adam Czyszczoń and Aleksander Zgrzywa

Politechnika Wrocławska, Division of Computer Science and Management,
Institute of Informatics. Wybrzeże Wyspiańskiego 27, PL50370 Wrocław, Poland
{adam.czyszczon,aleksander.zgrzywa}@pwr.wroc.pl
http://www.zsi.ii.pwr.wroc.pl/

Abstract. In this paper we address the problem of Web Service Retrieval by presenting the MapReduce approach to distributed and scalable indexing of Web Services. Services are modelled in vector space which enables obtaining ranked search results. Additionally, in order to reduce size of the inverted index and to decrease the search time, presented approach uses the Merged Weight Vector method. The research is supported with implementation of proposed approach by which we conducted experiments. The experimental results confirm that the proposed approach is effective and allows to reduce index construction time and Web Service retrieval time in a scalable manner.

Keywords: Web Service Retrieval, Web Service, MapReduce, distributed indexing.

1 Introduction

Our recent studies investigating the market of Web Services [1] indicate that by the end of 2013 its size will range between 27.6 and 37.1 billions of dollars. This concerns both SOAP and RESTful Web Services. We also project the total number of Web Services distributed over the Internet to range between 722.5 thousand and 136.3 million. Still the key problem for service–oriented web systems is finding services which satisfy information needs of potential users [2]. Additionally, the category-based Service Discovery methods which are fundamental to the Service Oriented Architecture (SOA) paradigm seem to be insufficient since they result in extreme large number of irrelevant services to investigate by the user [3].

The goal of this paper is to propose an effective and scalable Web Service distributed indexing approach to the Web Service Retrieval. The key problem in Web Service Retrieval is to choose the appropriate inverted index structure that enables fast and efficient retrieval of services while keeping its size relatively small. Web Services are composed of many elements and keeping information about all of them may lead to huge index size. In order to solve the above problems, our approach uses the MapReduce – introduced by Google programming model for processing and generating large data sets [4]. Additionally, presented

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 517–526, 2013.

approach uses the Merged Weight Vector [2] which allows to avoid comparing each of the parameters separately to the query by the searching mechanism and reduces the inverted index size. In order to enable ranking of search results, service components are modelled in vector space.

2 Related Work

The studies on Web Service Retrieval started with keyword-based approaches [5,6]. However, they were limited by low precision and recall. The concept of service retrieval that used the idea of inverted index was presented in [7,8,9]. However, the indexed services where treated as sets of terms ("bag of words") and therefore indexed at the service level. Indexing at operational level allow to achieve higher effectiveness [3]. Further changes to the similarity degree introduced in [10,11,12] resulted in higher retrieval effectiveness. Despite the fact the above study present very interesting results, none of them considered distributed and scalable indexing for Web Service Retrieval.

The MapReduce method was firstly used in 2003 to simplify construction of the inverted index for handling searches at *Google.com* [13] and proved its very high efficiency for this task. However, the basic concept of distributed indexing using this approach was presented in [14]. Authors proposed dividing the inverted index into term partitions according to the first letter. The mapper output were the term-documentID pairs divided into r segments, where each segment was assigned to one reducer.

The concept of Merged Weight Vector (MWV) was proposed in our previous research on parametric indexing [2] where we also presented the definition of a Web Service. The usage of Merged Weight Vector showed that it can reduce inverted index size while keeping effectiveness of Web Service Retrieval almost at the same level as the index without MWV. The MWV is calculated as the average weight of all service elements but using different weighting for different elements. Therefore, this method is more effective than the "bag of words" approaches.

3 Distributed Indexing

The task of Web Service Retrieval is to find relevant services from service collection satisfying given query, describing the special requirements of the user [3]. There are thousands or even millions of Web Services distributed over the Internet, and therefore the data collection is so big that the inverted index construction cannot be performed efficiently as a single process. To improve the efficiency of the index construction process, we use distributed indexing algorithms which allow to split the indexing process into parallel tasks distributed across nodes in a computer cluster and partition the index into separate parts. To solve such a problem, the distributed inverted index construction method described in this section is an application of MapReduce. However, before we explain it, we must firstly distinguish Web Service structural elements. Based on our previous research presented in [2], we consider a Web Service to be composed of quadruple of elements where the first three represent *parameters* which correspond to service name, description and version, and the fourth represents service *components* which are composed of six-tuple of following component elements: name, input value, output value, description, method, representation.

Secondly, based on our research carried out in [2], the service components are modelled in vector space to enable ranking of search results. Each parameter and component is represented as a vector. Vectors are composed of weights where each weight corresponds to a term from the "bag of words" of particular service parameter or component. Weights are calculated using one of the best known combination of traditional weighting schemes in information retrieval: *tf-idf.* To measure the relevance between parameter/component vectors the cosine function is used. However, to measure it the vectors have to be length-normalized by dividing their every element by their length. In result weights have the same order of magnitude. The inverted index composed of such a vectors is very complex and relatively big. To reduce its size and not to loose much retrieval effectiveness we use the Merged Weight Vector (MWV). The MWV of a Web Service is calculated as the average weight of all parameter and component vectors. In result, the final service vector is represented by the MWV. This assures quick index search and effective retrieval. Moreover, after a relevant service is found, the parameters or components which are relevant to the query can be later easily extracted.

MapReduce is a linearly scalable programming model. The computation problem is specified in terms of a map and a reduce function, and the underlying runtime system automatically parallelizes the computation across clusters of machines, handles machine failures, and schedules inter-machine communication to make efficient use of the network and disks [4]. The input data is split into fixed-size chunks and distributed across nodes executing mapping tasks. The map function processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function merges all intermediate values associated with the same intermediate key [13]. Reduce tasks are distributed by partitioning the intermediate key space into parts using a partitioning function. Based on the above we propose the following pseudo-code algorithms of the mapper and reducer functions:

Algorithm 1. Mapper Function

```
I: $\{serviceID, serviceData\}$. O: $\{term, \{serviceID, \vec{s}\}\}$.
MAPPER($serviceID, serviceData$):
  $D \leftarrow serviceDictionary \in serviceData$
  for $parameter$ or $component$ as $\alpha \in serviceData$ do
    $\vec{\alpha} \leftarrow [0_1, 0_2, 0_3, \ldots, 0_{|D|}]$
    for $term \in$ UNIQUETERMS($\alpha$) do
      $\vec{\alpha}_{D_{term}} \leftarrow$ CALCULATEWEIGHT($term$)
    end
    for $term \in D$ do
      emit $\{term, \{serviceID, \vec{\alpha}\}\}$
    end
  end
```

Algorithm 2. Reducer Function

```
I: $\{term, serviceVectors\}$. O: $\{term, postingsList\}$.
REDUCER($\{term, serviceVectors\}$):
1:  for $term$ do
2:      for $\{serviceID, \overrightarrow{s}\} \in serviceVectors$ do
3:          $\overrightarrow{s} \leftarrow$CALCULATEMWV($\overrightarrow{s}$)
4:          emit $\{term, \{serviceID, \overrightarrow{s}\}\}$
5:      end
6:  end
```

The mapper reads input split key/value pairs, where the key is *serviceID* and value is *serviceData* composed of: service dictionary (sorted unique terms that occur within the service content), and the set of service parameters and components. For every parameter/component denoted as α we create its vector $\overrightarrow{\alpha}$ with the number of elements equal to the number of unique terms in the service dictionary (lines 2-10). In the next step, we calculate the normalized *tf-idf* weight of every unique term in α and assign it to its corresponding position in the $\overrightarrow{\alpha}$ (lines 4-6). Lastly, for each term in service dictionary the intermediate key/value pairs are emitted to the partitioner. The keys are terms in service dictionary and values are the sets composed of services IDs and corresponding to them parameters/components vectors. The MapReduce framework sorts the mapper output by key and values.

The partitioner divides the intermediate keys into k term partitions according to the first letter of a term and distributes them over k reducers. In result the reducing tasks can run in parallel and the distributed index is partitioned into k partitions. For example if $k = 2$ the index is distributed over 2 nodes and divided into *a-m* and *n-z* partitions, according to the alphabet.

The reducer iterates over its corresponding intermetiate key/value pairs. The keys are represented by terms, and the values are represented by the set of service vectors. Every service vector is a pair composed of the *serviceID* and its corresponding parameter/component vectors which together represent the service vector $\overrightarrow{s}$. In line 3 we calculate the MWV for the service vector $\overrightarrow{s}$ and reassign it to $\overrightarrow{s}$. In line 4 the reducer emits the postings list for each term which is appended to a final output file for this reduce partition.

It must be also noted that for proposed algorithms it is assumed that the total number of parameters and components is known a priori (this results from the *tf-idf* weighting), and the service dictionary is already build. Therefore the indexing process requires some preprocessing, which can be done by a web crawler.

4 Experimental Results

Based on the approach presented in this paper we implemented the MapReduce Web Service Retrieval system by which we conducted experiments. The aim of the experiment was to measure the performance and effectiveness of proposed approach. Additionally, the Merged Weight Vector computation mechanism was compared to k-means and SOM (Self-Organising Maps) clustering algorithms and to the basic index structure with separate weights for every parameter/component. In order to evaluate the effectiveness, the following classical information retrieval measures were used: *precision*, *recall*, *F-measure* ($\beta = 1$).

The service collection of the experiment was composed of 778 SOAP Web Services with total number of 5140 components and 801 parameters collected by our implementation of Web Service Crawler. The resulting index dictionary was composed of 3499 terms. The crawler's destination host was: *xmethods.net* – a directory of publicly available Web Services, used by many researchers for service retrieval benchmarks. Because the methods of RESTful Web Services identification are still being improved, the experiment does not include any services of this class. However, this fact does not influence presented approach and resulting experiment.

The MapReduce implementation was running on a single computer node with CPU Intel(R) Core(TM) 2 Duo T7500 2x2,20GHz. The MapReduce environment was pseudo-distributed where each core acted as one single CPU node. For such a configuration maximally two tasks could be executed in parallel.

4.1 Overall Effectiveness

The experiment was carried out for following queries: *"temperature conversion"*, *"email validation"*, *"weather forecast"*, *"weather forecast service"*, *"currency exchange"*, *"demographics"*, *"new york"* and *"send sms"*, denoted as *Q1, Q2, . . . , Q8*. The retrieval effectiveness evaluation for above queries for the *MapReduce MWV* approach is presented in Fig. 1.

The effectiveness was very high for unique query terms contained by only few services. For popular terms the effectiveness was lower because among big group of retrieved services only few of them were relevant. For example for query

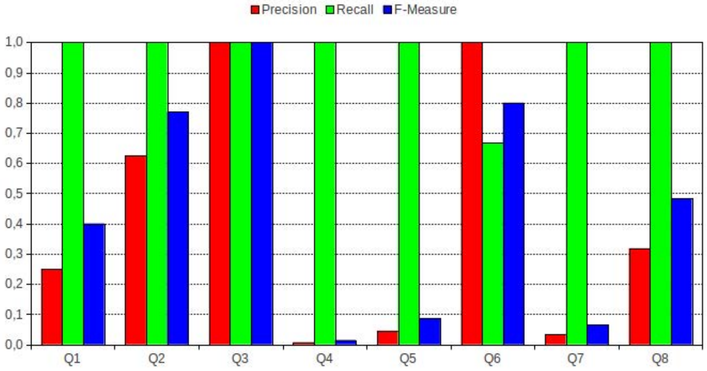


Fig. 1. Effectiveness of Web Service retrieval in MapReduce MWV index for eight different queries

"new york" only one service was relevant but the group of retrieved services was very big because many of them contain term "new". Similar result can be observed for queries "weather forecast" and "weather forecast service" where term "service" decreases the overall effectiveness. However, this drawback is not significant for ranked results where the most relevant ones are returned on top.

4.2 Average Indexing Time

In Fig. 2 we presented the average indexing time for six different configurations of MapReduce MWV jobs. The first four jobs produced single-partition index. The first one (*no_MapReduce*) acted as single process which is an equivalent to one map and one reduce task (not pararell). Jobs *MapReduce_1r*, *MapReduce_2r*, *MapReduce_4r* ran 2 mapping tasks and sequentially: 1 reduce task, 2 reduce tasks, and 4 reduce tasks. Configurations *MapReduce_2k* and *MapReduce_4k* also ran 2 mapping tasks but produced index divided into 2 and 4 partitions with 2 and 4 reduce tasks. As can be seen the MapReduce service approach shortens indexing time in all cases. Moreover, as the MapReduce indexing scales up by increasing the number of parallel tasks, the indexing time decreases. The exception are jobs with 4 reduce tasks (*MapReduce_4r* and *MapReduce_4k*) because they exceed the number of available nodes in presented pseudo-cluster and they needed to be queued.

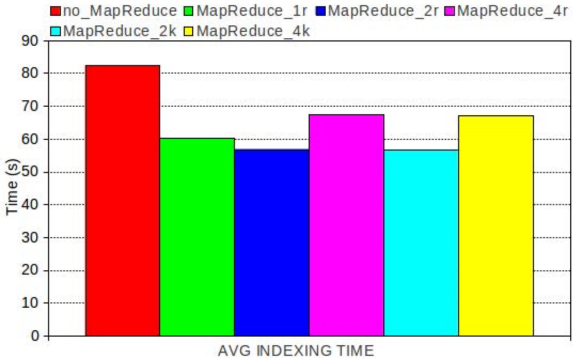


Fig. 2. Average service indexing time of six different MapReduce MWV indexing configurations. From the left: single process (no MapReduce), MapReduce with 1 reduce task, 2 reduce tasks, 4 reduce tasks, MapReduce with 2 index partitions (2 reduce tasks) and 4 index partitions (4 reduce tasks).

4.3 Average Indexing and Retrieval Time

As stated in previous section, partitioned index allows faster service retrieval than in the case of single-partition index. In Fig. 3 we illustrated the average retrieval time for queries $Q_{1..8}$ and average indexing time for three chosen MapReduce MWV indexing configurations – one producing single partition and two producing partitioned index. The first one (*MapReduce_2r*) had the shortest indexing time among all single-partition methods and thus it was presented for comparison. The shortest retrieval time is achieved using *MapReduce_4k* with index partitioned into 4 parts. Despite the fact this method had the worst indexing time, on a 4 node cluster it would be the best one. The *MapReduce_2k* utilized cluster resources most effectively and therefore its indexing performance was the best. However, its retrieval time was relatively small. The *MapReduce_2r* performed with slightly worse indexing time than *MapReduce_2k* and proved that single-partition index results have the longest retrieval time. To sum up, the experiment confirmed that by increasing the number of partitions the retrieval time is reduced. However, in order to achieve best indexing time the number of reduce tasks cannot exceed the number of available CPUs in the cluster.

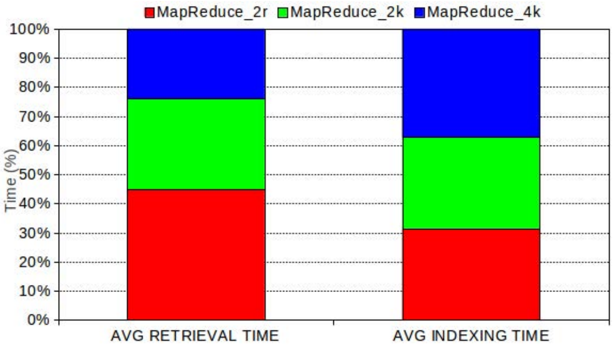


Fig. 3. Average Web Service retrieval and indexing time of three different MapReduce MWV indexing configurations. From the left: 2 reduce tasks, 2 index partitions, 4 index partitions.

4.4 Retrieval Effectiveness

Fig. 4 presents the effectiveness of Web Service retrieval for four index structures where each structure uses different method of MWV calculation. The *basic* structure represents index with weights for every parameter/component, *MapReduce_MWV* represents proposed in this paper MapReduce approach with standard MWV computation method, *KMeans_MWV* represents index structure with MWV computation conveyed using k-means clustering algorithm, and *SOM_MWV* computation conveyed using SOM with 10x10 map. The effectiveness measures of precision, recall and F-measure in Fig. 4a) are calculated as their mean values for queries $Q_{1..8}$. The overall effectiveness of presented index structures was very close. Despite in all structures the recall was the same, there was a small difference in precision which resulted in different values of F-measure as presented on Fig. 4b). As can be seen, the *MapReduce_MWV* index structure performed with the highest precision and overall effectiveness.

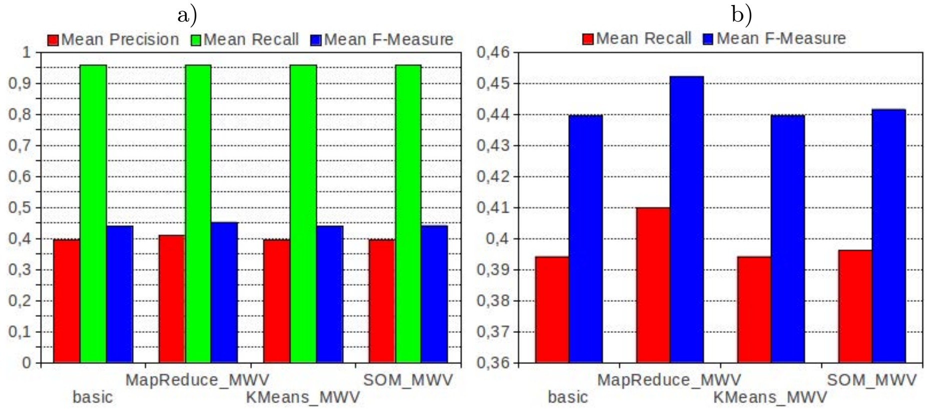


Fig. 4. The Web Service retrieval effectiveness for four MWV index structures. On the left side *(a)* three elements are considered: mean precision, mean recall and mean F-measure. The right side *(b)* highlights the differences in precision and F-measure for presented indexing methods.

In Fig. 5 we presented the Mean Average Precision (MAP) for queries $Q_{1..8}$ at 5, 10, 15 and 20 top positions. For every *k-th* position the MAP of *MapReduce_MWV* index structure was higher than for other structures. However, the *KMeans_MWV* and *basic* structures achieved their maximal Average Precision within top 10 results, whereas for *MapReduce_MWV* and *SOM_MWV* maximally 15 top results were needed. This shows that despite presented in this paper MapReduce MWV approach had the highest MAP, it not always returned all relevant services within the top ten results.

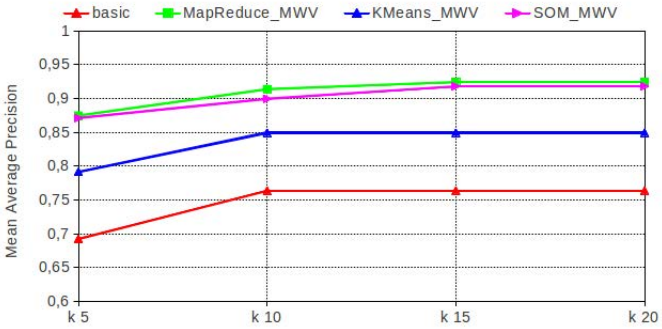


Fig. 5. The *top-k* Mean Average Precision of different MWV index structures

In Fig. 6 we presented the precision-recall curve for top 20 results. The recall and precision were calculated as their average values for queries $Q_{1..8}$. All MWV index structures achieved high mean recall (58%-71%) at very high mean precision level (88%-100%). The *MapReduce_MWV* and *SOM_MWV* structures performed closely to each other and were the only ones that returned relevant service at top 1 position for all queries. However, for every mean recall the MapReduce MWV structure had higher mean precision than its SOM equivalent. On the other hand, at certain point the effectiveness of *MapReduce_MWV* dropped below the *basic* and later *KMeans_MWV*. However, at this point the mean precision was relatively low.

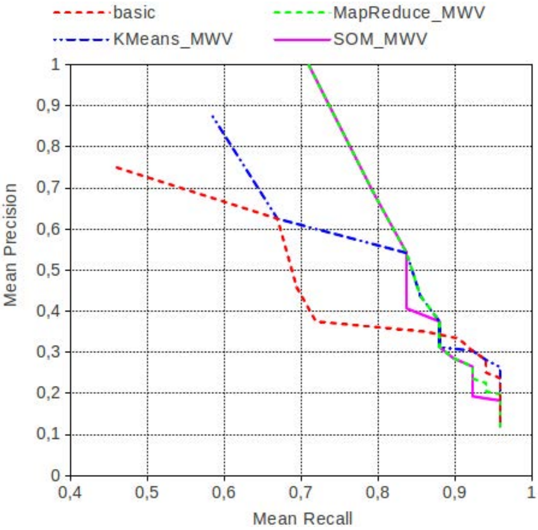


Fig. 6. The precision-recall curve of four MWV index structures

5 Conclusions and Future Work

The aim of presented research was to propose an effective and scalable approach of Web Service distributed indexing for Web Service Retrieval. The problem of Web Service indexing was to choose the appropriate index structure that enables fast and efficient retrieval of services while keeping the index size suitably small. Presented approach was based on the MapReduce method for parallel processing and generating large data sets. Here, an additional problem appeared that concerned the division of the Web Service indexing process into many parallel tasks.

Our research solved the above problems by presenting the mapper and reducer algorithms which split the input services, process the content of their parameters and components, and partition into separate term partitions distributed over the nodes in computer cluster. For service retrieval performance improvement and index size reduction the mapper used the Merged Weight Vector for computing the weights of all parameters and components of a service. Presented experimental results showed that the overall effectiveness of proposed approach was higher than in the basic index structure and higher than in other chosen structures with different MWV computation methods. Moreover, the experimental results confirmed that proposed method not only allows to significantly reduce inverted index size but also allows to reduce index construction time and Web Service retrieval time in a scalable manner. This means that as more nodes are added to the computer cluster the performance of index construction and service retrieval grows.

On the other hand, in proposed mapper algorithm we assumed that the total number of parameters and components is known in advance. For a web-scale mapper input, it is highly desired to use weighting scheme which does not require this information. Another improvement possibility results from the fact that since the resulting index structure is partitioned, the MapReduce can also be applied for index searching. Both mentioned problems are the subjects of our further research.

References

1. Czyszczoń, A.: Analiza rynku usług internetowych. In: Interdyscyplinarność Badań, Naukowych, Wrocław, Oficyna Wydawnicza Politechniki Wrocławskiej, pp. 199–204 (2012) (in Polish)
2. Czyszczoń, A., Zgrzywa, A.: The concept of parametric index for ranked web service retrieval. In: Zgrzywa, A., Choroś, K., Siemiński, A. (eds.) Multimedia and Internet Systems: Theory and Practice. AISC, vol. 183, pp. 229–238. Springer, Heidelberg (2013)
3. Peng, D.: Automatic conceptual indexing of web services and its application to service retrieval. In: Jin, H., Rana, O.F., Pan, Y., Prasanna, V.K. (eds.) ICA3PP 2007. LNCS, vol. 4494, pp. 290–301. Springer, Heidelberg (2007)
4. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. In: Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation, OSDI 2004, vol. 6, p. 10. USENIX Association, Berkeley (2004)
5. Wang, Y., Stroulia, E.: Flexible interface matching for web-service discovery. In: Proceedings of the Fourth International Conference on Web Information Systems Engineering, WISE 2003, pp. 147–156 (December 2003)
6. Wu, J., Wu, Z.: Similarity-based web service matchmaking. In: IEEE SCC, pp. 287–294. IEEE Computer Society (2005)
7. Wu, C., Chang, E.: Searching services "on the web": A public web services discovery approach. In: Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, SITIS 2007, pp. 321–328 (December 2007)
8. Song, H., Cheng, D., Messer, A., Kalasapur, S.: Web service discovery using general-purpose search engines. In: 2012 IEEE 19th International Conference on Web Services, pp. 265–271 (2007)
9. Atkinson, C., Bostan, P., Hummel, O., Stoll, D.: A practical approach to web service discovery and retrieval. In: IEEE International Conference on Web Services, ICWS 2007, pp. 241–248 (July 2007)
10. Alon, X.D., Dong, X., Halevy, A., Madhavan, J., Nemes, E., Zhang, J.: Similarity search for web services. In: Proc. of VLDB, pp. 372–383 (2004)
11. Zhuge, H., Liu, J.: Flexible retrieval of web services (2004)
12. Andrikopoulos, V., Plebani, P.: Retrieving compatible web services. In: 2011 IEEE International Conference on Web Services (ICWS), pp. 179–186 (July 2011)
13. Dean, J., Ghemawat, S.: Mapreduce: a flexible data processing tool. Commun. ACM 53(1), 72–77 (2010)
14. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)

Balanced Approach to the Design of Conversion Oriented Websites with Limited Negative Impact on the Users

Jarosław Jankowski

Faculty of Computer Science, West Pomeranian University of Technology, Szczecin, Poland,
jjankowski@wi.zut.edu.pl

Abstract. Web design requires consideration of many factors related to the usability, user experience and business objectives. In many cases conflict of interest can be observed and the desire can arise to improve the results represented by the conversions even at the expense of the user experience. This paper proposes a balanced approach which provides the ability to improve effects with a limited negative impact on users using interactive objects with adjustable levels of persuasion.

Keywords: website effectiveness, human-computer interaction, web design.

1 Introduction

The process of web design should offer such applications that can fulfil the business needs as well as the preferences of the users. For this purpose, it is important to analyse the usability and functionality of the system by using evolutionary methods [7], heuristics [4] and other approaches [23][19]. With the increasing competition in the field of e-commerce, web designers need to focus on more areas of web applications. Recent years have brought a wave of methods targeting the optimisation of effect-oriented websites, landing pages and email messages. A number of interactive objects are used to make the communication effective, including banners, multimedia elements, navigation systems and much more. Both verbal as well as visual persuasion together with the social engineering elements make a message interactive. The major goal of such messages is to increase the number of interactions within a website. The conversion factors represent performance which is measured by the number of effects in relation to the number of visitors. Conversion maximisation is obtained by giving up other parameters of the system. Usability assessments, however, can increase intrusiveness [16]. Intrusiveness is the psychological consequence or perception that results when the cognitive processes of the audience are disturbed [6]. When it comes to the elements of the website, intrusiveness would be the degree to which the content of the site influences the behaviour of the users and affects the experience of the users in negative ways as a result of some stimuli.

The main approach in this article is based upon providing an interface for end users to minimise the negative effects while browsing websites. Different elements of the website were analysed with varying levels of influence to observe behaviours while

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 527–536, 2013.

browsing. The measurements of the expected results as well as negative outcomes of the recipient and the selection of the design will be studied. The second section of the article focuses on the evolution of sites with respect to the effects-oriented systems and researched areas of the field. The third part explains the assumption of a balanced approach and the measurement of the effect of the interactive objects. The fourth section focuses on the results of the experiment in a real environment. Last but not the least; the fifth section includes an example of the process of searching for the balanced solutions and future research prospects.

2 Motivation and Related Work

A number of areas of human and computer interaction, like computer science, sociology and the psychology of human beings, are integrated in designing appealing websites [18]. The arrangement of the interaction components as well as other navigational interfaces including IA (Information Architecture) [13] and a number of other aspects of UCD (user-centred design) [22] [5] are considered significant. User-centred approaches are not the only things that are observed; content marketing for the conversion-oriented sites is considered the basis for the observation of evolution [20]. This trend has been identified by A. Schlosser who discussed the major components associated with the design of the website which in turn affect the results [17]. These effects are studied in a website design with respect to the actions of the users expected by the website operations [1]. With each element of the site analysed, various interactions can be measured. This might include the number of users who are interested in an offer, ratio or returning users, the number of visited pages, visit duration and the interaction of people in the site and, last but not least, the number of subscribers [15].

In the past, measures like the number of external links referring the target web page, site visit duration, and the number of times the site was accessed were only useful in providing an overview of the behaviour of the users and this didn't help the decision process much [9]. Behaviour analysis kept in the settings of the session and of dynamic modelling of navigation paths can offer a better and much more reliable measure of the actions of the user [10]. Comparison analyses can be made on the user's segmentation. M. Pearrow says that there several factors which impact upon the design, evaluation and usage of the product that is to be examined [14]. The perception of messages is one such component that can show the efficiency of whether the goals set for a site have been achieved or not [2]. The conceptual model of a website includes the combination of emotions as well as cognition for understanding behaviour on an online platform [12]. The optimisation process of a website should include these strategies. Call to action, animation, persuasion and other similar interactive objects must also be combined. Text, along with graphics, aims to 'call to action' so that the elements of persuasion can be used for encouraging a person to undertake some action for targeting the site.

When it comes to perception in the context of the Internet system, it refers to the observation of the conscious reaction of a sense organ to some stimulator. The receipt of the sensation, as well as reaction, depends on the electronic content supplied. Desensitisation and the perception of provided marketing content connected with sensory adaptation should be considered for making the design effective [3]. The users don't only view the whole content of the site in the process of communication; rather, interaction as a result of content scanning is also taken into account. If the limited ability of the users to process information is controlled, then it can help in grabbing the attention of users through methods like animations, effects, graphics and other visual searches [21]. With this, the intrusiveness of a site can be improved with techniques of internet marketing but this can lead to negative web user response [11]. Increased persuasion factors at a certain point do not increase the website's effectiveness as the information process ability becomes limited. When this situation arises a balanced approach is used to achieve the desired levels without influencing the users.

3 Approach Based on Modelling of Balanced Results

The structure of the modelling interface and the design of the information for achieving the desired results can influence the results substantially. The structure of the interactive elements is explained in this section which will help to locate the influenced levels and analyse how increased persuasion can change the results. The Web platform interfaces are the interactive objects that can integrate the components which can further affect the user and help in the generation of user interactions. When it comes to the system being focused on the direct effects, the major aim is to get such an influence that can be determined in a number of ways in line with the system's scope and the business model. The interactive object with separate components focuses on offering individual functionality and level of influence on a user [8]. Each object which has a set of components available and is determined by $E = \{E_1, E_2, \ldots, E_n\}$ and for every E_i there is a number of available variants $E_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,\, cnt(i)}\}$ where $cnt(i)$ defines the number of variants available for the i-th element. For every component $e_{i,j}$ can be given influence level $l_{i,j}$ which indicates the potential for persuasion of a user. For measuring the total selected variant influence, an aggregated interaction measure is used that includes the total partial levels. Total influence can be represented by an aggregated influence measure AI_i of object E_i consisting of k elements according to the following formula:

$$AI_i = \sum_{j=1}^{k} \left(l_{i,j} * w_i \right) \tag{1}$$

where $l_{i,j}$ stands for the influence level defined during the design process, w_j stands for the rank assumed for a given element based on the effect size from analysis, which defines the strength of influence in relation to other elements. Positive effects (response) and negative effects (e.g. blocking interface or exiting website) in relation to the aggregated influence of the website will be needed for calculating the balanced design. The results will focus on the balanced approach in which the total aggregated influenced measure will not influence the experience of the user when the results will be assumed by the designer. A few components as well as interactions will be registered in this whole interactive object of the group of system users' transformation process. A factorial experiment was used for the implementation in which all the variants were tested. This will help in providing information about the relationship between the persuasion level and influence of the consequences. The process of identification of the balanced design will then be designed on the basis of the results. It will further offer such a level of performance that will be acceptable with a less intrusive design.

4 Experimental Research

In this part, there are three stages of research which were identified, like engineering component construction, website integration and response data analysis. On the example website, the interactive object was situated with the primary purpose of diverting traffic from the main page to the other landing page. Two research goals were taken into account while undertaking the experiment. The first one was based on uncertain modelling and gathering collective knowledge in the interpretation models of web designing [8] and the second was associated with the balanced design. Fig.1 displays the structure of the interactive object based on the site's components with three sections i.e. S_1, S_2 and S_3.

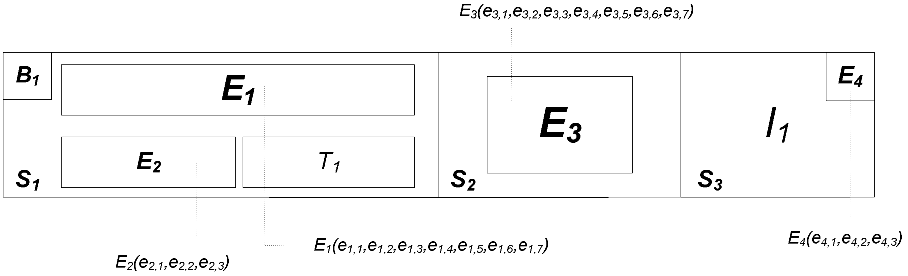

$E_3(e_{3,1},e_{3,2},e_{3,3},e_{3,4},e_{3,5},e_{3,6},e_{3,7})$
B_1
E_1
E_3
E_4
I_1
E_2
T_1
S_1
S_2
S_3
$E_1(e_{1,1},e_{1,2},e_{1,3},e_{1,4},e_{1,5},e_{1,6},e_{1,7})$
$E_4(e_{4,1},e_{4,2},e_{4,3})$
$E_2(e_{2,1},e_{2,2},e_{2,3})$

Fig. 1. The structure of an experimental interactive object

Elements $E = \{E_1, E_2, E_3, E_4\}$ are located in the objects area assigned to the sections. They represent different variations of the design. Two elements of design, T_1 and graphical element I_1, are not subject to versification. Element B_1 is basically an extra element that has two stages and it comes with the option to delete the interactive object from the screen of the user and create blocking response *BR*. The intrusiveness level is measured with it and it enables detection of the stage when the user wants to disable the content. The rest of the elements have got sets of possible design variants. Seven versions of text variants $\{e_{1,1}, e_{1,2}, e_{1,3}, e_{1,4}, e_{1,5}, e_{1,6}, e_{1,7}\}$ were chosen for $E_{1,}$ having integrated the call to action expression. They also have a varied level of influence l_i having incremental values i.e. $l_1 = 1$, $l_2 = 2$, $l_3 = 3$, $l_4 = 4$, $l_5 = 5$, $l_6 = 6$, $l_7 = 7$. They are the same for each element. E_2 has text variants $\{e_{2,1}, e_{2,2}, e_{2,3}\}$. They have a different level of influence on the incremental call to action. E_3 is the graphical button that is the call to action text in seven variants i.e. $\{e_{3,1}, e_{3,2}, e_{3,3}, e_{3,4}, e_{3,5}, e_{3,6}, e_{3,7}\}$. Variants $e_{3,1}$ and $e_{3,2}$ were stagnant, variant $e_{3,4}$ features a flashing animated text whereas the variant $e_{3,5}$ has the capability to flash it to $e_{3,7}$ level (which is the highest flashing frequency and potential intrusiveness). Last but not the least, E_4 features persuasion functions in three design variants. The next phase is based on a real experiment and the object was displayed 249,149 times on a testable web page. Each combination of elements was shown almost 282 times. 27,338 unique users were chosen for the generation of the message. 698 interactions were registered with a response *R* and *BR* response, where R is equal to the number of interactions/ impressions *100% and *BR* is equal to the number of blocking interactions/impressions *100%.

Table 1. Subsets of response values for selected variants

High and low *R* response instances					High and low *BR* response instances				
E_1	E_2	E_3	E_4	R[%]	E_1	E_2	E_3	E_4	BR[%]
7	2	2	3	0.2865	2	1	1	3	0.3115
2	1	7	3	0.2994	2	1	1	1	0.3115
6	2	5	2	0.2994	5	1	5	2	0.3115
3	2	5	2	0.3039	7	2	5	2	0.3134
5	1	3	1	0.3086	7	3	4	1	0.3144
5	2	6	1	0.3095	6	3	3	2	0.3164
3	2	1	3	0.3105	4	3	1	1	0.3184
7	1	1	2	0.3115	4	1	6	1	0.3184
		…						…	
	1	1	2	1.4652	2	2	3	3	1.0489
	2	2	3	1.4652	7	2	6	3	1.0909
[illegible]	1	7	2	1.4705	4	1	3	2	1.1029
[illegible]	2	5	1	1.6326	5	1	7	3	1.1152
[illegible]	2	6	3	1.7421	2	2	3	1	1.1320
[illegible]	1	6	2	1.7482	1	2	7	3	1.2861
[illegible]	2	3	1	2.2222	6	2	2	1	1.2931
[illegible]	1	6	3	2.4054	5	1	5	3	1.3201

R shows the projected interaction with the object through user clicks, whereas BR represents the blocking response through the clicks for removing the content (the user clicks the cross button at the header to close the window). The next phase included an analysis consisting of: the individual influence levels l_i and the average response value *R*. Another analysis was made for determining whether either combination of elements affected the results obtained. An aggregated influence measure AI_i represented specific project variants and ranks for each component were introduced.

5 Response Analysis towards the Selection of Balanced Designs

In the next phase the results were analysed by using methods related to response analysis within the factorial experiment. The analysis being conducted identified the importance of the elements E_3 and E_4 at levels $p(E_3) = 0{,}028959$ and $p(E_4) = 0{,}036837$ respectively. The importance of text information and messages on the indiscreetness is outlined by the results. There are situations in which the ligation of information function of interface can result, in which the attention of the user is diverted to the graphical elements. The Pareto chart (Fig. 2) shows the influence as well as dependences of the individual elements on the 2.19 and 2.09 input parameter values acquired. The chart also indicates certain levels. The increase in the value of the chosen element cannot be achieved with a decrease in other values. The response surfaces for the influential elements were observed with E_1 and E_2 at the minimal

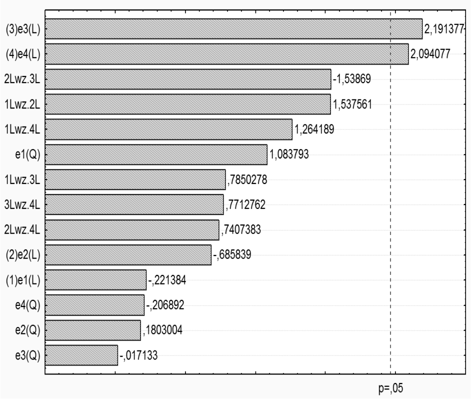


Fig. 2. Pareto chart for individual input parameters

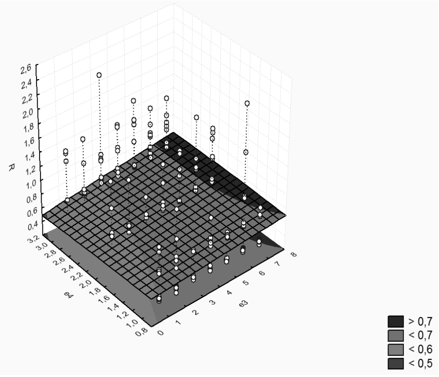


Fig. 3. Dependency of R on values of E_3 and E_4 with constant value of E_2 and E_1

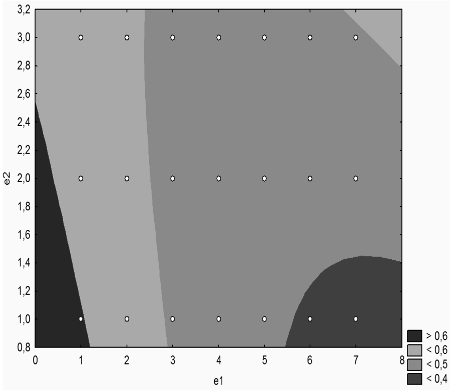


Fig. 4. Dependencies of E_1 and E_2 for minimum values of E_3 and E_4

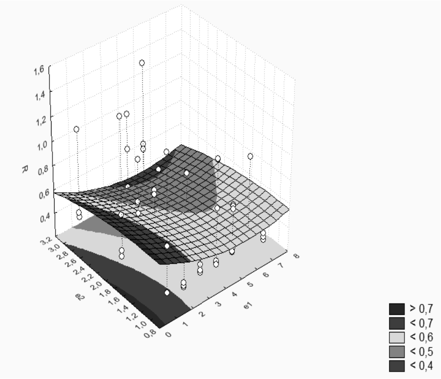


Fig. 5. Response for parameters E_1 and E_2 with assigned levels $E_3 = 1$ and $E_4 = 1$

level. Fig. 3 shows the response surface which represents the scope in which the system response R rises with the increase in the influence level of E_3 and $E_{4.}$ Fig 4 shows the analysis for text elements E_1 and E_2 having minimal values for E_3 and $E_{4.}$ All these charts show the minimum level of influence of the graphical elements at which the text structure influences the effects obtained. However, after using the ANOVA analysis, it has been estimated that with the increase in the graphical elements' invasiveness and the sample too, the level of text elements' significance falls drastically.

As animated elements were introduced, the users' attention begin to wane and this turned out to be a major hurdle in the research as it was influencing the text elements and other interactions. A subset having examples of messages with no animated elements was then analysed for the estimation of the influence of text messages. Fig. 5 shows the distribution of the response dependency of E_1 and E_2 at the minimal level of parameter E_3 with E_4 being turned off. It was not displayed that the service would be free of charge (E_4 represented it). When further analysis was made, it showed that E_4 increased the effects. Furthermore, E_3 being added to the background with a bigger contrast showed improvements in the influence of E_2. As a result, it was concluded that a text generated with E_2 can increase the reaction as well as the graphical influence. As the measurement data was not that big, it was easy to see the increase of the influence from $R = 0.49$ to $R = 0.62$. This means that the system response increased by 26 per cent with the increase of the influence levels. This also helps in determining the limit up to which the level of influence can fall, lowering intrusiveness so that the response level can be maintained at a level that would be much more appropriate. The increase is small, however. It shows the chances of the occurrence of the perception i.e. when the influential element is visible, the presence of other effects is not important. The influential measure of the design variant can be calculated. For this, the level of design is multiplied by the effect with 2.1913 for E_3, 2.0940 for E_4, 1.0837 for E_2 with –0.6858 and E_2 –0.2213. Each design variant will have an aggregated value. Fig. 6 shows that for the aggregated measure *AI* for

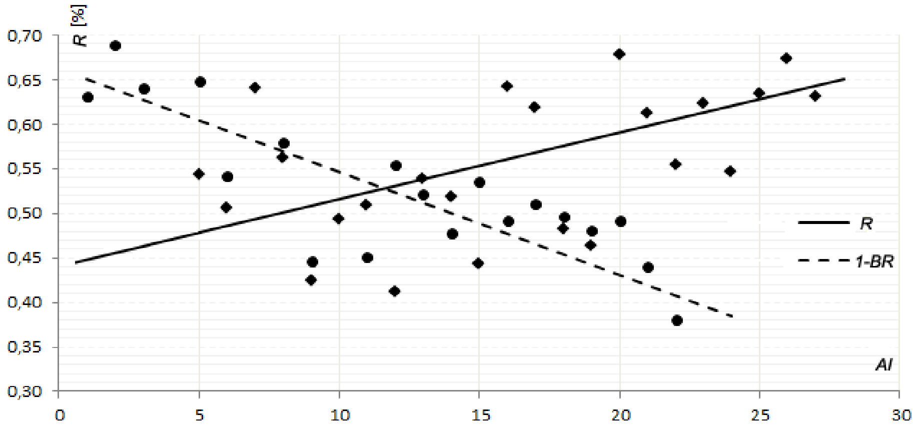


Fig. 6. Response rate *R* in relation to blocking actions *BR* and C region with balanced solutions

the design variant, *R* grows in a trend similar to the regression line from 0.45 to 0.65. The increase in *R* and in *AI* shows the projected intrusiveness of interface. The value *1-BR* is used in the chart for better illustration. It represents the increase in intrusiveness with a dotted line on the decreasing chart.

The results represent *R* in relation to the degree of invasiveness *1-BR*. The X axis shows the impact of *AI* whereas the Y axis shows the percentage of the no. of interactions per 100 exposures for *R* and *BR*. The regression line indications the rise in R with a solid line whereas the rise in *BR* is shown by a dotted line; both increase due to an increase in intrusiveness. The area in the middle of the chart represents the set of final balanced designs which can be used. With this approach, changes were reflected on the interactive environment which will further help in interpreting the dependencies. The solutions and the approach for building balanced interactive objects can also be used for designing the marketing message or other components of the website. The final solution depends on the range of difference in the elements being monitored and the designer's expectations. When the decomposition of objects is merged in the consistent parts, it becomes easier to observe the effects on invasiveness of the interactive elements. Due to the preferences of the designer of the website, the level that may be adopted can vary in each decision-making process. The research here concludes as well as proves that the increase in the influence levels that was actually caused by using the invasive elements in the message don't really help in achieving good results. Hence, the result can be based on such a design variant that delivers user experience as well as effects at a certain level. There is a limitation in the proposed method too, which calls for the introduction of measures of impact that can be calculated without any bias. In order to estimate the impact on the final effect, it is important to test a large number of combinations within the design options. If the designer wants to use an aggregated measure of impact then it will be necessary to conduct further studies and analyse the objects in open research directions so that more explorations can be conducted in future.

6 Conclusions

The focal point of this research is that interactive websites can be expected to deliver a premium end-user experience. The main purpose is to make a website user-friendly so that the website can be navigated with ease. This entire article describes the complete procedure of building vibrant items using the conceptual level by the application of certain elements and then analysing and observing the results acquired to deliver great websites. The approach that has been described will help in delivering an optimum solution to designers; however, it is still in the early stages of the research with plenty of room for improvement through further studies and investigation. The similarities among the influence elements show that the problem is quite complex. On the other hand, the results have been affected by numerous factors throughout the entire process of research and experiment. Analytical methods will be further used for the specification of the relationship. This will make the system much easier to understand and make adjustments to it according to the prevailing conditions. This whole approach can prove to be helpful in the development of the concept of superior end-user designs as proposed here. The levels of effects here are guaranteed but the results show that the excessive interface intrusiveness is limited. It cannot be extended to a level greater than its maximum. This research will prove to be very helpful for website designers although, as mentioned, it has room for improvements.

References

1. Barnes, S.J., Vidgen, R.: An integrative approach to the assessment of e-commerce quality. Journal of Electronic Commerce Research 3(3), 114–127 (2002)
2. Bernard, R.: Simple and Proven Ways Your Website Can Persuade Visitors (2010), http://www.renaldobernard.com
3. Burke, M., Gorman, N., Nilsen, E., Hornof, A.: Banner Ads Hinder Visual Search and Are Forgotten. In: Proceedings of Human Computer Interaction Conference 2004, pp. 1139–1142. ACM Press, New York (2004)
4. Flavian, C., Gurrea, R., Orus, C.: Heuristic evaluation of websites design for achieving the web success. International Journal of Services and Standards 5(1), 17–41 (2009)
5. Green, D., Pearson, J.M.: Development of a web site usability instrument based on ISO-9241-11. The Journal of Computer Information Systems 47(1), 66–72 (2006)
6. Ha, L.: Advertising clutter in consumer magazines: Dimensions and effects. Journal of Advertising Research 36(4), 76–84 (1996)
7. Ivory, M.Y., Hearst, M.A.: The state of the art in automating usability evaluation of user interfaces. ACM Computing Surveys 33(4), 470–516 (2001)
8. Jankowski, J.: Integration of collective knowledge in fuzzy models supporting web design process. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS, vol. 6923, pp. 395–404. Springer, Heidelberg (2011)
9. Kaplanidou, K., Vogt, C.: A structural analysis of destination travel intentions as a function of website features. Journal of Travel Research 45(2), 204–216 (2006)
10. Kelly, S.: Determining Effectiveness of Websites by Refining Transaction-Oriented Views through Supervised Competitive Clustering, Web effectiveness research group, Research report, The University of Georgia, Athens (2008)
11. Leggatt, H.: Intrusive ads (still) annoy consumers, BizReport (2008), http://www.bizreport.com/2008/07/intrusive_ads_still_annoy_consumers.html
12. López, I., Ruiz, S.: Explaining website effectiveness: The hedonic-utilitarian dual mediation hypothesis. Electronic Commerce Research and Applications 10(1), 49–58 (2011)
13. Morville, P., Rosenfeld, L.: Information Architecture for the World Wide Web. O'Reilly, Sebastopol (2006)
14. Pearrow, M.: Web Site Usability Handbook. Charles River Media, Rockland (2000)
15. Rayan, T.: How to Measure Website Effectiveness Using New Success Metrics (2008), http://ezinearticles.com
16. Rohrer, C., Boyd, J.: The rise of intrusive online advertising and the response of user experience research at Yahoo. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1085–1086. ACM Press, New York (2004)
17. Schlosser, A.: Converting web site visitors into buyers: How web site investment increases consumer trusting beliefs and online purchase intentions. Journal of Marketing 70(2), 133–148 (2006)

18. Sears, A., Jacko, J.: Human-Computer Interaction Handbook. CRC Press, London (2007)
19. Sharp, H., Rogers, Y., Preece, J.: Interaction Design: Beyond Human–Computer Interaction. John Wiley & Sons, Hoboken (2007)
20. Tarafdar, M., Zhang, J.: Analysis of critical website characteristics: A cross-category study of successful websites. The Journal of Computer Information Systems 46(2), 14–24 (2005)
21. Turatto, M., Galfano, G.: Color, form and luminance capture attention in visual search. Vision Research 40(13), 1639–1643 (2000)
22. Vredenburg, K., Isensee, S., Righi, C.: User-centered Design. Prentice Hall, New York (2002)
23. Zimmerman, D.E., Akerelrea, C.A.: Usability Testing: An Evaluation Process for Internet Communications. Wiley, Hoboken (2004)

Evaluation Framework for User Preference Research Implemented as Web Application

Aneta Bartuskova and Ondrej Krejcar

University of Hradec Kralove, Faculty of Informatics and Management,
Department of Information Technologies,
Rokitanskeho 62, Hradec Kralove, 500 03, Czech Republic
Aneta.Bartuskova@uhk.cz, Ondrej.Krejcar@ASJournal.eu

Abstract. This article presents an evaluation framework for user preference research. Web application was implemented to verify this framework. Interrelated influence of three main aspects of the website - usability, aesthetics and information part - is demonstrated in a performed experiment. Surprisingly, the perception of aesthetics was most significantly influenced by the quality of two remaining factors. The original hypothesis was, that the quality of aesthetics will influence the perception of usability and information part. The article also features a parallel comparison approach, which allows within-subject design and two-way performed manipulation of variables.

Keywords: evaluation framework, user preference, perceived usability, perceived aesthetics, paired comparison, parallel comparison.

1 Introduction

User preference is of the great importance in a competitive environment of various products, interaction systems and websites as well. There are many studies researching user preference and relations between the constructs like usability and aesthetics. Some of its limitations are addressed in this study. The main objective of this paper is presenting an evaluation framework for researching usability, aesthetics and information part as the main elements of the website and their impact on user preference. Presented framework was implemented into a web application to verify its functionality.

The type of tested websites was chosen for several reasons as small corporate websites. To the authors' best knowledge, there is very limited to none studies targeting corporate websites without integrated online shopping. Nevertheless small corporate websites represent a large part of the World Wide Web information space and combine the expectations of good aesthetics as well as usability and content. As an overall preference and a perception of usability and aesthetics are highly dependent on context [28] and also on mode in which a user approaches the system [8,15], concrete definition of tested website´s type is needed. A small corporate website was chosen for an accompanying experiment as one of the most common types of websites all over the world.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 537–548, 2013.

2 State of the Art

User preference reflects a user´s choice from several alternative websites and consequently a user´s decision about his behaviour on a chosen website [6]. User preference also signifies, how users evaluate a relative importance of interactive system´s properties [15]. While earlier studies suggest, that this influence is dependent more on usability of the website [1,2], many of the later studies put a greater value on website´s aesthetics [7-10,19,20]. Aesthetics qualities are also revealed much faster than usability facets and remain relatively stable, in conformity with first impression [16]. There are many studies, which investigate a connection between usability and aesthetics on websites or generally in human-computer interactions [11-15]. Previous studies have also shown that subjective evaluations of usability and aesthetics are correlated [5,8,9,14]. McCracken and Wolfe defined web design attributes as: content organization, visual organization, navigation system, colour and typography [24]. Finally, content or information quality is one of the key aspects in a website´s success [21]. Content´s characteristics can be defined as a quality and a quantity of provided information [28].

Usability can be considered as objective or subjective measure. Hornbæk divided usability measures of current studies into three groups: the measures of effectiveness, efficiency and satisfaction [27]. Lee and Koubek considered usability as two concepts: pre-use usability, which is perceived usability of the interface before use, and user performance as a result of user´s activities on the site [6]. Most widely used method for measuring usability is user testing, also an inspection and an inquiry [26]. Impact of usability is being researched in various ways, e.g. by manipulating information architecture [13] or disorganizing items and use of confusing labels [11].

Aesthetics of user interfaces was originally quite neglected in human-computer interaction research, it is however widely researched in various contexts today. An extensive study of Lavie and Tractinsky stated, that users´ perceptions of a webpage consist of two main dimensions, namely classical and expressive aesthetics [9]. A research of Robins and Holmes showed an influence of aesthetics on credibility and trust, dependent mainly on first aesthetics impression of the website [7]. That supports previous experiment of Lindgaard et. al., which specifies time needed for assessing a visual appeal of a website [16]. Schaik and Ling´s studies examined an impact of a context on perception of aesthetics [8] and relation of similar constructs - beauty and goodness [22]. Schenkman and Jönsson uncovered that beauty is the best predictor for an overall user judgement [10]. Study of van der Heijden introduced a new construct named "perceived visual attractiveness of the website", which influences also usefulness and ease-of-use (i.e. usability) [17]. Most of the recent research examines aesthetics facets in relation with actual or perceived usability [11-15]. Aesthetics of websites is usually manipulated through design features such as background or quality of pictures and decorative graphics. Manipulating aesthetics is difficult without affecting usability facets. An aesthetics manipulation can be limited as a result, allowing only minimal decorative changes or change of background [13,15].

3 Shortcomings of Previous Research

In order to examine an influence of usability and aesthetics facets independently, manipulation of one aspect cannot influence the other aspect. In many studies, implemented manipulation does not abide these requirements. In a recent study of Tuch et al., authors claim to manipulate aesthetics without an impact on usability by changing only background [13]. Nevertheless, change of background from plain light colour to a colourful texture is definitely changing perceived usability and presumably also actual usability. Contrast is a key element for legibility, which is a component of usability. In this case, contrast of a navigation area and individual items was changed significantly, therefore usability was changed as well. Perceived free space between navigation and content elements was also influenced by this manipulation. In Schaik and Ling´s research, aesthetics was manipulated by changing also content organization and navigation, which are usability aspects [8].

Another issue is a frequent use of between-subject design regarding independent variables aesthetics and usability. Since aesthetics can be measured only as perceived aesthetics, i.e. a subjective measure, and researching usability also contains subjective scales, there is a great risk of creating an unwished variability. This variability is caused by a different perception of participants, specifically different abilities in scaling with an equal volume. Variability within a group is then mixed with variability between groups, creating an unnecessary additional variability, which affects results of the study. Despite of this, many studies use between-subject variables - usability and / or aesthetics [5,12,25] - often in a mixed design, where time is the only within-subject variable [11,13]. Although within-subject designs should be preferred, there is sometimes reason for use of between-subject design. That is e.g. when participant should not have a previous knowledge of tested website, if this website is used in more alternate versions during the experiment. This could be solved by using „parallel“ instead of „serial“ comparison, as is discussed in a next section.

4 Evaluation Framework - Parallel Comparison Approach

According to an authors' opinion, there is a lack of “parallel” comparisons in recent research on usability, aesthetics and user preferences. As parallel is meant comparison method, which was applied e.g. in Schenkman and Jönsson research [10]. In their experiment, two calibrated identical monitors were used for each participant.

Recent research papers usually use „serial comparison“, as they design their experiments as series of screenshots or implemented websites. Each website is usually accompanied with additional tasks and evaluation inquiries before moving to another website [6,7,8,29]. This approach allows full focus on a presented website, but significant deviations can occur due to subjective nature of perception. In other words, we do not usually choose things in absolute terms, we rather focus on a relative advantage of one thing over another [4,18]. As we are evaluating things (and websites as well) in relation to others, it seems only logical to provide user with context for his choice - another website at the same time. This arrangement should provide for a coherent evaluation between websites, which are being compared.

There are of course limitations to this approach, most relevantly number of tested websites. Solution is an application of paired parallel comparison method, since two already create a context and comparison in pairs in the simplest. We can then assume, that most accurately evaluated will be the websites in pairs, so it is recommended to assign these pairs according to a purpose of particular study. We should aspire for parallel comparison whenever it is possible also because first impressions resonate over a long sequence of stimuli and affect following evaluations [16,18].

Experimental design of this study attempts to deliver an efficient combination of parallel and serial comparison for purpose of user´s evaluation of websites. Since a count of independent variables in this research is three (perceived aesthetics, perceived usability and perceived quality of content), that means eight combinations, if we assign two values to each variable (low and high). Recent research papers usually deal with only four combinations, since content aspect is omitted. It is a common knowledge that an experiment procedure should not take too long and a workload put on user should be minimized in order to get good-quality outcomes. Brooke emphasized a need for simplicity and speed, since the users, if working through many demanding tasks or long questionnaires, could be very frustrated towards the end. It is even likely that they would not complete them [3]. Every experiment should be therefore designed to gain most knowledge out of minimal possible volume of gathered data. Different approach is then needed, instead of comparing all possible combinations listed in [Table 1] between each other.

Table 1. Possible combinations of two values in three variables on one website

Website variation	Aesthetics	Usability	Content
Combination 1:	Low	Low	Low
Combination 2:	Low	Low	High
Combination 3:	Low	High	Low
Combination 4:	Low	High	High
Combination 5:	High	Low	Low
Combination 6:	High	Low	High
Combination 7:	High	High	Low
Combination 8:	High	High	High

In a situation of three independent variables with two possible values, instead of comparing every option with every other option, two approaches which allow simple statistical treatment can be identified. These two approaches are following real situations on the web, which can arise according to the expected quality of a website. Expected quality usually reflects an average quality of websites browsed by a user and subjective demands of that particular user.

Table 2. Outline of proposed approaches - negative and positive

Approach	Negative	Positive
Default quality	High in all variables	Low in all variables
Interpretation	User is expecting high level of quality, therefore is negatively influenced by lower quality in any aspect of the website	User is expecting low level of quality, therefore is positively influenced by higher quality in any aspect of the website

With regard to our variables (perceived aesthetics, perceived usability and perceived quality of content), one approach requires a comparison of three pairs of websites, which differ in one aspect´s quality from other website in the pair. We can demonstrate these approaches in [Table 3].

Table 3. Combinations for comparison with use of proposed approaches

	Positive approach			**Negative approach**		
Aspect	Aesthetics	Usability	Content	Aesthetics	Usability	Content
Pair 1 Site A	**High**	Low	Low	**Low**	High	High
Pair 1 Site B	Low	Low	Low	High	High	High
Pair 2 Site A	Low	**High**	Low	High	**Low**	High
Pair 2 Site B	Low	Low	Low	High	High	High
Pair 3 Site A	Low	Low	**High**	High	High	**Low**
Pair 3 Site B	Low	Low	Low	High	High	High

This comparison method is best used for researching an influence of one factor on perception of other factors. Few assumptions and hypotheses were made for a following experiment. Firstly, this study considers only post-use evaluation of a website, because we can assume an intention of visitors to use the site, and also their decision matters after use of the site. Secondly, we can assume that user´s evaluation and a choice between websites is subjective, therefore subjective evaluation is used in this study, not objective performance metrics. Finally, we can presume that a user will be more influenced by hedonic qualities than functional, as aesthetics is expected to be more apparent than usability issues. This is because websites of small businesses have usually a simple structure and no complex navigation or extensive forms. Impact of content is expected to be more apparent than usability but significantly less important than an aesthetics effect. Main hypothesis is, that a difference in a quality of website´s one aspect will likely affect an evaluation of website´s other aspects as well. Aesthetics is expected to most significantly influence perceived usability and perceived quality of content.

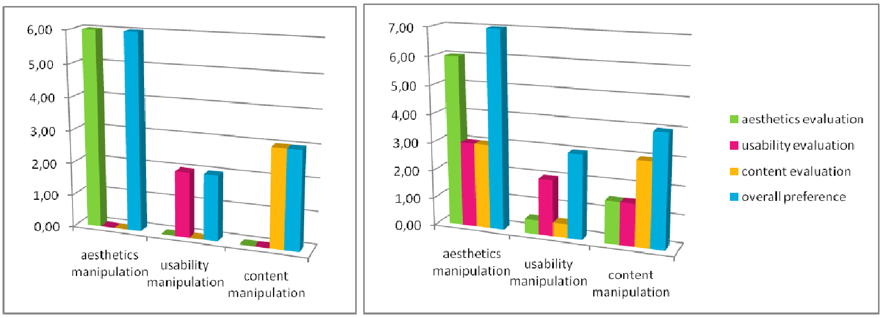


Fig. 1. Expected differences in evaluation without and with proposed interferences

5 Experimentation

5.1 Experimental Design

The experiment has a three-factor within-subject design, which benefits from a minimized variability in subjective evaluation scales of users. In this case, user´s subjectivity does not extend over results of other users, because every user participates in every task in this experiment. Independent variables are perceived aesthetics, perceived usability and perceived quality of content. Researched user´s preferences are subjective and therefore a subjective evaluation of perceived qualities is used in this study.

Design of an experiment is based on a paired comparison of different versions of the same website. Experiment is using a negative approach of the parallel comparison method, because the authors suppose that settings of this approach are closer to evaluation of real websites. Users were asked to perform a simple task on each of the websites in order to test its usability and gather an impression about its aesthetics and quality of content. Only post-use evaluation was researched in this experiment.

Procedure of an experiment simulated a real situation, when a user searches for information about competing companies on the internet. After finding official websites of these companies, a user decides about his preference according to aesthetics impression, ease of use and content quality of the websites. Presented framework presumes that a user does not know tested websites, as he is looking for a new piece of information. The experiment follows this assumption with use of websites, which were created only for purposes of the experiment.

The Likert scales were used for evaluation of aesthetics, usability, content and an inquiry about overall preference. First three inquiries were used for each version of tested website. 9-point Likert scales were used for these inquiries along with helpful labels "no", "rather no", "indifferent", "rather yes" and "yes". An inquiry about overall preference was used once for every pair of websites, three pairs in total according to presented evaluation framework. 17-point Likert scale, which allowed 9-point evaluation range of preferred website, was used for this inquiry.

Table 4. Evaluation inquiries translated from Czech

Researched factor	Evaluation inquiries
Aesthetics	Website is visually appealing for me.
Usability	Using the website is easy for me.
Content	Website´s content is of high quality.
Overall preference	Overall I prefer webpage...

5.2 Manipulation of Variables

Manipulation of variables was done by creating default-quality version of the website first and then enhancing or lowering the quality by slight but visible changes in interface, interaction and content.

Table 5. Two-way performed manipulation of variables

Aspect of the website	Manipulation towards low quality	Manipulation towards high quality
Aesthetics	higher JPEG compression rate and a slight blurring of images	shadow and gradient for enhancing menu bar
Usability	worse readability (lower contrast between text and background)	visible effect of menu items on hovering action (a good affordance)
Content	higher relevance of content	slight but noticeable spelling errors

5.3 Implementation

Evaluation framework was tested in technical implementation for user testing, which can be conducted online without laboratory environment and researcher´s supervision. This application was created exclusively for purposes of this experiment. Implemented web application was optimized for monitors with minimum width of 1024 pixels, which is c. 85% according to statistics in the Czech Republic [34]. It was also optimized for a majority of browsers, incompatible browsers were displaying warning, which recommended not to proceed in an experiment with this browser.

Tested websites were simplified in order to fit in limited space, which is less than half the width of the 1024px-width webpage. Nevertheless, essential elements were preserved and every model website was fully functional, able to interact with user without affecting the main window of the web application [31]. Functionality and usability of an application was successfully tested in a preliminary study.

Implementation of such an application required combining of several web technologies. Most significant used technologies are listed in [Table 6].

Table 6. List of selected technologies used in implementation

Technology	Purpose of application
JavaScript - jQuery UI	implementation of sliders, modal dialogs and progress bars
JavaScript - AJAX	changing content on sample websites without refreshing main webpage and changing URL address
HyperText Markup Language HTML 5	structure and content of websites, implementation of forms for gathering data from sliders
Cascade Sheets CSS	manipulation of layout and aesthetics features
Hypertext Preprocessor PHP	processing forms and communication with database
MySQL Database	storage of numeric data from respondent questionnaire, acquired by sliders mechanics

Principles of presented evaluation framework are not limited to the environment of web application. They can be applied e.g. in controlled environment with two calibrated monitors, as in Schenkman and Jönsson research [10]. In this case, an original size and complexity of websites could be preserved. However for fluent transition of user´s attention and coherence with accompanying inquiries, one monitor of sufficient fixed width should be a preferable choice for a testing environment.

5.4 The Procedure and Participants

Participants were provided with a website address, where the application for an experiment was implemented. First they were informed about estimated required time for an experiment (10-15 minutes according to preliminary study results) and on top of every webpage was progress bar available, in order to keep participants motivated. They were asked to follow the instructions and thanked for their participation. General information was then required about their gender, age and estimate of hours spent on internet per week. Blurred screenshot of an application with schematic descriptions was then presented as reference for use of the application, which was designed as easy and intuitive as possible. Reference about using a slider mechanism was added after preliminary study, which verified usability of application.

After clicking on a button "Start experiment", series of three more websites was presented, each of them consisting of two fully functional small websites and seven sliders, six referring to individual websites (aesthetics, usability and content inquiries), one of them referring to both of them (overall preference). Firstly however, modal window was presented with a task, different for each pair of websites. Content of each pair was changed too, along with accompanying pictures, without changing an overall level of impression. Pictures were downloaded from freepik database with licence for study and reference, original source http://www.zcool.com.cn/gfx/ZOTQ2MDQ=.html. After performing a task on both websites in the pair, a user was asked to evaluate these websites and then move on by using a button "Continue on next page". After third pair of websites, a user was thanked for his efforts.

Total number of 17 users participate in this experiment. Their average age was 35 years (SD = 17,2, range:19-64). Estimated time spent on internet per week was in average 16 hours (SD = 9,2, range: 3-30). Gender distribution was 6 men and 11 women. Each of the participants finished all evaluation inquiries.

6 Results

Applied evaluation framework confirmed existence of interferences regarding to perception of aesthetics, usability and information quality. Overall preference was best correlated with difference in evaluation of the manipulated aspect, except for a case of usability manipulation. Minimum possible value of difference is 0 points, maximum 8 points, which is given by 9-point Likert scales with values ranging from 0 to 8. The most significant differences were in aesthetics evaluation, in a case of an aesthetics manipulation - difference of 5,47 points in average, which confirmed an original hypothesis. Most significant in overall user preference was also aesthetics aspect with mean difference between high and low-quality version of 5,47 points.

Aesthetics was most significantly perceived by users differently, if quality differed in another website´s aspect. This discerns from an original hypothesis and shows quite a reverse effect. Average differences in aesthetics evaluation were 3,76 points in case of usability manipulation and 3,65 points in content manipulation, while aesthetics was the same in both cases and should therefore be null. Difference in evaluation of aesthetics was actually more pervasive than in usability in the case of usability manipulation. This can be explained by assumptions made earlier in this paper and cannot be generalized beyond small corporate websites.

Design of an inquiry with 17-point Likert scale, which was used for overall preference, allowed range from -8 to +8 points. This served as a simple control mechanism. If a user preferred a website with lower quality (in any aspect), he was probably not taking the experiment seriously. This way, negative numbers would appear in an overall preference evaluation. We are delighted to present results with no such deviations. Websites of better quality were always preferred by all participants in all comparisons. This presents a partial guarantee, that participants were taking the experiment responsibly, even without a supervisor.

Table 7. Table of average differences (SD) in evaluation scales

Manipulated aspect	**Differences in evaluation**			**Overall preference**
	aesthetics	usability	content	
Aesthetics	**5,47 (0,68)**	1,35 (1,01)	1,76 (1,27)	5,27 (1,26)
Usability	3,76 (1,19)	**2,24 (1,13)**	1,29 (1,04)	3,95 (1,31)
Content	3,65 (1,94)	2,59 (1,63)	**4,35 (1,74)**	4,75 (1,76)

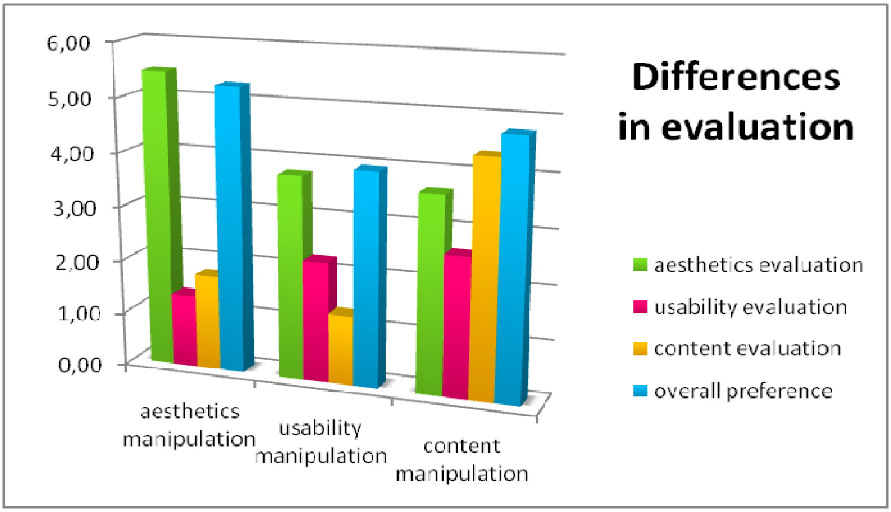


Fig. 2. Graph showing differences in evaluation scales

Paired t-test was performed on the data, which showed that each of the differences in scale evaluation is statistically significant with α=0,05. This is a frequent issue with comparing usability and other scales. More predicative is in this case confidence

Table 8. Confidence intervals for differences in evaluation

Manipulated aspect	Differences in evaluation			Overall preference
	aesthetics	usability	content	
Aesthetics	**4,79 - 6,15**	0,34 - 2,36	0,50 - 3,03	4,01 - 6,53
Usability	2,58 - 4,95	**1,11 - 3,36**	0,25 - 2,33	2,64 - 5,27
Content	1,71 - 5,59	0,96 - 4,21	**2,61 - 6,10**	2,99 - 6,51

interval around the difference as a practical significance [35]. When using this approach, we can be 95% confident that the actual difference between perception of aesthetics while aesthetics is manipulated ranges between 4,79 and 6,15 (while using 9-point Likert scale with maximum value of 8 points). Other values are in [Table 8].

7 Conclusions and Future Work

This study presented an evaluation framework for user preference research, which was implemented into a fully functional web application. Interferences of factors, which are not being manipulated and their perception is still affected by manipulation of other factors, were demonstrated in performed experiment. One of pivotal distinctions from other recent studies was including information quality into aesthetics-usability mix. Another asset is a two-way performed manipulation of variables by creating default-quality website and then lowering manipulated aspect in low-quality version and enhancing it in high-quality version of the website. Finally, the parallel comparison which allowed within-subject design of experiment was presented and verified in the framework used in experiment.

The authors confirmed in this study an essential role of aesthetics in user evaluation and preferences regarding websites. Although there is an agreement in recent research on aesthetics influence, its conditions are still subject of contradictions. The results of this study are similar to those of Tuch et al., which claimed that aesthetics did not affect perceived usability but low usability lowered aesthetics evaluation [13]. In the case of this study, changes in usability and also content quality reflected significantly on aesthetics evaluation. It can be argued that manipulation of variables was not entirely independent. However while borders between aesthetics and usability are not always clear, information quality is quite distinct from them. Since user evaluation [33] of aesthetics was altered by similar measure in cases of both usability and content manipulation, we can conclude that this effect wasn´t caused by eventual blending of performed manipulation.

This research has though several limitations, which should be addressed in future work. First, limited space for tested websites and resulting simplicity of websites could have caused lower ratings in general and less apparent effects in perceived usability. Second, users weren´t supervised to fulfil requested tasks and consequently they could not fulfil them, fulfil them carelessly or on only one of the compared websites. For this reason, a basic control mechanism has been implemented in the experimental design, and its results have been favourable. Despite of that, better supervision is needed for future experiments. Third, perception of users could have been unintentionally manipulated by choice of wording for the evaluation inquiries.

For these reasons, the experiment should be replicated with larger sample of participants in better controlled environment, to confirm its results. It could be also worth exploring a positive approach for further application of the framework. Presented evaluation framework principles are generally applicable, yet the current state of implemented application cannot be used with any chosen website. Generally usable web-based application [30], which implements the framework principles, should also be a subject of future work. Such an application could serve e.g. as a tool for comparing a website of particular company with websites of competition [32].

Acknowledgment. This work and the contribution were supported by: (1) the project No. CZ.1.07/2.2.00/28.0327 Innovation and support of doctoral study program (INDOP), financed from EU and Czech Republic funds; (2) project “Smart Solutions in Ubiquitous Computing Network Environments”, from the Grant Agency of Excellence, University of Hradec Kralove, Faculty of Informatics and Management, Czech Republic; (3) project “SP/2013/03 - SmartHomePoint Solutions for Ubiquitous Computing Environments” from University of Hradec Kralove.

References

[1] Keinonen, T.: Expected usability and product preference. In: Proceedings of the Conference on Designing Interactive Systems, Amsterdam, The Netherlands, pp. 197–204 (1997)

[2] Nielsen, J.: Usability Engineering. Academic Press, San Diego (1993)

[3] Brooke, J.: SUS: A Quick and Dirty Usability Scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry. Taylor & Francis, London (1996)

[4] Tversky, A., Simonson, I.: Context-Dependent Preferences. Management Science 39(10), 1179–1189 (1993)

[5] Tractinsky, N., Katz, A.S., Ikar, D.: What is beautiful is usable. Interacting with Computers 13(2), 127–145 (2000)

[6] Lee, S., Koubek, R.J.: The effects of usability and web design attributes on user preference for e-commerce web sites. Computers in Industry 61(4), 329–341 (2010)

[7] Robins, D., Holmes, J.: Aesthetics and credibility in web site design. Information Processing & Management 44(1), 386–399 (2008)

[8] van Schaik, P., Ling, J.: The role of context in perceptions of the aesthetics of web pages over time. International Journal of Human–Computer Studies 67(1), 79–89 (2009)

[9] Lavie, T., Tractinsky, N.: Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human–Computer Studies 60(3), 269–298 (2004)

[10] Schenkman, B., Jönsson, F.: Aesthetics and preferences of web pages. Behaviour and Information Technology 19(5), 367–377 (2000)

[11] Lee, S., Koubek, R.J.: Understanding user preferences based on usability and aesthetics before and after actual use. Interacting with Computers 22(6), 530–543 (2010)

[12] Hartmann, J., Sutcliffe, A., de Angeli, A.: Towards a theory of user judgment of aesthetics and user interface quality. Transactions on Computer-Human Interaction 15(4) (2008)

[13] Tuch, A.N., Roth, S.P., Hornbæk, K., Opwis, K., Bargas-Avila, J.A.: Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. Computers in Human Behavior 28(5), 1596–1607 (2012)

[14] Hassenzahl, M.: The interplay of beauty, goodness, and usability in interactive products. Human–Computer Interaction 19(4), 319–349 (2004)

[15] Ben-Bassat, T., Meyer, J., Tractinsky, N.: Economic and subjective measures of the perceived value of aesthetics and usability. ACM Transactions on Computer–Human Interaction (TOCHI) 13(2), 210–234 (2006)

[16] Lindgaard, G., Fernandes, G., Dudek, C., Brown, J.: Attention web designers: you have 50 milliseconds to make a good impression! Behaviour and Information Technology 25, 115–126 (2006)

[17] van der Heijden, H.: Factors influencing the usage of websites: the case of a generic portal in the Netherlands. Information & Management 40(6), 541–549 (2003)

[18] Ariely, A.: Predictably Irrational, Revised and Expanded Edition: The Hidden Forces That Shape Our Decisions, 2nd edn. HarperCollins (2010)

[19] Tuch, A.N., Bargas-Avila, J.A., Opwis, K.: Symmetry and aesthetics in website design: It's a man's business. Computers in Human Behavior 26(6), 1831–1837 (2010)

[20] Wu, O., Chen, Y., Li, B., Hu, W.: Evaluating the visual quality of web pages using a computational aesthetic approach. In: WSDM 2011, pp. 337–346 (2011)

[21] Lynch, P.J., Horton, S.: Web Style Guidelines, 2nd edn. Yale University Press (2001)

[22] van Schaik, P., Ling, J.: Modelling user experience with web sites: Usability, hedonic value, beauty and goodness. Interacting with Computers 20(3), 419–432 (2008)

[23] Roth, S.P., Schmutz, P., Pauwels, S.L., Bargas-Avila, J.A., Opwis, K.: Mental models for web objects: Where do users expect to find the most frequent objects in online shops, news portals and company web pages? Interacting with Computers 22(2), 140–152 (2010)

[24] McCracken, D.D., Wolfe, R.J.: User-centered Website Development: A Human-Computer Interaction Approach. Pearson Prentice Hall Inc., Upper Saddle River (2004)

[25] Sonderegger, A., Sauer, J.: The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. Applied Ergonomics 41(3), 403–410 (2010)

[26] Fernandez, A., Insfran, E., Abrahão, S.: Usability evaluation methods for the web: A systematic mapping study. Information and Software Technology 53(8), 789–817 (2011)

[27] Hornbæk, K.: Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies 64(2), 79–102 (2006)

[28] de Angeli, A., Sutcliffe, A., Hartmann, J.: Interaction, usability and aesthetics: what influences users' preferences? In: Proceedings of the 6th Conference on Designing Interactive Systems, University Park, PA, USA,

[29] Hartmann, J., Sutcliffe, A., de Angeli, A.: Investigating attractiveness in web user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, p. 396 (2007)

[30] Kasik, V., Penhaker, M., Novák, V., Bridzik, R., Krawiec, J.: User Interactive Biomedical Data Web Services Application. In: Yonazi, J.J., Sedoyeka, E., Ariwa, E., El-Qawasmeh, E. (eds.) ICeND 2011. CCIS, vol. 171, pp. 223–237. Springer, Heidelberg (2011)

[31] Liou, C.-Y., Cheng, W.-C.: Manifold Construction by Local Neighborhood Preservation. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 683–692. Springer, Heidelberg (2008)

[32] Choroś, K.: Further Tests with Click, Block, and Heat Maps Applied to Website Evaluations. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS (LNAI), vol. 6923, pp. 415–424. Springer, Heidelberg (2011)

[33] Longo, L., Kane, B.: A Novel Methodology for Evaluating User Interfaces in Health Care. In: 24th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2011, Bristol, England, June 27-30 (2011)

[34] TOPlist. TOPlist - globální statistika, `http://toplist.cz/global.html` (cit. April 04, 2013)

[35] Sauro, J., Lewis, J.R.: Quantifying the User Experience. Morgan Kaufmann (2012)

Investigating Feedbacks in Instructional Design Model Using Weak Tie Approach

Chia-Ling Hsu

Center for Teacher Education, Tamkang University
151, Ying-chuan Road, Tamsui, New Taipei City, 25137, Taiwan
clhsu@mail.tku.edu.tw

Abstract. This paper investigated the efficiency of the feedbacks which was analyzed with the weak tie approach, the integrating the grounded theory and the text mining technology. Using this approach, the rare and important chances may be found. These chances may improve the quality of the service learning activities and may enhance the pre-service teachers' self-efficacy. In this study, the ADDIE model was employed in developing the service activities. The feedbacks were the reflections from those who attended the service learning activities. Weak tie approach was engaged in order to find the chances which may make the service activities better. Moreover, the weak tie approach of analyzing feedback may enhance the pre-service teachers' self-efficacy. The significance of this paper is to propose the weak tie approach for enriching the feedback efficiency.

Keywords: curriculum and instruction, instructional design, teacher education, weak tie, knowledge management.

1 Introduction

Systemic instructional design models emphasize not only the objectives but also the feedbacks at each step. For example the Dick and Carey model, the ADDIE model proposed instructional design steps and indicated that the revision was important during the process[1][2]. Therefore, effectively using feedbacks information will improve the quality of the instruction. In other words, this study analyzed the data which may be used for the evaluation in traditional way. However, this study used the data for differently purposes. The reasons were, first, the importance of revision in instructional models. Second, the differential of feedback and evaluation would make clear.

Recently, the service learning courses or activities are the main indicator for building up the students characters of globalization and care for the disadvantaged groups in higher education. Students, who attend the service learning courses or activities, usually will participate in different events with different people. Through the experience of activities, students build up their character of caring people as well as extend their vision of globalization. Dr. Dewey emphasized that experiential learning accord because of students adding something and making contributions [3][4]. Dr. Lewin followed by the theory of Dewey and added the action research methodology to make the experiential learning as an integration of theory and practice [5][6]. Then, Dr. Piaget and Dr. Bruner caused a movement in curriculum and

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 549–561, 2013.

instruction for the experience-based design into the college level courses[7][8]. Therefore, designing an experience-based instruction is not a new topic. However, it is hard to practice in the classroom especially in higher education system where the instruction design is lecture for big audience.

The pre-service teacher program is in higher education for educating college students to be the secondary education teachers. Since the trend in higher education of service learning courses, the teacher program also recruits students practicing their teaching skills to service those children who are in remote educational district.

1.1 Research Motivation

Since Mark Granovetter (1973) brought up an idea of strong tie and weak tie, many scholars paid attention in weak tie for innovation, opportunity, chance or improving[9][10][11]. Many decision making researches applied Keeney's (1992) value focused Thanking [12]. In addition, journal writing is a popular technique for students to reflect their learning in service learning courses or activities. Bain et al. (2002) found that the journal writing was the better feedback for pre-service teachers to practice their teaching by self-analysis [13]. Therefore, this study used these techniques as tools to help students reflecting their teaching processing while they were practicing the teaching skills in servicing learning activities.

1.2 Research Purpose and Significance

The purpose of this study is to investigate the feedbacks from the reflections and observations of pre-service teachers for improving the quality of service learning activities. Moreover, the significance of this study is to enhance the pre-service teachers' self-efficacy by revising the instruction. In detail, the purposes below:

- Find the possible weak ties from the feedbacks in order to revise the instruction.
- Use instructional design model to improve the service learning activities.
- Investigate the pre-service teachers' self-efficacy with the weak tie approach service learning activities.

The significances of this study are enriching the researches in text mining as well as providing the service learning activities for pre-service teacher to enhance their self-efficacy. The benefit of the integrating text mining and grounded theory for finding chances in instructional design is to improve the quality of the service learning activities. As the results, the pre-service teachers' self-efficacy will be improved.

2 Related Literature

This study is based on the objective weak tie approach which was come up from different theories. The innovation and weak tie theories were the foundation. Then, the value focus thinking model was applied. On the other hand, the instructional design for the service learning activities was implemented for pre-service teachers. So, the literature review related to innovation, weak tie, value focus thinking, instructional design and service learning.

2.1 Innovation and Weak Tie

The technologies of finding weak tie are variety. Wang, Hong, Sung, and Hsu applied the KeyGraph technology to find the rare and important element [14]. The results indicated that although the statistics data showed no significant difference, the KeyGraph technology provided more information. Hsu and other educators also applied the KeyGraph technology in education setting. The results pointed out that the learners' scenario map would tell more information than the traditional statistics results. Huang, Tsai, & Hsu also applied the KeyGraph technology to exploring the learners' thinking [15]. Tsai, Huang, Hong, Wang, Sung, and Hsu [16] used KeyGraph technology and tried to find the chances in instructional activity.

The studies above were concerned in finding weak tie in text data. Some weak tie research was related to knowledge creation and transfer. Levin and Cross (2004) found that weak ties provide access to no redundant information [17].

2.2 Value Focus Thinking

Thinking about value was to decide what you want and then to figure out how you can get it. So, it was a nature way for alternative-focus thinking. Hsu, Hong, Wang, Chiu, and Chang (2009) used the VFT model in instruction design to improving the teaching quality [18].

2.3 Instructional Design and ADDIE Model

In the domain of curriculum and instruction are emphasizes not only designing an effective curriculum or instruction but also evaluation the students learning outcomes. No matter what approach the research used would lead to a reflection of instruction [19][20][21][22][23]. Marsh and Willis (1999) indicated that different school implemented the curriculum and instruction differently but it only a few models in total, the objective model, Countenance model, illuminative model and educational connoisseurship model [24].

Instructional design has two meanings. One is to provide teaching blueprints, and the other is to examine teaching and provide solutions. Accordingly, the practice of instructional design is to target specific learners, select specific approaches, contents, and strategies, and make an effective teaching policy (Smith & Ragan 1993) [25]. Given that effective teaching requires proper prior planning, there are several hypotheses about instructional design (Gagne, Briggs & Wager 1992): 1. Teaching is aimed at promoting individual learning. 2. Instructional design is divided into immediate design and long-term design. 3. Systematic instructional design has a great impact on individual development. 4. Instructional design should be produced systematically. 5. Instructional design involves the presentation and acquisition of knowledge on the basis of the learning theory and cognitive theory [26]. Hence, teaching materials should not be the presentation of what teachers know only. How to present teaching materials in order to facilitate learning must be taken into consideration as well.

Instructional design is often presented and explained through models. Currently, there are a variety of instructional design models, among which the ADDIE model is the most common one (Michael, Marlon & Roberto 2002) [27]. The ADDIE model includes five phases: analyze, design, develop, implement, and evaluate. Due to its simplicity and viability, the ADDIE model is heavily used in instructional design of e-learning (Hsu & Kuo 2000) [28]. The present study also used the ADDIE model to divide the process of the design and implementation of the web-based course into the above-mentioned five phases.

2.4 Service Learning

The service learning focuses on providing service as well as learning form the service activities [29]. However, Eyler (2000) indicated that the impact of service learning on college students needed to identify intellectual outcomes to embed into instructional design [30]. In order to assessing the outcome of service learning, Karayan and Gathercoal (20005) used a "ProfPort Webfolio System" for student service learning as teaching, learning, assessment and research tool [31]. The system somehow is part of traditional portfolio. So, how to improve the quality of service learning is back to how to investigate the reflection to making improving.

Journal writing is a popular technique for students to reflect their learning in service learning courses or activities. Bain et al. (2002) found that the journal writing was the better feedback for pre-service teachers to practice their teaching by self-analysis [32]. Therefore, this study used this technique as a tool to help students reflecting their teaching processing while they were practicing the teaching skills in servicing learning activities.

3 Research Method

The research method of this study, first, applied the instructional design model. Then, the grounded theory and text mining procedure was used. Finally, the pre-service teachers' self-efficacy was examining by the implantation the innovated service learning activities. Therefore, the steps of ADDIE model for developing the service learning activities was described. The technologies of the objective weak tie model in order to obtaining the chances for improving service learning activities were followed. The research procedure, participant, and instrument were described as well.

3.1 ADDIE Model for Developing the Service Learning Activities

The pre-service teachers practiced the service learning activities which were developed by ADDIE model. Each steps of the ADDIE model to develop the program was described below:

Analyze:
Selecting and analyzing the needs of the secondary schools in the disadvantaged remote education districts was the beginning. Therefore, interview with the directors of the secondary schools to understanding the environment of the schools and the characteristic of the junior high school students. Then, the pre-service teachers were voluntarily joint the service learning project.

Design:
Completing the analyzing stage, the pre-service teachers design the service learning activities according to the needs assessment from the analyzing stage. During the

design section, a training program was provided for the pre-service teachers to help them to design the activities. The break ice games, for example, helped not only increase the acquaintance among the pre-service teachers but also learn the role playing for them to prepare the activities design. As the results, they decided the topics were the global and domestics culture.

Develop:
According to the topics, understanding the global and domestic culture, the activities were constructed as figure 1. These activities included different academic area, such as, Social Science, Math, Chinese, English, and Japanese.

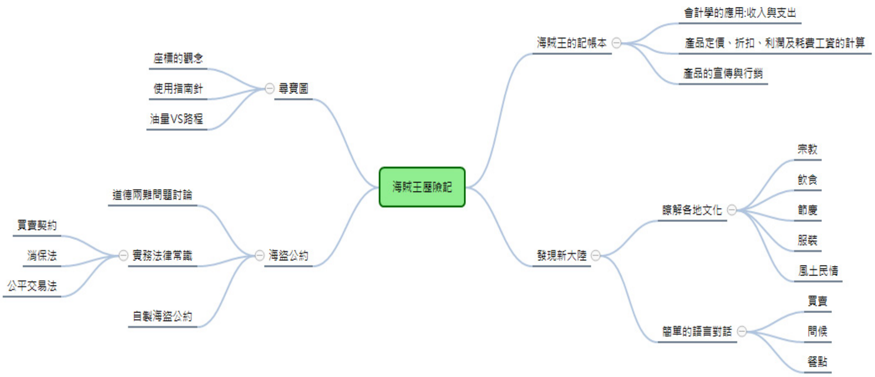


Fig. 1. The structure of the topics

Implement:
The implementation of the service learning activities was held in one weak section total for forty hours. The schedule was shown in table 1. At the end of activities of each day, the evaluating meeting was held.

Table 1. The schedule of the service learning activities

Day	Activities	Subjects	Objectives
Mon.	Introduction		Knowing each other
	Where are we?	Math	Understanding coordinates
	Who is right?	Social science	Decision making for the dilemma
	What will I do?	English	Role playing in English culture
Tue.	Where am I going?	Math	Using Magnetic needle
	What is right?(I)	Social science	Understanding law
	What will I do?	Japanese	Role playing in Japanese culture
	Is it balance?	Math	Calculating balance sheet
Wed.	Is it enough?	Math	Calculating the oil for the road
	What is right?(II)	Social science	Understanding law
	What do I say?(I)	English	English conversation
	Business Game(I)	Math	Understanding how to run business

Table 1. (*Continued.*)

Day	Activities	Subjects	Objectives
Thur.	Business Game(II)	Math	Understanding advertisement
	Court game	Social science	Understanding law
	What do I say?(II)	Japanese	Japanese conversation
	Show me	Social science	Creating merchandises
Fri.	Ground game	Social science	Integrating knowledge
	Flea market game	Social science	Integrating knowledge

Evaluate:
In the first stage of this research was based on the ADDIE model for developing the service learning activities. In this stage the program evaluation was held after the activities each day for one hour. Pre-service teachers discussed what the activities would be modified for the next day.

3.2 Objective Weak Tie Mode

The Objective weak tie model was based on the grounded theory with text mining technologies. Moreover, using KeyGraph technology would provide scenarios for possible chances. The KeyGraph was a kind of data-driven technology which meant that the scenarios were not driven by objectives. However, using value focus thinking model would reduce the effect of the KeyGraph technology. With the objectives and the interesting topics we added to the original data-driven information, a more meaningful scenario would be formed. Figure 2 presented the procedure of objective weak tie model. Each step was explained as follows.

Step 1: First, the reflective diaries of the pre-service teachers were collected. Then, clean the text data with CKIP system in order to identifying the syntactic from each sentence.
Step 2: Using the KeyGraph technology, the thresholds of key frequency and jaccard relation were determined. In this stage, the different tribes (subgroups) would be determined.
Step 3: In order to finding the meaningful chances from the KeyGraph map, the value focus thinking model was applied. In this stage, the chances would be considered to improve the activities.
Step 4: With the chances a new or innovated storyline would be clear in order to reform the activities.

3.3 Research Procedure

This study usied the weak tie approach in instructional design model in order to finding the chances for improving the activities. The chances finding would provide some suggestion for teacher education service learning course. Moreover, the junior high school students would benefit by the practical activities which will full fill the social justice in education. Therefore, the research procedure described below.

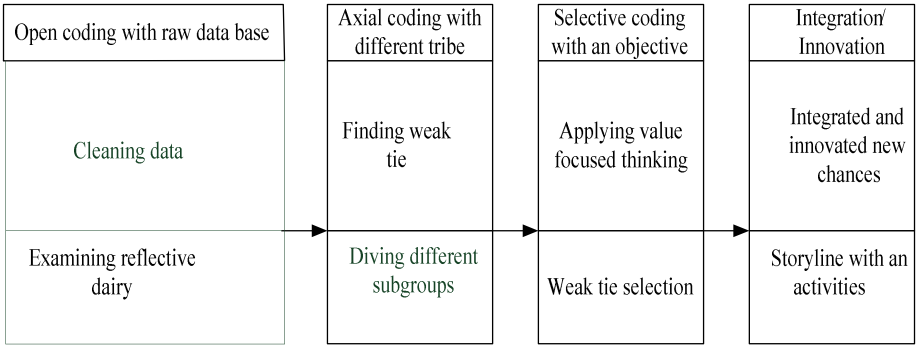

Open coding with raw data base
Cleaning data
Examining reflective dairy
Axial coding with different tribe
Finding weak tie
Diving different subgroups
Selective coding with an objective
Applying value focused thinking
Weak tie selection
Integration/ Innovation
Integrated and innovated new chances
Storyline with an activities

Fig. 2. The Objective Weak tie Model

First, this research was continuing on weak tie approach study that would discover or innovate a better teaching method for disadvantaged students for the social justice sake.

Second, the pre-service teachers were voluntarily joint the service learning research.

Third, the researcher as a supervisor helped the pre-service teachers develop the service learning curriculum. The ADDIE model was applied into the curriculum and instructional design. During these three weeks, the pre-service teachers finished the courses design. Then, for one week forty hours activities was implemented to those students who were in the disadvantaged remote education districts junior high school.

Forth, the pre-service teachers' reflections were analyzed. The objective weak tie model was applied for finding chances or innovating scenario.

Fifth, using the feedbacks from the objective weak tie model, the pre-service teachers refined and implemented the service learning activities.

Sixth, the videotaping, observation sheets, students' responds, teacher self-efficacy questionnaire were collected and analyzed.

3.4 Research Participants

The study was conducted with pre-service teachers and secondary high school students who were participated in "Educational disadvantaged remote district teaching service learning activities". Twenty three pre-service teachers and twenty students were enrolled. Although these pre-service teachers have some tutoring experience, they all had not the experience in a real class. These pre-service teachers' majors were variety, Math, English, Chinese, Japanese, and Business. The purposes of their attending were practicing their learning from teacher education program and increasing their teaching experiences.

3.5 Research Instrument

There were two text instruments, reflection and observation sheet. In addition, there were two questionnaires, the teachers self-efficacy and students courses motivation. Every pre-service teacher was asked to write a diary each day during the activities period of time. The pre-service teachers who taught the activities that day wrote reflections. The reflections were followed these guild lines:

- Content knowledge
- General pedagogical knowledge, with special reference to those broad principles and strategies of classroom management and organization that appear to transcend subject matter
- Curriculum knowledge, with particular grasp of the materials and programs that serve as 'tools of the trade' for teachers
- Pedagogical content knowledge, that special amalgam of content and pedagogy that is uniquely the province of teachers, their special form of professional understanding
- Knowledge of learners and their characteristics
- Knowledge of educational contexts, ranging from workings of the group or classroom, the governance and financing of school districts, to the character of communities and cultures; and
- Knowledge of educational ends, purposes, and values, and their philosophical and historical grounds

The pre-service teachers who did not teach that day observed the activities. The diary would include the teaching method, students-reaction, the teacher-student interaction and the reflections.

The questionnaire contained 30 questions. There were eleven reverse questions in the questionnaire, item 5, 13, 14, 15, 16, 19, 22, 24, 26, 28, 29. Pre-service teacher score the number from 1 to 10 for their feeling very uncomfortable to very comfortable for each item. After the principle components analysis by SPSS, the questionnaire was divided into 4 aspects: class management, teacher qualification, teaching strategies, and the educational theories. Therefore, the class management contained item 1, 2, 5, 6, 9, and 10. The teacher qualification contained item 4, 14, 15, 16, 17, and 21. The teaching strategies contained item 12, 13, 20, 22, 23, 27, and 29. At last, the educational theories contained item 18, 19, 24, 25, 26, 28, and 30.

The course motivation questionnaire was the ARCS questionnaire. The ARCS questionnaire contains 34 items. The Attention factor contains item 1, 4*, 10, 15, 21, 24, 26*, and 29. The items marked with * mean the inverse items. The Relevance factor contains item 2, 5, 8, 13, 20, 22, 23, 25, and 28. The Confidence factor contains item 3, 6*, 9, 11*, 17*, 27, 30, and 34. The Satisfaction factor contains item 7, 12, 14, 16, 18, 19, 31*, 32, and 33. The score is calculated for each item by 5 scale points, from non-agree to very agree

4 Results

The results from the objective weak tie indicated that the chances were adding stories and more animations. Figure 3 revealed the stories may be a chance from text mining technology. Figure 4 appeared that the animation may be the other chance from data analysis.

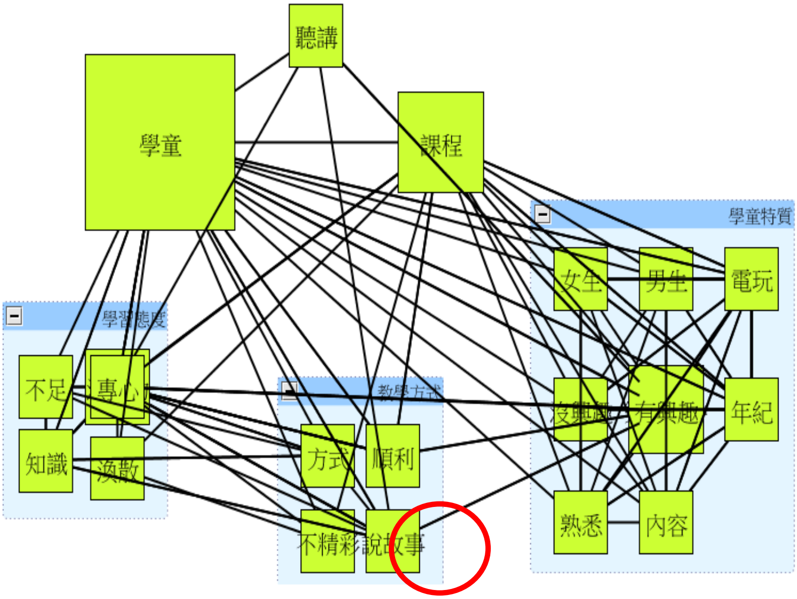


Fig. 3. The chance I form reflection data analysis

Observation Results

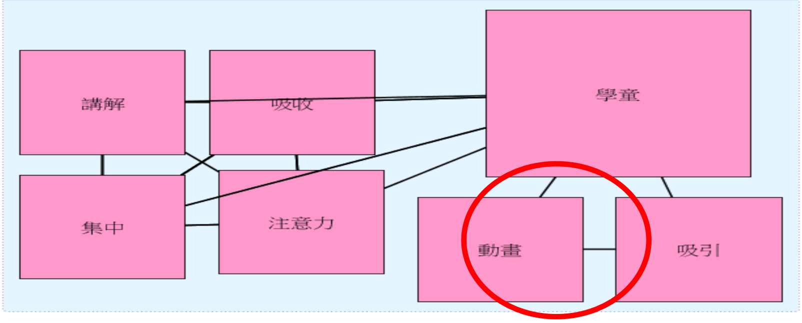


Fig. 4. The chance II from the observation sheet data analysis

The pre-service teachers self-efficacy was shown in table 1.

Table 2. The results of teacher self-efficacy

Implement I		advantage	p
class management	pre	**48.48**	.546
	aft	45.83	
teacher qualification	pre2	**20.65**	.735
	aft	20.26	
teaching strategies	pre	**39.17**	.021*
	aft	38.39	
educational theories	pre	32.91	.080
	aft	**34.70**	
Implement II		advantage	p
class management	pre	29.04	0.754
	aft	**31.14**	
teacher qualification	pre	45/79	0.149
	aft	**46.46**	
teaching strategies	pre	46.07	.000***
	aft	**55.75**	
educational theories	pre	35.79	0.513
	aft	**37.11**	

The students outcomes was measured by the ARCS questionnaire. The result was in table 3.

Table 3. The result of ARCS

	P	(I) group	(J) group	Avagee
Attention	**.011 ***	1	2	5.038
		1	3	3.871
		2	3	-1.167
Relevance	.140	1	2	3.194
		1	3	3.671
		2	3	.478
Confidence	.202	1	2	.975
		1	3	2.686
		2	3	1.711
Satisfaction	.094	1	2	3.429
		1	3	4.229
		2	3	.800

5 Conclusion and Suggestion

This study investigated the feedbacks using the objective weak tie model analyzing the data. The results indicated the way of analyzing feedbacks was efficience.

5.1 Conclusion

The results from the objective weak tie model indicated that the story and animation were rare and important factors. Using these feedbacks for improving the quality of the service learning activities and enhancing the self-efficacy of the pre-service teachers was effective.

5.2 Suggestions

The objective weak tie model found the rare and important chances that enhanced the pre-service teacher self0efficacy by practicing the good quality service learning activities. There are some suggestions to address. The objective weak tie model was based on the grounded theory and texting mining technologies. The method integrating the qualitative and quantitative methods to find an innovative scenario is a big challenge. The text mining technology is needed to put more efforts for analyzing text meaning correctly. More research is needed for future study.

Acknowledgments. We would like to thank many staffs and pre-service teachers in Tamkang University and those junior high school students who attained this project.

References

1. Dick, W., Carey, L., Birnbaum, R.: The Systemic Design of Instructions, 6th edn. Scott Foresman, Glenview (2004)
2. Peterson, C.: Bringing ADDIE to Life: Instructional Design at Its Best. Journal of Educational Multimedia and Hypermedia 12(3), 227–241 (2003)
3. Dewey, J.: Art as Experience. Capricorn Books, New York (1934)
4. Dewey, J.: Experience and Nature. Dover Publications, New York (1958); Clerk Maxwell, J.: A Treatise on Electricity and Magnetism, 3rd edn., vol. 2. pp. 68–73. Clarendon, Oxford (1892)
5. Lewin, K.: Field Theory in Social Sciences. Harper & Row, New York (1951)
6. Jacobs, I.S., Bean, C.P.: Fine particles, thin films and exchange anisotropy. In: Rado, G.T., Suhl, H. (eds.) Magnetism, vol. III, pp. 271–350. Academic, New York (1963)
7. Piaget, J.: Genetic Epistemology. Columbia University Press, New York (1970)
8. Bruner, J.: The Relevance of Education. W.W.Norton, New York (1971)
9. Granovetter, M.S.: The Strength of Weak Ties. American Journal of Sociology 78, 1360–1380 (1973)
10. Granovetter, M.S.: The strength of Weak Ties: A Network Theory Revisited. In: Marsden, P.V., Lin, N. (eds.) Social Structure and Network Analysis. Sage, Beverly Hills (1982)
11. Krackhardt, D.: The Strength of Strong Ties: The Importance of Philos in Organizations. In: Nohria, N., Eccles, R.G. (eds.) Networks and Organizations: Structure, Form, and Action. Harvard Business School Press, Boston (1992)
12. Keeney, R.L.: Value Focused Thinking – A Path to Creative Decision making. Harvard University Press, Cambridge (1992)
13. Bain, J.D., Mills, C., Ballantyne, R., Packer, J.: Developing Reflection on Practice Through Journal writing: impacts of variations in the focus and level of feedback. Teacher and Teaching: Theory and Practice 8(2), 171–196 (2002); Nicole, R.: Title of paper with only first word capitalized. J. Name Stand. Abbrev. (in press)
14. Wang, L.-H., Hong, C.-F., Hsu, C.-L.: Closed - ended Questionnaire Data Analysis. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4253, pp. 1–7. Springer, Heidelberg (2006)
15. Huang, C.J., Tsai, P.H., Hsu, C.L.: Exploring Cognitive Difference in Instructional Outcomes Using Text Mining Technology. In: Proc. of 2006 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2116–2120. IEEE Press, New York (2006)
16. Tsai, P.H., Huang, C.J., Hong, C.F., Wang, L.H., Sung, M.Y., Hsu, C.L.: Discover Learner Comprehension and Potential Chances from Documents. In: The 11th IPMU International Conference, Paris, France (2006)
17. Levin, D.Z., Cross, B.: The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in effective Knowledge Transfer. Management Science 30(1), 1477–1490 (2004)
18. Hsu, C.L., Hong, C.F., Wang, A.L., Chiu, T.F., Chang, Y.F.: Value Focused Association Map (VFAM) - An Alternative Learning Outcomes Presenting. In: Word Conference on Educational Multimedia, Hypermedia & Telecommunications, June 22-26, pp. 3270–3275, Hawaii, USA (2009)

19. Hsu, C.L., Chang, Y.F.: Study of the Relationship with the Media Material and the Students' Learning motivation. J. Educational Study 116, 64–76 (2003)
20. Hsu, C.L.: E-CAI Case Study. Educational Technology and Media 33, 28–35 (1997)
21. Hsu, C.L., Kuo, C.H.: Study of e-Learning Material Technology. In: 2000 e-Learning Theory and Practice Conference, pp. 61–65. National Chiao Tung University, Shin-Chu (2000)
22. Hsu, C.C., Wang, L.H., Hong, C.F., Sung, M.Y., Tasi, P.H.: The KeyGraph Perspective in ARCS motivation model. In: The 6th IEEE International Conference on Advanced Learning Technologies, pp. 970–974. IEEE Press, New York (2006)
23. Hsu, C.C., Wang, L.H., Hong, C.F.: Understanding students' Conceptions and providing Scaffold Teaching Activities. In: International Conference of Teaching and Learning for Excellence, Tamsui, pp. 166–175 (2007)
24. Marsh, C.J., Willis, G.: Curriculum alternative approaches, ongoing Issues. Prentice-Hall Inc., New Jersey (1999)
25. Smith, P.L., Ragan, T.J.: Instructional design. Macmillan, New York (1993)
26. Gagne, R.M., Briggs, L.J., Wager, W.W.: Principles of instructional design. Fort Worth, Harcourt Brace Jovanovich, TX (1992)
27. Michael, T., Marlon, M., Roberto, J.: The third dimension of ADDIE: A cultural embrace. TechTrends 46(12), 40–45 (2002)
28. Hsu, C.L., Kuo, C.H.: The Study of Web Material. In: The 2000 Conference on Web Learning Theory and Practice, Taiwan, pp. 61–65 (2000)
29. Furco, A.: Service - Learning: A Balanced Approach to Experiential Education. In: Expanding Boundaries: Service and Learning, Corporation for National Service, Washington, DC, pp. 2–6 (1996)
30. Eyler, J.S.: What Do We Most Need To Know about the Impact of Service-Learning on Student Learning? Michigan Journal of Community Service Learning, Special Issue, 11–17 (2000)
31. Karayan, S., Gathercoal, P.: Assessing Service-Learning in Teacher Education. Teacher Education Quarterly, 79–92 (2005)
32. Bain, J.D., Mills, C., Ballantyne, R., Packer, J.: Developing Reflection on Practice Through Journal writing: impacts of variations in the focus and level of feedback. Teacher and Teaching: Theory and practice 8(2), 171–196 (2002)

Collaborative Learning in Cultural Exchange: Exploring Its Current Trend and Feature

Ai-Ling Wang

English Department, Tamkang University
Tamsui, New Taipei City
Taiwan
wanga@mail.tku.edu.tw

Abstract. This research study aims at exploring what roles different activities practiced in cultural exchanges can play in an educational setting and what future trends in cultural exchanges are. Qualitative data were collected from the Educational Resources Information Center database and were analyzed based on the theoretical framework of grounded theory and text data mining.

Findings of the study showed that cultural exchanges can play the roles as catalyst of language learning, mediator of different linguistic and cultural groups, developer of learning communities, provider of authentic information, peace maker, and creator of multiple learning. On the other hand, the author envisioned and predicted the future of cross-cultural exchanges as an environment of multiple learning, a learning community with various modes of communication, participants with multiple language proficiency, and collaboration between/among students from different cultures. Hopefully, the research findings can provide language teachers who are conducting cultural exchanges with a way of thinking as to how a cross-cultural activity can be developed to satisfy students' needs and how a creative cross-cultural activity can be designed to ensure that students can gain the most benefits from it.

Keywords: grounded theory, text data mining, cultural exchanges, Teaching English to Speakers of Other Languages.

1 Introduction

As pointed out by Chuah [1], "the advancement of information and communication technologies (ICT) is posing more challenges for the delivery of education and training" (p. 37). One of the challenges might be searching for alternative learning and development solutions to support the delivery of learning. Chuah pointed out that lack of interactivity is the main obstacle that limits the development of e-learning and that the online content needs to be more interactive than mere page-turners.

Chuah [1] further pointed out that e-learning should provide learners with experiences that are practical and meaningful and considered experience a dynamic, complex, and subjective phenomenon. Citing from Pine and Gilmore, Chuah illustrated a model of experience and presents the four realms of experiences, namely being entertaining, educational, esthetic, and escapist. Being entertaining refers to

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 562–571, 2013.

experiences where individual is passively absorbing messages or information through various senses. Being educational refers to experiences where individual is actively absorbing messages or information through various senses. Being esthetic refers to experiences where individual is passively immersed in an event or environment but putting little or no effort on it. Being escapist refers to experiences where individual is actively involved in the event or environment. These four realms of experiences are defined based on two different dimensions, namely the level of participation (e.g. active vs. passive) and the kind of connection, or environmental relationship (e.g. absorption vs. immersion).

This paper aims at exploring how different cultural exchange activities may engage students in real learning and what roles cultural exchanges can play and what are the future trends of cultural exchange. In this study, the researcher first searched the online digital library of education and information and collect data relevant to cultural exchanges from the Education Resources Information Center (ERIC) database. The researchers then tried to look for patterns that characterize cultural exchanges and to envision a picture of the trend of future cultural exchanges. Data collected for the present study were analyzed based on principles of grounded theory and text data mining and on Chuah's [1] four realms of experiences as mentioned earlier.

2 Literature Review

Literature relevant to the present study, namely grounded theory and data/text mining, is reviewed in this section.

2.1 Grounded Theory

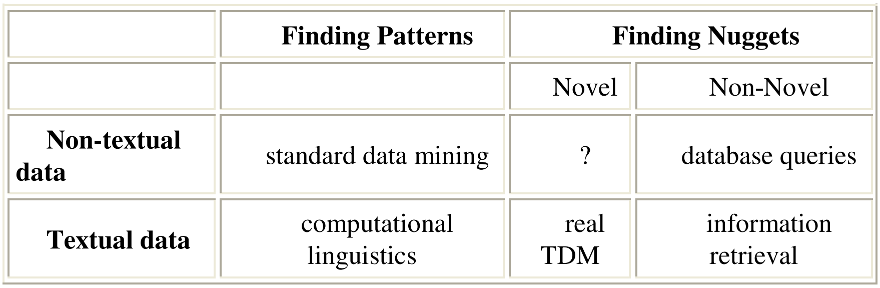
Mills, J., Bonner, A., and Francis, K. [2], in addressing constructivist grounded theory, argued that the theory actively repositioning the researcher as the author of a reconstruction of experience and meaning. They cited Lincoln and Cuba's words by saying that constructivism "is a research paradigm that denies the existence of an objective reality, 'asserting instead that realities are social constructions of the mind, and that there exist as many such constructions as there are individuals although clearly many constructions will be shared' (p. 2).

In this sense, constructivism emphasizes the subjective interrelationship between the researcher and participant, and the co-construction of meaning. Researchers are seen as "part of the research endeavor rather than objective observers, and their values must be acknowledged by themselves and by their readers as an inevitable part of the outcome" (p. 2).

2.2 Text Data Mining (TDM)

According to Hearst [3], text data mining is a nascent field, in comparison with the less nascent data mining, in that TDM practitioners tend to "mining for nuggets, rather than mining for ore extraction. Hearst further differentiated between text data mining and information access by saying that "standard procedure is akin to looking for needles in a needle-stack - the problem isn't so much that the desired information is not known, but rather that the desired information coexists with many other valid pieces of information" (p. 1). In the field of computational linguists, Hearst illustrated that associations or collocations found in computer corpus might not "likely to be what the general business community hopes for when they use the term text data mining" (p.3).

Hearst [3] further suggested that text categorization "does not lead to discovery of new information; presumably the person who wrote the document knew what it was about. Rather, it produces a compact summary of something that is already known" (p. 3). He illustrated the differences between data mining and text data mining applications as shown in the following table:



	Finding Patterns	**Finding Nuggets**	
		Novel	Non-Novel
Non-textual data	standard data mining	?	database queries
Textual data	computational linguistics	real TDM	information retrieval

In sum, Hearst [3] argued that text data mining is exploratory data analysis and suggested using text to form hypotheses and to uncover social impact.

3 Methodology

In this section, the researcher will first introduce the theoretical framework on which the present study is based. Then the researcher will describe how data were collected and analyzed. Finally, the researcher will present findings of the study.

3.1 Theoretical Framework

The present study mainly employs two research paradigms: grounded theory and data/text mining for several reasons: a) This study involves different types of cultural exchanges and different groups of students with different academic and cultural backgrounds, and data/text mining is capable of analyzing complex non-linear and interaction relationships [4]; b) Grounded theory developed by Glaser and Strauss is specific research methodology aiming at building theory from data, and it denotes theoretical constructs derived from qualitative analysis of data [5]. The methodology particularly fits in the nature of the present study because of the various data collected and the philosophical orientation of the researcher who believes that the world is created and recreated through interaction.

3.2 Data Collection

The researcher collected data from the Educational Resource Information Center (ERIC) online database dated between January 1, 2011 and December 31, 2011. Keywords such as *cross-cultural exchange* were used to elicit relevant cross-cultural exchange programs or activities.

3.3 Data Analysis

As Corbin and Strauss [5] argued, each analyst has his or her own repertoire of strategies for analyzing data, and they proposed different tools as analytic strategies. In the present study, the researcher employed what are the most stood-out strategies: the use of questioning and making comparisons. Different types of questions were first formulated as follows to get the researcher better understand the data collected and the ways the data can be organized.

Sensitizing Questions.

- What kinds of activities can be practiced in cultural exchanges?
- What are the activities most favored by students?
- What are the characteristics that popular and beneficial activities share?

Theoretical Questions.

- Do the same groups of students change their attitude toward cultural exchanges over time, collaborative partners, and activities?
- Do different groups of students (e.g. English majors, non-English majors, international students, etc.) favor different types of activity?
- Are there any connections between students and their favored activities?

Making Comparisons.

Another analytic tool used in this study is making comparison. Both constant comparison comparisons and theoretical comparisons were used in an attempt to discover general patterns as well as variation from the data.

Constant Comparisons. These are within-code comparisons. Incidents that are coded the same will be compared for similarities and differences. For example, in the case of this study, the researcher compared cross-cultural exchanges with the same objectives, such as language and cultural learning.

Theoretical Comparisons. To further constant comparison, theoretical comparisons will move comparison to a property and dimensional level. For example, the author compared the elements needed to be taken into consideration if the cultural exchange activity aims at language and cultural learning.

3.4 Analytical Findings and Discussions

After analyzing the data collected for the present study, the author would like to present findings in the order of sensitizing questions and theoretical questions as mentioned earlier.

Database search indicated that there may be different activities practiced in cross-cultural exchanges, including narrative inquiry, international student mentor, American Indian literature and history, Fulbright scholars' experience of teaching management, dialogues among English majors and pre-service English language teachers, team teaching in management, educators across the globe, cross-cultural partnership, and much more. In order to facilitate data analysis, the author first coded the data based on the major objective or objectives of the cultural exchange or investigation or evaluation of cultural exchanges.

From the online data collected for the study, the author roughly assigned the following codes, namely language and cultural learning, professional learning, learning in general, intercultural learning and global awareness, evaluation of exchange programs, and comparison of different culture or programs.

Cultural Exchange as Catalyst of Language Learning

Motivation and purposes of learning a foreign language depend on a lot of internal and external factors. For example, one might learn French as a foreign language because he is interested in French movies and another might be interested in learning German because he, as an engineer, needs to work with Germany engineers for some collaborative projects. How people can be benefited from learning a foreign language varies. However, people may not be aware of how a foreign language can bring to them until they really make use of the language.

Cultural exchanges "force" students to learning another language and provide students with an opportunity to make use of the language. They may develop a sense of accomplishment by using the language, and the sense of accomplishment, in turn, motivate students to keep improving their proficiency level of the language.

Cultural Exchange as Mediator of Different Linguistic and Cultural Groups

As modern technologies develop, communications between or among different cultural groups of students is no longer an issue. However, in order for cross-cultural communications to be meaningful and beneficial to students, organizers of a cross-cultural project need to well-organize the entire project, set the goals beforehand, and take each single detail into consideration. It does not quite make sense if a cultural exchange project is developed just for the sake of cultural exchange.

Another significant point underlying the nature of cultural exchanges is that outsiders always have a more objective and thoughtful view to see through an issue than those who are involved in the issue do. In his study, Stowe [6] studies English language instruction in Taiwan from an American's perspective in a broad and deep scope, and he pointed out some important issues that Taiwanese people were not aware of, such as putting too much emphasis on learning English from literary works, rather than teaching students more meaningful and useful sentence patterns that can be used in daily conversations.

Rajdev [7] reported the experience of American graduate students and professors participating in a service-learning project in India. What motivated the researcher to develop the project was that American society is becoming more diverse and it is common that there are different cultures and ethnic groups in the classroom. One of the objectives of the program was to prepare prospect teachers for their future career in a classroom with more and more culturally diverse students. In her findings, Rajdev stated that her graduate students found out that Indian schools conducted their teaching through a traditional approach. For example, they encouraged rote learning, used blackboard and chalk and paper and pencil, imparted knowledge through lecture and kept their children at the desks, and not took different learning styles into consideration.

Cultural Exchange as Developer of Learning Communities

In many cases, cultural exchanges end as the semester or academic year end. In the present study, however, the author found out that a long-term relationship and a learning community can be established if members of the community enjoy staying in the community and they felt they can be benefited from learning in the community. There are several examples found in the present study, such as Sayers's *De Orilla a Orilla*, which is an international sister-school learning network and Beauchamp's *Face to Faith*, which is an online discussion forum engaging students from all over the world.

Cultural Exchange as Provider of Authentic Information

In the information age, it is not an issue to get information of different cultures. However, no information can be more authentic and more meaningful than the information you receive and experience from direct contact with people from that culture. It is commonly agreed upon that learning a language and learning its culture cannot be separate. Teaching a language involves teaching its culture. The question is: how can a culture be taught? In a traditional language classroom, language teachers just transmit cultural messages orally or students may acquire some cultural message from the textbook.

In a cultural exchange activity, students are provided with an opportunity to have direct contact with their partners from a different culture. They do not *learn* the culture; rather, they *experience* the culture. Lee [8] defined *authenticity* in a more sophisticated way. She distinguished text authenticity from learner authenticity and argued that textually authentic materials are not necessarily learner authentic. Lee regarded authentic texts as those that are not written for teaching purposes, but for a real-life communicative purpose. She also explained what learner-authentic materials are by saying that learner-authentic materials are "motivating, interesting, and useful….[They] are mainly learner-centered and…can serve affectively to promote learners' interest in language learning (p. 324). In cultural exchanges, it can be easy for the student to get authentic information as long as the teacher can put the students in the first place in the course of developing a cultural exchange activity.

In the case of American students vs. Korean students, students' experiences with a one-year-and-a-half cross-cultural partnership showed that "the public face to the project allowed for partners to learn of the professional standards of the other cultural while representing their own.

Cultural Exchange as Peace Maker (Cultural Awareness)

Before having contact with each other, two culturally different groups of people may be ignorant of their partners' culture. Having direct contacts, culturally different groups of students may breakthrough the physical barrier and bridge the cultural gap between them. Most importantly, they will develop cultural awareness and learn to appreciate their own and others' culture.

Ellenwood and Snyders [9] argued that, based on social judgement theory, involvement has a major effect on attitude change. The social judgement theory holds that the more involved a person is with an issue, the more likely that an attitude change will occur. According to Ellenwood and Snyders, people may be aware that their negative attitudes toward people of difference have been misplaced if they are exposed to people from different cultures.

Sahin [10] even argued that foreign language learning contributes to world peace. He envisioned a theory called *Push and Pull Theory in Communication*, arguing that ignorance, which is the cause of wars and conflicts can be eliminated through the power of love and peace language. Sahin suggested that learning a foreign language at a very young age is possible.

Cultural Exchange as Creator of Multiple Learning

Thanks to the convenience of modern technologies and transportation, cultural exchanges can be easily organized. There is a trend that cultural exchanges are organized for multiple purposes, and, thus, students can learn different things in different cultural exchange activities. Jin [11] argued that Western and Asian people hold different beliefs about learning. Western people aim to cultivate the mind to understand the world, whereas Asian people prioritize the self to be perfected morally and socially. In a cultural exchange activity where Western people and Asian people meet, they can learn to appreciate different views and to look at things from different perspectives.

In the cultural exchange activities collected for this study, some of the activities aimed at different learning objectives. Fabregas Janeiro, Fabre, and Rosete [12] reported a faculty-led cultural exchange program, in which Mexican students visited Oklahoma State University in the U.S. for a short period of time. In the cultural exchange program, students not only learned a new language and developed their intercultural competences, but also learned some professional knowledge, including robotics, communication sciences, industrial engineering and biotechnology.

3.5 The Nuggets Found and the Future of Cultural Exchange

As mentioned earlier, in this study, among other things, the author intends to provide "nuggets", rather than just record down what have already existed. From the analytical findings stated in the previous chapter, the author would like to envision the future of cultural exchanges and try to provide some characteristics of the future trend. It can be expected that future cultural exchanges will reveal the following characteristics:

Multiple Learning

Cultural exchanges in the future can be expected to aim at multiple learning. Multiple learning here refers to learning different things at the same time and learning across disciplines. Prensky [13] pointed out that digital natives, who grew up with new technologies, have very different learning styles than those of the older generations. They go faster, have more random access to the information, and can learn different things at the same time. Prensky urged senior educators, aside from learning new stuff, to learn new ways to do old stuff. Given the characteristics of the new generations, organizers of a cultural exchange project have to take all the factors into consideration in the course of planning the project.

On the other hand, Holton [14], in discussing grounded theory, urged researchers to extensively study the methodology in tandem with experiencing the method in order to truly understand classic grounded theory. She also encouraged researchers to read and get involved in knowledge discovery of different disciplines. The point underlies Holton's claim was that there is no more clear-cut between two different academic disciplines. The trend of cultural exchanges may show that students learn from each other and learn across disciplines. Liner learning style can no longer satisfy learners in an era full of information. Participants of cultural exchanges expect to learn in a "hyperlink" mode, just like the way Web texts are organized.

A Learning Community with Various Modes of Communication

As modern technologies and transportation developed, cross-cultural communication in cultural exchange activities can be organized into different modes. For example, participants can interact on the web, through e-mailing, or at a video-conference, or they can pay physical visit and talk to each other face-to-face. One of the advantages of different types of communication is that cross-cultural interaction can be extended beyond the class time and the classroom, and, through more frequent contacts, a learning community and a long-term friendship can be developed.

A short-term meet in a cultural exchange program will not benefit the student a lot. Practitioners of cultural exchange programs in the future should keep a long-term objective in their mind when developing a program. Face to face communications at a video-conference or physically meet with each other may be insufficient for learning to occur or for friendship to develop. Written communications after a video-conference may help the development of a long-term relation and extended learning.

Multiple Language Proficiency

Multiple language proficiency refers to not only being able to speak different foreign languages but also being familiar with different accents of a language spoken by different people in the world. English has long been seen as a lingua franca in international communication. However, the development of modern technologies has contributed to the change of international communication. English is no longer the exclusive language used for international communication. People started learning each other's language. Even though English is used mostly for international communication, different accents and different expressions can be seen everywhere, even on English proficiency tests, such as TOEIC. The trend in international communication is evident in language proficiency tests. The governments in many countries offer language proficiency tests for different languages and for different purposes.

Learning more than one foreign language has also become a trend and bilingualism or multilingualism in the society, such as that of Singapore and Ottawa in Canada, is seen as successful cases. Many colleges or institutes in the world are offering courses for students to learn different languages. As international communication is getting more and more common, being able to speak your conversational partners' language can significantly bridge the gap between you and your partners.

Collaboration between/among Students from Different Cultures

To develop an equal and a friendly global community, cross-cultural collaboration can be a good way to achieve the goal. Based on findings of the study, the author can envision a prosperous future of cultural exchanges. The future is expected to see more and more team projects collaborated by culturally different groups of students, such as team teaching, joint research on a commonly concerned global issue, a language learning community participated by students from different linguistic and cultural backgrounds, and publications co-authored by students from different countries.

In the present study, the author explored in the sea of cultural exchanges and tried to discover something new, rare, and important. On the other hand, the author acted as a tie to connect what she has already known about cultural exchanges with what she is not yet familiar with in order to develop a more innovative model of cultural exchange.

References

1. Chuah, C.-K.P.: Experience Redesign: A Conceptual Framework for Moving Teaching and Learning into a Flexible E-learning Environment. In: Tsang, P., Kwan, R., Fox, R. (eds.) Enhancing Learning through Technology, pp. 37–50. World Scientific, Singapore (2007)
2. Mills, J., Bonner, A., Francis, K.: The development of constructivist grounded theory (2006), `http://www.ualberta.ca/~iiqm/backissues/5_1/pdf/mills.pdf`
3. Hearst, M.A.: Untangling text data mining (1999), `http://www.sims.berkeley.edu/~hearst`
4. Chye, K.H., Gervals, G., Kiu, Y.S.: Using technology in education: The application of data mining. In: Tsang, P., Kwan, R., Fox, R. (eds.) Enhancing Learning Through Technology, pp. 185–198. World Scientific, Singapore (2007)
5. Corbin, J., Strauss, A.: Basics of qualitative research: Techniques and procedures for developing grounded theory, 3rd edn. Sage, Los Angeles (2008)
6. Stowe, J.E.: English Language Instruction in the Schools in Transition: The Case of Taiwan in the 1980s. New York University, New York (1990)
7. Rajdev, U.: Educators across the Globe Collaborate and Exchange Ideas. Journal of International Research 7(2), 195–197 (1981)
8. Lee, W.Y.: Authenticity Revisited: Text Authenticity and Learner Authenticity. ELT Journal 49/4(2), 323–328 (1995)

9. Ellenwood, A.E., Synders, F.J.A.: Virtual Journey Couples with Face-to-face Exchange: Enhancing the Cultural Sensitivity and Competence of Graduate Students. Intercultural Education 21(6), 549–566 (2010)
10. Sahin, Y.: The Importance of the Foreign Language Learning Contributing to World Peace. US-China Education Review 8(5), 580–588 (2011)
11. Jin, L.: Cultural Foundations of Learning: East and West. Cambridge, New York (2012)
12. Fabregas Janeiro, M.G., Fabra, R.L., Rosete, R.T.: Developing Successful International Faculty Led Program. US-China Education Review B4, 375–382 (2012)
13. Prensky, M.: Digital Natives, Digital Immigrants. On the Horizon 9(5), 1–7 (2001)
14. Holton, J.A.: The Coding Process and Its Challenges. In: Bryant, A., Charmaz, K. (eds.) The Sage Handbook of Grounded Theory, pp. 265–289. Sage, London (2007)

Exploring Peer Scaffolding Opportunities on Experiential Problem Solving Learning

Min-Huei Lin[1], Ming-Puu Chen[2], and Ching-Fan Chen[3]

[1] Department of Information Management, Aletheia University, New Taipei City, Taiwan
au4052@au.edu.tw
[2] Graduate Institute of Information and Computer Education,
National Taiwan Normal University,Taipei, Taiwan
mpchen@ice.ntnu.edu.tw
[3] Department of Educational Technology, Tamkang University, New Taipei City, Taiwan
cfchen@mail.tku.edu.tw

Abstract. In the problem solving learning of database management and application course, whenever the novice students can follow and perform cannot ensure that they can really understand why they do so. The gaps between doing and knowing may hinder students from solving further database applications problems. So the study conducts the self-explanation prompts to elicit students' reflection-on-action of problem solving learning tasks. To explore students' reflective self-explanations, the opportunities for reducing the gaps between doing and knowing emerge and use it to design peer scaffoldings. To achieve appropriate peer tutoring scaffolding, students also need to report their social relationships and peers' interdependency. The aim of this study is to propose the experiential problem solving learning framework and double-spiral model to provide students with self-explanation prompts and analyze students' problem solving processes through the human-centered computing system. On the visualized information, teachers play the role of weak-tie to link clusters of students through identifying the weak-tie among clusters emerged from students' reflection and social relationships data. Based on the selected weak-tie, teachers can recommend peer tutors for students to support their problem solving learning.

Keywords: peer tutoring scaffolding, self-explanation, problem solving, weak-tie.

1 Introduction

The goal of teaching database management along with database applications is to prepare students to gather big data and transform these data into information and to solve authentic problems. Database applications are served as tools to help solving problems and making sound decision [1]. The novices are lack of domain-specific knowledge, so the doing and knowing about actions of problem solving learning are usually inconsistent. So it is necessary to explore the self-assessment of students about doing and the self-explanation of students about reflecting what they know and why they do so. Whenever the gaps between doing and knowing are emerged,

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 572–581, 2013.

teachers may have opportunities to design scaffolding for students to reduce the gaps. The present study will conduct peer tutoring scaffolding strategies to assist students in repairing misunderstanding and flaw mental models of database applications problem solving learning.

This study employed information technology, which is named as Human-Centered Computing System that is based on experts' recognition and concept of textual data, to collect the textual reflection data related to the tasks of the problem solving learning and textual social relationships data. Then, the experts' implicit knowledge (value-driven) is used to activate information technology so as to integrate data (data-driven) with a view to showing the knowing structures created by the problem solvers. Teachers play the role of weak-tie recognizer and identify the emerged weak-tie to design scaffolding strategies to support students performing tasks of problem solving learning in database management and applications.

2 Literature Review

2.1 Learning Database Application by Problem Solving

The essentials of teaching database management contain data modeling, database development, database applications and implementations of database information system projects. To learn the higher levels of database knowledge, the more abstract concepts which often make learners perform poorly are necessary. Therefore, various instructional strategies and methods have been suggested in order to reduce misconceptions and enhance meaningful learning, such as problem-based learning [1]. The goal of learning database applications is for students to applying the theoretical concepts to organize vast amounts of data quickly and convert the data into information. It is to say that database applications are tools that are used to help solve problems or make sound business decisions. When solving these types of problems, there usually is no clear path to a solution [9]. During the problem solving learning, students need to learn to integrate various database applications into solving various phases of problems. The links between learning and problem solving suggest that students can learn heuristics and strategies and become better problem solvers, and it is best to integrate problem solving with academic content. Through problem solving learning, content, skills, and attitudes could be integrated in the learning context in order to promote learning by problem solving. Therefore, problem solving could be employed as a prospective means for learning database management and applications. This study aims to investigate the conceptual and operational framework for scaffolding the problem solving learning process in database management and applications.

2.2 Linking Experiential Learning and Problem Solving

Kolb (1984) provides a descriptive model of the experiential learning process that shows how experience is translated through reflection into concepts. As shown in figure 1, the Kolb's experiential learning cycle emphasis the notion that learning experiences are developed through a four-stage cyclic learning process of concrete experience, reflective observation, abstract conceptualization, and active experimentation [7]. Students generate reflective observation from concrete experience and use this as a basis of assimilation to abstract conceptualization. They can then verify the acquired abstract knowledge through applying it in appropriate contexts. The problem solving strategies used in educational settings include authentic problem solving contextual situations, defining the problem, deciding what skills and resources are necessary to investigate the problem, and then posing possible solutions [5]. Authentic problem solving contextual situations provide a simplification of abstract concepts and facilitate learners in developing transferable skills. Understanding problem and exploring the curriculum strategies could be used to facilitate knowledge construction by generating possible solutions, and enhance knowledge consolidation by determining the best fit solution. Therefore, integrating problem solving with the experiential learning cycle provides students with richer learning experiences and enhances the development of database management skills.

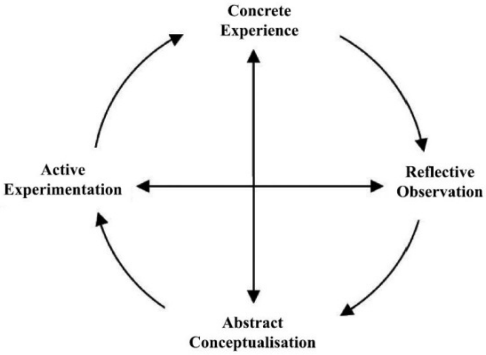


Fig. 1. The four stages of the cyclic learning process of experiential learning

2.3 Self-explanation Prompts as Scaffolding Strategies

Self-explanation refers to a reflective activity explaining to oneself a learning material in order to understand facts from the material or to repair misunderstanding during studying worked-out examples or reading exploratory texts [4]. It seems obvious that a student performs better at problem-solving tasks, generates inferences which facilitate conceptual understanding, and repairs flawed mental models as well when being encouraged to use the self-explanation strategy during learning [3]. Question prompts can guide student attention to specific aspects of their learning process [10], thereby helping students to monitor and evaluate problem-solving processes. The self-explanation prompts enabled students to reflect on which probability principles they used in solving the problems. When learners realize a gap between their current mental model and incoming information, they tend to infer new knowledge while explaining to themselves [3]. But conducting problem solving tasks requires more cognitive efforts such as identifying problems, testing hypotheses, and finding solutions, the problem solving performance will be examined by reflection on one's problem-solving process. This study examined the gaps between the students' self-assessment for resolving questions and self-explanation, and then recommended the peer tutors to bridge these gaps to scaffold students in promoting problem solving performance.

3 Methodology

3.1 The Experiential Problem Solving Framework

As shown in figure 2, the present study incorporated problem solving, question prompting and self-explanation activities with the experiential learning cycle (the blue cycle) to provide contexts for facilitating transformation of learning experiences from concrete experience to abstract conceptualization. Subsequently, teachers conduct reflection to discover the opportunities to scaffold students' problem solving learning by recommending peer tutors (the red cycle) by the Human-centered computing system in 3.3.

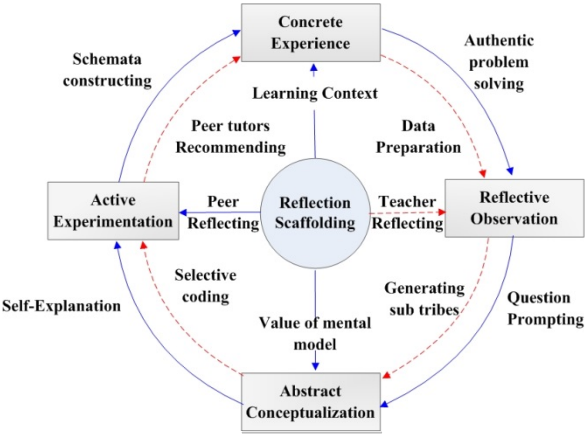


Fig. 2. The experiential problem solving learning and reflection scaffolding framework

Learning Contexts

Learning contexts provided problem-solving contextual situations for the students. With the help of the employed problem solving scenarios, real-life events (concrete experiences) which closely related to the target concepts were introduced to serve as contexts for problem solving. The theoretical concepts and practical skills of data modeling, database applications were embedded in the problem solving context of consistently transforming mysql database into sql server database through entity-relation model and sql scripts.

Question Prompts and Self-explanation

Question prompts provided opportunities for learners to clarify, consolidate, and elaborate concepts through matching prior knowledge and selecting decisions. Furthermore, in self-explanation activities, progressive challenges were employed by means of self-explanation prompts for task completion, levels of performance for students to challenge themselves in identifying correct concepts and non-examples, and upgrade their levels of performance. Through the progressive self-explanation process, students needed to consolidate and elaborate their learned concepts of data modeling and database applications through trial and error and reflection. The acquisition of cognitive skills was a ‘learning by doing’ process that translated declarative knowledge into procedural knowledge. Therefore, the transaction between a learner’s individual skill routine and its domain of application is thus developed iteratively by means of doing and reflecting process. The consolidated database knowledge of this phase also served as the prerequisite knowledge for further application and experimentation.

3.2 Intelligence for Design and Choice

In this study, the self-assessment and reflection from all students’ self-explanations are collected and rated according to some aspects about the critical points of database applications in advance. These ratings include the actions and reasons about why the decision making according to the declarative knowledge and procedural knowledge. Then, the scaffolding opportunities were explored by mining these collected data and the chance discovery framework is shown in figure 3 [8].

There are two “S” in the double-spiral analyses framework, the solid “S” is a value-driven process, it means the problem solving activities designed and conducted by a teacher, namely the trajectory of a teacher’s instruction about the tacit knowledge and value of the teacher in the domain-specific problem solving. The dotted “S” is a data-driven process, in that the actually comprehensive levels of knowledge and social relationships of students are analyzed by data mining and text mining the self-explanation and self-reported social relationships data, that is to say, it records the reactions and progresses of individuals or clusters of students’ problem solving under the problem solving learning guidance. The red section of the dotted “S” represents the actual developmental level of students that they really practice, internalize, and perform. After that, the cognition and meta-cognition of every cluster of students is analyzed through the intelligence activity – Human-Centered Computing System [4].

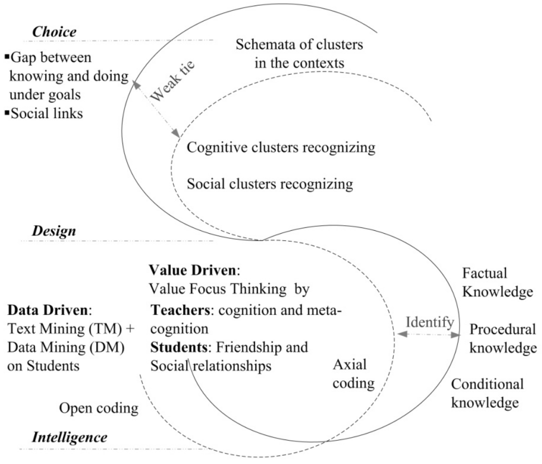

Choice
▪Gap between knowing and doing under goals
▪Social links
Weak tie
Schemata of clusters in the contexts
Cognitive clusters recognizing
Social clusters recognizing
Design
Value Driven:
Value Focus Thinking by
Teachers: cognition and meta-cognition
Students: Friendship and Social relationships
Data Driven:
Text Mining (TM) + Data Mining (DM) on Students
Identify
Factual Knowledge
Procedural knowledge
Conditional knowledge
Axial coding
Open coding
Intelligence

Fig. 3. Double-spiral analyses framework

3.3 Human-Centered Computing System for Recognizing the Reflections for Doing and Knowing

The Human-centered computing system has been constructed by employing grounded theory and text mining techniques, it can be used to assist teachers to collect and analyze the self-reflection for problem solving learning and social relationships of students to obtain the current problem-solving schema that they have built in their long-term memory. When their cognitive structures and social relationships have been visualized and clustered, teachers may have chance to find the weak-tie between the teacher and clusters of students, or the weak-tie between clusters of students, and then design the effective scaffolding through recommending peer tutors for the specific cluster. The detailed process is listed below.

***Step 1*: Data preprocess**

1-1H [1]) Teachers define the key words and concept terms according to content knowledge and problem solving activities.

1-1C [2]) Teachers specify a time period and course unit, and then retrieve the required students' reflection learning data and social relationships data from the database of the learning management system.

1-2H) Teachers recognize the students' reflection learning data according to their domain knowledge, and the social relationships data according to the value of peers' learning, and then eliminate meaningless words, tag words, and append the concept labels with these words.

***Step 2*: Terms co-occurrence analysis (open coding)**

The system processes the open-coding analysis, which contributes to the visualization by using data association analysis. The visual image helps the teacher to focus on the extractable various knowledge structures which are obtained from the data analysis.

2-1C) The association value of two terms can be computed as formula 1. It is based on their co-occurrence in the same sentence.

$$assoc(W_i, W_j) = \sum_{s \in D} \min\left(|W_i|_s, |W_j|_s \right) \tag{1}$$

where W_i and W_j are the ith word and the jth word; s denotes a sentence and is a set of words; D is a set of sentences and includes all the students reflection learning data and social relationships data. $|W_\mathrm{i}|_s$ and $|W_\mathrm{j}|_s$ denote the frequency of words W_i and W_j occurred in the sentence s.

2-2C) The result of co-occurrence analysis will be visualized to co-occurrence association graph.

2-1H) The co-occurrence association graph can help teachers to recognize the concepts and categories inside it, and stimulate teachers preliminary understand the association of students' cognitive clusters appeared from learning data.

***Step 3*: To develop various cognitive clusters (Axial coding)**

Based on the characteristics of knowledge and friendships, "importance" , "urgency" and "preference", teachers extract the data to create the cognitive clusters and social clusters from learning data. The process is illustrated as follows.

[1] 'H' means that actions in such a step are performed by human.

[2] 'C' means that the actions in such a step are performed by computer.

3-1H) Teachers need to decide what categories of knowledge ($w_{knowledge}$) are.
3-1C) Process the categories of knowledge from the first to the last one.
3-2C) Categories of knowledge ($w_{knowledge}$) are used to sift out the meaningful sentences. This variable is used to confirm the research topics and to remove irrelevant sentences with a view to narrowing down the data range and to sift out valid sentences.
3-3C) To integrate all the valid sentences related to knowledge terms, and create integrated association diagrams, then go to 3-1C.

***Step 4*: To decide the weak-tie based on value focus thinking**

4-1H) Based on their domain knowledge and students' social relationships, teachers identify the key factors and weak-tie to recommend peer tutors for achieving appropriate scaffolding.

4 Case Study

4.1 The Context for the Database Application Problem Solving

In this study, students are asked to solve the problem in the context of transferring the mysql database to the sql server database through the entity-relationship model and SQL query scripts. While students finished the task, they took self-assessment according the rubrics and reflected by self-explanation. These textual data were processed in the human-centered computing system for analyzing subsequently. The tasks designed in the problem solving learning are listed in the Figure 4.

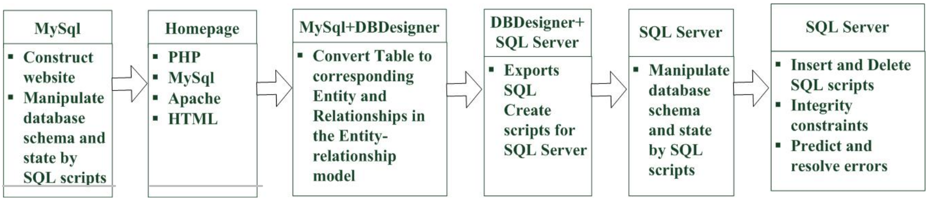

MySql
▪ Construct website
▪ Manipulate database schema and state by SQL scripts
Homepage
▪ PHP
▪ MySql
▪ Apache
▪ HTML
MySql+DBDesigner
▪ Convert Table to corresponding Entity and Relationships in the Entity-relationship model
DBDesigner+ SQL Server
▪ Exports SQL Create scripts for SQL Server
SQL Server
▪ Manipulate database schema and state by SQL scripts
SQL Server
▪ Insert and Delete SQL scripts
▪ Integrity constraints
▪ Predict and resolve errors

Fig. 4. Problem solving learning tasks

4.2 Data Resource

The experimental data has been collected from students since February, 2013, who have experiences in learning database management for one semester in September, 2012. They have the basis for programming languages, algorithms, flowchart, entity-relationship model and diagrams, mysql and php homepage design.

4.3 Human-Centered Computing Phase: Extract the Gaps between Doing and Knowing for Various Clusters

Ratings. According to the rubrics of the rating system, the self-explanations of all students are rated, and all students are dispatched to one of the four quadrants (figure 5). As shown in figure 5, the first quadrant means that the doing and knowing of students are almost consistent, students in the second quadrant may know the principles but lack the experiences of practice, students in the fourth quadrant can only do but do not know why, the students in the third quadrant have no ability to do and do not know the related knowledge. In the third, fourth, and fifth tasks during problem solving, most of the students said that they can do, but they cannot explain what the meaningfulness is about their action. Few students can do well and explain completely. However the mental models of students in the second and fourth quadrants are close to the ones which students in the first quadrant have. So it is necessary to investigate the self-explanation of students in each quadrant through HCCS to visualize the cognitive structure in each quadrant. Also the social relationships reported by students have to be visualized through HCCS. Through these two kinds of association graphs, the weak-tie between various quadrants may be emerged, and the opportunity to implement dynamic scaffolding by recommending peer tutors may be accomplished.

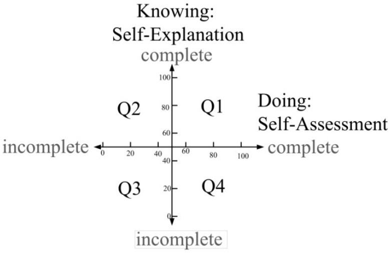


Fig. 5. Distribution of students in four quadrants

Data Analysis for Reflections and Social Relationships. In the figure 6, what actions the students have confidence to do has been selected. And teacher coded the students' self-explanation according to the meanings and concepts of every action. There are tasks about the converting mysql table to entity-relationship diagram, generating SQL create scripts for SQL Server from entity-relationship diagram, and insertion operating in the SQL server. The related declarative and procedural knowledge of every student is visualized by the HCCS and teachers may detect and recognize the weak-tie between various clusters to design the scaffolding strategy to repair the misunderstanding and flaw mental models. According to the step 1 and 2 described in section 3.2, after the data preprocess and open coding are completed, some key aspects of problem solving are recognized and used for the axial coding, that they are declarative and procedural knowledge about each tasks (step 3).

As shown in figure 6, the mental model of student S1 is almost complete about doing and knowing unless few partial misunderstanding. Students S4 and S5 are lack the knowledge about transferring entity-relationship diagram to the SQL server database schema, and also the knowledge about operations of data records for the SQL server. Through figure 7, the S1, S4 and S5 have close social distance, so the teacher may recommend S1 to S4 and S5 to support them repair their flaw mental models. S2 may be recommended to S6. Therefore while the step 4 in section 3.2 is entered, the weak-tie used to link the two clusters of students may be built. For one task, the peer scaffolding may be established by recommending the students in the first quadrant to the ones in the fourth quadrants. Through the observation and communication of similar models, students may have higher self-efficacy and perform problem solving tasks better.

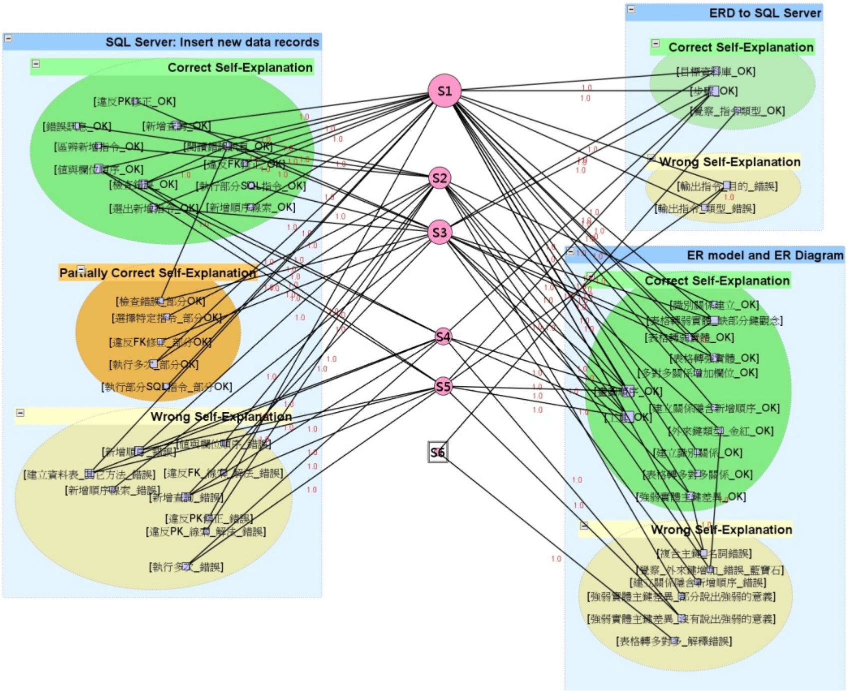


Fig. 6. The association graph of self-explanation of students for some problem solving tasks – in Q1 and Q4

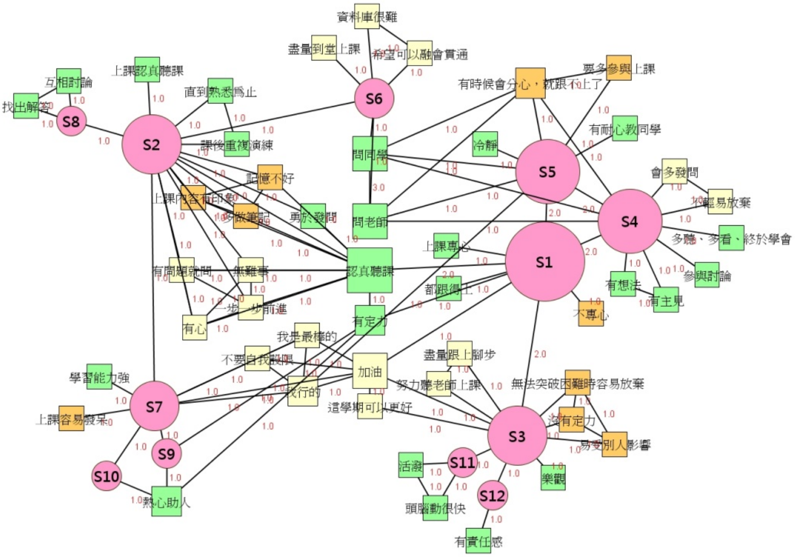


Fig. 7. The association graph of self-reported social relationships

5 Conclusions and Suggestions

The scaffoldings are of dynamic natures, and the competence of problem solving is important in educational settings. When students engage in problem solving learning, the consistency of doing and knowing is important, when doing and knowing are consistent, students may have ability to solve further application problem. But the novices often can follow and do actions, but lack the domain-specific knowledge and meta-cognitive knowledge to reason why the actions they can do, most of them feel difficult in problem solving. In this study, the HCCS and the identification of weak-tie indeed provide teachers the opportunities to identify the gaps of doing and knowing of students and the interdependent social relationships among students to implement effective peer tutoring scaffoldings. After students repair their misunderstanding and flaw mental models, the scaffoldings will be faded. The effectiveness of peer tutoring scaffolding strategies will be evaluated and the improvement of gaps between doing and knowing for problem solving learning also be assessed in the future study.

References

1. Chen, C.: Teaching problem solving and database skills that transfer. Journal of Business Research 63(2), 175–181 (2010)
2. Chen, C.H.: Promoting college students' knowledge acquisition and ill-structured problem solving: Web-based integration and procedure prompts. Computers and Education 55(1), 292–303 (2010)
3. Chi, M.T.H., de Leeuw, N., Chiu, M.-H., La Vancher, C.: Eliciting self-sxplanations improves understanding. Cognitive Science 18, 439–477 (1994)
4. Chi, M.T.H., Bassok, M., Lewis, M.W., Reimann, P., Glaser, R.: Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science 13(2), 145–182 (1989)
5. Duch, B.J., Groh, S.E., Allen, D.E.: The power of problem-based learning: A Practical "How-To" for Teaching Undergraduate Courses in Any Discipline. Stylus Publishing, Sterling (2001)
6. Hong, C.F.: Qualitative Chance Discovery – Extracting competitive advantages. Information Sciences 179, 1570–1583 (2009)
7. Kolb, D.A.: Experimental learning. Prentice-Hall, Englewood Cliffs (1984)
8. Lin, M.-H., Chen, C.-F.: Scaffolding Opportunity in Problem Solving – The Perspective of Weak-Tie. In: Nguyen, N.T., Trawiński, B., Katarzyniak, R., Jo, G.-S. (eds.) Adv. Methods for Comput. Collective Intelligence. SCI, vol. 457, pp. 71–81. Springer, Heidelberg (2013)
9. Pretz, J., Naples, A.J., Sternberg, R.J.: Recognizing, defining, and representing problems. In: Davidson, J.E., Sternberg, R.J. (eds.) The Psychology of Problem Solving. Cambridge University Press, New York (2003)
10. Rosenshine, B., Meister, C., Chapman, S.: Teaching students to generate questions: A review of the intervention studies. Review of Educational Research 66(2), 181–221 (1996)

An Insight into an Innovative Group-Trading Model for E-Markets Using the Scenario of a Travel Agent

Pen-Choug Sun and Feng-Sueng Yang

Department of Information Management, Aletheia University No. 32, Zhenli Street,
Tamsui Dist., New Taipei City, 25103, Taiwan, R.O.C.
{au1159,fsyang}@mail.au.edu.tw

Abstract. A Core Broking Model (CBM) has been proposed, which is a core-based model and uses physical brokers to resolve group-trading problems in e-markets. Some solution concepts are adopted in this model, so that three essential factors can be considered, namely incentive compatibility, distributed computing, and less computational complexity. A fees system including the commission for the brokers was suggested to the model. An example illustrates the innovative process of group-trading in the model and shows that the CBM successfully creates a profitable situation for customers, providers and brokers in e-markets.

Keywords: Brokers, Coalition, Core, E-Markets, Stability.

1 Introduction

Researches related to coalition problems in e-markets have been extensively studied in both economics and Computer Science [1, 2]. In Computer Science, coalitions are formed in a precise structure to maximise the overall expected utility of participants and formation algorithms of less computational complexity are prescribed [3]. On the other hand, the economics literature traditionally stresses the incentives of each selfish participant in a coalition [1]. The traders are self-interested to join a coalition only when it is to their own advantage [4]. A coalition with stability is when every member has the incentive to remain in the coalition. The earliest proposed concept related to stability notions was called the stable set [5]. For many years, it was the standard solution concept for cooperative games [6]. However, subsequent works by others showed that a stable set may or may not exist [7], and is usually quite difficult to find [8].

The core assigns to each cooperative game the set of profits that no coalition can improve upon [9]. It has become a well-known solution concept because the core provides a way to find a stable set and gives that set incentive compatibility. However, the core is incapable of dealing with large coalitions at least due to three problems: no stable set, high computational complexity in a large coalition and difficulty in obtaining information needed for locating a core [10]. The growth of Internet e-commerce is so rapid [11] that most companies have included e-markets into their business models. However, it is difficult to apply the core in e-markets [10].

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 582–592, 2013.

Building an online group-trading model can be a real challenge because incentive compatibility, distributed computing, and less computational complexity have to be considered at the same time [10]. A new core-based model for e-markets: Core Broking Model (CBM) has been built. Besides providers and buyers, brokers play important roles in the trading process in it. A fees system was set up for it, so that the providers and the buyers contribute the commission for the brokers. An example illustrates the process of the CBM applying to real-world e-markets. This is demonstrated through the case study in the group-trading project using the scenario of a travel agent to show the benefits of providers, customers and brokers.

2 Core Broking Model (CBM)

The CBM involves joint-selling of multiple goods in e-marketplaces, offering volume discount for group-buying coalitions in the e-marketplaces. Several providers are involved in transactions of bundle selling, while, on the other hand, many buyers form coalitions for the amount discount in the e-markets. It inherits two useful techniques from the core concept, but makes six precise improvements to have incentive compatibility, distributed computing, and less computational complexity than the core has. The CBM is composed of core brokers, projects, providers' coalitions, a Core Broking System (CBS), e-markets, market brokers and buyers' coalitions. Fig. 1 represents the structure of the CBM and gives an overview of the model. It shows that core brokers initiate projects, which involve multiple providers, on the CBS website and recruit market brokers to form a team to work on a session of group trading. The market brokers list the project on the appropriate shopping sites and form buyers' coalitions there.

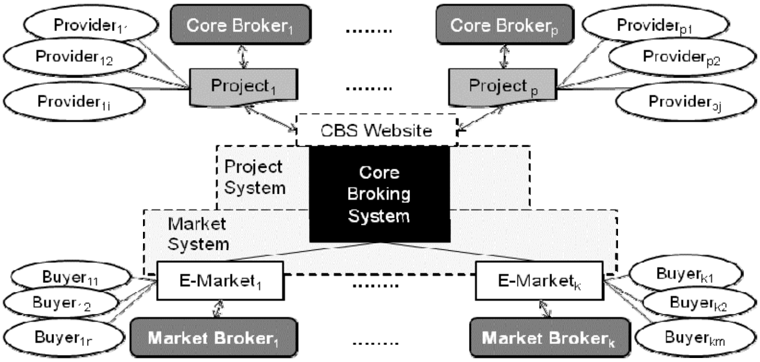


Fig. 1. The Structure of the CBM

Brokers make possible the collaboration between the members of coalitions in online group-trading. There are core-brokers and market-brokers. The core-brokers encourage providers to perform joint-selling to increase the 'competitive advantage' [12]. They act as the representatives for the coalition of providers, and a market-broker is the representative for a coalition of buyers. The core-broker is like a project manager. On the other hand, the market-brokers are like salesmen in the CBM.

The core-broker provides all the necessary information to the market-brokers for them to promote the product and market it. Commissions for online broking sites seem to be charged differently, therefore the fees for online shopping sites need to be investigated in order to set up a fees system for the new model. A survey has been made based on some popular marketplaces selected by AuctionBytes [13]. There are three types of fees which sellers are charged commonly: online store fees, insertion fees and final value fees. An online store fee is paid monthly by a seller who opens a store on the sites. An insertion fee is calculated at the time an item is listed. A final value fee is charged when an item is sold. In the CBM, a suggested session fee of £30 is paid by core-brokers every time they enter a listing for a session on a project on the site. An online store fee of £24.50 is a suggested monthly fee for market-brokers. A final value fee is 7% of the final selling value and is divided into two portions. The market-broker takes 4% and the core-broker gains 3%. A handling fee is suggested to be 10% of the extra discount, after each of the brokers has processed the orders.

3 An Illustration

Core-broker Ben created group-trading project S1: 'Summer Time around the Midlands' by integrating the products from the three providers offering inexpensive hotel rooms and low car rentals for economical travel in the Midlands. The purpose of the project is to enable sessions of bundle selling by integrating the resources of the providers. By offering wholesale discounts, customers may form groups in order to purchase items. Coupons can be chased and sent to the providers on them and exchanged into hotel rooms or car for the buyers to travel around the Midlands in the UK.

It would be unlikely to collect a real-world data of a group-trading project. A simulation system was therefore implemented and is used to generate random data needed for the illustration. The details of the products from three providers are listed in Table 1.

Table 1. Details of Products

Supplier ID	Name	Product ID	Retail Price	Stock	Cost
P1	Bob	Ca	98.29	48	25.15
		Cb	63.38	36	16.66
		Cc	47.04	42	12.56
		Ra	82.88	40	23.21
P2	Tom	Ca	96.29	40	27.62
		Cb	62.93	29	18.09
		Cc	46.64	31	11.29
		Cd	26.28	30	7.46
		Ra	87.27	38	21.39
P3	Ken	Ca	96.26	49	24.72
		Cb	68.98	34	20.70
		Cc	45.78	37	11.71
		Cd	26.89	26	6.46

Table 2. Products in Project S1

Project ID	Supplier	Product ID	Retail Price	Stock	Cost
S1	Ben	Ca	96.98	137	25.72
		Cb	65.17	99	18.47
		Cc	46.50	110	11.92
		Cd	26.56	56	7.00
		Ra	85.02	78	22.32

Table 2 shows the new contents of the project providing table. The price of each product is the averages price and so is the cost. In this way, project S1 looks to customers as if there was only one supplier, who is Ben himself. He works out the volume discount for the products he wants to include in the proposal. The project price list in Table 3 is compiled.

Table 3. Price Lists in Project S1

Project ID	Product ID	Range No	Minimum Amount	Discount
S1	Ca	1	2	5%
		2	5	10%
		3	10	20%
		4	20	30%
		5	50	40%
	Cb	1	5	5%
		2	10	10%
		3	20	20%
		4	50	30%
	Cc	1	10	5%
		2	20	10%
		3	50	20%
	Cd	1	20	5%
		2	50	10%
	Ra	1	3	5%
		2	8	10%
		3	15	18%
		4	25	25%
		5	45	35%

The system flow chart of the model is shown in Fig. 2. There are four stages in the CBM, namely commencing, gathering, combining and closing. After a project is initiated, a group trading session can be started. Each session has a starting and an ending date. The suggested duration for a session is usually one week. Ben registered himself and project S1 on the CBS site and uploaded three documents: project specs, product descriptions and price lists to the site. He then listed S1 and began a session of group-trading.

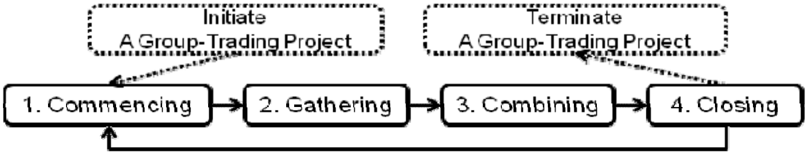


Fig. 2. The System Flow Chart of the CBM

3.1 Commencing

The Core-broker performs two tasks here: recruit market-brokers and start a session of group-trading. Ben recruits three market-brokers. After giving the information about the project to the market-brokers for their websites, he starts a session of the group-trading project. Session 01 of project S1 begins on the morning of 22 August 2013 at 9:00 and will end at 23:59 on 29/08/2013. Three market-brokers, Paul, Tim and Phil, are assigned by Ben to the project. Three ID: MK21, MK30 and MK37 have been given to them respectively. They are in charge of gathering customers locally, from their own geographical areas. The project that the core-broker has initiated now has global proportions. In Table 4, each record shows the status of the respective market-broker. The field headed 'total' is the total payment of the market order of a market-broker. The received payment, which the market-broker has paid, is stored in the *received* field.

Table 4. Trading Records

Project ID	Session	Broker ID	Order ID	Order Date	Order Time	Total	Received
S1	01	MK21	OP10121	29/08	18:42:56		
S1	01	MK30	OP10130	29/08	09:24:50		
S1	01	MK37	OP10137	29/08	13:52:43		

3.2 Gathering

There are four steps in this stage: setup websites; accept orders; produce market orders and submit market orders. The market brokers list goods for a session in the group-trading project and perform online group-buying transactions in their shopping websites.

Table 5. Original Details in Tim's Customers' Orders

Order ID	Product ID	Quantity Ordered	Expected Discount	Customer ID	Actual Discount
O081101	Cb	2	10%	C92	0%
O081101	Cc	17	20%	C92	10% (5%)
O081101	Cd	14	10%	C92	5% (0%)
O081101	Ra	11	18%	C92	10%
O081109	Cc	9	10%	C34	0%
O081109	Cd	4	10%	C34	0%
O081109	Ra	6	25%	C34	5%
O081207	Cc	5	20%	C92	10% (0%)
O081207	Cd	6	10%	C92	5% (0%)
O081211	Cb	5	30%	C108	5%
O081211	Ra	8	35%	C108	10%

Martin, Steve and John are Tim's customers and their customer IDs are C34, C92 and C108. Orders O081109 and O081211 are from C34 and C108. Steve's two orders, O081101 and O081207 have different shipping addresses. Because he places his needs in two orders, so the actual discount must reflect the total amount of every product. The data of 4 orders is randomly generated by the simulation system. Table 5 shows the details of the orders including the actual discounts for every order line. Steve orders 20 units of product Cd altogether and gets 5% discount. He expects to get 10% discount out of product Cd. His intention is shown in the expected discount field. A market-broker can only be allowed to give Ben one market order for one session.

Table 6. Details in Tim's Market Order

Order ID	Product ID	Quantity Ordered	Expected Discount	Actual Quantity	Actual Discount
OP10130	Cb	7	30%		5%
OP10130	Cc	31	20%		10%
OP10130	Cd	24	10%		5%
OP10130	Ra	25	35%		25%

Tim's market order has his ID, MB30 on it. It does not include any information from Tim's customers. This effectively protects the information of his customers. In Table 6, the field headed 'quantity ordered' stores the total quantities of the products in the orders that have been combined by the market-brokers. In Table 5, the discounts of product Ra are not greater than 10%. The quantity of Ra in Table 6 is 25. So the actual discount becomes 25%. Combining the orders is beneficial because every product in the market bigger chance to gain better discounts.

The 'date/time ordered' field stores the actual submission time of a market order. After the current coalitions have been transformed into market orders, the actual submission time will be recorded in them as the order date/time. Tim submitted his order at 9:24:50 on 29/08/2013, which is the critical time in the event of a stock shortfall. At this point, he does not have any way of knowing whether the stock is enough for all the market-brokers, so the best policy for him is to submit the market order as early as possible.

The expected discounts of product Ra are 18%, 25% and 35% originally. The expected discount for product Ra in the market order is the largest discounts, that is, 35%. The quantity of the order line in the market order will be set to zero by the core-broker, when its expected discount is higher than its actual discount. The consequences of this will be quite serious, because now nobody in the group that the market-broker has assembled will get the product. It is the market-brokers' duty to make sure that every member of their group gets the goods they want in the trading session. A warning report lists all the market orders which contain over-high expected discounts.

Table 7. Details in Tim's Customers' Orders

Order ID	Product ID	Quantity Ordered	Expected Discount	Customer ID	Actual Discount
O081101	Cb	2	5%		0%
O081101	Cc	17	10% (20%)		10%
O081101	Cd	14	5% (10%)		5%
O081101	Ra	11	18%		10%
O081109	Cc	9	10%		0%
O081109	Cd	4	5% (10%)		0%
O081109	Ra	6	25%		5%
O081207	Cc	5	10% (20%)		10%
O081207	Cd	6	5% (10%)		5%
O081211	Cb	5	5% (30%)		5%
O081211	Ra	8	25% (35%)		10%

Tim has negotiated with his customers and tried to cut down their expected discounts. Table 7 shows the order detail table from customers after Tim has fulfilled his duty. The new order detail table of Tim reveals excellent results after the negotiation. The details of the revised market order are given in Table 8.

Table 8. Details in Tim's Market Order

Order ID	Product ID	Quantity Ordered	Expected Discount	Actual Quantity	Actual Discount
OP10130	Cb	7	5%		5%
OP10130	Cc	31	10%		10%
OP10130	Cd	24	5%		5%
OP10130	Ra	25	25%		25%

3.3 Combining

The purpose of this stage is to obtain higher discounts by going through the following steps: check stability; rank orders; combine orders; calculate discounts and deliver notices. The details of the market orders from the three market-brokers are shown in Tables 8, 9 and 10, which are used to illustrate how the core-broker processes the orders, as the four steps are followed through. Ben checks the stability of the coalitions by making sure that the current coalition's total benefit is larger than the total benefit of its every subset. It turns out that the coalition, which combines the three market orders from the market-brokers, is stable, because it has the largest benefit, i.e. £6321.38. He then combines these market orders into a single large order after the orders are sorted in ascending order according to the dates on them, because the principle that the CBM uses in dealing with the orders is first come first served. The actual quantity ordered is checked against the quantity in stock to take account of any shortfall. The quantity of product Cd is 78, but the stock of product Cd is only 56. Ben may try to contact the providers and request a further supply, but in this case, Ben did not succeed. The highest 35% is then put into the expected discount field in product Ra.

Table 9. Details in Phil's Market Order

Order ID	Product ID	Quantity Ordered	Expected Discount	Actual Quantity	Actual Discount
OP10137	Cb	12	20%		10%
OP10137	Cc	15	20%		5%
OP10137	Cd	16	10%		0%

Table 10. Details in Paul's Market Order

Order ID	Product ID	Quantity Ordered	Expected Discount	Actual Quantity	Actual Discount
OP10121	Cb	10	30%		10%
OP10121	Cd	28	10%		5%
OP10121	Ra	36	35%		25%

Table 11. Possible Core

Order ID	Product ID	Quantity Ordered	Expected Discount	Actual Quantity	Actual Discount
OP101ALL	Cb	29	5%	7	5%
OP101ALL	Cc	56	10%	31	10%
OP101ALL	Cd	78	10%	78	10%
OP101ALL	Ra	61	35%	56	35%

Table 12. Final Details in Phil's Market Order

Order ID	Product ID	Quantity Ordered	Expected Discount	Actual Quantity	Actual Discount
OP10137	Cb	12	20%	0	0%
OP10137	Cc	15	20%	0	0%
OP10137	Cd	16	10%	16	10%

Table 13. Final Details in Tim's Market Order

Order ID	Product ID	Quantity Ordered	Expected Discount	Actual Quantity	Actual Discount
OP10130	Cb	7	5%	7	5%
OP10130	Cc	31	10%	31	10%
OP10130	Cd	24	5%	24	10%
OP10130	Ra	25	25%	25	35%

Table 14. Final Details in Paul's Market Order

Order ID	Product ID	Quantity Ordered	Expected Discount	Actual Quantity	Actual Discount
OP10121	Cb	10	30%	0	0%
OP10121	Cd	28	10%	16	10%
OP10121	Ra	36	35%	36	35%

The order lines of order OP101ALL need to be adjusted if their expected discount is higher than their actual discount. Table 11 is the combined order after being adjusted. Ben works on the actual discounts for each market-broker, which are in Tables 12, 13 and 14. Table 15 shows the total payments for each market-broker. By observing above four totals, one can see how important the role of a good market-broker is. Tim's coalition has bought the biggest number of items and has the highest total payment. Ben delivers the notices to the market-brokers normally via e-mail and awaits payment from them. The payment from a customer includes the handling fees and the final value fees for the market-broker and the core-broker. Ben receives payments from the market-brokers after the latter collect money from their clients. Ben has no way to contact the customers.

Table 15. Trading Data with Totals

Project ID	Session	Broker ID	Order ID	Order Date	Order Time	Total	Received
S1	01	MK37	Phil	29/08	08:42:56	382.46	0
S1	01	MK30	Tim	29/08	09:24:50	3686.65	0
S1	01	MK21	Paul	29/08	13:52:43	2372.87	0

3.4 Closing

There are three steps in this stage: prepare invoices; purchase coupons and close transactions. The market-brokers prepare invoices for the customers, when the notice has been received. According to customer Martin's order, market-broker Tim works out an invoice for him. Martin can settle the total payment of £824.40 via Tim's PayPal account.

Table 16 shows the amounts of every product to each provider which are distributed fairly by using Shapley value. According to this, Ben works out orders to providers. The amount that Ben receives for each product includes the final value fee and the handling fee. The payment for each item to the providers needs to exclude the expenses, which are the brokers' commission and the PayPal fee. In order to calculate the providers' net profits, the cost of each product is given. The money is transferred into Ben's bank accounts and coupons are sent out in exchange. The market-brokers need fill in the customers' names on the coupons and pass them on to the customers. The customers will claim the products from the providers. Finally, Ben pays the providers and closes the transactions to end the current session of group trading. The process is repeated until Ben decides to terminate the project and drop the list from the CBS website.

Table 16. Shares of Items

Product ID	Bob	Tom	Ken
Cb	3	2	2
Cc	11	10	10
Cd	0	30	26
Ra	23	22	16

Table 17. Benefits of Customers and Providers

Product ID	Discount of Tim's Customers	Discount of Phil's Customers	Discount of Paul's Customers	Provider Bob's Profit	Provider Tom's Profit	Provider Ken's Profit
Cb	0.00	22.81	0.00	117.74	75.64	70.42
Cc	0.00	144.15	0.00	277.55	265.02	260.82
Cd	42.50	63.74	42.50	0.00	423.78	393.28
Ra	0.00	744.28	1071.76	614.51	627.83	533.56
Total	**42.50**	**974.98**	**1114.25**	**1009.80**	**1392.26**	**1258.07**

Table 17 shows the benefits for the customers and providers. Both the discounts and profits have excluded the commissions for the brokers. The commissions for the core and market-brokers in this group-trading session are given in Table 18. Tim's handling fees are shown here, but in the real-world, would only be disclosed to anybody except the market-brokers and their clients. Ben has no way to know the handling fees. Although Ben seems to have more information than the other participants, the customers' orders and personal information will always remain hidden from him.

The rates of commission for the brokers are fixed throughout the illustrations. In practice, they are fixed for the duration of the session, but the handling fee from customers and the final value fee from the providers are definitely negotiable. The session fee and the online store fee are also subject to negotiation in real-world websites. Some of the expenses incurred by the brokers may be not considered in the

Table 18. Commissions for Brokers

Product ID	Market Broker Phil's Commission	Tim's Handling Fee	Market Broker Tim's Commission	Market Broker Paul's Commission	Core Broker Bens Profit
Cb	0.00	4.84	17.34	0.00	12.62
Cc	0.00	0.53	51.89	0.00	37.79
Cd	15.30	24.88	22.95	15.30	48.28
Ra	0.00	28.95	55.29	79.62	148.63
Total	**15.30**	**59.20**	**147.47**	**94.91**	**247.32**

examples, such as a £30 session fee for Ben to list the session of the group-trading project on the site. Every participant seems to get a reasonable benefit out of the group-trading session in the project.

4 Conclusion

The above section provides a deeper insight into the functioning of the CBM and to demonstrate the applicability of the model to real-world markets. The model appears to bring good benefits for all the participants in the group-trading session. The customers are meant to have better discounts there than they are in a tradition market for they obtain volume discounts by ganging up with other buyers. The providers may earn more profit from more customers attracted by high discounts. The brokers gain the commissions they deserve. The CBM seems to create a win-win-win situation for all the participants.

A comparison between the results of the CBM and a traditional market shows that the way to find a core in the new model is superior to the latter's, in terms of three criteria: the use of distributed computing, the degree of computational complexity and incentive compatibility [13]. The outputs from a simulator demonstrate that the model can attract customers and deal with online group-trading problems in a large coalition [13]. An evaluation of the techniques applied in the CBM was made showing that all of them have produced the desired results effectively and efficiently [14]. The market-brokers distribute the items here among his clients by using FCFS. They can use Shapley value instead. An agreement may also be reached amongst the customers on the distribution over the conflicting issues through a multiple stage negotiation process [15]. The other assumptions of the CBM have been listed [16] for the model to function properly. The main contribution of this research is the CBM itself, but three additional issues have emerged which also made a contribution to knowledge in this field: (a) the advantages and problems of the core (b) a stability check for a coalition and (c) the use of brokers in group-trading.

There will be two major targets for future research. One target is to create more incentives for participants. Another target will be to expand the CBM by including particular e-markets and selling a great diversity of products and services on them. To select suitable products for bundle selling would be a practice way for the model. Involving customers in the path of new products development is effective for providers to offer successful new products [17]. However, it is difficult for sellers to collaborate with buyers together [18]. The CBM involving brokers is an innovative way and it may make sellers and buyers cooperate together. This should bring out a good market result because core-brokers understand what providers can offer and market-brokers know what customers want [19].

References

1. Moulin, H.: Cooperative Microeconomics: A game theoretic Introduction. Princeton University Press, Princeton (1995)
2. Sandholm, T., et al.: Coalition structure generation with worst case guarantees. Artificial Intelligence 111, 209–238 (1999)
3. Ferguson, D., Nikolaou, C., Yemini, Y.: An Economy for Flow Control In Computer Networks. In: Proceedings of the 8th Infocom, Ottawa, Ontario, Canada, April 23-27, pp. 100–118. IEEE Computer Society Press, Los Alamitos (1989)
4. Shehory, O., Kraus, S.: Feasible Formation of coalitions among autonomous agents in non-super-additive environments. Computational Intelligence 15(3), 218–251 (1999)
5. Neumann, J., Morgenstern, O.: Theory of Games and economic Behaviour. Princeton University Press, Princeton (1944)
6. Owen, G.: Game theory. Academic Press, New York (1995)
7. Lucas, W.F.: A game with no solution. Bulletin of the American Mathematical Society 74, 237–239 (1968)
8. Lucas, W.F.: von Neumann-Morgenstern stable sets. In: Aumann, R.J., Hart, S. (eds.) Handbook of Game Theory 1, pp. 543–590. Elsevier, Amsterdam (1992)
9. Gillies, D.: Some theorems on n-person games. Unpublished PhD thesis, Princeton University (1953)
10. Sun, P., et al.: Extended Core for E-Markets. In: Isaias, P., White, B., Nunes, M.B. (eds.) Proceedings of IADIS International Conference WWW/Internet 2009, Rome, Italy, November 19-22, pp. 437–444. IADIS Press (2009)
11. Forrester Research, Inc. ‘Forrester Forecast: Double-Digit Growth For Online Retail in the US and Western Europe’, CAMBRIDGE, Mass (2010), http://www.forrester.com/ER/Press/Release/0,1769,1330,00.html (February 3, 2011)
12. Porter, M.: Competitive advantage: Creating and Sustaining Superior Performance. Free Press, New York (1985)
13. Sun, P., et al.: A Core Broking Model for E-Markets. In: Proceedings of the 9th IEEE International Conference on e-Business Engineering (ICEBE 2012), Zhejiang University, Hangzhou, China, September 9-11, pp. 78–85. IEEE Press (2012a)
14. Sun, P., et al.: Evaluations of A Core Broking Model from the Viewpoint of Online Group Trading. In: Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM 2012), Hong Kong Convention and Exhibition Centre, Hong Kong, December 10-13, pp. 1964–1968. IEEE Press (2012)
15. Chao, K., Younas, M., Godwin, N., Sun, P.: Using Automated Negotiation for Grid Services. IJWIN 13, 141–150 (2006)
16. Sun, P.: A Core Broking Model for E-Markets. Unpublished PhD thesis, Coventry University (2011)
17. Gordon, M., et al.: The Path to Developing Successful New Products. MIT Sloan Management Review Press (2009)
18. Athaide, G., Zang, J.: The Determinants of Seller-Buyer Interactions During New Product Development in Technology-Based Industrial Markets. The Journal of Product Innovation Management 28(suppl. 1), 146–158 (2011)
19. Cooper, R.G.: Predevelopment activities determine new product success. Industrial Marketing Management 17(2), 237–248 (1988)

Exploring Technology Opportunities in an Industry via Clustering Method and Association Analysis

Tzu-Fu Chiu[1], Chao-Fu Hong[2], and Yu-Ting Chiu[3]

[1] Department of Industrial Management and Enterprise Information, Aletheia University, Taiwan, R.O.C.
chiu@mail.au.edu.tw
[2] Department of Information Management, Aletheia University, Taiwan, R.O.C.
cfhong@mail.au.edu.tw
[3] Department of Information Management, National Central University, Taiwan, R.O.C.
gloria@mgt.ncu.edu.tw

Abstract. To explore the technology opportunity is essential for a company and an industry so that the company can consider the allocation of R&D investments and the industry can observe the developing directions of rare topics. Patent data contains plentiful technological information from which it is worthwhile to extract further knowledge. Therefore, a research framework for exploring the technology opportunity has been formed where clustering method is employed to generate the clusters, similarity measurement is adopted to identify the variant patents, and association analysis is used to recognize the focused rare topics. Consequently, the clusters were generated and named, the variant patents were found, the focused rare topics were recognized, and the technology opportunities for companies were discussed. Finally, the variant patents and the technology opportunities would be provided to assist the decision makers of companies and industries.

Keywords: technology opportunity, IPC-based clustering, similarity measurement, association analysis, patent data, thin-film solar cell.

1 Introduction

In order to explore the technology opportunities in an industry, it is needed to overview the homogeneity and heterogeneity of the patents in the whole dataset and to separate them into different clusters, and then to measure the similarity of an individual patent in its cluster to identify the variant patents. Based on these variant patents, it is reasonable to find the rare topics and technology opportunities in the industry. For attempting to do so, a clustering method, IPC-based clustering, is employed to generate the clusters; the similarity measurement is adopted to identify the variant patents; and the association analysis is used to figure out the focused rare topics in the industry. Meanwhile, as up to 80% of the disclosures in patents are never published in any other form [1], it would be worthwhile for researchers and practitioners to explore the possible technology opportunities upon the patent

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 593–602, 2013.

database. Therefore, a research framework will be formed to identify the variant patents, to recognize the rare topics, and to figure out the technology opportunities for the company and industry.

2 Related Work

As this study is aimed to explore the technology opportunities for companies using patent data, a research framework needs to be constructed via a consideration of clustering method, similarity measurement, and association analysis. Therefore, the related areas of this study would be technology opportunity exploration, patent data, thin-film solar cell, similarity measurement, and association analysis.

2.1 Technology Opportunity Exploration

Technology opportunity analysis (TOA) draws on bibliometric methods, augmented by expert opinion, to provide insight into specific emerging technologies [2]. TOA performs value-added data analysis, collecting bibliographic and/or patent information and digesting it to a form useful to the research or technology managers, strategic planners, or market analysts. TOA can identify the following topics: component technologies and how they relate to each other; who (companies, universities, individuals) is active in developing those technologies; where the active developers are located nationally and internationally; how technological emphases are shifting over time; and institutional strengths and weaknesses as identified by research profiles. TOA had been applied in the thin film transistor - liquid crystal display (TFT-LCD) area [3] and in the blue light-emitting diode (LED) area [4].

In this study, a research framework, based on clustering method, similarity measurement, and association analysis, will be formed for exploring the technology opportunity using patent data.

2.2 Patent Data and Thin-Film Solar Cell

A patent is similar to a general document, but includes rich and varied technical information as well as important research results [5]. Patents can be gathered from a variety of sources, such as the Intellectual Property Office in Taiwan (TIPO), the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO), and so on. A patent document contains numerous fields, such as: patent number, title, abstract, issue date, application date, application type, assignee name, international classification (IPC), US references, claims, description, etc.

Solar cell, a sort of green energy, is clean, renewable, and good for protecting our environment. It can be mainly divided into two categories (according to the light absorbing material): crystalline silicon (in a wafer form) and thin films (of other materials) [6]. A thin-film solar cell (TFSC), also called a thin-film photovoltaic cell (TFPV), is made by depositing one or more thin layers (i.e., thin film) of photovoltaic material on a substrate [7]. In 2009, the photovoltaic industry production increased by more than 50% (average for last decade: more than 40%) and reached a world-wide production volume of 11.5 GWp of photovoltaic modules, whereas the thin film segment grew faster than the overall PV market [8]. Thin film is the most potential segment with its highest production growth rate in the solar cell industry, and it would be appropriate for academic and practical researchers to contribute efforts to this technology.

2.3 IPC-Based Clustering

Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both [9]. Classes, or conceptually meaningful groups of objects that share common characteristics, play an important role in how people analyze and describe the world. Clusters are potential classes and cluster analysis is the study of techniques for automatically finding classes [9]. Moreover, an IPC (International Patent Classification) is a classification derived from the International Patent Classification System (supported by WIPO) which provides a hierarchical system of symbols for the classification of patents according to the different areas of technology to which they pertain [10].

IPC-based clustering, proposed by the authors, is a modified clustering method which utilizes the professional knowledge of the patent office examiners (implied in the IPC field) to tune the clustering mechanism and to classify the patents into a number of clusters effectively [11]. It mainly includes the following five steps: generating the IPC code groups, generating the centroids of IPC code groups, producing the initial clustering alternatives, producing the refined clustering alternatives, and selecting the optimal alternative.

2.4 Similarity Measurement

Similarity measurement is a way to measure the likeness between two objects (e.g., documents, events, behaviors, concepts, images, and so on). The methods for measuring similarity vary from distance-based measures, feature-based measures, to probabilistic measures [12]. In distance-based measures, there are Minkowski family, intersection family, inner product family, Shannon's entropy family and so on [13]. Here, the shorter the distance is, the bigger the similarity will be. Among distance-based measures, the Euclidean distance is one the most popular methods, which can be defined as in Equation 1, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are vectors with l features (i.e., x_{ik} and x_{jk}) [14].

$$dis(\mathrm{x}_i, \mathrm{x}_j) = \sqrt{\sum_{k=1}^{l} (x_{ik} - x_{jk})^2} \tag{1}$$

In this study, the similarity measurement will be used to calculate the distance between the individual patent and its centroid of cluster so as to find out the variant patents, which deviates from the mean of cluster with no less than 2 standard deviations.

2.5 Association Analysis

Association analysis is a useful method for discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of association rules or co-occurrence graphs [9]. An event map, a sort of co-occurrence graphs, is a two-dimension undirected graph, which consists of event clusters, visible events, and chances [15]. An event cluster is a group of frequent and strongly related events. The occurrence frequency of events and co-occurrence between events within an event cluster are both high. The co-occurrence between two events is measured by the Jaccard coefficient as in Equation (2), where e_i is the *i*th event in a data record (of the data set *D*). The event map is also called as an association diagram in this study.

$$Ja(e_i, \mathrm{e}_j) = \frac{Freq(e_i \cap e_j)}{Freq(e_i \cup e_j)} \tag{2}$$

In this study, the association analysis will be adopted to generate the association diagram of variant patents for rare topics identification.

3 A Research Framework for Exploring the Technology Opportunity

As this study is attempted to observe the technology opportunity in thin-film solar cell, a research framework for exploring the technology opportunity, based on IPC-based clustering, similarity measurement, and association analysis, has been developed and shown in Fig. 1. It consists of five phases: data preprocessing, cluster generation, similarity measurement, rare topic recognition, and new findings; and will be described in the following subsections.

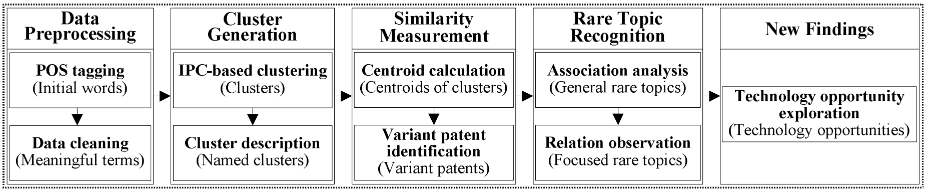


Fig. 1. A research framework for exploring the technology opportunity

3.1 Data Preprocessing

In first phase, the patent data of thin-film solar cell (during a certain period of time) will be downloaded from the USPTO [16]. For considering an essential part to represent a complex patent data, the Title, Abstract, Assignee, and Issue Date fields are selected as the objects for this study. Afterward, two processes, POS tagging and data cleaning, will be executed to clean up the source textual data.

(1) POS Tagging: An English POS tagger (i.e., a Part-Of-Speech tagger for English) from the Stanford Natural Language Processing Group [17] will be employed to perform word segmenting and labeling on the patents (i.e., the abstract field). Then, a list of proper morphological features of words needs to be decided for sifting out the initial words.

(2) Data Cleaning: Upon these initial words, files of n-grams, stop words, and synonyms will be built so as to combine relevant words into compound terms, to eliminate less meaningful words, and to aggregate synonymous words. Consequently, the meaningful terms will be obtained from this process.

3.2 Cluster Generation

Second phase is designed to conduct the cluster generation via IPC-based clustering and cluster description so as to obtain the clusters of thin-film solar cell.

(1) IPC-Based Clustering: In order to carry out the homogeneity analysis, an IPC-based clustering is adopted for separating patents into clusters. It consists of five steps: (a) to distribute patents into IPC code groups, (b) to generate centroids for every IPC code group, (c) to produce a series of initial clustering alternatives, (d) to produce a series of refined clustering alternatives, and (e) to select an optimal alternative.

(2) Cluster Description: The above clusters will be named using the description deriving from its original IPC. Each named cluster will then be utilized in the following phases.

3.3 Similarity Measurement

Third phase, including centroid calculation and variant patent identification, is used to calculate the cluster centroid, to measure the distance between each patent and its cluster centroid, and then to find out the variant patents.

(1) Centroid Calculation: A cluster centroid is calculated by average all patents of a cluster as in Equation (3) for identifying the variant patents in the next step.

$$cen(\mathrm{x}_i) = (1/l) \cdot \left(\sum_{k=1}^{l} x_{ik}\right) \tag{3}$$

(2) Variant Patent Identification: By referring to Equation (1), the distance between a patent and its cluster centroid is calculated as in Equation (4). The distance of each patent is examined by comparing it with the mean of distance of its cluster; and a variant patent will be identified if that patent deviates from the mean of cluster with no less than 2 standard deviation (SD) as in Equation (5).

$$dis(\mathrm{x}_i, \mathrm{cen}) = \sqrt{\sum_{k=1}^{l} (x_{ik} - cen_k)^2} \tag{4}$$

$$(dis_i - \overline{\mathrm{x}}) \geq 2 \cdot \mathrm{SD} \tag{5}$$

3.4 Rare Topic Recognition

Fourth phase, containing association analysis and relation observation, is used to draw the association diagram and to find out the focused rare topics.

(1) Association Analysis: An association diagram will be drawn via the variant patents, so that a number of rare topics can be generated. These rare topics will be named using the domain knowledge.

(2) Relation Observation: According to the relations between general rare topics and issue years as well as companies, the focused rare topics will be recognized. As the variant patents in recent years are more likely to be the clues of technology opportunity, the time frame is divided into three periods of time: earlier (1999 to 2002), middle (2003 to 2006), and later (2007 to 2010). Therefore, the general rare topics in the middle and later periods will be identified as the focused rare topics.

3.5 New Findings

In last phase, technology opportunity exploration will try to figure out the technology opportunities based on the clusters, variant patents, and focused rare topics.

Technology Opportunity Exploration: According to the named clusters, variant patents, and focused rare topics, the technology opportunity will be explored. The possible technology technologies will be recognized. Both the focused rare topics and possible technology technologies will be provided to facilitate the decision-making of managers and stakeholders.

4 Experimental Results and Explanation

The experiment has been implemented according to the research framework. The experimental results would be explained in the following five subsections: result of data preprocessing, result of rare patent retrieval, result of cluster analysis, result of notable rare patent recognition, and result of emerging technology recognition.

4.1 Result of Data Preprocessing

As the aim of this study was to explore the emerging technology via patent data, the patents of thin-film solar cell were the target data for the experiment. Mainly, the Title, Abstract, Assignee, and Issue Date fields were used in this study. 213 issued patents during year 1999 to 2010 were collected from USPTO, using key words: "'thin film' and ('solar cell' or 'solar cells' or 'photovoltaic cell' or 'photovoltaic cells' or 'PV cell' or 'PV cells')" on "title field or abstract field". The POS tagger was then triggered to do data preprocessing. Consequently, the patents were cleaned up and the meaningful terms were obtained.

4.2 Result of Cluster Generation

After executing the programs of IPC-based clustering, the eleven clusters of thin-film solar cell were generated. The Cluster-ID, Num. of patents (number of comprising patents), and IPC description were listed in Table 1. For example, the cluster with IPC: H01L031/18 was consisting of 78 patents and described as "Processes or apparatus specially adapted for the manufacture or treatment of these devices or of parts thereof". These clusters would be used in the following similarity measurement and association analysis.

Table 1. 11 clusters of thin-film solar cell via IPC-based clustering

No	Cluster-ID	Num. of patents	IPC description
1	H01L031/18	78	Processes or apparatus specially adapted for the manufacture or treatment of these devices or of parts thereof
2	H01L031/06	7	characterized by at least one potential-jump barrier or surface barrier
3	H01L031/00	18	Semiconductor devices sensitive to infra-red radiation, light, electromagnetic radiation of shorter wavelength, or corpuscular radiation and specially adapted either for the conversion of the energy of such radiation into electrical energy or for the control of electrical energy by such radiation; Processes or apparatus specially adapted for the manufacture or treatment thereof or of parts thereof; Details thereof
4	H01L021/00	27	Processes or apparatus specially adapted for the manufacture or treatment of semiconductor or solid state devices or of parts thereof
5	H01L027/142	14	Energy conversion devices (including semiconductor components sensitive to infra-red radiation, light, electromagnetic radiation of shorter wavelength or corpuscular radiation and specially adapted either for the conversion of the energy of such radiation into electrical energy or for the control of electrical energy by such radiation)
6	H01L031/042	7	including a panel or array of photoelectric cells, e.g. solar cells
7	H01L031/052	15	with cooling, light-reflecting or light-concentrating means
8	H01L031/0336	18	in different semiconductor regions, e.g. Cu2X/CdX hetero-junctions, X being an element of the sixth group of the Periodic System
9	H01L031/032	7	including, apart from doping materials or other impurities, only compounds not provided for in groups H01L31/0272 to H01L31/0312
10	H01L031/075	12	the potential barriers being only of the PIN type
11	H01L021/762	10	Dielectric regions (making of isolation regions between components)

4.3 Result of Similarity Measurement

Using similarity measurement, the variant patents were generated via comparing the distance of each patent with the mean of distance of its cluster. The patents with differences (after comparison) no less than 2 standard deviations were identified as variant patents. For examples, three variant patents in Cluster H01L021/00 were Patent 07252781 (= 1.7016), 07288617 (= 1.7016), and 07795452 (= 1.8775) with the differences up to more than 2 SD (= 0.2211), while SD was 0.22113 and the mean of cluster was 1.1784. There was no variant patent in Cluster H01L031/00 as the distance of each patent was approximately even (with small differences), while SD was 0.0706 and the mean of cluster was 0.9401. Graphs of Cluster H01L021/00 and Cluster H01L031/00 were shown in Fig. 2. The result of all variant patents in eleven clusters was summarized in Table 2. The notable directions which possessed the variant patents were: Cluster H01L031/18, H01L021/00, H01L027/142, H01L031/042, H01L031/075, and H01L021/762.

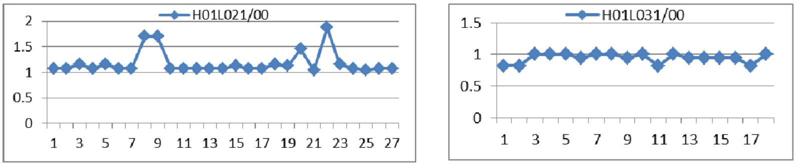


Fig. 2. Variant patents of Cluster H01L021/00 and Cluster H01L031/00

Table 2. The result of all variant patents in eleven clusters

No.	Cluster ID.	Num. of Patent	Variant Patent	Year	Company (Country)
1	H01L031/18	78	07641937	2010	In-Solar Tech Co., Ltd. (KR)
			07811633	2010	In-Solar Tech Co., Ltd. (KR)
2	H01L031/06	7	--	--	--
3	H01L031/00	18	--	--	--
4	H01L021/00	27	07252781	2007	Merck Patent GmbH (DE)
			07288617	2007	Merck Patent GmbH (DE)
			07795452	2010	Nissan C. Industries, Ltd. (JP)
5	H01L027/142	14	06013870	2000	Angewandte S.-ASE GmbH (DE)
			06274804	2001	Angewandte S.-ASE GmbH (DE)
6	H01L031/042	7	06537845	2003	University of Delaware (US)
7	H01L031/052	15	--	--	--
8	H01L031/0336	18	--	--	--
9	H01L031/032	7	--	--	--
10	H01L031/075	12	06124545	2000	Angewandte S.-ASE GmbH (DE)
			06133061	2000	Canon Kabushiki Kaisha (JP)
			07122736	2006	Midwest Research Institute (US)
11	H01L021/762	10	06133112	2000	Canon Kabushiki Kaisha (JP)
			06534383	2003	Canon Kabushiki Kaisha (JP)

4.4 Result of Rare Topic Recognition

Through association analysis, an association diagram of 13 variant patents was drawn using the meaningful terms of Abstract field of the variant patents as in Fig. 3. In the diagram, 9 general rare topics were identified and named according to the domain knowledge, including light-absorbing-layer, vapor-deposition, organic-laser-diode, liquid-coating-composition, SOI-substrate, film-deposition, SiGe-deposition, hot-wire-method, and electrode-layer.

By inserting the Issue Year and Assignee fields into the association diagram, nine companies were linked to the rare topics, such as In-Solar Tech Co., Ltd (KR), Nissan Chemical Industries, Ltd (JP), Merck Patent GmbH (DE), and so on, also shown in Fig. 3.

In order to recognize the focused rare topics, the relations between general rare topics and issue years as well as companies were observed. The relation criteria were that: the one with issue years in the later period of time (2007~2010) and with two companies was the most focused rare topic; the one with issue years in the later period of time (2007~2010) and with one companies was the more focused rare topic; and the one with issue years in the middle period of time (2003~2006) and with one company was the common focused rare topic

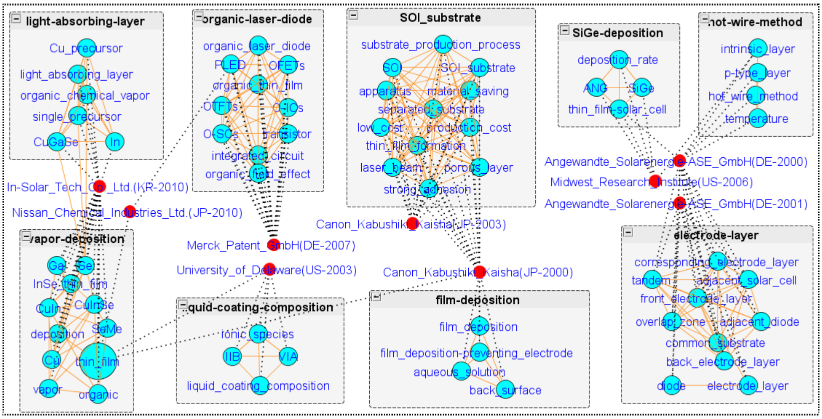


Fig. 3. An association diagram with rare topics and linked companies

4.5 Result of Emerging Technology Recognition

According to the association diagram with rare topics, linkages of issue years and companies, as well as relation criteria, the focused rare topics were recognized: the most focused rare topic were vapor-deposition and organic-laser-diode; the more focused rare topic was light-absorbing-layer; the common focused rare topic were liquid-coating-composition, SOI-substrate, and SiGe-deposition. The more possible technology opportunities were as follows.

(1) Vapor-deposition topic: It linked to In-Solar Tech Co., Ltd (KR) and Nissan Chemical Industries (JP); and to year 2010.
(2) Organic-laser-diode topic: It linked to Nissan Chemical Industries (JP) and Merck Patent GmbH (DE); and to year2007 and 2010.
(3) Light-absorbing-layer topic: It linked to In-Solar Tech Co., Ltd (KR); and to year 2010.

5 Conclusions

The research framework for technology opportunity exploration has been formed and applied to thin-film solar cell using patent data. The experiment was performed and the experimental results were obtained. Eleven clusters of thin-film solar cell during 1999 to 2010 were found via IPC-based clustering. Thirteen variant patents were identified through similarity measurement. Nine rare topics were also identified through similarity measurement. Among 9 rare topics, 6 focused rare topics were recognized via relation criteria. The more possible technology opportunities were vapor-deposition, organic-laser-diode, and light-absorbing-layer. The variant patents and the possible technology opportunities on thin-film solar cell would be helpful for managers and stakeholders to facilitate their decision-making.

In the future work, the research framework may be joined by some other methods such as co-authorship analysis or citation analysis so as to enhance the validity of experimental results. In addition, the data source can be expanded from USPTO to WIPO or TIPO in order to explore the technology opportunity on thin-film solar cell widely.

Acknowledgments. This research was supported by the National Science Council of the Republic of China under the Grants NSC 99-2410-H-156-014.

References

1. Blackman, M.: Provision of Patent Information: A National Patent Office Perspective. World Patent Information 17(2), 115–123 (1995)
2. Porter, A.L., Detampel, M.J.: Technology opportunities analysis. Technological Forecasting and Social Change 49(3), 237–255 (1995)
3. Yoon, B.G., Park, Y.T.: A systematic approach for identifying technology opportunities: keyword-based morphology analysis. Technological Forecasting and Social Change 72(2), 145–160 (2005)
4. Li, X., Wang, J., Huang, L., Li, J., Li, J.: Empirical research on the technology opportunities analysis based on morphology analysis and conjoint analysis. Foresight 12(2), 66–76 (2010)
5. Tseng, Y., Lin, C., Lin, Y.: Text Mining Techniques for Patent Analysis. Information Processing and Management 43, 1216–1247 (2007)
6. Solarbuzz, Solar Cell Technologies (2010), `http://www.solarbuzz.com/technologies.htm`
7. Wikipedia, Thin film solar cell (2010), `http://en.wikipedia.org/wiki/Thin_film_solar_cell`
8. Jager-Waldau, A.: PV Status Report 2010: Research, Solar Cell Production and Market Implementation of Photovoltaics. JRC Scientific and Technical Reports (2010)
9. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison Wesley, Boston (2006)
10. WIPO: Preface to the International Patent Classification (IPC) (October 30, 2010), `http://www.wipo.int/classifications/ipc/en/general/preface.html`
11. Chiu, T.-F., Hong, C.-F., Chiu, Y.-T.: A Proposed IPC-based Clustering and Applied to Technology Strategy Formulation. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part II. LNCS, vol. 7197, pp. 62–72. Springer, Heidelberg (2012)
12. Scholarpedia, Similarity measures (2013), `http://www.scholarpedia.org/article/Similarity_measures`
13. Cha, S.H.: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of Mathematical Models and Methods in Applied Sciences 1(4), 300–307 (2007)
14. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press (2007)
15. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-Occurrence Graph Based on Building Construction Metaphor. In: Proceedings of the Advanced Digital Library Conference (IEEE ADL 1998), pp. 12–18 (1998)
16. USPTO: the United States Patent and Trademark Office (2010), `http://www.uspto.gov/`
17. Stanford Natural Language Processing Group, Stanford Log-linear Part-Of-Speech Tagger (2009), `http://nlp.stanford.edu/software/tagger.shtml`

Hidden Topic Models for Multi-label Review Classification: An Experimental Study

Thi-Ngan Pham[1,2], Thi-Thom Phan[1], Phuoc-Thao Nguyen[1], and Quang-Thuy Ha[1,*]

[1] Vietnam National University, Hanoi (VNU), College of Technology (UET),
144, Xuan Thuy, Cau Giay, Hanoi, Vietnam
[2] The Vietnamese People's Police Academy, Hanoi, Vietnam
{nganpt.di12,thompt_53,thaonp_53,thuyhq}@vnu.edu.vn

Abstract. In recent years, Multi-Label Classification (MLC) has become an important task in the field of Supervised Learning. The MLC tasks are omnipresent in real-world problems in which an instance could belong to different classes simultaneously. In this paper, we present a method for MLC using the hidden topic method to enrich data features and using mutual information for feature selection. Our experiments on classifying user reviews about one thousand Vietnamese hotels showed the efficiency of the proposed approach.

Keywords: feature selection, hidden topic model/LDA, multi-label classification (MLC), mutual information, opinion mining/sentiment analysis.

1 Introduction

Multi-Label Classification (MLC) is the task of assigning each of the given data to a set of predefined classes or categories in which an instance can belong simultaneously to different classes [14, 20]. The MLCs are omnipresent in real-world problems. For example, a document can belong to more than one category in the text categorization [9], an image may belong to multiple semantic categories in the image classification [18]. Other domains for which multi-label classifications have proved useful also include protein function classifications [6] and classifications of music into emotion [13]. A recent survey on the progress of multi-label classification and its uses in different applications can be found in [10, 12, 20].

The popular approach to multi-label classifications bases on transforming the problem into one or more single-label classifications [11]. The most simple transformation method is the binary relevance (BR) which consists of different classifiers for different labels. In other words, the original problem is transformed into n two-class single-label classifications, where n is the number of possible labels. One of the major drawbacks of BR is it may exclude the dependence between labels.

[*] Corresponding author.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 603–611, 2013.

Label powerset (LP) is a different transformation method which does consider this dependence. This simple method works as follows: It considers each unique set of labels that exists in a multi-label training set as one of the classes of a new single-label classification task. The number of classes created this way being potentially huge. Some multi-label dimensionality reduction methods had been proposed such as the MDDM (Multi-label Dimensionality reduction via Dependence Maximization) method in [19].

In this paper, we present an experimental study on the MLC models implementing on the domain of user reviews about Vietnamese hotels. We considered reviews on five aspects: Location and Price, Staff Service, Hotel Condition/Cleanliness, Room Comfort/Standard and Food/Dining. Furthermore, a user review may orientate to positive or negative or even both for each above aspect. The problem is to determine all of labels for each user review.

Because user reviews are short documents we proposed a solution of using the hidden topic model – LDA to enrich features for classification. We also used mutual information (MI) to achieve feature selection in multi-label classification problem. We performed several experiments on Vietnamese reviews of about one thousand Vietnamese hotels to evaluate the solution.

The rest of this article is organized as following. The proposed model is described in section 2. This section will give more details about the process of multi-label classification using the combination of LDA and MI in enriching and selecting features. Experiments and results are presented in Section 3. Recent studies related to our work are introduced in Section 4. Conclusions are showed in the last section.

2 A Hidden Topic Model for Multi-label Review Classification

Our proposed model for Multi-Label Classification bases on using the Hidden Topic Probability Model is described in Figure 1. Sample dataset for training and evaluating the multi-label classifier is crawled from one thousand websites containing hotel reviews in Vietnamese.

In the preprocessing step, abstract features are added by using the hidden topic probability model. After that, a feature selection method based on MI is applied to improve the features for the classifier.

In the last step, a multi-label classifier is built based on a binary transformation method. The classifier will be applied for new reviews.

2.1 Determining the Hidden Topic Probability Model

Because user reviews are short documents we have used a hidden topic model to enrich features for classification. The Hidden Topic Probability Model has been determined by the same process described in [16] but different in the sources for the universal dataset.

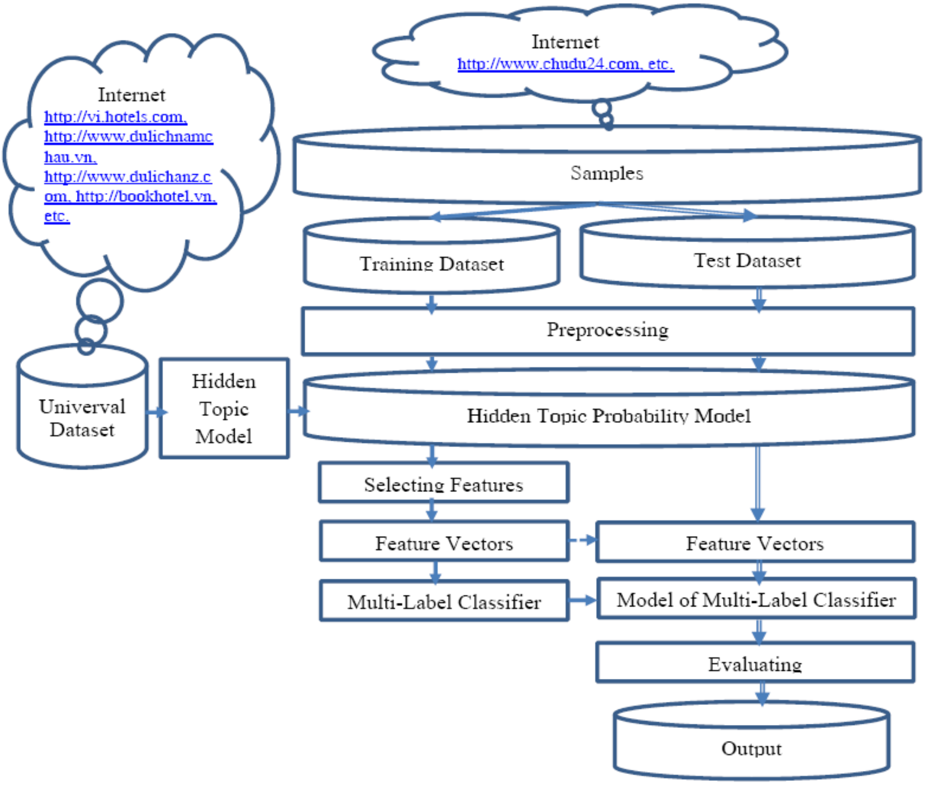

Internet
http://www.chudu24.com, etc.
Internet
http://vi.hotels.com, http://www.dulichnamchau.vn, http://www.dulichanz.com, http://bookhotel.vn, etc.
Samples
Training Dataset
Test Dataset
Preprocessing
Univerval Dataset
Hidden Topic Model
Hidden Topic Probability Model
Selecting Features
Feature Vectors
Feature Vectors
Multi-Label Classifier
Model of Multi-Label Classifier
Evaluating
Output

Fig. 1. A Hidden Topic Model for Multi-Label Review Classification

Most of webpages from Vietnamese websites on tourism and hotels such as: http://vi.hotels.com, http://www.dulichnamchau.vn, http://www.dulichanz.com, http://bookhotel.vn, http://www.dulichvtv.com, http://chudu24.com and so on have been crawled. After cleaning the data from the Internet, the universal dataset with more than twenty two thousand documents have been determined.

Then, we applied a method similar to the method described in [16] using the GibbsLDA++ tool [22]. Because of only five aspects of hotel were considered, the number of hidden topics in the model should be small. That is why three hidden topic probability models with 15, 20, and 25 hidden topics have been built and evaluated.

2.2 The Preprocessing

The preprocessing has been implemented on reviews crawled from the website containing hotel reviews in Vietnamese (http://chudu24.com). Three sub-steps have been done in the Preprocessing step.

In the Data Standardization sub-step, Vietnamese words in the misspelling form have been switched into the corresponding valid spelling form, such as "**than thien**" or "**thân thịên**" has been switched into "**thân thiện**" (friendly).

In the Tokenized Segmentation sub-step, all the review sentences have been segmented, such as the sentence **"Nhân viên phục vụ ở đây rất thân thiện."** (The waiters are very friendly) has been segmented into **Nhân_viên | phục_vụ | ở_đây | rất | thân_thiện.**

In the representation sub-step, the TF.IDF feature vector *(tfidf(d,1), tfidf(d,2), …., tfidf(d,n))* for each review ***d*** has been determined, where the dimension ***n*** be the number of separate keyword in the domain vocabulary and *tfidf(d,i)* be the value TF.IDF of the review ***d*** on the keyword ***i*** in the domain vocabulary.

2.3 Using the Hidden Topic Probability Model

The TF.IDF feature vector should be extended by adding hidden topic features determined by using the hidden topic probability models. We have applied the GibbsLDA++ tool [22] to determine the probabilities for all of the reviews. Table 1 illustrates the relationship between the hidden topic probabilities and the content of a review.

Table 1. The Hidden Topic Probabilities for a review

Topic	Probability	Representative words in the review
1	0.924	Nhân viên/staff, phục vụ/service, nhiệt tình/fervour, thân thiện/friendly,…
2	0.001	Ngon/delicous, món/food, nhiều/much, …
3	0.002	Đẹp/beauty, rộng/commodious, thoáng/well-aired,…
4	0.065	Gần/near, trung tâm/central, ví trí/location, đường/road, ….
….	…	…..

Denote *p(d,j)* be the probability that a review ***d*** belongs to the hidden topic ***j*** (*j=1...k*; *k* be the number of hidden topics). The vector (*p(d,1), p(d,2), …., p(d,k)*) is called the hidden topic feature vector of the review.

The feature vector of each review ***d*** is the combination between its TF.IDF feature vector and its hidden topic feature vector *(tfidf(d,1), tfidf(d,2), …., tfidf(d,n), p(d,1), p(d,2), ..., p(d,k))*.

2.4 Selecting Features Based on the Mutual Information

The feature vector of reviews should be optimized by using some feature selection methods. In case of multi-label classification tasks, feature selection methods based on using mutual information (MI) are useful.

In this work, the multi-label classification task is transformed into multi-class single-label classification tasks. For each single-label classification task, a forward/backward feature selection algorithm based on mutual information is employed to choose the "optimal" feature subset. In the forward search strategy, it begins with an empty set of features and first selects the feature whose MI with the class vector is the highest. Then, sequentially, the algorithm selects the feature not yet selected whose addition to the current sub-set of selected features lead to the set having the highest MI with the output. Other search strategy could also be considered such as backward elimination, which starts with the set of all features and recursively removes them one at a time. The procedure can be ended when a predefined number of features have been chosen. The feature set for the multi-label classification is the combination of the "optimal" feature subsets.

2.5 The Model of Multi-label Classifier

In our model, we have built two levels of classifiers [9, 10]. In the first level, five classifiers have been built to label a review to one or more of five hotel aspects including Location and Price, Staff Service, Hotel Condition/Cleanliness, Room Comfort/Standard and Food/Dining mentioned above. For each aspect, the review would be assigned to the positive label or/and the negative label by two second level classifiers. In total, our Multi-label classification model includes five first-level classifiers and ten second-level classifiers.

For example, the review "**Phòng sạch sẽ rộng rãi, cung cách phục vụ cũng tốt. Tuy nhiên, tôi không vui với cách anh phục vụ đưa khăn ở ngoài bãi biển**" ("Room is clean, the service is good. However, I am not happy with the way a waiter put out the towel in beach") will be recognized as Room Comfort/Standard: Positive ("Room is clean"), Staff Service: Positive ("the service is good"), and Staff Service: Negative ("I am not happy with the way a waiter put out the towel in beach").

3 Experiments and Results

3.1 The Datasets

We have taken experiments on the set of customers' comments about one thousand hotels in Vietnam. About 3700 sentences have been retrieved from website http://www.chudu24.com – a famous website about Vietnam tourism, in which training data set was about 3200 sentences and testing data set was about 500 sentences.

We have also collected data of the articles, introductions, comments about hotels in Vietnam. These data have been the input for the process of LDA hidden topic model.

3.2 Experiments

The process of experiment is described as follows.

- Processing data: Preprocessing data, creating learning data for the classification model, creating data for the LDA model and converting data into vectors.
- Creating function to select features: Using method of MI to select feature set.
- Building classification function: Using method of transformation binary classifier.
- Evaluating reputation of 1000 hotels using the most optimal model.

In order to evaluate the effect of the solution using the hidden topic model and feature selection, three experiments have been done:

- Experiment 1 (denoted by the TF.IDF case): Classifying with TF.IDF features only (the baseline case);
- Experiment 2: Classifying with the combination of TF.IDF features and the hidden topic features. We considered three cases of LDA with 15 hidden topics (denoted by TF.IDF + LDA_15 topics case), LDA with 20 hidden topics (denoted by TF.IDF + LDA_20 topics case), and LDA with 25 hidden topics (denoted by TF.IDF + LDA_25 topics case);
- Experiment 3 (denoted by the TF.IDF + LDA_20 topics + Feature Selection case): Classifying with the combination of TF.IDF features and the hidden topic features (TF.IDF + LDA_20 topics case) and using feature selection based on MI.

A 5-fold cross-validation based on the Precision, the Recall and the F1 has been applied. The Precision indicates the percentage of system positives that are true instances, the Recall indicates the percentage of true instances that the system retrieves, and the F1 is a combination of the two measures as follows:

$$F1 = \frac{2 \times Pr\,ecision \times Re\,call}{Pr\,ecision + Re\,call}$$

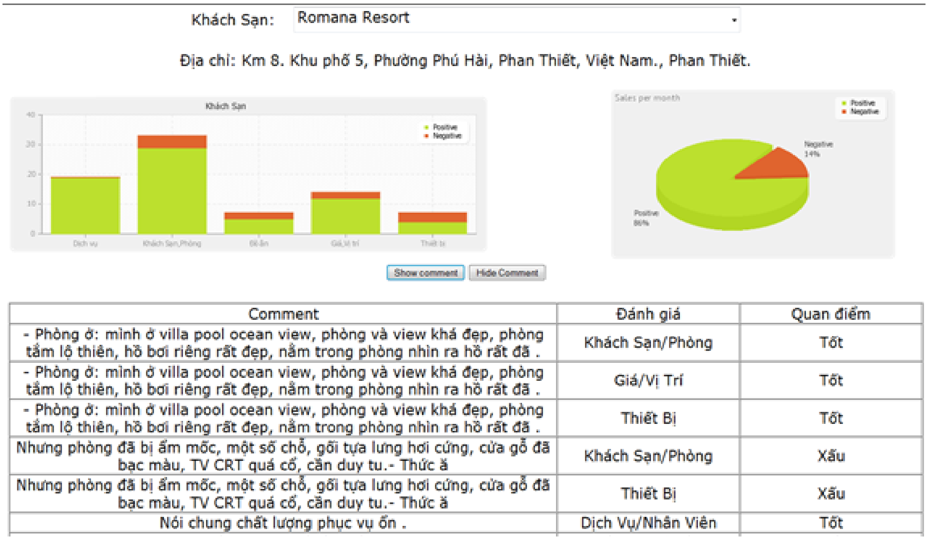
The results of the experiments are described in the Table 2. The experiments show that solutions to enrich features based on hidden topic probability model and to select features based on MI give 1% improvement in F1 to the performance of the MLC.

Table 2. The results of the experiments

Average of 5-folds valuation	Precision	Recall	F1
TF.IDF	0.6764	0.7025	0.6804
TF.IDF + LDA_15 topics	0.6798	0.7056	0.6842
TF.IDF + LDA_20 topics	0.6827	0.7125	0.6883
TF.IDF + LDA_25 topics	0.6793	0.7075	0.6844
TF.IDF + LDA_20 topics + Feature Selection	0.6835	0.7108	0.6890

Figure 2 shows a part of the experimental results on reviews about the Romana Resort located in Phan Thiet province. In our system, there are two options to display the results. In the default case (Show Comment), the summarization with users' comments is displayed and in the other case (Hide Comment), only the summarization is displayed. Figure 2 shows the results in the default case, in which only five first comments were displayed above and the remaining can be seen by using the scroll button.

According to customers' reviews, the Romana Resort is a good hotel. In all of five aspects, the number of positive opinions is greater than the number of negative opinions, especially on the Staff Service (19 positive, 0 negative) and Room Comfort/Standard (30 positive, 4 negative).



Khách Sạn: Romana Resort

Địa chỉ: Km 8. Khu phố 5, Phường Phú Hải, Phan Thiết, Việt Nam., Phan Thiết.

Comment	Đánh giá	Quan điểm
- Phòng ở: mình ở villa pool ocean view, phòng và view khá đẹp, phòng tắm lộ thiên, hồ bơi riêng rất đẹp, nằm trong phòng nhìn ra hồ rất đã .	Khách Sạn/Phòng	Tốt
- Phòng ở: mình ở villa pool ocean view, phòng và view khá đẹp, phòng tắm lộ thiên, hồ bơi riêng rất đẹp, nằm trong phòng nhìn ra hồ rất đã .	Giá/Vị Trí	Tốt
- Phòng ở: mình ở villa pool ocean view, phòng và view khá đẹp, phòng tắm lộ thiên, hồ bơi riêng rất đẹp, nằm trong phòng nhìn ra hồ rất đã .	Thiết Bị	Tốt
Nhưng phòng đã bị ẩm mốc, một số chỗ, gối tựa lưng hơi cứng, cửa gỗ đã bạc màu, TV CRT quá cổ, cần duy tu.- Thức ă	Khách Sạn/Phòng	Xấu
Nhưng phòng đã bị ẩm mốc, một số chỗ, gối tựa lưng hơi cứng, cửa gỗ đã bạc màu, TV CRT quá cổ, cần duy tu.- Thức ă	Thiết Bị	Xấu
Nói chung chất lượng phục vụ ổn .	Dịch Vụ/Nhân Viên	Tốt

Fig. 2. Sentiment Analysis on Romana Resort's reviews

We consider the review: "**Phòng ở: mình ở villa pool ocean view, phòng và view khá đẹp, phòng tắm lộ thiên, hồ bơi riêng rất đẹp, nằm trong phòng nhìn ra hồ rất đã**" ("**Room facilities: I was in the villa-pool ocean view, the room and the view are quite good, with an out-door bathroom, and beautiful private pool. Laying inside room, I can experience amazing lake-view**") which is showed in the three top rows in the table of Figure 2. In the table, the review is recognized into three aspects of Room Comfort/Standard (Row 1), Location and Price (Row 2), and Hotel Condition/Cleanliness (Row 3). The review also gives the positive opinion (**Tốt/Good**) in each of the three aspects.

4 Related Works

This work shows an experimental study on the MLC models on Vietnamese reviews about hotels. There are a lot works on the field of MLC such as [6, 9-14, 18, 20]. In our MLC model, we have applied the simple solution of using binary transformation method [9, 10] to build a multi-label classifier. This approach may lead to the inefficiency of our model as expectation. We will consider to apply another multi-instance multi-label learning [20] for improving our model.

Hidden topic probability models [1, 2] are useful to improve the semantic meaning of the text representations, such as [1, 8, 16]. Daniel Ramage el al [8] proposed Labeled LDA model to learn user word-tag correspondences directly. A multi-label text classifier based on their Labeled LDA model was implemented. Experiments showed the Macro-average of F1-Score reach about 0.40 and the Micro-average of F1-Score reach about 0.52. Xuan-Hieu Phan et al [16] showed a well-done work to apply hidden topic probability models on Vietnamese short documents (ad, search snippets and so on) with very large scale. In this work, we have focused on the domain of Vietnamese short documents – user reviews on hotels however the universal dataset was still small. Some characteristics of this domain should be more investigated.

By Manoranjan Dash and Huan Liu [3] as well as Gauthier Doquire and Michel Verleysen [5], feature selection has been much considered since 1970s and much work has been done. Hector Menendez et al [21] presented a straightforward strategy to reduce the dimension of the attributes orienting to clustering analysis. Many works are concerned with feature selection based on the mutual information for the multi-label classification problem such as [4, 5, 7, 15, 17, 19]. In our work, a greedy feature selection procedure based on multidimensional mutual information has been conducted. Gauthier Doquire and Michel Verleysen [4] proposed a two-aspect solution for feature selection in multi-label classification based on mutual information. On the aspect of transforming the multi-label problem into a single-label one, a pruned problem transformation method has been applied. By experiments, authors commented that the interest of their approach over the method based on the $\chi 2$ statistic. In our work, the combinative features by adding the hidden topic probability features with the TF.IDF features had been improved based on selecting the good feature set for multi-label classifiers. This is the first time the approach of combining the hidden topic model LDA to enrich features and the method of feature selection based on mutual information to select optimal features for the classification has been done.

5 Conclusion

This paper shows an experimental study on the MLC models on Vietnamese reviews. There are some solutions to improve the features for MLC. Firstly, the feature set has been enriched by adding the hidden topic features. Secondly, the combination feature set has been improved by using a feature selection method based on mutual information. Some experiments have been implemented and gave 1% improvement in the performance of the MLC. The "universal dataset" in domain of user reviews about Vietnamese hotels with small size may be not enough for an effective hidden topic probability model.

The work should be upgraded by some skillful solutions. Firstly, the universal dataset for hidden topic model should be extended. Secondly, the method to select features for MLC should be modified. Lastly, advanced solutions for building multi-label classifiers should be considered.

Acknowledgments. This work was supported in part by Vietnamese MOET Grant B2012-01-24 and VNU-UET Grant CN-12.01.

References

1. Blei, D.M.: Probabilistic topic models. Commun. ACM (CACM) 55(4), 77–84 (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research (JMLR) 3, 993–1022 (2003)
3. Dash, M., Liu, H.: Feature Selection for Classification. Intell. Data Anal (IDA) 1(1-4), 131–156 (1997)

4. Doquire, G., Verleysen, M.: Feature Selection for Multi-label Classification Problems. In: Cabestany, J., Rojas, I., Joya, G. (eds.) IWANN 2011, Part I. LNCS, vol. 6691, pp. 9–16. Springer, Heidelberg (2011)
5. Doquire, G., Verleysen, M.: A Comparison of Multivariate Mutual Information Estimators for Feature Selection. In: ICPRAM 2012, pp. 176–185 (2012)
6. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: NIPS 2001, pp. 681–687 (2001)
7. Novovicová, J., Malík, A., Pudil, P.: Feature Selection Using Improved Mutual Information for Text Classification. In: SSPR/SPR 2004, pp. 1010–1017 (2004)
8. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: EMNLP 2009, pp. 248–256 (2009)
9. Rousu, J., Saunders, C., Szedmák, S., Shawe-Taylor, J.: Kernel-Based Learning of Hierarchical Multilabel Classification Models. Journal of Machine Learning Research 7, 1601–1626 (2006)
10. Silla Jr., C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov. (DATAMINE) 22(1-2), 31–72 (2011)
11. Tsoumakas, G., Katakis, I.: Multi-Label Classification: An Overview. IJDWM (JDWM) 3(3), 1–13 (2007)
12. Tsoumakas, G., Katakis, I., Vlahavas, I.P.: Mining Multi-label Data. In: Data Mining and Knowledge Discovery Handbook 2010, pp. 667–685 (2010)
13. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-Label Classification of Music into Emotions. In: ISMIR 2008, pp. 325–330 (2008)
14. Tsoumakas, G., Zhang, M.-L., Zhou, Z.-H.: Introduction to the special issue on learning from multi-label data. Machine Learning 88(1-2), 1–4 (2012)
15. Gómez-Verdejo, V., Verleysen, M., Fleury, J.: Information-Theoretic Feature Selection for the Classification of Hysteresis Curves. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 522–529. Springer, Heidelberg (2007)
16. Phan, X.-H., Nguyen, C.-T., Le, D.-T., Nguyen, L.-M., Horiguchi, S., Ha, Q.-T.: Classification and Contextual Match on the Web with Hidden Topics from Large Data Collections. IEEE Transactions on Knowledge and Data Engineering 23(7), 961–976 (2011)
17. Zhang, M.-L., Pena, J.M., Robles, V.: Feature selection for multi-label naive Bayes classification. Information Sciences 179, 3218–3229 (2009)
18. Zhang, M.-L., Zhou, Z.-H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition (PR) 40(7), 2038–2048 (2007)
19. Zhang, Y., Zhou, Z.-H.: Multilabel dimensionality reduction via dependence maximization. TKDD 4(3) (2010)
20. Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., Li, Y.-F.: Multi-instance multi-label learning. Artif. Intell. 176(1), 2291–2320 (2012)
21. Menendez, H., Bello-Orgaz, G., Camacho, D.: Features selection from high-dimensional web data using clustering analysis. In: WIMS 2012, 20 (9 pages) (2012)
22. Phan, X.-H., Nguyen, C.-T.: GibbsLDA++ (2007), `http://gibbslda.sourceforge.net/`

Influence of Twitter Activity on College Classes

Ilkyu Ha, Jason J. Jung, and Chonggun Kim*

Department of Computer Engineering, Yeungnam University, Rep. of Korea
ilkyuha@ynu.ac.kr, j2jung@gmail.com, cgkim@yu.ac.kr

Abstract. With the rapid adoption of smartphones, Twitter has become a major mobile SNS. In spite of the increasing use of Twitter services by students and instructors in colleges, very little empirical studies that concern the impact of social media on the student learning environment have been published. In this paper, a Twitter utilization experiment is conducted for college classes to understand the effects on college student engagement and grades. The participants of experimental study comprise students of 2 college classes during one semester and only voluntary students attend this Twitter-based experiment. Network of Twitter using pattern such as community building, collaboration relationship and information sharing among students are analyzed. Some results about the relationships of seat positions, following patterns and grades among students have been studied. Through understanding of the student's netwoking pattern, we find that the possibility of Twitter as a mobile educational tool.

Keywords: Twitter, tweet, college class, Twitter activity, education.

1 Introduction

Twitter, a popular mobile microblogging tool, had been widely used to obtain information and communicate with peers. With the increasing use of SNS(Social Networking Service) in campus, some pioneering studies using Twitter in college have been conducted [1-2]. Although Facebook is the most popular social networking method for global college students for mutual understanding[3], Twitter is more amenable for public and fast diffusion dialogue than Facebook because Twitter is a mobile microblogging platform [3-4].

In spite of the increasing use of Twitter services by students and instructors in colleges, very little empirical studies that concern the impact of social media on the student learning environment have been published. In this paper, we conduct a Twitter utilization experiment for students of 2 classes at YU(Yeungnam University). Some meaningful data are collected and some interesting results are analyzed.

2 Related Works

2.1 Twitter Studies Related with College Education

Little research has focused on the relationship between social media and student engagement in higher education institutes. For example, Junco et al. [1] present the

[*] Corresponding author.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 612–621, 2013.

effect of Twitter on college student engagement and grades through a semester-long experimental study. It provides experimental evidence that Twitter can be used as an educational tool to help engage students and to mobilize faculty toward a more active and participatory role. Grosseck and Holotescu [2] show that Twitter proved to be an effective tool for professional development and for collaboration with students. Twitter can change the studying form of the college courses and used as a tool for a good pedagogy. Ebner et al. [3] show that microblogging can help students to be partially and virtually present and to be part of a murmuring community, the members of which can work on a specific problem without any restrictions of time and place. Stepanyan et al. [5] summarize the analyses of participant interaction within the Twitter microblogging environment and show that the higher scoring participants have more followers and follow others more. Several other studies [6-10] show that the use of Twitter helps to enhance and facilitate student engagement in the educational environment as a pedagogical tool. Four other studies [11, 12, 14 and 17] show that the use of Twitter in higher education aids students learning and Twitter has a potential to introduce the learning community into higher education. Thoms [13] and Thoms [16] show that Twitter can be integrated into a Course Management System (CMS) in an online course community. Ullrich et al. [15] analyze student interaction patterns and trends of network dynamics in the Twitter microblogging environment.

2.2 Research Questions and Purpose of the Study

The following questions are defined in order to attain some analytic results through this experimental study.

1. Is Twitter a helpful tool for education?
2. Is there any relationship between the networking of students based on Twitter and their grade? What are the relationships such as seat position and following relationship, seat position and grade, information sharing pattern and grade.
3. What are the characteristics of influential students in Twitter using group.

3 The Object and Methods

3.1 Experimental Group

As experimental groups, we select 2 college classes and conduct the experiment for students in these classes in autumn semester 2012. One is data communication(DC) sophomore class of the Dept. of Computer Engineering and another is computer programming(CP) freshman class of Electronic Engineering of YU.

3.2 Experiment Preparation(Pre-Survey)

Prior to experiment, we conduct a pre-survey to investigate the current using state of the students with a questionnaire composed of 12 questions. The result of the pre-survey is shown in Table 1. Courses DC and CP has 58 and 47 students, respectively. The student's age ranges from 19 to 27 years. Although all of the students use SNS services, the ratio of the student who use Twitter service is very small. Table 1 shows reasons why Twitter is not the favorite SNS. Facebook is selected as a favorite SNS.

Table 1. Results of the pre-survey

Q: Reasons for not using Twitter		
(a)	Difficult to join	1.3%
(b)	Not interesting	17.2%
(c)	Inconvenient	5.6%
(d)	Prefer other SNS	63.5%
(e)	Others	12.4%
Q: The most used SNS		
(a)	Facebook	70.4%
(b)	KAKAO	21.6%
(c)	Twitter	6.5%
(d)	Cyworld	1.5%

3.3 Experimental Method

All of the students participate voluntarily in this experiment. An introduction and guidance for using Twitter are presented to all students before they decide whether to participate in this experiment. Advanced consent for Twitter using experiment in class is also obtained. Students and professor use Twitter services in the following educational situations, we get the meaningful data in this environment:

- Students can ask to the professor some questions at anytime in the classroom or outside of classroom by sending tweet. A professor can reply with an answer to the question and re-tweet it to all members of the experimental group. And the professor can post tweets about class information such as canceling a class, supplementary lecture.
- A professor can suggest a discussion topic for simple and short discussion of students, the students can post their opinion on Twitter and a professor can monitor the tweets and participate in the discussion. For exchanging message, making a following relation with class members is necessary.

Through these following relationships and tweeting pattern among students, we can get some meaningful data that can show various relationships in Twitter using environment as well as how is a student's integrity, who is a influential and active student, etc. And students' seat positions in the classroom are investigated to find relationships between student's seat position and their grade. The investigation is conducted 7 or 8 times during a semester.

3.4 Area Division of Classrooms

Fig. 1 and Fig. 2 show classroom layouts of experimental classes. Seats were divided into 3 areas according to the distance from screen or professor.

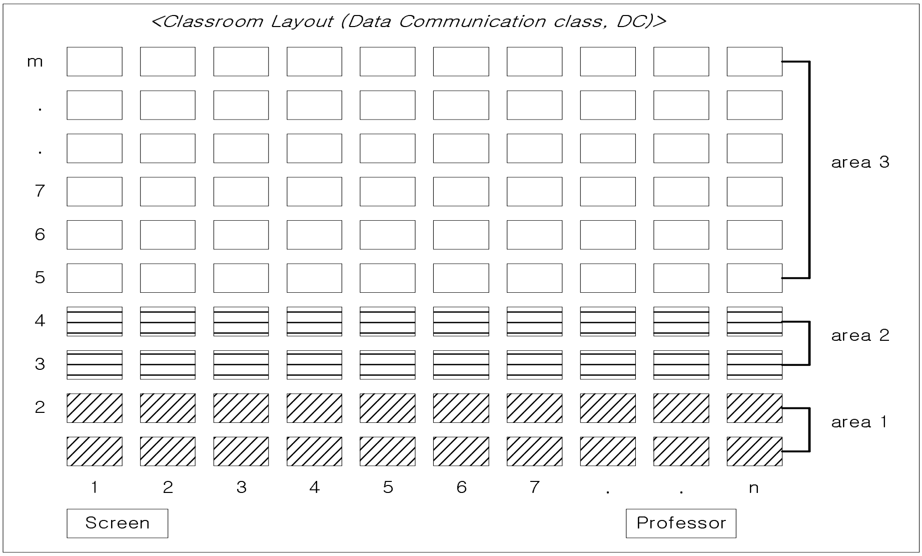


Fig. 1. Classroom layout of course DC

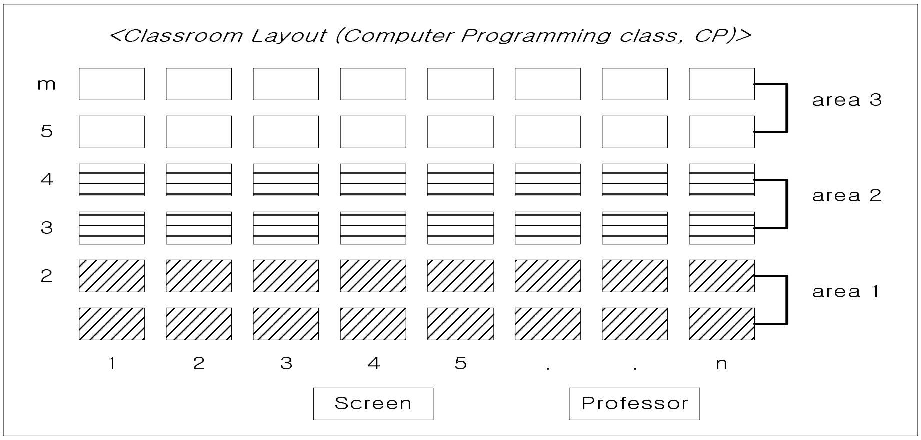


Fig. 2. Classroom layout of course CP

3.5 Average Seat Position

The seat position data of each student is calculated to get the average seat position. The average seat position of a student is calculated by using equation (1). We can also derive a variance and a standard deviation of the positions by using equations (2) and (3).

$$ave(X, Y) = (\sum_{i=1}^{n} Xi / n, \sum_{j=1}^{m} Yj/m) \quad (1)$$

$$Vx = \sum_{i=1}^{n} (Xi - ave(X))^2 / n ,$$

$$Vy = \sum_{j=1}^{m} (Yj - ave(Y))^2 / m \qquad (2)$$

$$SDx = \sqrt{Vx} \; , \; SDy = \sqrt{Vy} \; . \qquad (3)$$

4 Experimental Results

In course CP, 29 students of total 47 students participate in this experiment. In course DC, 31 students of total 48 students participate in.

4.1 Seat Position and Twitter Activity

The standard deviation value of the seat positions is not large as shown in Fig. 3 and Fig. 4. This means that most of the students have their preferred seating position in the clases.

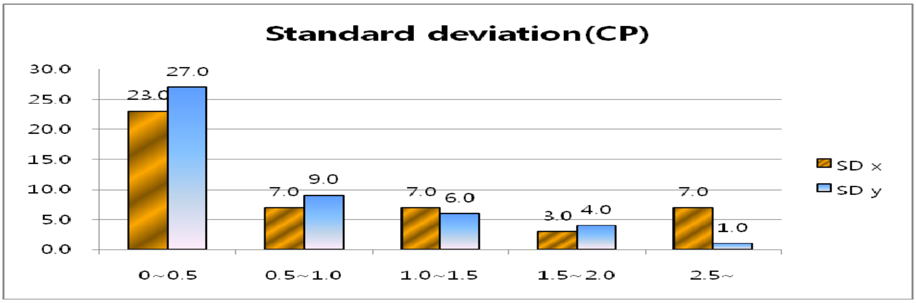


Fig. 3. Standard deviation of CP

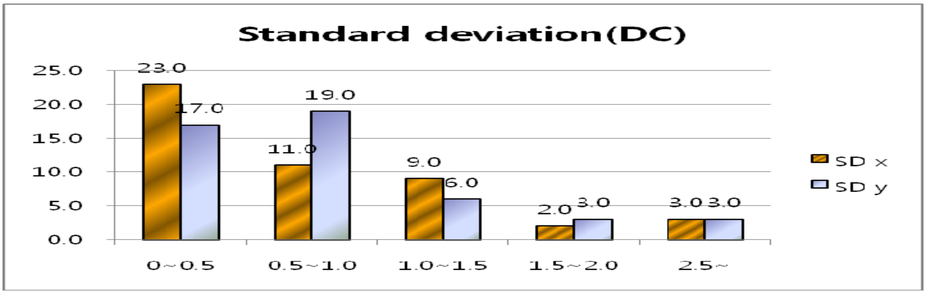


Fig. 4. Standard deviation of DC

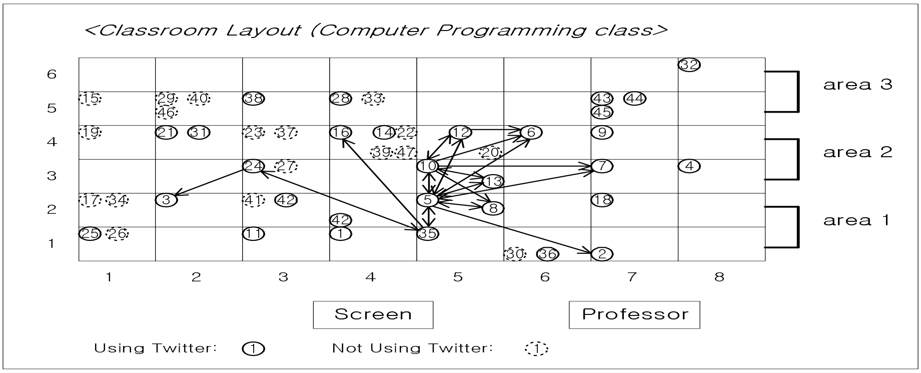


Fig. 5. Seat positions and followings in classroom of CP

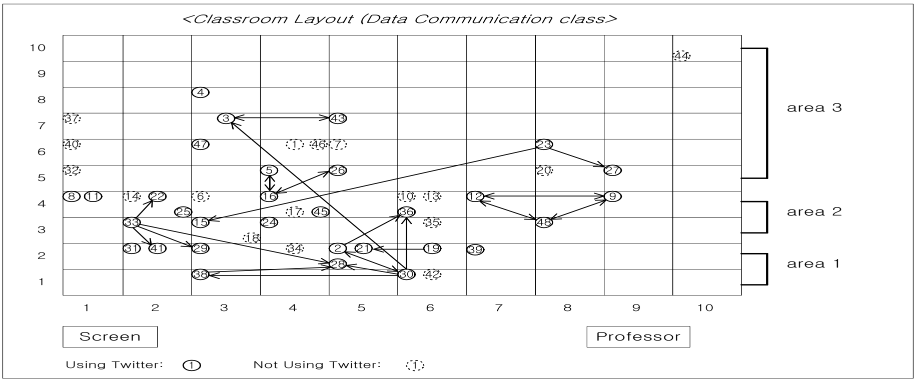


Fig. 6. Seat position and followings in classroom of DC

Fig. 5 and Fig. 6 show the average position of each student who participates in the experiment. A number in a circle means a student's intrinsic number(i.e. identification number) in the class. A solid circle represents a student who participates in the experiment and a dotted line circle represents a student who does not participate in that. An arrow means a relationship of following or follower in Twitter communications. For example, the student of number 24 in Fig. 5 is following the student of number 3, and the student of number 24 is followed by the student of number 35.

In Fig. 5 and Fig. 6, we can estimate influential students such as the students of numbers 5, 10, 12, 35 of Fig. 5 and the students of numbers 30, 28, 33 of Fig. 6.

4.2 Relationship between Followings and Grades

Fig. 7 and Fig. 9 show the relationships between the number of followings and grades of students. Fig. 8 and Fig. 10 show the relationships between the number of followers and grades of students. We can see the fact that the students got a good grade have more followers as well as follow to other students.

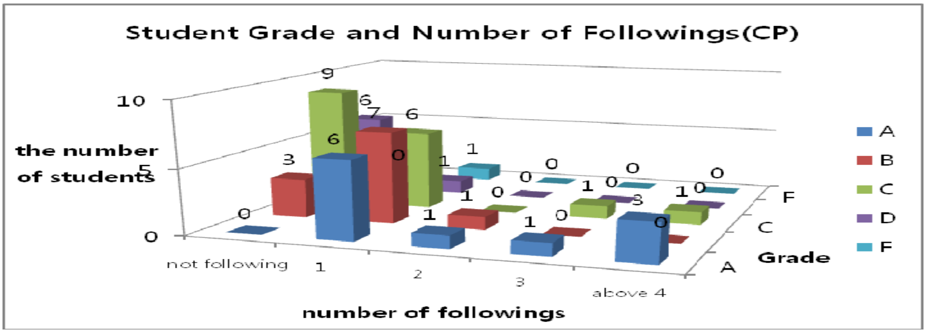


Fig. 7. Number of followings and grades

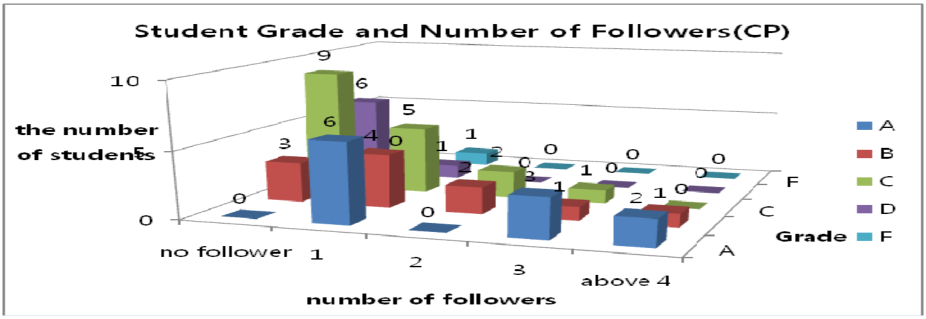


Fig. 8. Number of followers and grades

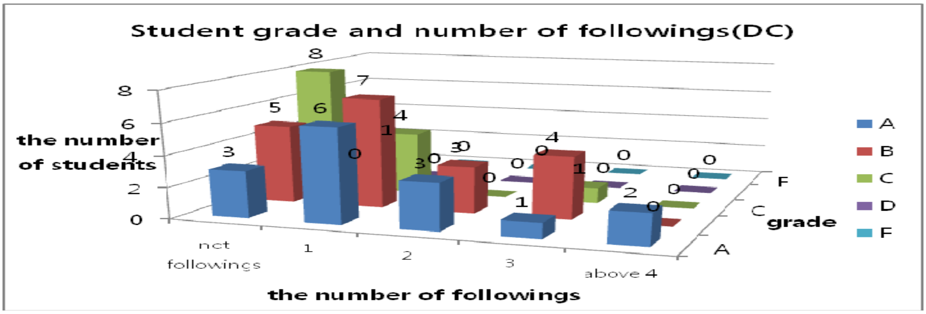


Fig. 9. Number of followings and grades

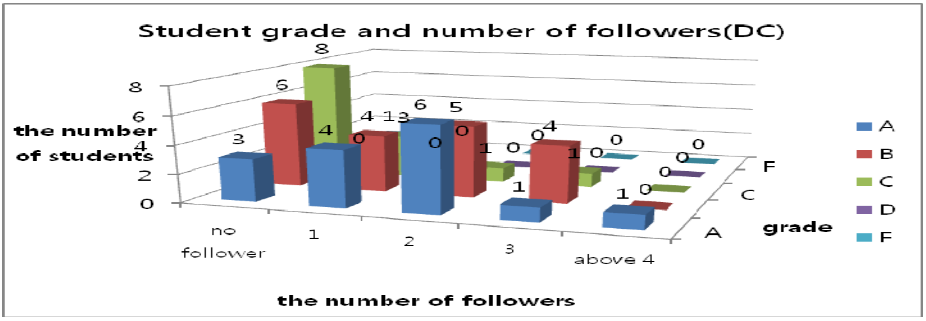


Fig. 10. Number of followers and grades

4.3 Post-survey

A post-survey is conducted for the students of the two experimental groups to get their opinions about using Twitter as an educational supporting tool. The questionare is composed of 10 questions with 5-point Likert type[18] answers as shown in Table 2 (5: excellent, 4: good, 3: normal, 2: poor, 1: worst).

The result of the post-survey shows that most of the students in the experimental group are normal satisfied with using Twitter in class. Especially, they shows relatively positive opinions to question Q4 and Q7.

Table 2. Post-survey questions and the survey result

No	Questions	5 level choices(%)					Average (level)
		5	4	3	2	1	
1	Twitter is useful to study this course	4	22	42	18	14	2.83
2	Using Twitter make me more interest in this course	3	18	45	23	10	2.79
3	Using Twitter helped me to adapt this course	4	27	38	22	9	2.95
4	Using Twitter helped me to communicate with my professor	9	36	30	14	10	3.19
5	Using Twitter helped me to communicate with class mates	1	9	38	34	18	2.42
6	Using Twitter helped me to get information of this course	5	30	36	18	10	3.01
7	Twitter is a desirable thing for class	9	35	38	6	12	3.23
8	I will recommend Twitter as a lecture support system	8	17	27	38	10	2.74
9	Twitter is better than other SNS services for class	5	21	48	16	10	2.95
10	Twitter should be used in future classes	4	21	35	27	13	2.75

5 Conclusions

Through understanding Twitter using pattern of college students in the class for the experimental groups, we studied to find the possibility of Twitter as a useful educational tool. Professor and students involved in experimental groups in 2 classes used Twitter in various educational environments, and Twitter using pattern data among student was collected and analyzed. The following and follower relationship among students were examined, and the average seat positions and grades of them were studied as influencing factors. The following results were obtained from the experiment and analysis:

- Seat positions have some relationship with the following and follower relationship. Students sitting around a student have a higher possibility to make a following and follower relationship with the student.
- Students who get a good grade have more followers and followings.
- Influential students have better grades than students who do not participate in the experiment or do not actively participate in Twitter communications.
- Seat position and grade have some relationship. Students sitting close to a professor or a projection screen have a good grade than the student sitting far away from that.

Therefore, Twitter can be used as a useful tool in educational environment. For future study, more experiments are needed to conduct in various educational environments with various SNS services.

Acknowledgement. This research was supported by the Yeungnam University research grant in 2012.

References

1. Junco, R., Heibergert, G., Loken, E.: The effects of Twitter on college student engagement and grades. J. of Computer Assisted Learning, Jour. Of Computer Assisted Learning 27(2), 119–132 (2011)
2. Grosseck, G., Holotescu, C.: Can we use Twitter for educational activities. In: Proc. of the International Scientific Conference eLearning and Software for Education (2008)
3. Ebner, M., Lienhardt, C., Rohs, M., Meyer, I.: Microblogs in higher education – a chance to facilitate informal and process-oriented learning. Computer & Education 55, 92–100 (2010)
4. McFedries, P.: All A-Twitter. IEEE Spectrum (2007)
5. Stepanyan, K., Borau, K., Ullrich, C.: A Social Network Analysis Perspective on Student Interaction within the Twitter Microblogging Environment. In: Proc. of ICALT 2010, pp. 70–72 (2010)
6. Rinaldo, S., Tapp, S., Laverie, D.: Learning by Tweeting: Using Twitter as a Pedagogical Tool. Jour. of Marketing Education 33(2), 193–203 (2011)
7. Lowe, B., Laffey, D.: Is Twitter for the Birds? Using Twitter to Enhance Student Learning in a Marketing Course. Jour. of Marketing Education 33(2), 183–192 (2011)

8. Junco, R., Elavsky, C., Heiberger, G.: Putting twitter to the test: Assessing outcomes for student collaboration, engagement and success. British Jour. of Educational Technology (2012)
9. Kader, A.: Using Twitter to More Actively Engage Students in the Learning Process
10. Clarke, T., Nelson, C.: Classroom Community, Pedagogical Effectiveness, and Learing Outcomes Associated with Twitter Use in Undergraduate Marketing Courses. Jour. for Advancement of Marketing Education 20(2) (2012)
11. Kassens-Noor, E.: Twitter as a teaching practice to enhance active and informal learning in higher education: The case of sustainable tweets. Active Learning in Higher Education 13(1), 9–21 (2012)
12. Wakefield, J., Warren, S., Alsobrook, M.: Learning and Teaching as Communicative Actions: A Mixed-Methods Twitter Study. Knowledge Management & E-Learning: An International Journal 3(4), 563–584 (2011)
13. Thoms, B.: Student Perceptions of Microblogging: Integrating Twitter with Blogging to Support Learning and Interaction. Jour. of Information Technology Education: Innovation in Practice 11 (2012)
14. Veletsianos, G.: Higher education scholar's participation and practices on Twitter. Jour. of Computer Assisted Learning 28, 336–349 (2012)
15. Ullrich, C., Borau, K., Stepanyan, K.: Who Students Interact With? A Social Network Analysis Perspective on the Use of Twitter in Language Learning. In: Wolpers, M., Kirschner, P.A., Scheffel, M., Lindstaedt, S., Dimitrova, V. (eds.) EC-TEL 2010. LNCS, vol. 6383, pp. 432–437. Springer, Heidelberg (2010)
16. Thoms, B.: Integrating Blogging and Microblogging to Foster Learning and Social Interaction in Online Learning Communities. In: Proc. of 2012 Hawaii International Conference on System Sciences, pp. 68–77 (2012)
17. Andrade, A., Castro, C., Ferreira, S.A.: Cognitive communication 2.0 in Higher Education: to tweet or not to tweet? The Electronic Journal of e-Learning 10(3), 293–305 (2012)
18. Likert, R.: A technique for the measurement of attitudes. Archives of Psychology 22, 1–55 (1932)

Extracting Collective Trends from Twitter Using Social-Based Data Mining

Gema Bello[1], Héctor Menéndez[1], Shintaro Okazaki[2], and David Camacho[1]

[1]Departamento de Ingeniería Informática. Escuela Politécnica Superior. Universidad Autónoma de Madrid
C/Francisco Tomás y Valiente 11, 28049 Madrid, Spain
[2]Department of Finance and Marketing Research. College of Economics and Business Administration. Universidad Autónoma de Madrid
C/Francisco Tomás y Valiente 5, 28049 Madrid, Spain
{gema.bello,hector.menendez,shintaro.okazaki,david.camacho}@uam.es

Abstract. Social Networks have become an important environment for Collective Trends extraction. The interactions amongst users provide information of their preferences and relationships. This information can be used to measure the influence of ideas, or opinions, and how they are spread within the Network. Currently, one of the most relevant and popular Social Network is Twitter. This Social Network was created to share comments and opinions. The information provided by users is specially useful in different fields and research areas such as marketing. This data is presented as short text strings containing different ideas expressed by real people. With this representation, different Data Mining and Text Mining techniques (such as classification and clustering) might be used for knowledge extraction trying to distinguish the meaning of the opinions. This work is focused on the analysis about how these techniques can interpret these opinions within the Social Network using information related to IKEA® company.

Keywords: Collective Trends, Social Network, Data Mining, Classification, Clustering, Twitter.

1 Introduction

Data Mining techniques have become an important field with several applications over the last few years[13]. Some of these applications have been oriented to Social Networks which contain a lot of information about their users, specially preferences, opinions and ideas[3]. Using this data, different companies have focused their marketing strategies on the influence of their products in their potential clients[3].

Currently, one of the most popular Social Networks is Twitter [1]. This Network allows its users to communicate between them using text string of 140 characters. It becomes a Collective Intelligence where the users generate an emergency information source through their comments about different topics. Twitter

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 622–630, 2013.

has several APIs to extract the information provided by the users, which offer new research challenge in different science fields[18].

Document clustering techniques can be applied for efficient organization, navigation, retrieval, and summary of huge volumes of text documents [19,9,15]. These methods can automatically organize a document corpus into clusters or similar groups which allow the knowledge extraction about user behaviour. The clustering techniques were designed to find hidden information or patterns in a dataset. They are based on a blind search in an unlabelled data collection, grouping the data with similar properties in clusters without the necessity of labelled data or human supervision. The topic detection problem can be considered as a special case of the document clustering, therefore, these techniques can be used over the textual messages provided by Twitter to extract the conversation topics and then detect collective trends from the data.

This work has been oriented on the identification of the types of comments which are provided from the users about the quality of a concrete company, in this case IKEA® . The present method can be applied to understand Twitter sentiment trends regarding companies, extracting the community mood based on a small set of tweets gathered at an instant of time. It shows how different classification and clustering techniques can be used to extract this information from the Social Networks. Finally, a comparative study of these techniques that have been applied to message text collected is presented.

The rest of the paper is structured as follows: Section 2 shows the Related Work and presents the classification and clustering techniques used during the analysis. Section 3 explains the metrics used for the model validation phase of the analytical process. Section 4 is focused on the experiments which have been carried out and their results. Finally, the last section presents the Conclusions and Future Work.

2 Related Work

Techniques of Collective Intelligence have been used in several fields. These techniques are based on the intelligence emerged by the groups which compete or collaborate in an environment. It has applications to Biology, Psychology and Computer Networks, amongst others. This work is focused on trends extraction, similar to [3] where Data Mining techniques are applied to extract information of users from electronic commerce. This information is related to ideas, preferences and behaviours of the users and the interest of the users where they are trying to find products according to similar users preferences and opinions.

Data Mining classical techniques have been applied in the Collective Trends extraction process. Different classification and clustering methods have been compared trying to find the best approach. Following subsections introduce the techniques used in this work.

2.1 Classification Techniques

The data classification techniques which have been used are the following:

- **C4.5 trees**: C4.5 [17] technique is the most classical technique in data classification. It divides the data linearly using limits in the attributes and generates a decision tree. The division is chosen using a metric such as the data entropy.
- **Naive Bayes**: The Naive Bayes (NB) [8] classifier considers each feature independent to the rest of the features. Each of them contributes to the model information. It is based on Bayes Probability Laws.
- **K-Nearest Neighbours**: K-Nearest Neighbour algorithm (KNN) [6] classifies an element according to its neighbours. Depending on the K value, it considers the K-nearest neighbours and estimate the value of the data instance which is not classified.
- **Support Vector Machines**: Support Vector Machines (SVM) [5] usually changes the dimension of the search space through different kernel functions trying to improve the classification through a hyperplane separation of the data instances in the expanded space.

2.2 Clustering Techniques

Document clustering has been studied intensively because of its wide application in areas such as Web Mining [19], Search Engine and Information Retrieval [9,15]. This technique allows the automatic organization of documents into clusters or groups [7]. Documents within a cluster have high similarity in comparison to one another, but are very dissimilar to documents in other clusters [12]. The grouping is based on the principle of maximizing intra-cluster similarity and minimizing inter-cluster similarity [2,14].

In this paper K-Means which is a partitioning clustering algorithm, is applied to obtained the clusters or topics of the Tweets extracted from Twitter. It is a simple and well known algorithm for clustering [11]. All items are represented as a set of numerical features, and the number of resulting clusters (k) must be fixed before the algorithm has been executed. Then the algorithm randomly chooses k points in vector space such as the initial cluster centers. Afterwards, each item is assigned to the closer center using the distance measure chosen. After that, for each cluster a new center is calculated by averaging the vectors of all items assigned to it. The process of assigning items and recalculate centers is repeated until the process converges or a number of iterations is completed.

3 Model Validation Metrics

The validation metrics which have been used to measure the quality of the classification algorithms are Precision, Recall and F-Measure. These metrics are defined as follow [16]:

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$F - Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Where tp represents true-positives, fp represents false-positives and fn represents false-negatives.

Precision is used to measure the situation when an instance which does not belong to the class set is classified as part of the class set. **Recall** measures the situation when an instance is rightly classified according to its class. The **F-measure** is a metric which balances these measures.

4 Experiments

This section describes the experiments carried out in this work. The first part describes the data extraction process. The second part explain the data preprocessing processes which have been used during the analysis. The third part shows the experimental setup. Finally, the last part is focused on the results obtained and their interpretation.

4.1 Data Extraction

The data which have been analysed in this work comes from Twitter. Twitter is a Social Network where people usually publish information about personal opinions. It is divided in two kind of users behaviour: follower and following. As a follower, the user receives information of people which is followed by him, and as a following, the user information is sent to its followers. The information that the users share is called Tweets. Tweets are sentences limited by 140 characters which can contain information about personal opinions of the users, photos, links, etc. A user can also re-tweet the information of other users and share it.

The information extracted for the analysis are 100 comments about IKEA®. The comments have been extracted from '11-02-2013 15:24' to '18-02-2013 15:25', all comments come from different users (there are 100 users), the comments have been taken from Spain and the language is Spanish. These comments have been classified by marketing experts in four categories:

1. **Exclusion**: Those comments which are provided by companies to advertise their products. The class corresponds with the 8% of the total tweets.
2. **Satisfaction**: Positive information of the users about a product. The class corresponds with the 31% of the total tweets.
3. **Dissatisfaction**: Negative information of the users about a product. The class corresponds with the 29% of the total tweets.
4. **Neutral**: Neutral information of the users about a product. The class corresponds with the 37% of the total tweets.

4.2 Data Preprocessing

Due to the different techniques need different preprocessing, two methods have been used according to the nature of the process applied: classification and clustering.

Data Preprocessing for Classification The Preprocessing process consists in some typical steps oriented to simplified the text information. In this case, the preprocessing has been divided in three steps:

1. Eliminate Stop-Words and special characters of the sentences.
2. Generate a term-document matrix with the keywords.
3. Use a feature selection technique to choose the most relevant words for the analysis and reduce the search space.

The original term-document matrix is formed by 747 attributes. The Feature Selection technique used is the Correlation-based Feature Subset Selection [10] combined with an Exhaustive Search. The final term-matrix has 15 attributes: 'bien', 'millones', 'todo', '#publicidad ', 'bonita', 'estás', 'hacer', 'pues', 'quiero', 'toca', 'has', 'llevo', 'más', 'saben', 'solo'.

Data Preprocessing for Clustering. A usual model for representing the content of document or text is the vector space model. In this vector space model each document is represented by a vector of frequencies of remaining terms within the document [9]. The term frequency (TF) is a function of the number of occurrences of the particular word in the document divided by the number of words in the entire document. Other function usually use is the inverse document frequency(IDF), typically, documents are represented as a TF-IDF feature vectors. With this data representation a document represents a data point in d-dimensional space where d is the size of the corpus vocabulary.

Text documents are tokenized transforming them in TF-IDF vectors. This step has included stop words removal and stemming on the document set. Besides a log normalization is applied to cleaning up edge data cases and then the TF-IDF vectors are generated which are used for further clustering process.

4.3 Experimental Setup

The experiments have been carried out using the classification and clustering algorithms describes in Sections 2.1 and 2.2. The parameters and metrics selection can be found in Table 1.

All the classification algorithms have been validated using a 10-Cross Fold validation process.

Table 1. Parameters and metrics selection for the techniques

Algorithm	Parameters	Metric
Naive Bayes	-	-
C4.5	Confidence factor = 0.25 Min. Number objects = 2	Information Entropy
SVM	$\sigma = 0.1$	RBF[1]
K-Nearest Neighbour	$K = 5$	Euclidean Distance
K-Means	$K = 3 \ldots 5$	Euclidean Distance

4.4 Experimental Results

Table 2 shows the results of the analysis applying the classification techniques defined above. The metrics used are Precision, Recall and F-Measure.

The first technique, NB, obtains the best results for the classification according to the F-measure metric. It has problems to classify the first class (Exclusion) which are those comments introduced by companies, however the rest of the techniques obtain worse results. Satisfaction and Dissatisfaction obtains generally good results for NB, although SVM achieves similar results. For the Neutral class all the algorithms have good results, therefore this class is easier to identified for classifiers. The highest Precision value is achieved by C4.5 for Dissatisfaction and the highest Recall value is achieved by SVM and NB for Neutral. Since the best F-measure values (the balanced metric of Precision and Recall) for the classes are achieved by NB, it is considered the best classifier of this analysis.

There are some details which also should be mentioned related to the classification analysis:

- The Exclusion class is difficult to distinguish in almost all the cases.
- The Neutral class is clearer separated from the others.

The clustering results have shown that the best K-value for the K-means algorithm is 5. These results are concluded from both the number of discriminate classes and the F-measure metric. For each K-value: 3-means and 4-means only discriminate two classes from the original analysis (Neutral and Satisfaction) while 5-means also separates Dissatisfaction. The F-measure shows that 3-means has the best value for Neutral class and 5-means has the best value for Satisfaction, however, both F-measure value results are closed using these algorithms. Hence, since 5-means can distinguish the Dissatisfaction class, it has been chosen as the best clustering results.

Analysing the number of clusters related to each class (in 5-means results), there are several aspects which are remarkable:

- The Neutral and Satisfaction classes, which are the most predominant, can be separated in two sub-trends per class. It means that a more detailed analysis of these trends would perform a better separation of the users opinions.

[1] Radial Basis Function [4]: this metric is defined by $e^{-\sigma||u-v||^2}$

Table 2. Results of the application of the different models using Precision, Recall and F-measure metrics for validation

Technique	Class	Cluster Num	Precision	Recall	F-Measure
NB	Exclusion	-	0.5	0.25	**0.333**
	Neutral	-	0.61	0.973	**0.75**
	Satisfaction	-	0.652	0.484	**0.556**
	Dissatisfaction	-	0.692	0.391	**0.556**
KNN	Exclusion	-	0.2	0.25	0.222
	Neutral	-	0.605	0.703	0.65
	Satisfaction	-	0.5	0.387	0.436
	Dissatisfaction	-	0.5	0.478	0.489
C4.5	Exclusion	-	0	0	0
	Neutral	-	0.45	0.973	0.615
	Satisfaction	-	0.714	0.323	0.444
	Dissatisfaction	-	1	0.174	0.296
SVM	Exclusion	-	0	0	0
	Neutral	-	0.621	0.973	0.758
	Satisfaction	-	0.571	0.516	0.542
	Dissatisfaction	-	0.769	0.435	0.556
3-Means	Exclusion	0	-	-	-
	Neutral	1	0.385	0.921	0.543
	Satisfaction	2	0.667	0.100	0.174
	Dissatisfaction	0	-	-	-
4-Means	Exclusion	0	-	-	-
	Neutral	2	0.387	0.780	0.491
	Satisfaction	2	0.714	0.0876	0.154
	Dissatisfaction	0	-	-	-
5-Means	Exclusion	0	-	-	-
	Neutral	2	0.402	0.802	**0.509**
	Satisfaction	2	0.833	0.113	**0.196**
	Dissatisfaction	1	0.5	0.0435	**0.080**

– The Exclusion class is undistinguishable in all cases. It means that this class should not be considered as a trend in the Tweets.

Comparing classification an clustering techniques, there are several things that are concluded: classification techniques obtains better results than clustering techniques, it is a consequence of the nature of the methods, clustering is a blind process while classification is a supervised process. However, applying the clustering techniques, a higher number of trends is obtained which allows a more detailed analysis of the conversations. Also clustering does not need a previous human-labelling process which is really problematic for huge datasets.

The clustering techniques have similar results distinguishing the Neutral class (which is the predominant class) as the classification methods. Also, they are not able to distinguish the Exclusion class. Hence, Exclusion should not be considered as a trend.

5 Conclusions and Future Work

This work has shown the application of Data Mining methods to extract Collective Trends from Twitter. A human-labelled dataset, extracted from Tweets of different users about IKEA®, has been used for the analysis. Clustering and classification techniques have been applied to extract the trends of users opinions and also to compare their results.

The different techniques have proved to be useful for this kind of analysis. However, these techniques are not enough to distinguish the classes. Classification techniques have achieved better results than clustering techniques, however, clustering techniques do not need to have the predefined classes for their application which is more useful for larger datasets. In addition, clustering techniques also provides more detailed information about the trends. It suggests that a clustering technique should be helpful for the initial human-labelling process.

Future work will be focused on the combination of both, classification and clustering techniques, to improve the trends identification using a previous clustering process to guide the human-labelling work. In addition, a more complete clustering study might be applied using more complex techniques to make a deeper trend study.

Acknowledgments. The preparation of this manuscript has been supported by the Spanish Ministry of Science and Innovation under the following projects: TIN2010-19872, ECO2011-30105 (National Plan for Research, Development and Innovation) and the Multidisciplinary Project of Universidad Autónoma de Madrid (CEMU-2012-034). The authors thank Ana Daz and Mercedes Rozano for the manual classification of the comments.

References

1. Twitter web site (2013), `http://twitter.com`
2. Ahonen-Myka, H.: Mining all maximal frequent word sequences in a set of sentences. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM 2005, pp. 255–256. ACM, New York (2005)
3. Bruckhaus, T.: Collective intelligence in marketing. In: Casillas, J., Martínez-López, F.J. (eds.) Marketing Intelligent Systems Using Soft Computing. STUDFUZZ, vol. 258, pp. 131–154. Springer, Heidelberg (2010)
4. Buhmann, M.D.: Radial Basis Functions. Cambridge University Press, New York (2003)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995)
6. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
7. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: 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, SIGIR 1992, pp. 318–329. ACM, New York (1992)

8. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier under zero-one loss. Mach. Learn. 29(2-3), 103–130 (1997)
9. Frakes, W.B., Baeza-Yates, R.A. (eds.): Information Retrieval: Data Structures & Algorithms. Prentice-Hall (1992)
10. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD thesis, University of Waikato, Hamilton, New Zealand (1998)
11. Hartigan, J.A., Wong, M.A.: A K-means clustering algorithm. Applied Statistics 28, 100–108 (1979)
12. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis, 9th edn. Wiley-Interscience (March 1990)
13. Larose, D.T.: Discovering Knowledge in Data. John Wiley and Sons (2005)
14. Li, Y., Chung, S.M., Holt, J.D.: Text document clustering based on frequent word meaning sequences. Data Knowl. Eng. 64(1), 381–404 (2008)
15. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
16. Powers, D.M.W.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia (2007)
17. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
18. Trung, D.N., Jung, J.J., Lee, N., Kim, J.: Thematic analysis by discovering diffusion patterns in social media: An exploratory study with tweetScope. In: Selamat, A., Nguyen, N.T., Haron, H. (eds.) ACIIDS 2013, Part II. LNCS, vol. 7803, pp. 266–274. Springer, Heidelberg (2013)
19. Zamir, O., Etzioni, O.: Web document clustering: a feasibility demonstration. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1998, pp. 46–54. ACM, New York (1998)

A Descriptive Analysis of Twitter Activity in Spanish around Boston Terror Attacks

Álvaro Cuesta, David F. Barrero, and María D. R-Moreno

Universidad de Alcalá, Departamento de Automática
Crta. Madrid-Barcelona, Alcalá de Henares, Madrid, Spain
alvaro.cuestac@gmail.com, david@aut.uah.es, mdolores@aut.uah.es

Abstract. On April 16, 2013 two bombs detonated in the Boston Marathon, with the dramatic result of three people killed and more than 180 people injured. The strong social impact that this event produced in the public opinion has been impressed on the social networks that stored opinions, comments, analysis, pictures and other materials. The availability of this large amount of information and computational capability to process it, provides a new way to study social behaviors. In particular, understanding the social network responses to the Boston terror attack can give us some clues to understand its impact on the society. Among the social networks available on Internet, *Twitter*, given its open nature, provides amazing research opportunities. This paper presents our first step building a Twitter analysis tool in Spanish. We illustrate this approach introducing a description of the activity generated on Twitter around the sentence "Maratón de Boston" (which means "Boston Marathon" in Spanish) along one week after the terror attack. This result will be used to implement a sentiment analysis tool in Spanish. During the observed time frame we have observed little social iteration and a high number of retweets in the Spanish-speaker twitter community.

Keywords: Social networks, terrorism, Boston, Twitter analysis.

1 Introduction

Without any doubt, social networks platforms, or simply social networks, since its inception over a decade ago, are revolutionizing the way people can communicate, thanks to the new way of sharing interests and activities in different areas such as politics, economy, religion, etc. A social network is a social structure made up of a set of actors (i.e. individuals or groups) and a complex set of relationships between these actors. The social network perspective provides a clear way of analyzing the structure of the whole social entities [1]. The study of these structures allows us to identify local and global patterns, locate influential entities, and examine network dynamics.

We can extract a lot of information about what is going on specific events, but simply grouping and mining the information, not only about social relationships in terms of network theory, but also about the social impact and the general

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 631–640, 2013.

opinion of their members, as well as where the opinions are produced. Although social connections through Internet has emerged in the 80s, it is after the creation of Facebook in 2004 when its use has skyrocketed. Although this netwoork's goals was initially to help students in a Faculty/University get to know each other and keep in touch, it has exceeded all expectations and nowadays it allows people from different countries stay connected and share information.

Among the large number of social networks available on the Internet, there is one that contains a collection of characteristics that makes it specially interesting from the research perspective: Twitter. It is a platform that gathers people's opinions about a wide spectrum of topics, and most of them are open, allowing a fascinating field of research. Not surprisingly, there is a large corpus of literature devoted to Twitter analysis [2] such as applying Twitter to, among other tasks, events detection [3] or public health [4].

It allows their members to send and read text-based messages up to 140 characters long, known as *tweets*. Although users follow each other in a graph-like manner, these follows serve only for subscription purposes, while *tweets* remain publicly available online for anyone to read or reply. The vast amount of rapidly-changing data found both on the Internet and specifically on social networks has led to a growing desire of knowledge extraction without manual intervention. The popular nature of these services is ideal for the discovery of trends and mass-opinion. The discovery and systematic analysis of knowledge is useful for both individuals and organizations.

Twitter analysis has been based on a complete set of techniques that allow, for instance, the early detection of trending topics [5]. One important and powerful mathematical tool used on Twitter analysis is Graph Theory [6,7] and the associated graph metrics [8], that let a deeper analysis of the relationships found on Twitter. Another relevant type of analysis applied to Twitter is sentiment analysis, that tries to quantify the emotional response to a given topic [9,10,11], or clustering [12,13].

Then, the purpose of the article is to describe the first step to develop a framework to automatically gather data from Twitter streams for further analysis. This tool will integrate sentiment analysis in Spanish, and to this end we captured a data stream in Spanish related to the Boston terror attack. This framework has been tested extracting tweets that contained the sentence in Spanish "Maratón de Boston" (wich means "Boston marathon") along one week, and performing a basic analysis of some elemental statistics. The goal is to characterize the activity in Spanish that the Boston terror attack generated on Twitter.

The paper is structured as follows. Next section describes the architecture of the application we have developed to gather and analize Twitter activity. Then, we report the data acquisition process followed by the description of the captured data about the Boston terror attack. The paper finishes with some conclusions and future work.

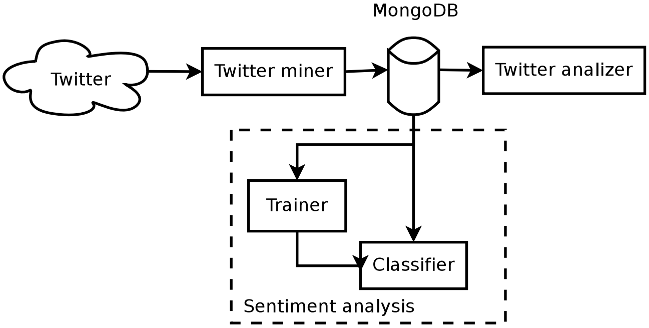


Fig. 1. Architecture of the application with its three subsystems: Twitter miner, Twitter analyzer and sentiment analysis

2 Architecture Description

Twitter offers several ways to access its data. Under the perspective of data analysis we should stress two: The Search API and the Streaming API. The first one is used to query Twitter about its content such as users or a keyword. This API imposes a limitation to the number of queries that a user is allowed to query (200 per hour), and therefore doing data-intensive analysis through the Search API is complicated. On the contrary, the Streaming API provides a real-time stream of tweets through an HTTP connection. Of course, this API also poses some limitations but even with those limitations it is easy to extract large amounts of data.

We have developed a Twitter analysis tool based on the Twitter Stream API. There are many programming languages that provide interfaces to this API: Java, C++ or R, just to mention some of them. In our case, given its simplicity, we have selected Python for data extraction. Data analysis, which is described in the next section, is performed with R because of its powerful statistical and advanced plotting functionalities.

The stream that Twitter provides can be filtered by a string, which might be a plain word, a hashtag (a keyword that begins with '#'), or a user name (which begins with '@'). A serious limitation in the Stream API is that Twitter only allows us to filter for each IP address, difficulting the data extraction in case there were need to gather data from several filters.

Roughly speaking, we can distinguish two steps on Twitter analysis: Data extraction and data analysis. In our tool those tasks are implemented by three independent (but related) subsystems, illustrated in Fig. 1, and described next:

1. **Twitter miner.** This subsystem is in charge of extracting data from Twitter using the Streaming API. Once data is extracted, Twitter miner stores data locally for further processing. These two tasks (extraction and storage) are performed in real-time.

2. **Twitter analyzer.** It is a collection of R scripts that access the database and generates reports with a collection of descriptive statistics and graphics such as time series.
3. **Sentiment analysis.** Subsystem based on a classifier aimed to analize sentiments linked the Twitter stream. The trainer takes a collection of tweets, labels them according to the criteria of a human supervisor and then construct a classifiers. This subsystem is not discussed in this paper.

Some words should be dedicated to storage. Twitter miner keeps tweets and their associated metainformation such as geolocalization, owner or timestamp in MongoDB, which is a non-SQL database. It provides increased performance in comparison to other classical SQL driven databases such as MySQL or MariaDB. For massive data storage without the need of complex queries to the database and advanced functionalities, MongoDB seems a better choice. Once the application is deployed in a server, data extraction may begin.

3 Data Extraction and Dataset Description

Data acquisition began shortly after the terror attack. The first bomb detonated on April 16, 2013, at 2:50 p.m., local east coast time, and our data extraction begun at 00:43 GMT+2, four hours after the first detonation. The time frame to capture data was exactly one week, from April 16 to April 23, which seems a reasonable amount of time to obtain a general perspective. The low activity observed at the end of this time window supports this decision.

In order to capture tweets related to the terror attack, we filtered the tweets containing the sentence "Maratón de Boston". Of course, the attack generated

Table 1. Statistics of the dataset used in this study. The dataset was created gathering tweets filtered with the sentence in Spanish "Maratón de Boston" (Boston matathon) along one week, from April 16 2013, 00:43 GMT+2.

	Count
Tweets	28,894
Retweets	12,864
Tweets without retweets	16,030
Users	24,990
Words without stop words	345,179
Geolocated tweets	255
Mentions	1,223
Responses	852
Mentions (not responses)	371

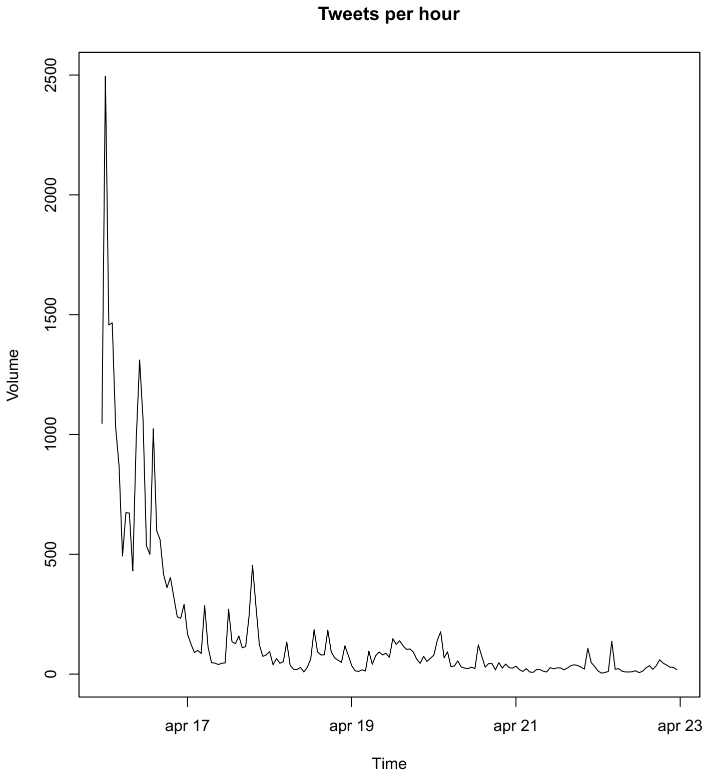


Fig. 2. Number of tweets (original and retweets) grouped by hour containing the sentence in Spanish “Maratón de Boston”. The time window goes from April 16 2013, 00:43 GMT+2, to April 23 2013, just one week.

a large number of hashtags on Twitter, such as “#marathon” or “#bombing”, however, most of these hashtags are written in English and we needed tweets in Spanish to develop the sentiment analysis subsystem. A logical choice would have been “Boston”, but this word is used in several languages such as English, Spanish or French, just to mention some of them, and we wanted to filter tweets in Spanish. Geolocaliation is not a good solution since the Spanish-speaker population is widely dispersed in Europe and America. Therefore we filted using the sentence in Spanish “Maratón de Boston”.

Table 1 summarizes the dataset of tweets we captured [1]. The overall amount of information stored is 105MB, which contains the tweets and associated metainformation such as its owner, timestamp and geolocalization data.

As Table 1 shows, along one week just after the terror attack, 24, 990 Twitter users generated or retweeted 28, 894 tweets, which yields an average value of 1, 15 tweets per user. The 44% of the tweets (exactly 12, 864) are actually retweets.

[1] This dataset is freely available upon request to the authors.

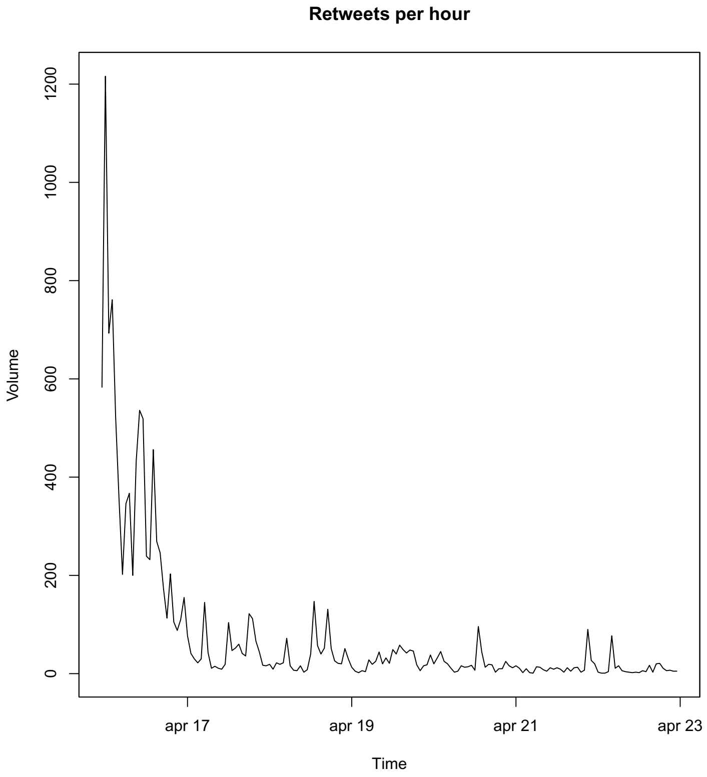


Fig. 3. Number of retweets grouped by hour containing the sentence in Spanish "Maratón de Boston" (Boston Marathon). The time window goes from April 16 2013, 00:43 GMT+2, to April 23 2013, just one week.

Therefore a remarkable amount of activity on Twitter are just retweets. More surprisingly the number of mentions is as low as $1,223$, and the mentions that are responses to a previous mention only 852, the 5% of the tweets[2]. This fact suggests that the social activity on Twitter around the Boston terror attack is not as social as one could expect. An explanation to this might be found in the nature of the tweets: They might have been used just to express feelings instead of communicating with other users.

4 Descriptive Analysis

The evolution in time of the activity on Twitter might arise valuable information. This section is devoted to study this evolution through a set of time series that plot the amount of activity on Twitter. We do not go through more elaborated

[2] We have not considered retweets in this computation; in that case, the value would have been even lower.

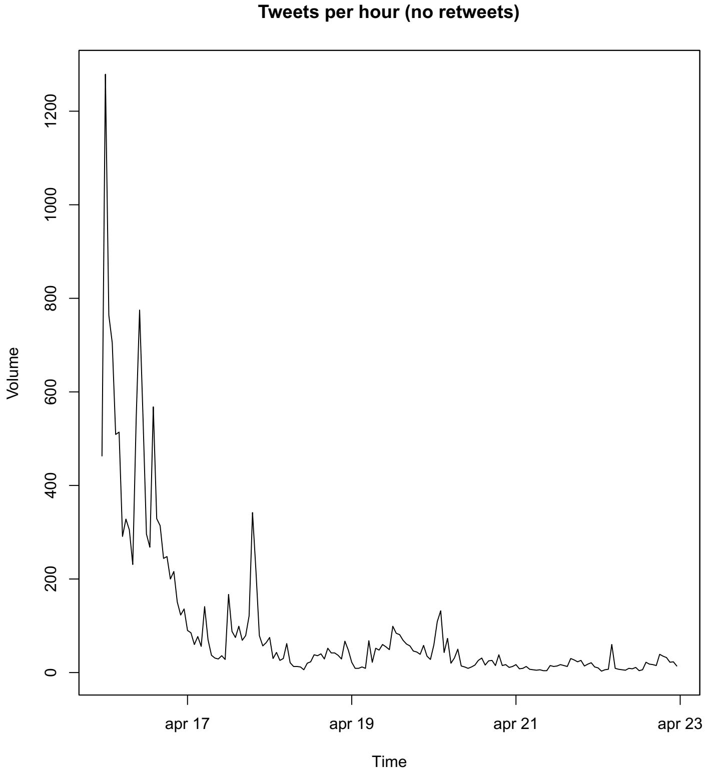


Fig. 4. Number of original tweets (retweets are excluded) grouped by hour containing the sentence in Spanish “Maratón de Boston”. The time window goes from April 16 2013, 00:43 GMT+2, to April 23 2013, just one week.

techniques such as sentiment or graph analysis. In order to provide a proper granularity of the data, all the figures in this section plots data grouped by hour.

Figure 2 reports the number of tweets captured along one week from the terror attack, measured in number of tweets per hour. It is clear that most activity is focused short after the terror attack. After one day, at April 17 we can observe that the activity lowers dramatically, and then it remains almost constant. However, there are activity oscillations that decay with time. A more detailed view of days 19 and 20 with the tweets grouped by second (not shown) shows two peaks that coincide in time with the pursuit and detention of the terror attack suspects. This high frequency peak is filtered by the grouping used in the figure. Reading random samples of the tweets located in those peaks reveal that, actually, many of them are related to the suspects pursuit.

To complement the perspective provided by Fig. 2, which includes original tweets and retweets as well, Fig. 3 reports a time serie that only contains retweets.

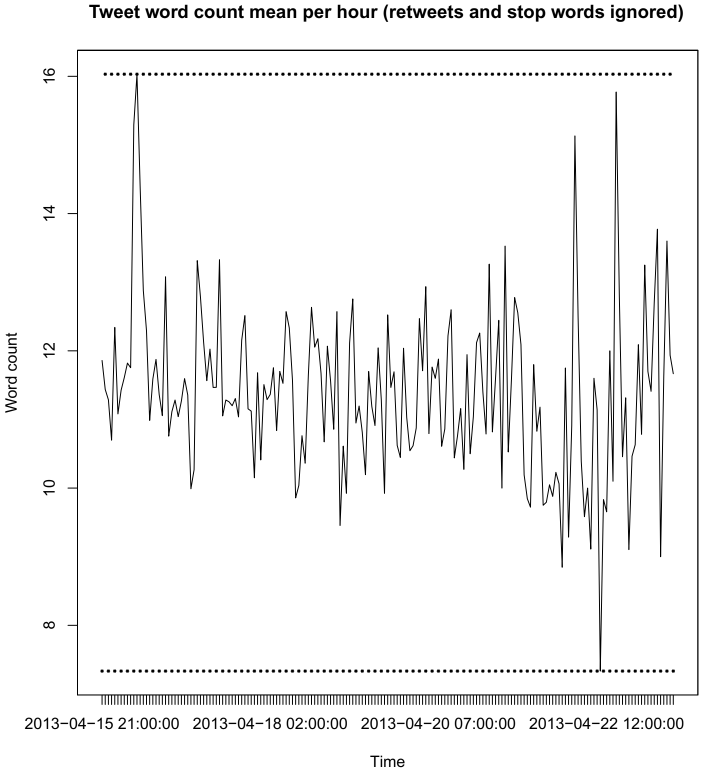


Fig. 5. Number of tweets without retweets grouped by minutes containing the sentence "Maratón de Boston". The time window goes from April 16 2013, 00:43 GMT+2, to April 23 2013.

As in the previous case, we observe that most retweets are originated just after the terror attack, and their number shrinks quite fast to remain almost constant with some small oscillations after April 17 and without those oscillations after April 19. One week after, the number of retweets containing the word "Boston" is very low. This behavior shows how the interest of Twiter users after the event decreases fast with time.

Figure 4 reports the number of tweets, excluding the retweets. Its behavior is very similar to the retweets. As in the previous case, we observe higher activity in the days close after the terror attack. If we compare Figs. 3 and 4 it turns out that much activity generated on Twitter around Boston terror attacks were retweets, which average roughly half of the activity on Twitter. It is consistent with the results shown in Table. 1.

Finally, Fig. 5 reports the tweets word count. To provide a fair measure, the figure excludes the retweets and stop words. The figure shows that, in average, the tweets contain around 12 words, and this value remains almost constant in the whole time serie. There are very few tweets with a small number of words. Twitter limits the number of characters to 140, and therefore there is obviously a higher limit of the number of words.

5 Conclusions and Future Work

In this paper we have briefly described the activity generated on Twitter around the Boston terror attacks. We have shown how the activity on Twitter is concentrated along one day after the attack with a rapid decay. The percentage of retweets along all the time serie, which is one week long, is quite high, around 44%, therefore much of the activity were not original tweets and many users just forwarded tweets. Notably, the amount of social iterations on Twitter was pretty low, less than 5% of the tweets were responses to a previous mentions, suggesting that users did not use Twitter as a communication channel, but rather as a platform to transmit a message. The dramatic event that motivated this dataset provides a good testbed for sentiment analysis.

Acknowledgements. This work was partially supported by the Spanish CDTI project COLSUVH, leaded by the *Ixion Industry and Aerospace* company.

References

1. Wasserman, S., Galaskiewicz, J.: Advances in social network analysis: Research in the social and behavioral sciences. SAGE Publications, Incorporated (1994)
2. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65. ACM (2007)
3. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
4. Paul, M., Dredze, M.: You are what you tweet: Analyzing twitter for public health. In: Fifth International AAAI Conference on Weblogs and Social Media, ICWSM 2011 (2011)
5. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, p. 4. ACM (2010)
6. Hayes, B.: Graph Theory in Practice: Part I. American Scientist 88(1), 9–13 (2000)
7. Hayes, B.: Graph Theory in Practice: Part II. American Scientist 88(2), 104–109 (2000)
8. Luciano, Rodrigues, F.A., Travieso, G., Boas, V.P.R.: Characterization of complex networks: A survey of measurements. Advances in Physics 56(1), 167–242 (2006)
9. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684. ACM (2011)

10. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology 63(1), 163–173 (2012)
11. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), vol. 2010 (2010)
12. Bello-Orgaz, G., Menendez, H., Camacho, D.: Adaptive k-means algorithm for overlapped graph clustering. International Journal of Neural Systems 22(05), 1–19 (2012)
13. Menendez, H., Bello-Orgaz, G., Camacho, D.: Extracting behavioural models from 2010 fifa world cup. Journal of Systems Science and Complexity 26(1), 43–61 (2012)

Causal Dependence among Contents Emerges from the Collective Online Learning of Students

Roberto Araya and Johan Van der Molen

Centro de investigación Avanzada en Educación (CIAE), Universidad de Chile, Chile
robertoaraya@automind.cl

Abstract. Countries are regularly upgrading K12 curricula. This is a major challenge, involving the knowledge and experience of experts on teaching and experts on the subject matters. But to teach a curriculum it is also critical to know the causal dependencies between contents during the learning process: how the students' previous performance in each content influences their future performance in each one of them. This critical empirical information is not provided in the curriculum. However, nowadays with the massive online activity of teacher and students, patterns among contents can be detected. Applying machine learning algorithms on the trace of more than half a million mathematical exercises done by 805 fourth graders from 23 courses in Chile we have identified graphs with causal dependencies among contents. These graphs emerge from the collective activity of teachers and students. They implicitly take into account logical relations, teachers' practices as they follow the curriculum, and students' learning processes.

Keywords: Educational process mining, curricular design, lesson planning.

1 Introduction

On a typical K12 curriculum, at each grade level, there are several basic contents that students should learn. Curriculum design teams propose the contents and a sequence of them. However the learning of the contents not only depends on the logical hierarchy and structure of the subject knowledge. Their learning is also influenced by psychological and social phenomena [7], [9], [16]. Curricula typically do not specify any measure of the accuracy of the proposed network of interrelations between contents. It is usually not clear if the new upgraded curriculum is easier to learn than the previous one. Furthermore, in the design process there is no direct input from the collective activity of teachers following the curriculum and the collective activity of millions of students learning the contents. This activity produced by a massive swarm of students and teachers pursuing a very specific and well defined goal, could be very helpful to find causal relation between the contents.

In order to integrate the knowledge and practices of the teachers and the learning process of the students, it must be considered the nature of both activities. It is not effective to survey teacher opinion neither to survey students' views. What is needed

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 641–650, 2013.

is to analyze students' activity while solving the exercises posed by teachers, and search for features that help to predict future students' performances. Therefore, it is critical to include students' activity in the wild, while learning. The learning process is far from a completely logical phenomenon, nor is a monotonic increasing phenomenon where students are slowly improving. While learning there are not only mistakes but also frequent regressions to previous strategies that are more inefficient. The process is being described as superposing waves [20]. Even when no errors are made the frequency of old inefficient strategies persists and only their frequency of use slowly decreases but with sporadic upsurges. There are also inhibitions and interruption of strategies, inertia from previous activities [21], fundamental misconceptions that are very difficult to overcome [9], [24], [26], etc.

The educational process mining considers several sources of information. On the one hand there is the curriculum. This is a set of contents, grouped in strands. Also, there are thousands of exercises. Each exercise can be attached to one or two contents. These are the main contents that are required to be able to solve the exercise. This way, each time a student does an exercise its response time and performance feed the corresponding contents. There is also the teacher planning. Each session the teacher selects a set of exercises to be solved by the students. Teachers interpret the curriculum, and based on their own experiences, teaching strategies and students' background, every week they plan what to teach and select the exercises to be solved by the students. There could be big differences on the plan between teachers, as well as in the exercises selected.

Our goal is to explore the possibility of automatically detect causal dependencies between contents. This type of dependencies can help teachers and curriculum designers to schedule contents throughout the year. One of the main challenges is that the data obtained varies considerable across courses since each teacher does different scheduling and does his own selection of exercises. Related works are mainly interested on predicting each student performance. For example, [27] proposes student progress indicators and attempt to mathematically model the causal relationship of both performance and non-performance attributes with a recurrent linear model. [22] uses factorization models and Support Vector machines to predict student performances. [12] proposes a model of student improvement with a logistic regression that includes knowledge components. All these studies are focused on student performance. They do not address the problem of inferring relations between the knowledge components. An important difference in this paper is that the object of study is the contents of the curriculum. We are interested in the causal dependencies between contents, independently of the students and teacher strategies. We pretend to be able to inform curriculum designers what the actual causal relations between contents are, so that this information could help them in future upgrades of the curriculum. We also pretend to inform teachers the causal relation between contents in order to help them schedule contents more effectively.

In section two we describe the data generation process. In section three we define the causality metric and the basic algorithms used. In section four we present the results obtained from the more than half a million exercises done by 805 students from 23 fourth grade courses belonging to 13 different schools.

2 Data Sources and Datasets

In this section we review the data generation process. We monitored twenty three fourth grade courses from thirteen Chilean urban, low socio economic status (SES), and high risk elementary schools. Each course has its own teacher that teaches all subjects. Twice a week each class attends a computer lab. The lab is run by a lab teacher who is specifically in charge of operating a platform of mathematic online exercises. The lab teacher also knows the mathematical contents and has teaching experience in the content. The lab teacher and the teacher both together select and assign exercises, and both help the students. The online platform contains approximately 2,000 mathematic exercises per grade. The exercises are of diverse types and they are more than just multiple choice. For example, some exercises ask to click a position on the number line, where in principle there are infinite alternatives. Other exercises require writing results of calculations. Others require reviewing a list of steps or arguments and click the step that is wrong or click indicating that all is well. There are other exercises that require multiple clicks describing an ordered sequence of positions on the screen, such as the sequence required to make a graph. As recommended in [24], there are routine exercises intended to help memorize (like multiplication tables) and non-routine. There are exercises intended to encourage the use of metaphors [2], [11], [18], [19] such as addition and subtraction are translations, especially visual representations. Other exercises promote the use of ecologically valid formats [10], such as the use of natural frequencies to estimate probabilities and use of exchanges and temporal frequencies to compare fractions. There are exercises in the form of puzzles, Sudoku puzzles and mazes. There are also exercises that promote mathematical modeling skills [17], [4], such as detect and describe patterns, etc. Furthermore, if the teacher believes there is a lack of certain types of exercises, he can bring his own exercises and ask to be incorporated. This is a very dynamic data base of exercises, constantly growing. It is very important to consider that the exercises are not completely tested before being used. However while using the platform with their students teachers can mark exercises as wrong or with bad rating in order to be reviewed by the platform maintenance team. Each exercise is assigned a strand and one or more contents from the curriculum. The contents are assigned by the teacher together with a member of the platform maintenance team, based on what is needed to solve the exercise. Each week the lab teacher coordinates with the teacher of the course the type of exercises to be assigned to the students for the week. It is very important to observe that the scheduling of exercises and contents varies considerable across courses and teachers. The data analyzed in this paper comes from 23 different courses from 13 different schools. Teachers independently interpret the curriculum on their own way and plan their own strategies and sequence. On this high risk schools they also include several contents from previous grades (third grade and even second grade) according to their assessment of the knowledge of their students. Therefore, students from a course work different concepts repeatedly and sometimes they work mainly content A, next week B, then maybe A again, then C, and some weeks they work with a mix contents. Furthermore, the sequence can be completely different in another course. The data is pure passive observational data. There are not planned interventions or any experimental manipulations.

During the session the teacher monitors on a tablet or a netbook computer the progress of students. An early alert system, at every moment, lists students who are having more difficulties. The early alert system uses the performance during the session, the number of exercises successfully solved, the number of attempts per exercise before solving them, the number of exercises opened and closed without trying to solve them, and the time already spent on the exercise currently being solved by the student. In this way the teacher knows which students need personal attention and in what specific exercise. All this information is recorded for future statistics. The early alert system also detects if there are exercises that are producing high difficulty to the whole class. This way the teacher can freeze the system and explain the required concepts. The platform is designed to drive the progression of the entire class as a whole, and not to leave students alone. It has facilities to promote the cooperation and support of students that are ahead of their peers. Students who finish early and with good performances are automatically assigned as part of the support team for the session. The early alert system starts assigning them to help peers with difficulties. Student being assisted by peers assesses the quality of the support. With this information the teacher and the lab teacher can work to improve the understanding of the contents and the communication skills between students.

3 Models of Causality between Contents

To improve curriculum design we must not only get to know the overall impact of the curriculum, but also to know in more detail the interaction and impact on learning across the curriculum components. The challenge is to measure for each content, how learning depends on the prior learning of contents of the curriculum. According to the American Statistical Association [23], "Understanding the causal effects of instructional regimes is central to applied research in education." This requires careful consideration of causation, not just correlation between learning concepts. It is not enough just to know that the performances of students in two concepts are generally similar, both are good or both are wrong. It is necessary to try to find out how performance in a given content causes or facilitates performance in other content.

This means that for any moment, it is necessary for each content C to find which of the other contents (or possibly itself) is such that its past performance or related non-performance information influences the future performance in C. For example, in the month of June (in the southern hemisphere June is in the middle of the school year), which of the contents and their respective performances and nonperformance information in the months prior to June can better predict the students' performance on fractions from June to August? Then it is crucial to verify if the pattern found it is also valid in July instead of June, and do so in many other months. In the predictions, there are two types of errors. On one side there are good future performances over the next three months that are predicted as bad ones and, on the other hand, there are bad futures performances on the next three months that are predicted as good ones. For each pair month-student, we define what it is a good performance, both for assessment tests and for tasks with on line feedback. In assessment tests, we define a pair student-month as having good future performance if in the next three months the student gets 50% or more of correct answers on exercises containing the content. In tasks with feedback, we consider a pair student-month has a good performance if in the next three months the student answered correctly at the first attempt on more than half of the exercises containing the content. For a pair month-student we define a bad performance if it is not a good one.

To make a prediction we examine different historic variables associated to every one of the contents. For example, score on the assessment questions, number of attempts before getting correct answer on questions with feedback, response times in assessment questions, response time in tasks with feedback, number of exercises done per month, time since last exercise made that contained the content, etc. In total, for each one of the 19 contents there are 18 historic variables. For each historic variable we compute the Kolmogorov-Smirnov (KS) discrimination metric: the percentage of future good performance cases with correct predictions and from it we deduct the percentage of cases with future bad performance that were incorrectly predicted. Since KS depends on the cut point where it is predicted bad future performance for values of the variable less than the cut point, we search for the cut point that produces the best KS. This best KS is the KS of the variable. In other words, KS is the maximum difference between the cumulative true positive and cumulative false positive rate. KS is always a number between 0 and 100. A value of 100 means no error in the forecast. All future good performances were predicted as good, and all future bad performances were predicted as bad. Conversely, if this measure is zero, then it means that the prediction ability is nil. For example, half of the good cases are predicted as good and half of the bad cases are predicted as good. This discrimination measure is known as the Kolmogorov-Smirnov distance [8]. It is important to note that there are many other measures of discrimination used (Gini, purity, etc). However, empirical comparative studies [14], [15] show that different metrics do not produce very different results.

It is also important to note that this measure incorporates causality and it is not just a measure of correlation. Here the causal linkage is marked by temporality. While it can sometimes occur by chance that a good (bad) future performance in a given content precedes a good (bad) performance in another, it is very rare that this occurs repeatedly by chance in different months of the year and with hundreds of students from different schools. If so happens, then it means that there is some degree of causality and it is not due to chance. Anyway, it is very important to ensure the stability and robustness of the findings. This is accomplished by removing a sample of student performance and use only the rest of the data to find the patterns. The removed sample is the validation database where all KS will be independently computed.

It is also very important to note that each student does around 40 exercises per week, but it is highly probable that in one month he does no items from a given content. That is why we have defined variables comprising what has happened in three months intervals. Therefore we have very short time series of only 5 periods: June 1st, July 1 st, August 1st, September 1st, October 1st. These very short series limit the use of methods suitable for long time series.

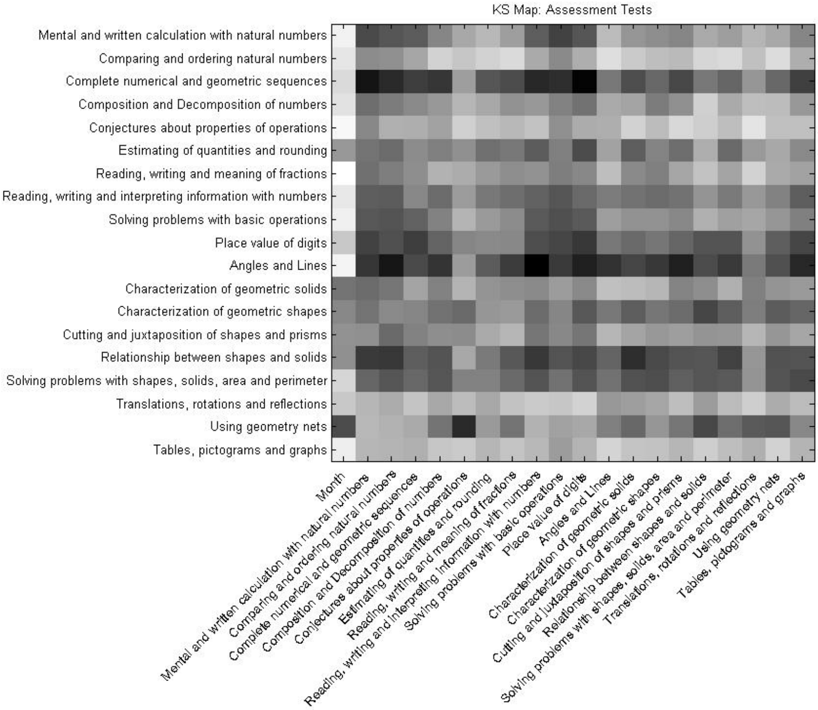
4 Results

The results of the algorithm based on the computation of KS are shown in Table 1 and in the graphs of Figure 1. Table 1 shows for any content in one of the rows, the KS measure of the best historic variable for each content of the columns. For example, future performance in "*mental and written calculation with natural numbers*" is best predicted by past performance or nonperformance information in "*mental and written calculation with natural numbers*" and by "*solving problems with basic operations*". This is not a symmetric table, since if previous performance and nonperformance information in content A predicts future performance in content B, it doesn't necessarily follows that previous performance and nonperformance information in B predicts future performance in A.

Figure 1 shows a graph with the causal network between contents generated from Table 1 using Graphviz 2.28.0, an open source graph visualization software. It is a visual way of representing the findings, part of what is called Visual Data Analytics (U.S. Department of Education, [25]). The lines between content are directed, they have an arrowhead pointing in the direction of causality. For each content the algorithm selected the historic content with highest KS as its only father in the graph. Also, higher KS are represented as thicker lines. This figure shows the central role of the Numbers strand whose contents are green. Such content predict performance in geometry whose contents are in light blue, and in Statistics and Probability whose contents are in red. This network has the metrics evaluated in a validation set, using student-month data different from those used for the selection of variables and the construction of the network. It is observed that the future performance over the next 3 months in the content "*solving problems with basic operations"* is predicted by the same content. This is obtained with the variable *score_rate_test* a variable that measure historic performance in assessments without feedback on this content. This variable is shown on the graph to inform the particular variable of the content responsible of the prediction. The positive sign indicates that better historical performances on tests in the content (measured by *score_rate_test*) predict better future performance. This means that there is a positive causality. The associated KS is 35.3 with a standard deviation of 2.3.

It may well be that a variable of a given content can improve its prediction when is used together with other variables of the same content. To account for the possibility of using several historic performance or nonperformance variables that belong to a given content, we examine two kinds of models. One built by a decision tree and other with support vector machines. For example, if the variable *score_rate_test* is used with up to 2 other variables that belong to the same content in order to build a decision tree, one also achieves a KS of 35.3. This is indicated with the number below the first KS. And if the variable *score_rate_test* is used with the variable with second best KS to build a support vector machine using the WEKA software [5], it is obtained a KS of 28.0 with standard deviation of 2.4 points. This is indicated in the third row. Moreover future performance over the next three months in "*mental and written calculation with natural numbers*" is best predicted by "*solving problems with basic*

Table 1. List of KS. For each content listed in the rows the KS of each content shown in the columns, expressed in grades of gray (white for small KS and black for high KS).



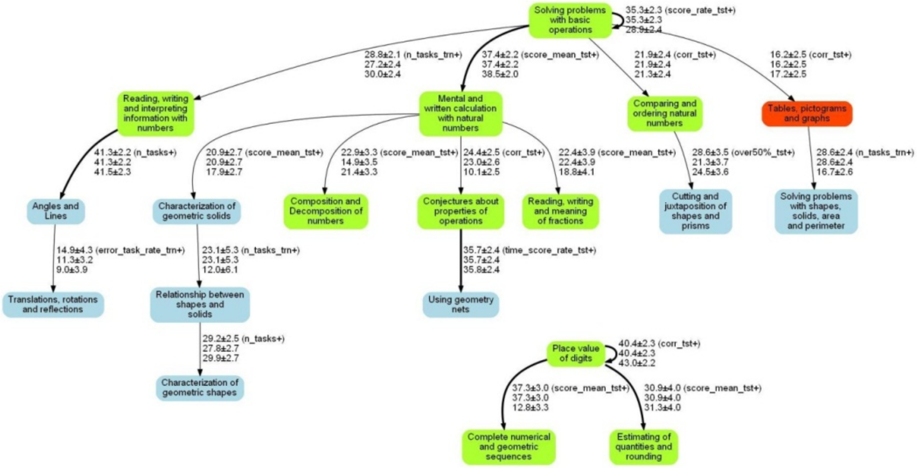

Solving problems with basic operations
35.3±2.3 (score_rate_tst+)
35.3±2.3
28.9±2.4
28.8±2.1 (n_tasks_trn+)
27.2±2.4
30.0±2.4
37.4±2.2 (score_mean_tst+)
37.4±2.2
38.5±2.0
21.9±2.4 (corr_tst+)
21.9±2.4
21.3±2.4
16.2±2.5 (corr_tst+)
16.2±2.5
17.2±2.5
Reading, writing and interpreting information with numbers
Mental and written calculation with natural numbers
Comparing and ordering natural numbers
Tables, pictograms and graphs
41.3±2.2 (n_tasks+)
41.3±2.2
41.5±2.3
20.9±2.7 (score_mean_tst+)
20.9±2.7
17.9±2.7
22.9±3.3 (score_mean_tst+)
14.9±3.5
21.4±3.3
24.4±2.5 (corr_tst+)
23.0±2.6
10.1±2.5
22.4±3.9 (score_mean_tst+)
22.4±3.9
18.8±4.1
28.6±3.5 (over50%_tst+)
21.3±3.7
24.5±3.6
28.6±2.4 (n_tasks_trn+)
28.6±2.4
16.7±2.6
Angles and Lines
Characterization of geometric solids
Composition and Decomposition of numbers
Conjectures about properties of operations
Reading, writing and meaning of fractions
Cutting and juxtaposition of shapes and prisms
Solving problems with shapes, solids, area and perimeter
14.9±4.3 (error_task_rate_trn+)
11.3±3.2
9.0±3.9
23.1±5.3 (n_tasks_trn+)
23.1±5.3
12.0±6.1
35.7±2.4 (time_score_rate_tst+)
35.7±2.4
35.8±2.4
Translations, rotations and reflections
Relationship between shapes and solids
Using geometry nets
29.2±2.5 (n_tasks+)
27.8±2.7
29.9±2.7
Characterization of geometric shapes
Place value of digits
40.4±2.3 (corr_tst+)
40.4±2.3
43.0±2.2
37.3±3.0 (score_mean_tst+)
37.3±3.0
12.8±3.3
30.9±4.0 (score_mean_tst+)
30.9±4.0
31.3±4.0
Complete numerical and geometric sequences
Estimating of quantities and rounding

Fig. 1. Network of causal influence among contents that emerges from on line assessment exercises, with KS measured on the validation data base

operations" and the KS is 37.4 with standard deviation of 2.2. KS around 40 are encouraging for a first study. In these cases, approximately half of future performances are classified as good, and around 70% of good and 70% of bad performances are predicted correctly. This is done using only previous performance and nonperformance log data, and from one content at a time. That is, this is a prediction that is made without using the history of a mix of contents. Additionally, it is not used characteristics of each student, or information about particular type of exercises, or information about what the teacher did recently. Improvements can be expected with a more exhaustive study with different algorithms and type of data.

5 Discussion

According to a recent McKinsey report [13], in the education sector the intensity of use of information technology is low, the culture of use of information for decision-making is weak, and the availability of information is scarce. However, with the use of online systems, teachers and student are generating very valuable information where different important patterns emerge. We have examined the trace of the collective activity of 23 teachers and 805 students during a year. The analysis of the trace of over half a million exercises made on line shows the importance and centrality of the number strand. Historical performance not only in numbers predicts future performance in the contents of numbers, but numbers also is an excellent predictor of performance in the other two strands. Furthermore, within the number strand the contents with more causal impact on other contents (higher KS) are "*mental and written calculation with natural numbers", "solving problems with basic operations",* and *"place value of digits"*. In different networks built both in assessment exercises and in exercises with feedback, these three components have strong causal influences to various other contents. This finding is consistent with the diagnosis of experts as they designed the Common Core Standards in the U.S. (Common Core State Standards Initiative, [6]), whose strategy is to focus on core ideas and in particular the contents of numbers.

One limitation of this study is the low number of courses analyzed. However, with 23 courses of 13 vulnerable urban schools, it already emerges a first clear picture of the impact of the different components of the curriculum in fourth grade. In the future we will compare the patterns obtained with the generated from the on line activity in the years to come. It would also be important to expand the measure to more teachers, more schools and other grade courses, and to examine whether there are differences between types of schools. It would also be interesting to compare teaching strategies, as new capabilities of the system will record use of videos, metaphors, and subjective quality of peer support. Another very important aspect to study is to analyze the causal interactions of the mathematics curriculum with the curriculum of other subject matters

Every time a teacher selects contents and exercises, and students do the corresponding exercises in on line assessments or in on line tasks with feedback and eventual support from peers, they are adding valuable input to compute the relation between contents. This is a truly collective activity that no teacher or expert is able to generate alone. The causal graphs that emerge as result of this collective work can have important consequences to better plan classes and upgrade the curriculum.

Acknowledgements. Conicyt – PIA Project CIE-O5 Center for Advanced Research on Education and Fondef grant TE 10i001.

References

1. Araya, R., Gigon, P.: Segmentation Trees: A New Help for Building Expert Systems and Neural Networks. In: COMPSTAT 1992, vol. 1, pp. 119–124 (1992)
2. Araya, R., Calfucura, P., Jiménez, A., Aguirre, C., Palavicino, M., Lacourly, N., Soto-Andrade, J., Dartnell, P.: The Effect of Analogies on Learning to Solve Algebraic Equations. Pedagogies: An International Journal. Special Issue The teaching of Algebra 5(3) (2010)
3. Araya, R., Van der Molen, J.: A mayor ejercitación en línea en matemáticas más aumenta el SIMCE. Revista Chilena de Educación Matemática RECHIEM (2010)
4. Araya, R.: Introducing Mathematical Modeling Skills in the Curriculum. In: Mathematical Modeling Course in Mathematics Curriculum: Some Best Practices in APEC Economies (2012), http://publications.apec.org/publication-detail.php?pub_id=1362 (retrieved)
5. Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., et al.: WEKA manual for version 3-6-0, pp. 3–6. University of Waikato, Hamilton (2008)
6. Common Core State Standards Initiative: Common Core State Standards for Mathematics (2011), http://www.corestandards.org/the-standards/mathematics (retrieved)
7. Consortium for Policy Research in Education: School Improvement by Design: Lessons from a Study of Comprehensive School Reform Programs (August 2009)
8. Friedman, J.H.: A recursive partitioning Decision Rule for Nonparametric Classification. IEEE Transactions on Computers 26(4), 404–408 (1977)
9. Geary, D.: Educating the evolved mind: Conceptual foundations for an evolutionary educational psychology. Information Age Publishing, Charlotte (2007)
10. Gigerenzer, G.: Adaptive thinking: Rationality in the real world. Oxford University Press, Oxford (2000)
11. Hadamard, J.: The Mathematician's Mind. Princeton University Press, New Jersey (1945)
12. Koedinger, K.R., McLaughlin, E.A., Stamper, J.C.: Automated student model improvement. In: Proceedings of the Fifth International Conference on Educational Data Mining (2012)
13. McKinsey Global Institute: Big data: The next frontier for innovation, competition, and productivity (2011)
14. Michie, D., Spiegelhalter, D., Taylor, C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood (1994)
15. Mingers, J.: An Empirical Comparison of Selection Measures for Decision Tree Induction. Machine Learning 3 (1989)
16. National Mathematics Advisory Panel, Report of the Task Group on Instructional Practices (2008)
17. OECD: The PISA 2003 – Assessment Framework (2003)

18. The National Academies: Knowing What Students Know: The Science and Design of Educational Assessment (2001)
19. Richland, L.E., Zur, O., Holyoak, K.J.: Cognitive supports for analogies in the mathematics classroom. Science 316(5828), 1128–1129 (2007)
20. Siegler, R.: Children's thinking. Prentice Hall, Upper Saddle River (1998)
21. Siegler, R., Araya, R.: A computational model of conscious and unconscious strategy discovery. In: Kail, R.V. (ed.) Advances in Child Development and Behavior, pp. 1–42. Elsevier, Oxford (2005)
22. Thai-Nghe, N., Drumond, L., Horváth, T., Schmidt-Thieme, L.: Using Factorization Machines for Student Modeling (2012), `http://www.ismll.uni-hildesheim.de/pub/pdfs/Nguyen_factmod_2012.pdf` (retrieved)
23. The American Statistical Association: Using Statistics Effectively in Mathematics Education Research (2007)
24. U.S. Department of Education: Improving Mathematical Problem Solving in Grades 4 Through 8 (2012)
25. U.S. Department of Education: Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics. (draft for public comment) Office of Educational Technology (2012)
26. Lehn, V.: Mind Bugs: the Origins of Procedural Misconceptions. MIT Press (1990)
27. Yang, F., Li, F.W.B., Lau, R.W.H.: Fuzzy Cognitive Map Based Student Progress Indicators. In: Leung, H., Popescu, E., Cao, Y., Lau, R.W.H., Nejdl, W. (eds.) ICWL 2011. LNCS, vol. 7048, pp. 174–187. Springer, Heidelberg (2011)

Social Network Analysis and Data Mining: An Application to the E-Learning Context

Camilo Palazuelos, Diego García-Saiz, and Marta Zorrilla

Dept. of Mathematics, Statistics, and Computer Science, University of Cantabria
Avenida de los Castros s/n, 39005, Santander, Spain
{camilo.palazuelos,diego.garcia,marta.zorrilla}@unican.es

Abstract. With the increasing popularity of social networking services like Facebook or Twitter, social network analysis has emerged again. Discovering the underlying relationships between people—as well as the reasons why they arise or the type of those interactions—and measuring their influence are examples of tasks that are becoming to be paramount in business. However, this is not the only field of application in which the use of social network analysis techniques might be appropriate. In this paper, we expose how social network analysis can be a tool of considerable utility in the educational context for addressing difficult problems, e.g., uncovering the students' level of cohesion, their degree of participation in forums, or the identification of the most influential ones. Furthermore, we show that the correct management of social behavior data, along with the use of the student activity, helps us build more accurate performance and dropout predictors. Our conclusions are drawn from the analysis of an e-learning course taught at the University of Cantabria for three consecutive academic years.

Keywords: social network analysis, data mining, e-learning.

1 Introduction

Social network analysis (SNA), which consists in generating patterns that allow identifying the underlying interactions between users of different platforms, has been an area of high impact in the last years. The appearance of social networking services, such as Facebook or Twitter, has caused a renewed interest in this area, providing techniques for the development of market research using the activity of the users within those services.

However, SNA techniques do not just concentrate on social networks, but also focus on other fields, such as marketing (customer and supplier networks) or public safety [7]. One of the fields in which they are also applied is education [11]. Thanks to SNA, it is possible to extract different parameters from the student activity in online courses, e.g., the students' level of cohesion, their degree of participation in forums, or the identification of the most influential ones. This kind of analyses might be helpful for teachers to understand their students' behavior, and as a consequence, help them to get better results.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 651–660, 2013.

SNA is also useful for generating new data as attributes, which can be subsequently applied data mining techniques to obtain student behavior patterns. In the educational field, there is a well defined area called *educational data mining* [12]. Building accurate performance and dropout predictors, which help teachers to prevent students from failing their subjects, is one of the main problems tackled in this area. For this purpose, classification techniques, by means of prediction models, are usually applied to uncover the students' behavior, e.g., amount of time dedicated to accomplish certain tasks or activity in forums, that results in a pass, a fail, or a dropout. For the issue of prediction, SNA provides a new useful framework that might improve the accuracy of those models. In this paper, in which we analyze an e-learning course taught at the University of Cantabria (UC) for three consecutive academic years, we show how SNA helps to uncover behavior patterns and build models that predict the performance and dropouts of students accurately.

The paper is organized as follows. Section 2 provides background and reviews the state of the art in the use of SNA in the educational field. Section 3 describes the characteristics of the academic course under study and presents the datasets generated for the experiments. Section 4 discusses the results obtained. Finally, Sect. 5 summarizes the most important conclusions of our work and draws our future lines of research.

2 Background and Related Work

SNA is the methodical study of the relationships present in connected actors from a social point of view. SNA represents both actors and relationships in terms of network theory, depicting them as a *graph* or *network*, where each *node* corresponds to an individual actor within the network, e.g., a person or an organization, and each *link* symbolizes some form of social interaction between two of those actors, e.g., friendship or kinship. Although social networks have been studied for decades [13,14], the recent emergence of social networking services like Facebook or Twitter has been the cause of the unprecedented popularity that this field of study has now.

Despite the fact that there are antecedents in the SNA literature that push back its origin to the end of the nineteenth century, Romanian psychiatrist Jacob Moreno—who, in the early 1930s, became interested in the dynamics of social interactions—is widely credited as the founder of SNA. In his seminal book [8], Moreno established the foundations of *sociometry*, a field of study that later became SNA. Since then, an extraordinary variety of SNA techniques has been developed, allowing researchers to model different types of interactions, e.g., movie actors [15] or sexual contact networks [6], and giving solution to very diverse problems, e.g., detection of criminal and terrorist patterns [7] or identification of important actors in social networks [9,10].

In order to estimate the prominence of a node in a social network, many centrality measures have been proposed. The research devoted to the concept of centrality addresses the question “Which are the most important nodes in a social network?” Although there are many possible definitions of importance, prominent nodes are supposed to be those that are extensively connected to other nodes. Generally, in social networks, people with extensive contacts are considered more influential than those with comparatively fewer contacts. Perhaps, the most simple centrality measure is the *degree* of a node, which is the number of links connected to it, without taking into consideration the direction of the links. If we take into account that direction of the links, a node has both *indegree* and *outdegree*, which are the number of incoming and outgoing links attached to it, respectively. There are more complex centrality measures, such as the *betweenness* [4] of a node, which is equal to the number of shortest paths from all nodes to all others that pass through such a node, as well as *authorities* and *hubs* [5]; a node is an authority if its incoming links connect it to nodes that have a large number of outgoing links, whereas a node is a hub if its outgoing links connect it to nodes that have a large number of incoming links.

From a node-level point of view, centrality measures constitute a very useful tool for the inference of the importance of nodes within a network. Due to their own nature, some of them, e.g., betweenness, cannot be trivially calculated, so that *network-level metrics*, which can be computed more easily and provide helpful information by considering the network as a whole, can be used for complementing the aforementioned centrality measures. One of these network-level metrics is the *density* of the network, which measures the number of links within the network compared to the maximum possible number of links. The *diameter* of the network is also a useful network-level metric; it is defined as the largest number of nodes that must be traversed in order to travel from one node to another. Other meaningful network-level metric is the *number of connected components* of the network, i.e., the number of subnetworks in which any two nodes are connected to each other by paths without taking into consideration the direction of their links. Finally, the last metric to be mentioned is *reciprocity*; it occurs when the existence of a link from one node to another triggers the creation of the reverse link.

There are some applications of SNA to the educational field. Brewe et al. [2] used a multiple regression analysis of the Bonacich centrality for evaluating the factors that influence participation in learning communities, e.g., students’ age or gender, and Crespo and Antunes [3] proposed a strategy to quantify the global contribution of each student in a teamwork through adaptations of the PageRank algorithm. SNA can also provide relevant information that can be used in educational data mining tasks, such as predicting performance or dropouts. For instance, Bayer et al. [1] used the following centrality measures for the prediction of dropouts: degree, indegree, outdegree, and betweenness. They came to the conclusion that these measures improve the accuracy of classification models in comparison with the sole use of demographic and academic attributes, e.g., students’ age, gender, or number of finished semesters.

3 Data Characterization

Our case study uses data from a virtual course hosted on Blackboard/WebCT, entitled "Introduction to Multimedia Methods." This course was taught for three consecutive academic years (2007–2008, 2008–2009, and 2009–2010) at the UC. It was designed by means of web pages and included some video tutorials, Flash animations, and interactive elements. Students had to complete four exercises and an ordinary final exam (not online). In particular, the forum—used, in this paper, for building the social network of interactions between instructors and students—was mainly used for making questions about the organization of the course, its contents, and deadlines by students, as well as answering students' doubts and make announcements by the instructor. The average number of students enrolled in the course was 70, of which more or less the half followed the course up to the end, whereas the rest dropped out. The students' profile was diverse, coming from different degrees, such as Computer Science, Mathematics, and even History.

We created three datasets with the student activity data from Blackboard and social behavior from the social networks analyses performed with *ORA. The attributes related to the student activity are: (i) total time spent in the course, (ii) total number of sessions performed, (iii) average time spent and number of sessions performed per week, and (iv) number of messages read and written in email and forums. The SNA attributes chosen are the following centrality measures: (i) degree, (ii) indegree, (iii) outdegree, (iv) betwenness, (v) authority, and (vi) hub, as well as (vii) *top3* (if a node is ranked in the top 3 nodes in some of the previous centrality measures, *top3* is true; otherwise, it is false), and (viii) percentage in *top3*, i.e., the number of *top3*s that are set to true for a certain node. The three datasets have 194 instances, i.e., one instance per student, being the difference between them the values of the class attribute. For each student, `performance.dat` indicates whether he or she passed the subject, `dropout.dat` whether he or she dropped out the course, and `mixed.dat` whether he or she passed, failed, or dropped out.

4 Experimentation

The methodology followed in this paper includes the tasks listed below:

1. Extraction and transformation of the student activity data from Blackboard;
2. Generation of social networks with *ORA using the questions and answers present in forums;
3. Analysis of social attributes for the educational field;
4. Selection of meaningful social attributes for prediction;
5. Development of classifiers for predicting performance and dropouts;
6. Discussion and conclusions.

Figure 1 shows graphically the steps performed in our case study.

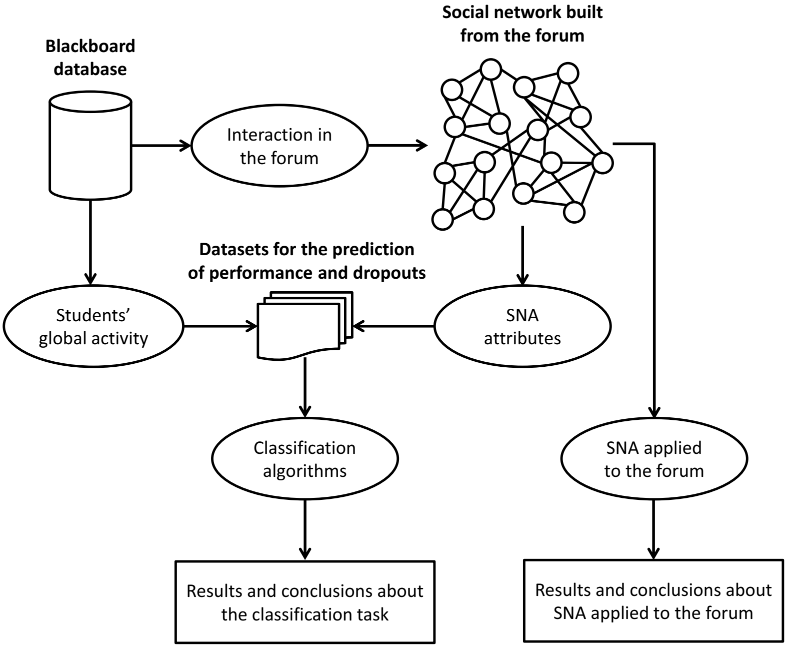

Blackboard database
Interaction in the forum
Social network built from the forum
Datasets for the prediction of performance and dropouts
Students' global activity
SNA attributes
Classification algorithms
SNA applied to the forum
Results and conclusions about the classification task
Results and conclusions about SNA applied to the forum

Fig. 1. Process of the experimentation

Figure 2 depicts the network of interactions between the instructors and the students of the course "Introduction to Multimedia Methods" taught in 2008–2009 at the UC. In this course, we found a single connected component and a diameter of 11, as well as low values of density (0.07) and reciprocity (7% of the links were reciprocal). In the previous and the subsequent courses (2007–2008 and 2009–2010), we found three and one connected components, and diameters of 19 and 16, respectively, as well as low values of density (0.05 and 0.06) and reciprocity (12% and 8%). A possible explanation for the low values detected of both density and reciprocity is that the instructor answered to the questions in the forum faster than students, preventing these from helping each other.

As can be observed, the node with more links is that corresponding to the main instructor. This means that, in this course, most interactions in the forum occur between the instructor and the students, whereas it is less frequent that those interactions occur between the students themselves. Thus, the forum is mainly used in two different ways, as the instructor pointed out: (i) students make questions—about the contents or the organization of the course—that should be answered by the instructor and (ii) the instructor makes important announcements. This kind of interactions are better showed in a graphical way by using SNA, making it easy for the instructor to interpret them, i.e., if the forum was used for a concrete activity, it could be helpful to ensure its good performance.

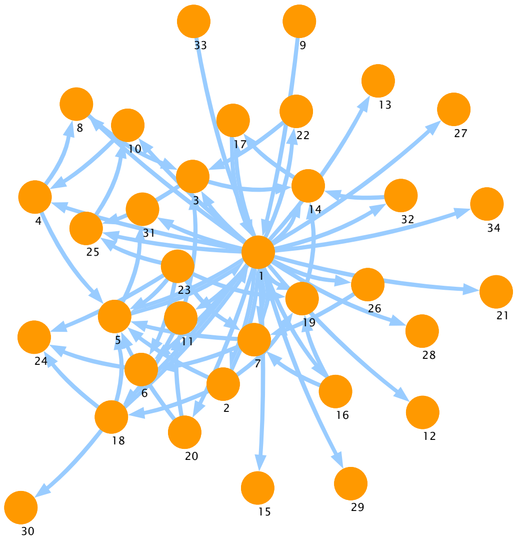


Fig. 2. Network of interactions between the instructor and the students of the course "Introduction to Multimedia Methods" taught in 2008–2009 at the UC

These conclusions can be better understood by analyzing the node centrality values exposed in Table 1. On the one hand, the instructor (node 1) has the highest values of degree and outdegree. Moreover, the difference in outdegree between the instructor and the second and the third ranked users is very high. This also happens to the betweenness and hub centrality measures. Thus, we can conclude that the instructor is the user that answered the great majority of messages posted by the students in the forum. On the other hand, the highest indegree and authority values correspond to nodes 2, 3 and 6. These students are the users that posted more messages in the forum. As a matter of fact, these three students scored the best in the course. Thus, with this analysis, we can conclude that students with a high number of interactions in the forum are likely to get good scores, a fact to be analyzed using data mining techniques. The instructor can have a better understanding of the students' behavior and improve their participation in the course. Similar results were obtained with the other two academic years, 2007–2008 and 2009–2010.

Firstly, we studied which SNA attributes were more relevant for the classification of the three datasets described in Sect. 3. We first used the attribute selection algorithm named *CfsSubSetEval* provided by Weka, which selects the best attributes for all classifiers. As a result, the most important attribute is *top3*, i.e., there is not other attribute that is considered most important than *top3*. The same was corroborated using association rules. We run the Apriori algorithm with the `mixed.dat` dataset as input—discretized with *PKIDiscretize*—and one of the rules showed that if *top3* is true, then students pass the subject with a confidence of 70%. On the one hand, by using *ClassifierSubSetEval* as attribute

Table 1. Rankings of top 3 nodes for different centrality measures

	Top 1		**Top 2**		**Top 3**	
	Node ID	Value	Node ID	Value	Node ID	Value
Degree	1	166	3	39	5	35
Indegree	3	36	5	33	6	17
Outdegree	1	157	2	6	23	6
Betweenness	1	505	17	151	14	142
Authority	3	0.93	5	0.76	6	0.41
Hub	1	1.41	23	0.03	2	0.03

selection algorithm with J48 as base classifier, we found that for this concrete classifier, the most important attribute is not only *top3*, but also the betweenesss and authority centrality measures. On the other hand, by using naïve Bayes as base classifier, the most important attributes were the degree, authority and hub centrality measures. Thus, we conclude that, for different models, different attributes are the most important ones for the prediction task.

Next, we proceeded to analyze whether SNA metrics might be suitable to predict performance and dropouts. For this task, we built six classifiers with the three datasets described in Sect. 3, but only using social data, i.e., without considering activity data. The classification algorithms chosen for this purpose were: J48, random forests, naïve Bayes, Bayesian networks, JRip, and Ridor. As can be observed, these algorithms follow different paradigms: J48 is based on trees, random forests are based on trees and meta-learning, naïve Bayes and Bayesian networks follow a Bayesian approach, and finally, JRip and Ridor are based on association rules. The implementations of these algorithms are those offered by Weka and they were run with the default parameters. Table 2 shows the average accuracy achieved by each algorithm using 10-fold cross-validation.

Table 2. Accuracy percentage in `performance.dat` and `dropout.dat`

	`performance.dat`	`dropout.dat`
J48	71.10	71.00
Random Forests	71.90	73.10
Naïve Bayes	70.32	71.00
Bayesian Networks	71.10	74.61
JRip	70.10	73.58
Ridor	68.00	74.01

As can be seen, the accuracy achieved for predicting dropouts (`dropout.dat` dataset) is higher than 70% in all cases, reaching a value of nearly 75% with Bayesian networks. Regarding performance, only the Ridor algorithm obtained an accuracy below 70%, being the rest of results better than this. So, in the light of the results, we can draw that SNA attributes can reasonably predict student performance and dropouts separately.

Finally, we compared the accuracy achieved by these classifiers—when run with the `mixed.dat` dataset as input—for assessing how SNA attributes contribute to improve the prediction task. Firstly, we built classifiers by only considering activity data, and before that, by using the whole data set, i.e., activity and social data. Likewise, we also tested whether using discretized SNA attributes would improve the accuracy. The discretization was performed by the *PKIDiscretize* algorithm offered by Weka. Table 3 shows the accuracy achieved in each case.

Table 3. Accuracy percentage in `mixed.dat`

	Activity Attributes	**Act. and SNA Attributes**	**Act. and SNA Discr. Attributes**
J48	77.20	79.79	80.31
Random Forests	78.75	80.31	80.31
Naïve Bayes	65.29	65.29	65.29
Bayesian Networks	80.83	81.87	76.68
JRip	83.42	77.20	80.83
Ridor	78.23	79.79	78.75

The results show that 4 of the 6 classifiers improved their accuracy when SNA attributes were used, achieving a significant improvement of 2.59% with J48. That improvement was higher, 3.11%, when discretized SNA attributes were used. Naïve Bayes got neither better nor worse accuracy, and only JRip got worse, but the models generated by this algorithm did not use SNA attributes, so that these results might be worse due to the randomness of the 10-fold cross-validation process.

In short, we can conclude that SNA attributes are useful for improving both students' performance and dropout prediction. Figure 3 depicts one of the classification models generated by J48 using the `mixed.dat` dataset with SNA attributes without discretization. Here, we can observe that the improvement in the classification task is due to the use of two SNA centrality measures: indegree and authority.

5 Conclusions

In this paper, we applied SNA techniques for the analysis of the interactions between the students of a course taught at the UC for three consecutive academic years. Furthermore, we used data mining methods for the prediction of the students' performance and dropouts. SNA in the educational field can be a powerful framework for the analysis of the students' social behavior and the relationships between them and their instructors. Forums are one of the tools in which this kind of analysis can be applied.

Moreover, we found that SNA can be useful for instructors to better understand how students use the forum, concluding that they often use this tool for

```
average_time_per_week <= 63: DROPOUT
average_time_per_week >  63
|   number_of_messages_written_in_the_forum <= 0
|   |   average_number_of_sessions_per_week <= 3
|   |   |   number_of_messages_read_in_the_forum <= 52: PASS
|   |   |   number_of_messages_read_in_the_forum >  52: DROPOUT
|   |   average_number_of_sessions_per_week >  3
|   |   |   total_time <= 1962: PASS
|   |   |   total_time >  1962
|   |   |   |   total_time <= 2000: DROPOUT
|   |   |   |   total_time >  2000: PASS
|   number_of_messages_written_in_the_forum >  0
|   |   number_of_messages_written_in_the_forum <= 8
|   |   |   normalized_indegree <= 0.013
|   |   |   |   number_of_messages_read_in_the_forum <= 87: PASS
|   |   |   |   number_of_messages_read_in_the_forum >  87
|   |   |   |   |   number_of_email_messages_read <= 24
|   |   |   |   |   |   number_of_messages_written_in_the_forum <= 4
|   |   |   |   |   |   |   indegree <= 1
|   |   |   |   |   |   |   |   normalized_authority <= 0.029
|   |   |   |   |   |   |   |   |   number_of_email_messages_read <= 7: FAIL
|   |   |   |   |   |   |   |   |   number_of_email_messages_read >  7: PASS
|   |   |   |   |   |   |   |   normalized_authority >  0.029: PASS
|   |   |   |   |   |   |   indegree >  1: FAIL
|   |   |   |   |   |   number_of_messages_written_in_the_forum >  4: PASS
|   |   |   |   |   number_of_email_messages_read >  24: FAIL
|   |   |   normalized_indegree >  0.013: DROPOUT
|   |   number_of_messages_written_in_the_forum >  8: PASS
```

Fig. 3. A J48 model for the `mixed.dat` dataset

making questions about the contents or the organization of the course, but it is seldom utilized for answering the questions posted by other students, which are responded by the instructor. As a matter of fact, we showed that the students who posted more questions and were answered the most are indeed the same students that, at the end of the course, achieved higher scores. This fact is better reflected with data mining analysis.

Regarding the classification of the students' performance and dropouts, we can conclude that, for both goals, SNA attributes are very useful since classification models of different kinds of classifiers achieve accuracy values over 70% with them. Moreover, using SNA attributes, along with the attributes containing the students' behavior, produces an improvement in models with respect to only using the former. In some cases, this improvement is higher than 2% of accuracy.

In a near future, we wish to extend our work in many directions. Firstly, we would like to test the effectiveness of our approach by using more e-learning courses with different characteristics. Our intention is to experiment on courses with a higher number of students, even though the number of instances in educational data mining datasets is usually limited. Also, we will attempt to apply the classification models presented in this paper to predict the hypothetical students' performance in early stages of the courses. Finally, we would like to consider other SNA metrics and algorithms, e.g., PageRank, as well as other data mining techniques aside from classification, e.g., association or clustering.

Acknowledgments. This work is partially supported by the Ministry of Education, Culture, and Sport of the Government of Spain, under grant Beca de colaboración (2012–2013), and by the University of Cantabria, under a Ph.D. studentship (2011–2015).

References

1. Bayer, J., Bydzovská, H., Géryk, J., Obšıvac, T., Popelınský, L.: Predicting Dropout from Social Behaviour of Students. In: Proceedings of the 5th International Conference on Educational Data Mining, pp. 103–109 (2012)
2. Brewe, E., Kramer, L., Sawtelle, V.: Investigating Student Communities with Network Analysis of Interactions in a Physics Learning Center. Physical Review Special Topics–Physics Education Research 8(1), 010101 (2012)
3. Crespo, P., Antunes, C.: Social Networks Analysis for Quantifying Students' Performance in Teamwork. In: Proceedings of the 5th International Conference on Educational Data Mining, pp. 234–235 (2012)
4. Freeman, L.: A Set of Measures of Centrality based on Betweenness. Sociometry 40(1), 35–41 (1977)
5. Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604–632 (1999)
6. Klovdahl, A., Potterat, J., Woodhouse, D., Muth, J., Muth, S., Darrow, W.: Social Networks and Infectious Disease: The Colorado Springs Study. Social Science & Medicine 38(1), 79–88 (1994)
7. Krebs, V.: Mapping Networks of Terrorist Cells. Connections 24(3), 43–52 (2002)
8. Moreno, J.: Who Shall Survive? Beacon House (1934)
9. Palazuelos, C., Zorrilla, M.: FRINGE: A New Approach to the Detection of Overlapping Communities in Graphs. In: Murgante, B., Gervasi, O., Iglesias, A., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2011, Part III. LNCS, vol. 6784, pp. 638–653. Springer, Heidelberg (2011)
10. Palazuelos, C., Zorrilla, M.: Analysis of Social Metrics in Dynamic Networks: Measuring the Influence with FRINGE. In: Proceedings of the 2012 EDBT/ICDT Workshops, pp. 9–12 (2012)
11. Rabbany, R., Takaffoli, M., Zaïane, O.: Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques. In: Proceedings of the 4th International Conference on Educational Data Mining, pp. 21–30 (2011)
12. Romero, C., Ventura, S.: Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 40(6), 601–618 (2010)
13. Scott, J.: Social Network Analysis: A Handbook. SAGE Publications (2000)
14. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences. Cambridge University Press (1994)
15. Watts, D., Strogatz, S.: Collective Dynamics of Small-world Networks. Nature 393(6684), 440–442 (1998)

Efficient n-Gram-Based String Matching in Electronic Testing at Programming

Adam Niewiadomski and Adio Akinwale

Institute of Information Technology
Lodz University of Technology, Lodz, Poland
Adam.Niewiadomski@p.lodz.pl

Abstract. The purpose of this study is to explore the grammatical properties and features of string matching based on n-grams techniques and to apply them to electronic testing at programming languages. Because of the intensive and extensive availability of internet in the academic environment, there exists a real need for computational procedures within artificial intelligence methods that support assessment of examination questions for uniformity and consistency. There are many computer-aided assessment packages, mainly designed for single and multiple choice tests, but they are not suitable for electronic testing at programming languages. n-grams based string matching is being successfully applied to document retrieval and other natural language processing applications. Generalized n-grams matching during substring processing tends to be time-consuming since there are N^2 n-grams extracted where n is the length of a (sub)string. The choice of selecting parameter n in n-grams approximately is an important task since the large size of n leads to polynomial growth and its small size may shorten search time significantly. As the result, some new string matching methods based on n-grams are proposed for the improvement of generalized n-grams. Experiments are conducted with the method using programming language codes as both pattern and text matching. The results are compared to chosen existing methods. We found the obtained results very promising and suggest that the proposed methods can be successfully applied to electronic testing at programming languages as an intelligent support for teachers involved in e-learning processes.

1 Introduction

The origin of students test electronic assessments in schools and professional institutions would be naturally traced to the deployment of the internet and intranet. Electronic test would remove all human errors manually recorded and create opportunity for students to access their results on-line in real time. To achieve this, computational intelligence methods are required. For example, a testing system may be equipped with modules for automatic assessments of programming codes which would reduce significantly the time for manual checking by tutors to save time. It is feasible to computerize automatic assessment of programming languages using n-gram-based string matching methods, since it

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 661–670, 2013.

has been successfully applied in document retrieval systems such as spelling checking, word variants, text retrieval, pattern matching, hypertext and natural languages (for literature references, see Section 2).

The main principle in n-grams method is the assumption that strings the structures of which are highly similar has a high probability of having the same meaning. In this method, $n \in \mathcal{N}$ represents the number of symbols determined to be extracted as a substring from particular string. If $n = 1$, the method is referred as unigram, and $n = 2$ as bigram, $n = 3$ as trigram, $n = 4$ as tetragram, $n = 5$ as pentagram, $n = 6$ as hexagram, etc. The degree between two strings is measured by comparing numbers of selected n-grams from both strings. The more common n-grams of both strings, the higher degree of matching.

The rest of the paper is organized as follows: Section 2 reviews some literature references in different methods of involving n-grams in computations. Some of known string matching methods based on n-grams, e.g. Dice coefficient and generalized n-grams, are given in Section 3. Our new proposals of n-grams-based string matching methods are described in Section 4, especially two new methods: odd-grams and sumSquare-grams, as far as two original modifications of classic n-grams, are explained in details. Section 5 briefly refers to the experiment run to prove the efficiency of newly described string matching, and some of their characteristics are compared (e.g. running time). Finally, conclusions are drawn in Section 6.

2 Literature References

One of first use of n-grams is dated from Second World War and it was used by cryptographers. Adamson also described bigrams as a method for conflating terms [1]. Angel verified that n-grams could be used in spelling error detection and correction [2]. Damerau specified four categories of spelling errors in which n-grams have been successfully used for detection [3]; he introduced a similarity measure of counting differences between two words. Pfeifer combined Damerau method with other similarity metric to rank search results. He also used n-grams based similarity metric for various phonetic matching algorithms [4]. Using classification scheme, Zamora showed that trigram analysis provides a viable data structure for identifying misspelling and transposed characters [5]. Damashek stated that frequency of occurrence of n-grams patterns can be used for identifying the language of an item [6]. Schuegraf stated that trigram have been used in information retrieval for text compression and manipulation of index length [7]. Celko described an algorithm that considered n-grams during translation [8].

A typical example, acceleration of general string searching has been accomplished using n-grams signatures by Harrison in 1971 [9]. Church and Gale proposed smoothing techniques to solve the problematic of zero-frequency of n-grams that never occurred in a corpus [10]. Kuhn and De Moris suggested the weighted n-grams model which precisely approximates the n-grams length based on their position in the context [11]. Niesel and Woodland used the variable length n-grams model which changes the n size of n-grams depending on the text being manipulated so that better overall system accuracy is achieved [12]. Assale et. al. addressed an n-grams based signature method to detect computer viruses [13]. n-grams methods have proven to be useful in a variety of tasks ranging from comparison of two texts to the quantification of degrees of homology in genetic sequence. The method is widely used for solving problems in different areas such as operation research, computer science, biology, etc. Arsmah had used n-grams to grade mathematic texts [14]. Ukkonen used the sum absolute different between corresponding numbers of n-grams occurrence in each string for approximate string matching [15]. Barman-Cedeno and Rosso used n-grams to determine if a given text is plagiarized from the pool of METER corpus [16]. In order to determine if a text is a candidate of being plagiarized from the document, they compared the corresponding set of n-grams. Lin computed a TF-IDF-like score for each n-grams that was used for ranking the results in order to find out which n-grams are informative [17]. Li combined n-grams with other similarity metrics such as Jaccard similarity, cosine similarity, edit distance, etc for approximate string queries [18]. Prahlad [19] stated that n-grams was used for intrusion detection whereby the system relies on substring match of network traffic or host activities with normal patterns or attack patterns.

Niewiadomski used generalized n-grams matching for automatic evaluating test examination and used the method also to evaluate electronic language test using German language as a case study [20]. Akinwale and Niewiadomski present some basic ideas on examination methods at programming languages, based on n-grams in [21]. Besides, methods of automated evaluation of tests and other methods of string matching based on n-grams are described in [22,23].

3 On Some Known Methods of String Matching

We consider a relation of similarity x_1 *and* x_2 which is written as $x_1 \sim x_2$. These similarity relations are subject to reflexive and symmetry and may not be necessarily be transitive. In this case, relation R on $\mathcal{X}$ is called the relation of neighbourhood if R is reflexive on $\mathcal{X}$ and R is symmetrical on $\mathcal{X}$. A similarity relation R for a given domain $\mathcal{X} \times \mathcal{X}$ maps each pair of elements in the domain to an element in a closed interval $[0, 1]$.

nary Fuzzy relation R is a *neighbourhood relation* iff its membership function $\mu_R\colon \mathcal{X} \times \mathcal{X} \to [0, 1]$ fulfills conditions of:

$$\mu_R(x, x) = 1 \text{ (reflexivity)} \tag{1}$$

$$\mu_R(x, y) = \mu_R(y, x) \text{ (symmetry)} \tag{2}$$

for each $x, y \in \mathcal{X}$. We are interested in fuzzy neighbourhood relations relations at a syntactic level. Neighbourhood relationship is also referred as: non-sup-min transitive similarity relation, tolerance relation, proximity relation, partial pre-order relation, resemblance relation, approximate equality relation, etc.

There are a variety of functions to measure the similarity of sets, Jaccard measure, overlapping measure, cosine amplitude, correlation coefficient, etc. These and other similarity measures are inter-related and recent researches have been combined them to improve the performances of string processing for different applications [24].

Basic *n*-Grams Method allows us to compute similarity of two strings s_1 and s_2 as follows:

$$sim_n(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=1}^{N-n+1} h(i) \tag{3}$$

where

$h(i) = 1$ if n-element subsequence beginning from position i in s_1 appears in s_2 ($h(i) = 0$ otherwise);
N – the length of the string;
$N - n + 1$ – number of n-element subsequences in s_1.

The n-grams method is usually applied for $n = 3$, the so-called *trigrams*. Hence, Eq. (3) takes the form of:

$$sim_3(s_1, s_2) = \frac{1}{N-2} \sum_{i=1}^{N-2} h(i) \tag{4}$$

Generalized *n*-Grams Matching. Generalized n-grams matching was introduced by Niewiadomski [20]. The algorithm matches an answer string to a template string. The matched strings are denoted as s_1, s_2 and $N(s_1) = N(s_2) = N$ is the length of the string. Hence,

$$sim(s_1, s_2) = f(n_1, n_2) \sum_{i=n_1}^{n_2} \sum_{j=1}^{N-n+1} h(i,j) \tag{5}$$

where $f(n_1, n_2) = \frac{2}{(N-n_1+1)(N-n_2+2)-(N-n_2+1)(N-n_1)}$ denotes the number of possible substrings not shorter than n_1 and not longer than n_2 in $s_1, h(i,j) = 1$ if an i-element-long substring of the string s_1 starting from j-th position in s_1 appears (at least) once in s_2 (otherwise $h(i,j) = 0$). If all substrings from one argument of comparison are found in the other, the final similarity degree is evaluated as 1 which is interpreted as the identity of s_1 and s_2 [20].

Dice Coefficient. In evaluating one term against another term, Dice coefficient is chosen because it is popular and widely used in analogous text of retrieval systems. This measure takes into account the length of terms. The coefficient values varies between zero and one. If two terms have no characters in common then the coefficient value is zero. On the other hand, if they are identical, the coefficient value will be one [24]. For two strings s_1 and s_2, the Dice coefficient is measured as

$$sim(s_1, s_2) = \frac{2(n\text{-gram}(s_1 \cap s_2))}{n\text{-gram}(s_1) + n\text{-gram}(s_2)} \tag{6}$$

where (n-gram($s_1 \cap s_2$)) is the number of bigrams found in both strings s_1 and s_2. n-gram(s_1) is the number of bigrams in strings s_1 and n-gram(s_2) is the number of bigrams in string s_2.

For example, if s_1 is "*night*" with bigrams of (ni, ig, gh, ht) and s_2 is "*naght*" with bigrams of (na, ag, gh, ht), then the Dice coefficient is $\frac{2 \cdot 2}{4+4} = \frac{4}{8} = 0.50$.

4 New Methods of String Matching

4.1 The Odd-Grams Method

Odd-grams was inspired by the generalized n-grams matching which takes $n(n-1)/2$ substrings for processing before measuring the performance. The odd-grams would take half substrings of generalized n-grams matching for processing the performance which would still reduce the running time. For the method, the matched strings are denoted as s_1, s_2 and $\max(N(s_1), N(s_2)) = N$ which is the maximum length of string s_1 and s_2.

If N is odd then for $N\colon = \lceil \frac{N}{2} \rceil$

$$sim(s_1, s_2) = \frac{1}{N^2} \sum_{i=N}^{N} \sum_{j=1}^{N-i+1} h(i,j) \tag{7}$$

else (i.e. if N is even)

$$sim(s_1, s_2) = \frac{1}{N^2 + N} \sum_{i=N}^{N} \sum_{j=1}^{N-i+1} h(i,j) \tag{8}$$

For example, if s_1 and s_2 have the same strings of "PRIMARY", the substrings are shown in table 1. The Odd-grams is calculated as follows: N is odd then $\lceil \frac{N}{2} \rceil = \lceil \frac{7}{2} \rceil = 4, sim(s_1, s_2) = \frac{1}{N^2} \sum_{i=N}^{N} \sum_{j=1}^{N-i+1} h(i,j) = \frac{1}{4^2} \cdot \frac{7+5+3+1}{1} = \frac{16}{16} = 1.$

4.2 sumSquare-Grams

Likewise odd-grams, sumSquare-gram is inspired by the generalized n-gram matching which processing time is quadratic for every n-grams in the query string with every n-grams in a line. While similarity measures of n-grams are easy to generate and manage, they do require quadratic time and space complexity and therefore ill-suited to both odd-gram and sumSquare-grams which work in quadratic. Odd-grams and sumSquare-grams methods are expected to write their results into similarity measure between a pair of submissions (pattern matching and text matching). Given pattern matching and text matching i and j, s_{ij} will be near to 1 if both patterns are considered identical and near to 0 if they are very dissimilar. That is, odd-gram and sumSquare-grams are normalized to fall within the interval [0, 1]. Similarly, similarity measure of odd-gram and sumSquare-grams are expected to be symmetric, that is the equality $s_{ij} = s_{ji}$ is

expected to hold for every i, j. For the sumSquare-grams, the matched strings are denoted as s_1, s_2 and $\max(N(s_1), N(s_2)) = N$ which is the maximum length of string s_1 and s_2.
$N\colon = \lfloor\sqrt{N}\rfloor$
$M =$ times-to-jump $= N - 1$
$P =$ first-jump $= N^2 - (N-1)^2 = 2N - 1$

$$sim_{sq}(s_1, s_2) = \frac{6}{N(N+1)(2N+1)} \sum_{i=1}^{P} \sum_{j=1}^{M} h(i, j) \tag{9}$$

4.3 Bigrams into Bi-n-Grams

Generalized n-grams matching is normally used to derive bi-n-grams where n represents two letters for each statement in programming codes. The formula is as follows:

$$sim(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=0}^{N-n+1} h(i) = \frac{1}{N-2+1} \sum_{i=0}^{N-2+1} h(i) = \frac{1}{N-1} \sum_{i=0}^{N-1} h(i) \tag{10}$$

4.4 Trigrams into Tri-n-Grams

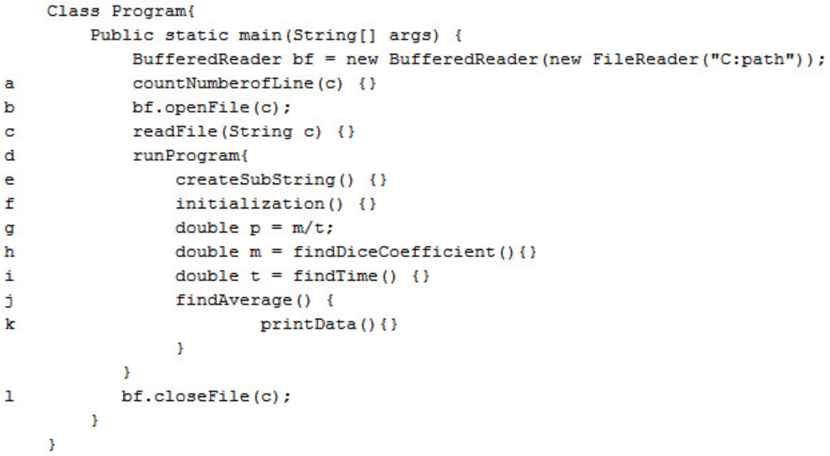
Tri-n-grams is also derived from generalized n-grams matching as follows where n represents three letters for each statement in programming codes.

$$sim(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=0}^{N-n+1} h(i) = \frac{1}{N-3+1} \sum_{i=0}^{N-3+1} h(i) = \frac{1}{N-2} \sum_{i=0}^{N-2} h(i) \tag{11}$$

5 Experiment and Results

All the methods introduced in Section 4 are implemented using JAVA Programming Language. The created software is run on Intel Pentium 2.10 GHz dualcore CPU and 1.00 GB of RAM, under a 32-bit Windows Vista. Each line of code of a programming language is assigned to a unique letter. The combination of these unique letters forms string codes. Figure 1 illustrates one of the sample codes and unique letters assigned. In learning programming languages, students are expected to write a program that solves a task, line by line. To test the knowledge of computer science students, they are requested to study the program line by line and arrange these lines in a correct sequence. Sequences of unique letters are formed by students and represent texts for matching, while the correct unique letters are formed by the tutor and represent patterns for matching.

For example, program lines presented in Figure 1, the correct answer is **b c l a d f e h i g j k** while the first five unique sequences generated by five students are as follows:



```
        Class Program{
            Public static main(String[] args) {
                BufferedReader bf = new BufferedReader(new FileReader("C:path"));
a               countNumberofLine(c) {}
b               bf.openFile(c);
c               readFile(String c) {}
d               runProgram{
e                   createSubString() {}
f                   initialization() {}
g                   double p = m/t;
h                   double m = findDiceCoefficient(){}
i                   double t = findTime() {}
j                   findAverage() {
k                           printData(){}
                    }
               }
l              bf.closeFile(c);
            }
        }
```

Fig. 1. Sample program code, the similarity of which is evaluated by the proposed string matching methods

1. **b a c d f e h i g j k l**
2. **b a c l d f e h i g j k**
3. **b c a l d e f h i g j k**
4. **b c a l d e f h i g j k**
5. **b c l a d e f h i g j k**

These data are read by the system and running times and performance-to-price for Niewiadomski generalized n-grams matching, Dice coefficient, odd-grams, sumSquare-grams, bi-n-grams and tri-n-grams as illustrated in Figure 2. The "performance-to-price" coefficient (P2P) is evaluated as:

$$\mathrm{P2P} = \frac{\text{similarity value}}{\text{running time}} \tag{12}$$

The value of running time has been converted to milliseconds. Figure 3 shows the performance of each method with respect to running time and performance-to-price of computing string matches. For the efficient evaluation of each method with respect to the degree of running time and performance-to-price, we are able to get 302 pattern matches from 100 programming codes through students assignments which are not necessarily the same as the correct answers.

Looking at Figure 2, the running time which includes the time of creating the substrings n-grams for each method, shows that Dice coefficient, Odd-grams and new methods are more or less the same although odd-grams computational time is a bit lower than Dice coefficient. Figure 3 illustrates the total performance-to-price of the two known and four new methods using 302 patterns and test matches.

As it is depicted in these two figures, the results achieved by the odd-grams method are not significantly better than Dice coefficient, but they are encouraging in comparison to generalized n-grams. It is also noticed in the experiment

Table 1. Analysis of the results using sumSquare-grams, Dice coefficient, bi-n-grams and Tutors' grading scores

Methods	70-100 (A)	60-69 (B)	50-59 (C)	45-49 (D)	40-44 (E)	0-39 (F)	Total
Dice	50	48	81	14	20	89	302
sumSquare	61	133	94	4	2	8	302
bi-n-gram	111	160	29	2			302
Tutor	58	90	88	39	18	9	302

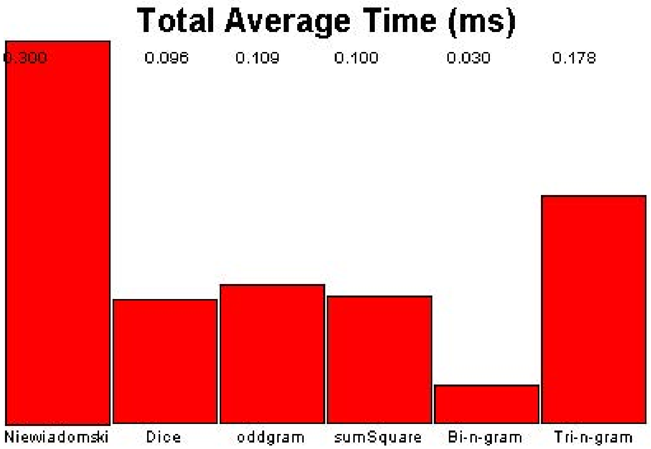


Fig. 2. Total average of running times of 302 pattern matches

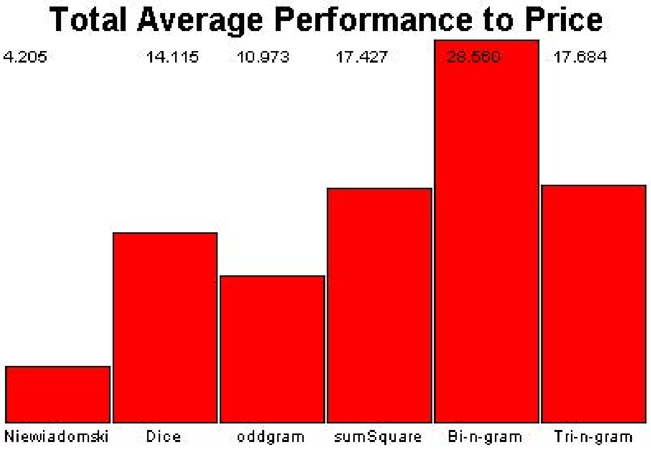


Fig. 3. Total average of performance-to-price of 302 pattern matches

that the number of substrings in Dice coefficient is fixed and does not decrease, as other numbers of substring, i.e. odd-grams and generalized n-grams, do.

6 Final Conclusions

In this paper, the new methods of odd-grams and sumSquare-grams as far as modifications of bigrams and trigrams are proposed to improve the performance of string matching based on n-grams. The measures are then applied to electronic testing at programming languages. The computational process of the methods drastically reduces a large amount of storage by using half nested loops to compare n-grams in the pattern match with every n-gram in the string being matched. The experimental results indicate that bi-n-grams is the best choice among the six methods using two-letter-strings for each line of code which has improved the efficiency of n-grams analysis. The results achieved by the odd-grams method is not better than Dice coefficient values but the running times with Dice coefficient are highly encouraging and better than generalized n-grams matching. The results obtained from sumSquare-grams method are very close to the results of experts, and it indicates that it can be successfully used in electronic testing at programming languages.

The investigated methods are also compared with two existing string matching methods: Dice coefficient and generalized n-grams. The outcome of the results indicate that odd-grams can be used to measure similarity between pattern matches generated by the students and text matches generated by tutor in electronic test at programming languages. The results achieved by the odd-grams method are not higher than Dice coefficient values but the running times of Dice coefficient are highly encouraging and better than those of the generalized n-grams.

References

1. Adamson, G.W., Boreham, J.: The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Information Storage and Retrieval 10, 253–260 (1974)
2. Angel, R., Frennd, G., Willet, P.: Automatic spelling corection using a trigram similarity measure. Information Processing and Management 19(4), 255–261 (1983)
3. Damerau, F.: A technique for computer detection and correction of spelling errors. Communication of ACM 7, 171–176 (1964)
4. Pfeifer, U., Poersh, T., Fubr, N.: Retrieval effectiveness of proper name search methods. Information Processing and Management 32(6), 667–679 (1996)
5. Zamora, E., Pollack, J., Zamora, A.: The use of trigram analysis for spelling error detection. Information and Management 17(6), 305–316 (1981)
6. Damashek, M.: Gauging similarity with n-gram language-independence categorization of text. Science 276, 845–848 (1995)
7. Schuegraf, E.J., Heaps, H.S.: Selection of equifrequent word fragments for information retrieval. Information Storage and Retrieval 9, 697–711 (1973)
8. Celko, J.: A Sql programming. Morgan Kaufman Publishers (1995)

9. Harrison, M.: Implementation of the substring test by hashing. Communication of the ACM 14(12), 777–779 (1971)
10. Church, K.W., Gale, W.A.: A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. Computer Speech Language 5(1), 19–54 (1991)
11. Kuhn, R., De Mori, R.: A cache-based natural language model for speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(6), 570–583 (1990)
12. Niesler, T.R., Woodland, P.C.: A variable-length category-based n-gram language model. In: Proceedings, IEEE ICASSP, pp. 164–167 (1996)
13. Abou-Assaleh, T., Cercone, N., Keselj, V., Sweidan, R.: N-gram based detection of new malicious code. In: COMPSAC 2004 Proceedings of the 28th Annual International Computer Software and Applications Conference - Workshops and Fast Abstracts, vol. 02, pp. 41–42 (2004)
14. Ibrahim, A., Abu-Bakar, Z.: Automated grading of linear algebraic equation using n-gram method. Master's thesis, Pensyarah matmatik FTMSK, Kampus, Cawangen, Kuala Pilar (2005)
15. Ukkonen, E.: On approximate string matching in fct. Science (1983)
16. Barrón-Cedeño, A., Rosso, P.: On automatic plagiarism detection based on *n*-grams comparison. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 696–700. Springer, Heidelberg (2009)
17. Lin, J., Yeh, J., Ke, H., Yang, W.: Learning to rank for information retrieval using genetic programming. In: Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval, Amsterdam, Netherland (2007)
18. Li, C., Lu, J., Lu, Y.: Efficient merging and filtering algorithms for approximate string searches. In: ICDE, pp. 257–266 (2008)
19. Prahlad, F., Lee, W.: Q-gram matching using tree models. IEEE Transactions on Knowledge and Data Engineering 18(4), 433–447 (2006)
20. Niewiadomski, A.: Methods for the linguistic summarization of data: application of fuzzy sets and their extensions. Akademicka Oficyna Wydawnicza EXIT, Warszawa (2008)
21. Akinwale, A.T., Niewiadomski, A.: New similarity measures in electronic tests at programming languages. In: IIIrd Conference TEWI "Technologia Edukacja Wiedza Innowacja", Łódź Poland, July 03 (2012)
22. Niewiadomski, A.: Interval-valued data structures and their application to e-learning. In: Vojtáš, P., Bieliková, M., Charron-Bost, B., Sýkora, O. (eds.) SOFSEM 2005. LNCS, vol. 3381, pp. 403–407. Springer, Heidelberg (2005)
23. Niewiadomski, A., Kryger, P., Szczepaniak, P.S.: Fuzzy Comparison of Strings in FAQ Answering. In: Abramowicz, W. (ed.) Proceedings of the 7th Business Information Systems, Kwietnia 21-23, pp. 355–362. Poznań (2004)
24. Buckles, B.P., Petry, F.E.: Information theoretic characterization of fuzzy relational databases. IEEE Transaction Systems Man Cybernet. 13(1), 74–77 (1983)

Using Fuzzy Logic for Recommending Groups in E-Learning Systems

Krzysztof Myszkorowski and Danuta Zakrzewska

Institute of Information Technology Lodz University of Technology,
Wolczanska 215, 90-924 Lodz, Poland
{kamysz,dzakrz}@ics.p.lodz.pl

Abstract. Performance of Web-based learning environment depends on the degree it is adjusted into needs of virtual learning community members. Creating groups of students with similar needs enables to differentiate appropriately the environment features. Each new student, who joins the community, should obtain the recommendation of the group of colleagues with similar characteristics. In the paper, it is considered using fuzzy logic for modeling student clusters. As the representation of each group, we assume fuzzy numbers connected with learner attributes ranked according to their cardinality. Recommendations for new students are determined taking into account similarity of their dominant features and the highest ranked attributes of groups. The presented approach is investigated, taking into considerations learning style dimensions as student attributes. The method is evaluated on the basis of experimental results obtained for data of different groups of real students.

Keywords: recommender systems, fuzzy logic, groups modeling.

1 Introduction

Web-based learning environments should be differentiated according to the preferences of all students from the virtual learning community. Students' attitude towards using the software may depend on the degree it is tailored into their needs. The contents and information organization of courses as well as learner acceptance of interface may play significant role in educational system performance. Adjusting the system into individual needs of each student may be difficult and costly, in the case of their big amount. As the solution, students may be divided into groups of similar preferences and the system may be differentiated according to their needs [1].

The performance of educational environment depends on the inner similarity between members of each group and accuracy of the model. Each learner joining the community should be assigned into the closest cluster, what guarantees the similar characteristic features for all the group members. To obtain consistency inside groups, models of good quality are required. As attribute values characterizing the cluster, there may be assumed the average values for all the cluster members or the attribute values that characterize majority of students

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 671–680, 2013.

in the cluster. Both of the approaches may not reflect the features of all cluster members. In the paper [2] the concept of group representation was introduced. It was developed in [3], where group representation in the probabilistic form was defined and accordingly Bayes classifier was used to build group recommendations. In the current paper we use models based on fuzzy logic, where the degree of the membership is taken into account. The proposed method is examined for student attributes based on their learning style dimensions. It is validated, on the basis of experiments, done for real students' clusters.

The paper is organized as follows. The related work is depicted in the next section. The recommender system is described in Section 3. Then, fuzzy logic for building recommendations, starting from group modeling, is presented. Section 5 focuses on application of the method into attributes based on learning style dimensions. In the following section some experimental results are presented and discussed. Finally, concluding remarks and future research are outlined.

2 Related Work

Technologies for adaptive group formation was mentioned by Brusilovsky and Peylo [4] as one of the most important supporting tools in intelligent collaborative learning. Christodoulopoulos and Papanikolaou [5]discussed several factors that should be considered while assigning learners into groups. They investigated the choice of the best algorithms for creating student clusters. A survey of the state-of-the art in group recommendation was presented in [6]. Masthoff [7] described using of group modeling and group recommendation techniques for recommending to individual users. A collaborative Bayesian network-based group recommender system has been proposed in [8]. Bayesian networks were also applied by García et al. [9] to detect student learning styles on the basis of their behaviors.

Several researchers examined possibilities of using fuzzy logic for student modeling. Most of them were connected with e-learning systems. Authors used fuzzy sets to describe reality by means of linguistic terms, which are close to human nature. Such approach was applied by Hawkes and Derry [10], who defined linguistic labels associated with membership functions and use them in the process of informal fuzzy reasoning, for the purpose of students' evaluation. The intelligent system, using fuzzy logic for evaluation and classification of student performance on the basis of the structure of the observed learning outcome, was presented in [11]. The system enables estimation of students' knowledge in several theme sections. Then the final level is implemented. In [12] for evaluation of intelligent learning systems the authors proposed two tests for creation student fuzzy models. They considered the use of fuzzy sets to specify the relevance or learning intensity of cognitive elements and fuzzy rules that establish or modify those fuzzy sets. Lascio et al. [13] applied an algebraic fuzzy structure in a multimedia tutoring system. The structure was used for modeling user's states during navigation. Inference engines based upon fuzzy logic were developed in a didactic methodology described by Monova-Zheleva et al. in [14].

A number of works is devoted to adapting and personalizing of the learning process to the students' needs. In the paper [15] authors applied fuzzy logic for defining a fuzzy ontology in which relationships between objects, attributes and classes were described by means of fuzzy relations. They elaborated a fuzzy ontology discovery algorithm to extract the concept maps representing students' knowledge structure. Fazlollahtabar and Mahdavi applied a neuro-fuzzy approach for obtaining an optimal learning path [16]. Characteristics of students were inferring on the base of teachers' opinions expressed by means of linguistic terms. Xu et al. [17], in turn, built fuzzy models on the basis of learning activities and interaction history. As the main goals of their applications there should be mentioned personalization and adaptivity features of educational systems as well as recommendation possibilities.

3 Recommender System

Let us assume that students are described by N attributes of nominal types S_i, $i = 1,...,N$. A tuple ST representing a student is of the form:

$$ST = (st_1, st_2, ..., st_N), \quad st_i \in DOM(S_i) \ , \tag{1}$$

where $DOM(S_i)$ stands for the domain of S_i. Attribute S_i may take on m_i nominal values, $i = 1,...,N$, where m_i stands for cardinality of $DOM(S_i)$:

$$m_i = card(DOM(S_i)) \ . \tag{2}$$

To find group characteristics we will use the group representation in a probabilistic form, defined in [3]:

Definition 1. *Let GS be a group of objects described by vectors of N components of nominal type, each of which of M different values at most. As the group representation GSR we will consider the set of column vectors $gsr_i, i = 1, ..., N$ of m_i components, representing attribute values of GS objects where $M = max$ $\{m_i : i = 1, ..., N\}$. Each component of a vector $gsr_i, i = 1, ..., N$ is calculated as likelihood $P_{i,j}, i = 1, ..., N; j = 1, ..., m_i$ that objects from GS are characterized by the certain attribute value:*

$$P_{i,j} = card(\{ST \in GS : ST(S_i) = s_{i,j}\})/card(GS), \quad s_{i,j} \in DOM(S_i) \ , \tag{3}$$

and is called the support for the respective attribute value in GS.

The above definition allows to look at the group features comprehensively. The biggest supports indicate dominant attribute values in groups. Probabilistic form of the group representation allows to determine matching degrees of attributes of new students and classify them to appropriate groups.

Let a tuple $NST = (ns_1, ns_2, ..., ns_N)$, $ns_i \in DOM(S_i)$, represent a new student. Let us consider a group of students GS represented by GSR. The matching degree, denoted as C_i, of NST to GS, for the attribute S_i is computed as the proportion of students ST belonging to the group, $ST \in GS$, such that $ST(S_i) = ns_i$ to the size of the group. Thus, it is equal to the corresponding component of gsr_i, $i = 1, ..., N$ in the group representation. The total matching degree C for the group is a minimal value of C_i, $i = 1, ..., N$:

$$C = \min_i C_i \ . \tag{4}$$

Maximal value of C indicates the group that should be recommended for students.

4 Fuzzy Logic for Building Recommendations

The recommendation procedure does not take into account closeness relationships which may be associated with attribute domains. This can lead to the improper choice of the group.

Example 1. Let us consider a matching degree C_i with respect to the attribute S_i for two groups of students $GS1$ and $GS2$. Suppose $DOM(S_i) = \{a, b, c, d, e\}$. Let $gsr_i(GS1)$ and $gsr_i(GS2)$ be vectors of the group representations of $GS1$ and $GS2$, respectively. Suppose

$$gsr_i(GS1) = \begin{bmatrix} P(S_i = a) = 0 \\ P(S_i = b) = 0.45 \\ P(S_i = c) = 0.1 \\ P(S_i = d) = 0.45 \\ P(S_i = e) = 0 \end{bmatrix} \text{ and } gsr_i(GS2) = \begin{bmatrix} P(S_i = a) = 0.4 \\ P(S_i = b) = 0 \\ P(S_i = c) = 0.2 \\ P(S_i = d) = 0 \\ P(S_i = e) = 0.4 \end{bmatrix} .$$

Let $NST(S_i) = c$, where NST denotes a tuple representing a new student. Thus matching degrees for $GS1$ and $GS2$ equal 0.1 and 0.2, respectively. However, if the neighboring values are close to one another the evaluation of matching degrees should be changed. Intuitively, if c is highly close to b and d, then a higher matching degree for GS1 may be expected.

Assumption of sharp boundaries between elements of attribute domains impose a unique qualification to the corresponding category. In the paper we extend this approach by introduction imprecision to the definition of the group representation. To make it possible one has to apply tools for describing uncertain or imprecise information. One of them is the fuzzy set theory.

Definition 2. *Let X be a universe of discourse. A fuzzy subset A in X is defined as a set of order pairs:*

$$A = \{< x, \mu_A(x) >: x \in X, \mu_A : X \to [0, 1]\} , \tag{5}$$

where $\mu_A(x)$ denotes a membership function of A.

In the presented approach we use the cardinality concept. For a finite fuzzy set A the cardinality is defined as:

$$card(A) = \sum_{x \in X} \mu_A(x) \ , \quad X = \{x_1, x_2, ..., x_N\} \ . \tag{6}$$

If $\mu_A(x) = 1$ or $\mu_A(x) = 0$ for every $x \in X$, the cardinality of A is equal to the number of its elements.

Closeness relationships among elements of the attribute domain are represented by a closeness relation R which is defined as a mapping $DOM(S) \times DOM(S) \rightarrow [0, 1]$, such that

$$R(x_j, x_j) = 1 \text{ and } R(x_j, x_k) = R(x_k, x_j), \text{ where } x_j, x_k \in DOM(S) \ . \tag{7}$$

Let R_i, $i = 1, ..., N$, denotes a closeness relation associated with domain of the attribute S_i. Let $DOM(S_i) = \{s_{i,1}, s_{i,2}, ..., s_{i,M}\}$. Thus, elements of $DOM(S_i)$ are linguistic terms represented by the following fuzzy sets $FS_{i,j}$:

$$FS_{i,j} = \left\{ < x, \mu_{FS_{i,j}}(x) >: x \in DOM(S_i), \ \mu_{FS_{i,j}}(x) : DOM(S_i) \rightarrow [0, 1] \right\} , \tag{8}$$

where $\mu_{FS_{i,j}}(x) = R_i(s_{i,j}, \ x)$, $i = 1, ..., N$; $j = 1, ..., M$. According to (8) one can define fuzzy sets of students $FST_{i,j}$ with corresponding values of attributes:

$$FST_{i,j} = \left\{ < ST, \mu_{FST_{i,j}}(ST) >: ST \in GS, \mu_{FST_{i,j}}(ST) : GS \rightarrow [0, 1] \right\} \ . \tag{9}$$

The membership function of $FST_{i,j}$ is as follows:

$$\mu_{FST_{i,j}}(ST) = \mu_{FS_{i,j}}(ST(S_i)) \ . \tag{10}$$

Definition 3. *Let GS be a group of objects described by N attributes S_i, i =1,..., N, of nominal type, each of which of M different values at most. Let the attribute domains be associated with closeness relations R_i. As the fuzzy group representation FGSR we will consider the set of column vectors $fgsr_i$, i = 1, ..., N of m_i elements, where $M = \max \{m_i : i = 1, ..., N\}$. Elements of $fgsr_i$ represent attribute values of GS objects expressed by means of linguistic terms represented by fuzzy sets (8). Each element of $fgsr_i$, i = 1, ..., N is calculated as likelihood $P_{i,j}$, i = 1, ..., N; j = 1, ..., m_i that objects from GS are characterized by the linguistic term $s_{i,j}$ and is called the support for the respective attribute value in GS:*

$$P_{i,j} = card(FST_{i,j})/card(GS) \ , \tag{11}$$

where $FST_{i,j}$ is a fuzzy set defined by (9).

Let a tuple NST represent a new student. Let us consider a group of students GS represented by $FGSR$. Matching degrees of NST to GS for respective attributes equal to the corresponding elements of $FGSR$. The total matching degree is expressed by formula (4).

Let us assume, that there are NG student groups, then the whole process of recommendation building will take place in the following way:

```
[Input]:A set of $NG$ groups $GS_k$, containing students
        described by $N$ nominal attributes;
        a tuple $NST$ representing a new student;
Step 1: For each group $GS_k, k = 1, 2, ..., NG$ find its
```

```
        representations $FGSR_k$ according td Definition 3;
Step 2: For the student $NST$ find the group $GL_{rec}$
        with the maximal value of the matching degree (4);
Step 3: Recommend $GL_{rec}$ to the student.
```

5 Students Characterized by Learning Styles

For the purpose of the evaluation of recommendations' quality, we will consider a model based on dominant learning styles [4]. We will apply Felder & Silverman [18] model, which has often been indicated as the most appropriate for the use in computer-based educational systems. According to the authors learning styles are described by means of 4 attributes which indicate preferences for 4 dimensions from among excluding pairs: active vs. reflective (L_1), sensing vs. intuitive (L_2), visual vs. verbal (L_3), and sequential vs. global (L_4) or balanced if the student has no dominant preferences. Attribute values belong to the set of odd integers from the interval [-11, 11]. These numbers describe the grade of traits represented by respective attributes. They are determined on the base of the questionnaires which are filled by students. Each student, who filled ILS questionnaire [19], can be modeled by a vector SL of 4 integer attributes:

$$SL = (sl_1, sl_2, sl_3, sl_4), \quad sl_i \in \{-11, -9, -7, -5, -3, -1, 1, 3, 5, 7, 9, 11\} \ . \qquad (12)$$

Negative values of sl_1, sl_2, sl_3, sl_4 mean scoring for active, sensing, visual or sequential learning styles, respectively. Positive values of them indicate scoring for reflective, intuitive, verbal or global learning styles. Values -5,-7 or 5,7 mean that a student learns more easily in a learning environment which favors the considered dimension; values -9,-11 or 9,11 mean that learner has a very strong preference for one dimension of the scale and may have real difficulty learning in an environment which does not support that preference.

For creating of the fuzzy group representation we will define in the interval [-11, 11] fuzzy sets FL_j with the following membership functions:

$$\mu_{FL_1}(x) = -(x+7)/4, \ \text{ if } x \in \{-11, -9\}, \qquad (13)$$

$$\mu_{FL_{12}}(x) = (x-7)/4, \ \text{ if } x \in \{9, 11\} \qquad (14)$$

and for j = 2, 3, ... , 11

$$\mu_{FL_j}(x) = \begin{cases} (x - 2j + 17)/4, & \text{if } x \in \{2j-15, 2j-13\} \\ -(x-2j+9)/4, & \text{if } x \in \{2j-13, 2j-11\} \end{cases} . \qquad (15)$$

Each set FL_j contains exactly one value which fully belongs to it. The membership grades of neighboring values are equal to 1/2. The degree of closeness between fuzzy sets FL_j and FL_{j+1} (referred to as the height of their intersection) also equals 1/2.

Let fuzzy sets FL_j represent linguistic terms $l_{i,j}$ corresponding to values of the attribute L_i. We assume that l_1 corresponds to -11, l_2 corresponds to -9 and in general l_j corresponds to $2j$ - 13. Thus, the domain of L_i is associated with the closeness relation R_i such that

$$R_i(l_{i,j}, l_{i,j+1}) = 1/2, \;\; R_i(l_{i,j}, l_{i,j+k}) = 0 \text{ if } k > 1, \;\; l_{i,j}, l_{i,j+k} \in DOM(L_i) \quad (16)$$

The membership functions of fuzzy sets of students $FSL_{i,j}$ with corresponding values of attributes L_i are as follows:

$$\mu_{FSL_{i,1}}(SL) = \begin{cases} 1, \text{ if } SL(L_i) = -11 \\ 1/2, \text{ if } SL(L_i) = -9 \end{cases}, \quad (17)$$

$$\mu_{FSL_{i,12}}(SL) = \begin{cases} 1, \text{ if } SL(L_i) = 11 \\ 1/2, \text{ if } SL(L_i) = 9 \end{cases} \quad (18)$$

and for j = 2, 3, ... , 11

$$\mu_{FSL_{i,j}}(SL) = \begin{cases} 1, \text{ if } SL(L_i) = 2j - 13 \\ 1/2, \text{ if } SL(L_i) = 2j - 15 \\ 1/2, \text{ if } SL(L_i) = 2j - 11 \end{cases}. \quad (19)$$

According to Definition 3 the fuzzy group representation takes the form of the matrix and may be defined as :

Definition 4. *Let GL be a cluster containing objects with data determined by (12). As the fuzzy group representation we will consider the matrix* $FGLR = [fglr_{i,j}]$, $1 \le i \le 4, 1 \le j \le 12$, *where the columns represent attributes from SL model and the rows values of attributes. Each component of FGLR is calculated as likelihood that students from GL are characterized by the certain linguistic term from SL model and is called the support for the respective SL attribute value in GL:*

$$fglr_{i,j} = card(FSL_{i,j})/card(GL) \;, \quad (20)$$

where $FSL_{i,j}$ *is a fuzzy set with the membership function defined by (17 - 19).*

Let $jmax_i$ denotes the index of the maximal component of $fglr_i$. As the fuzzy group representative we will consider four sets Rep_i, $1 \le i \le 4$, consisting of 3 elements, $Rep_i = \{rep_{i,1}, rep_{i,2}, rep_{i,3}\}$, such that

$$rep_{i,1} = l_1, \;\; rep_{i,2} = l_2, \;\; rep_{i,3} = l_3, \;\; \text{if } jmax_i = 1, \quad (21)$$

$$rep_{i,1} = l_{10}, \;\; rep_{i,2} = l_{11}, \;\; rep_{i,3} = l_{12}, \;\; \text{if } jmax_i = 12 \quad (22)$$

and for $jmax_i$ = 2, 3, ... , 11

$$rep_{i,1} = l_{jmax_i - 1}, \;\; rep_{i,2} = l_{jmax_i}, \;\; rep_{i,3} = l_{jmax_i + 1}. \quad (23)$$

For the new student NSL = (nsl_1, nsl_2 , nsl_3 , nsl_4), and each group GL_k, k =1,..., NG we can define a recommendation error Err_k as follows:

$$err_{k,i} = \begin{cases} 1 & \text{if } nsl_i \notin Rep_i \\ 0 & otherwise \end{cases}, \quad (24)$$

$$Err_k = \sum_{i=1}^{4} err_{k,i}. \quad (25)$$

6 Experiments

The goal of the experiments was to evaluate the performance of the proposed recommendation technique taking into account numbers of considered groups and their sizes. The evaluation was done by comparison of classification results obtained by the system, with belongings to the groups which match students the best taking into account recommendation error defined by (25).

The tests were done for two different datasets of real students' attributes representing their dominant learning styles as was presented in SL model (see (12)). The process of collecting that kind of data was described with details in [2]. The first set contains data of 194 Computer Science students from different levels and years of studies, including part-time and evening courses. That data was used for building groups of similar students. The second set contains data of students, who were to learn together with their peers from the first dataset and whose data was used for testing the recommendation efficiency. The set consists of 31 data of students studying the same master's course of Information Systems in Management.

The groups were created as clusters of disparate structures and sizes, by application of different techniques. For the purpose of the experiments, there were considered clustering schemes built by three well known algorithms: partitioning - K-means, statistical - EM and hierarchical Farthest First Traversal (FFT) [20]. In the case of the first algorithm 2 different distance functions: Manhattan and Euclidean were taken into account. Such approach allows to check the proposed technique for schemes consisting of groups of different similarity degrees and different structures. During the tests, data was divided, by each of the algorithm, into schemes of 3, 6 and 7 clusters to enable comparison of the technique performance depending on the number of groups considered for recommendations. To exclude groups which do not contain members of characteristics similar to the considered student, we assumed that the group cannot be recommended to the student if the support for any of his attribute values is less or equal to the given threshold.

During quantitative analysis recommendation errors Err defined by (25) were calculated and compared. The results showed that the majority of the students obtained the best recommendations. Dependency between clustering schema and the percentage of properly assigned recommendations has not been noticed, however accuracy of the recommendations measured by (25) increased together with the growth of the threshold value. The detailed results of quantitative analysis for different group structures and the threshold equal to 0.5 are presented in Table 1. The first two columns present clustering method and the number of clusters, where KM1 and KM2 denote K-means algorithm with Euclidean and Manhattan distance functions respectively. Next columns show percentage of students: for whom better suggestions can be done (there exists at least one group, for which Err is less than for the recommended one), for whom recommendation error was equal respectively to 0,1,2,3 and 4.

Qualitative analysis of the fuzzy group representations showed that more accurate recommendations were obtained when the biggest support values were concentrated around fuzzy group representatives. Special attention was done to

Table 1. Quantitative analysis for different group structures

Schema	Cl. no	Better ch.	Err=0	Err=1	Err=2	Err=3	Err=4
KM1	3	3.23%	16.13%	25.81%	38.71%	19.35%	0%
	6	6.45%	25.81%	41.94%	16.13%	6.45%	9.68%
	7	12.90%	19.35%	32.26%	38.71%	6.45%	3.23%
KM2	3	25.81%	12.90%	19.35%	45.16%	22.58%	0%
	6	3.23%	22.58%	41.94%	19.35%	12.90%	3.23%
	7	3.23%	19.35%	45.16%	12.90%	19.35%	3.23%
EM	3	3.23%	19.35%	35.48%	29.03%	12.90%	3.23%
	6	9.68%	12.90%	35.48%	25.81%	19.35%	6.45%
	7	6.45%	16.13%	35.48%	35.48%	12.90%	0%
FFT	3	3.23%	12.90%	35.48%	25.81%	22.58%	3.23%
	6	9.68%	19.35%	29.03%	35.48%	6.45%	9.68%
	7	9.68%	22.58%	22.58%	41.94%	9.68%	3.23%

students, for whom recommendation errors were the highest. In the considered data set, for almost all the clustering schemes there was distinguished the same student of the error equal to 4. Similar cases should be indicated as outliers.

7 Concluding Remarks

In the paper, fuzzy logic for building group recommendations for students was considered. The technique was examined in the case of students described by dominant learning styles. Experiments done for datasets of real students and different group structures showed that for the majority of the students the system indicated the best possible choice of colleagues to learn together.

The proposed method of recommendation building can be applied by educators during the process of course management as well as organization of joint activities for student groups of similar features.

Future research will consist in further investigations of the recommendation tool, examination of other attributes and including to recommendations student historical activities as well as making group creating process more dynamic, by adding new learners each time the recommendation is accepted.

References

1. Gonzalez-Rodriguez, M., Manrubia, J., Vidau, A., Gonzalez-Gallego, M.: Improving accessibility with user-tailored interfaces. Appl. Intell. 30, 65–71 (2009)
2. Zakrzewska, D.: Student groups modeling by integrating cluster representation and association rules mining. In: van Leeuwen, J., Muscholl, A., Peleg, D., Pokorný, J., Rumpe, B. (eds.) SOFSEM 2010. LNCS, vol. 5901, pp. 743–754. Springer, Heidelberg (2010)

3. Zakrzewska, D.: Building Group Recommendations in E-Learning Systems. In: Nguyen, N.T. (ed.) Transactions on CCI VII. LNCS, vol. 7270, pp. 144–163. Springer, Heidelberg (2012)
4. Brusilovsky, P., Peylo, C.: Adaptive and intelligent web-based educational systems. International Journal of Artificial Intelligence in Education 13, 156–169 (2003)
5. Christodoulopoulos, C.E., Papanikolaou, K.A.: A group formation tool in an e-learning context. In: 19th IEEE ICTAI 2007, vol. 2, pp. 117–123 (2007)
6. Boratto, L., Carta, S.: State-of-the-Art in Group Recommendation and New Approaches for Automatic Identification of Groups. In: Soro, A., Vargiu, E., Armano, G., Paddeu, G. (eds.) Information Retrieval and Mining in Distributed Environments. SCI, vol. 324, pp. 1–20. Springer, Heidelberg (2010)
7. Masthoff, J.: Group Recommender Systems: Combining Individual Models. In: Ricci, F., et al. (eds.) Recommender Systems Handbook, pp. 677–702. Springer Science+Business Media (2011)
8. Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Rueda-Morales, M.A.: Managing uncertainty in group recommending processes. User Modeling and User-Adapted Interaction 19(3), 207–242 (2009)
9. García, P., Amandi, A., Schiaffino, S., Campo, M.: Evaluating Bayesian networks' precision for detecting students learning styles. Comput. Educ. 49, 794–808 (2007)
10. Hawkes, L.W., Derry, S.J.: Advances in local student modeling using informal fuzzy reasoning. Int. J. Hum.-Comput. St. 45, 697–722 (1996)
11. Vrettaros, J., Vouros, G., Drigas, A.S.: Development of an intelligent assessment system for solo taxonomies using fuzzy logic. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 901–911. Springer, Heidelberg (2007)
12. de Arriaga, F., El Alami, M., Arriaga, A.: Evaluation of Fuzzy Intelligent Learning Systems. In: Méndez-Vilas, A., González-Pereira, B., Mesa González, J., Mesa González, J.A. (eds.) Recent Research Developments in Learning Technologies, FORMATEX, Badajoz, Spain (2005)
13. Lascio, D., Gisolfi, A., Loia, V.: Uncertainty processing in user modeling activity. Information Sciences 106, 25–47 (1998)
14. Monova-Zheleva, M., Zhelev, Y., Mascitti, I.: E-learning, E-practising and E-tutoring: An Integrated Approach. Methodologies and Tools of the Modern (e-) Learning. Information Science and Computing, Supplement to International Journal "Information Technologies and Knowledge" 2(6), 84–90 (2008)
15. Lau, R., Song, D., Li, Y., Cheung, T., Hao, J.: Towards A Fuzzy Domain Ontology Extraction. IEEE Transactions on Knowledge and Data Engineering 21(6), 800–813 (2009)
16. Faziolahtabar, H., Mahdavi, I.: User/tutor optimal learning path in e-learning using comprehensive neuro-fuzzy approach. Educational Research Review 4(2), 142–155 (2009)
17. Xu, D., Wang, H., Su, K.: Intelligent student profiling with fuzzy models. In: HICSS 2002, Hawaii (2002)
18. Felder, R.M., Silverman, L.K.: Learning and teaching styles in engineering ducation. Eng. Educ. 78, 674–681 (1988)
19. Index of Learning Style Questionnaire, `http://www.engr.ncsu.edu/learningstyles/ilsweb.html`
20. Han, J., Kamber, M.: Data Mining. Concepts and Techniques, 2nd edn. Morgan Kaufmann Publishers, San Francisco (2006)

QORECT – A Case-Based Framework for Quality-Based Recommending Open Courseware and Open Educational Resources

Monica Vladoiu, Zoran Constantinescu, and Gabriela Moise

UPG University of Ploiesti, Romania
{monica,zoran}@unde.ro, gmoise@upg-ploiesti.ro

Abstract. More than a decade has passed since the start of the MIT OCW initiative, which, along with other similar projects, has been expected to change dramatically the educational paradigms worldwide. However, better findability is still expected for open educational resources and open courseware, so online guidance and services that support users to locate the appropriate such resources are most welcome. Recommender systems have a very valuable role in this direction. We propose here a hybrid architecture that combines enhanced case-based recommending (driven by a quality model tenet) with (collaborative) feedback from users to recommend open courseware and educational resources.

Keywords: open courseware (OCW), open educational resources (OERs), quality model, case-based reasoning, recommendation system.

1 Introduction

More than a decade has passed since the start of the MIT OCW initiative, which, along with other similar projects, has been expected to change dramatically the educational paradigms worldwide. However, despite the huge opportunities offered by open education, traditional textbooks and readings, and intranet educational resources are still here, dominating the majority of teaching and learning venues of Higher Education institutions even though all students are effectively online. Greater adoption of OERs both within formal and informal education seems to be impeded by four issues: *discoverability, quality assurance, bridging the last mile, and acquisition* [1]. Modern search engines generally do an ill job when searching for educational content because they are not tailored with this purpose, focusing mainly on content and metadata, and, moreover, they lack what it takes to locate the proper educational resource that is suited for a specific user's goal, that builds up on her prerequisites (for example, learner's previous knowledge), and that provide for making the next step towards her goal (e. g. mastering of a certain concept). For the time being, there is no quality assurance mechanism that could provide support for (1) *learners and instructors* in their quest for reaching the most appropriate educational resources for their specific educational needs in any particular context, neither for (2) *faculty or institutions* that are or want to become involved in this movement, and they may be concerned about the challenges or interested in the gains of this process, nor for (3) *developers* who need guidelines for designing and building such educational resources, nor for (4)

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 681–690, 2013.

educational resources' *evaluators* [2, 3, 4]. In many OCW/OER repositories educational content exists only immersed in context and without a significant effort this content cannot be both sorted out from its initial environment (becoming truly reusable and remixable) and entangled within a new educational context, bridging the last mile. Acquisition is also difficult, taking into account all the fears of OCW/OERs providers (faculty, teachers, educational resources designers etc.): lack of credit, of copyright control over derivative works, and so on. Therefore, better findability is expected for open courseware and OERs, so online guidance and services that support users to locate the appropriate ones is beneficial as related work shows [5, 6, 7, 8, 9].

Recommender Systems (RSs) are a sort of information filtering systems that either try to predict whether a particular user would like a given item (*prediction problem*), or try to identify a set of N items that will be of interest to a certain user (*top-N recommendations*). Various kinds of recommendation approaches that rely on various paradigms are available: content-based (item features, user ratings), collaborative (similar ratings from similar users), case-based (content-based case based reasoning), demographic user profiles, knowledge based, and hybrid [10-15]. When using recommender systems in e-educational contexts (some authors call that Technology Enhanced Learning - TEL), the object of recommendation may be a learning resource, a learning activity, a peer learner, a mentor, and so on [7, 8, 9, 16]. Moreover, the recommendation goal is usually complex, e.g. the RS may suggest a set of alternative learning paths throughout a mixture of educational resources, in various forms (learning sequences, hierarchies of interacting learning resources), and the recommendation must be done within a meaningful pedagogical paradigm that reflects user instructional goal, specific interests, the context of use etc., and that helps him accomplish his instructional goal and objectives [16, 17].

In this paper we propose a hybrid approach that combines enhanced case based recommending (driven by a quality model tenet) with (collaborative) feedback from users to recommend OCW and OERs within a unified framework. The structure of the paper is as follows: the next section presents our case-based architecture for recommending OCW and OERs, detailing the quality model and the case-based reasoning process, the third section includes the related work pointing on its main unsolved issues, and the last section shows some conclusions and future work ideas.

2 Case-Based Architecture for Recommending OCW and OERs

In this section we present our approach of a case based recommendation system for OCW/OERs. Case Based Reasoning (CBR) is a very well-known artificial intelligence technique, which has already proven its effectiveness in numerous domains. The fundamental concept in CBR is that similar problems have similar solutions, and, therefore, solving a new problem is done by analyzing the solutions to previous, similar problems [18]. The solutions offered as an outcome of the CBR cycle rely on previous cases stored in the case base, and the system is able to learn continuously by adding new cases to the case base. In its more general form, CBR relies on the *k-Nearest Neighbors* (kNN) algorithm, which core is a similarity function that will be used to find *k* previous cases similar to the new (target) case. Assessing similarity at the case level (or between the target query and candidate case) is based on combining the individual feature level similarities for the relevant features. Usually, a weighted sum metric such as that shown in Eq. 1 is used, in which the similarity between some target query, t and some candidate case c, is the weighted sum of the individual similarities between the corresponding features of t and c, namely t_i and c_i. Each weight encodes the relative importance of a particular feature in the similarity evaluation and each individual feature similarity is calculated according to a similarity function that is defined for that feature, $sim_i(t_i, c_i)$ (shown in Eq. 2 in our case, where $d_i = |c_i - t_i|$). The value of the similarity score is between 0 and 1, and the more the two cases *t* and *c* are similar, the more the similarity score gets closed to 1 [19, 20, 21].

$$similarity(t,c) = \frac{\sum_{i=1}^{n} w_i * sim_i(t_i, c_i)}{\sum_{i=1}^{n} w_i} \quad (1) \qquad sim_i(d_i) = \begin{cases} 1, \ d_i \leq 1 \\ 2 - d_i, 1 \leq d_i \leq 2 \\ 0, \ d_i \geq 2 \end{cases} \quad (2)$$

2.1 The Quality Model

We present briefly here the quality criteria for quality assurance of OCW/OERs, which have been introduced and presented in detail in [2], and put to work and refined further elsewhere [3-4]. These criteria can be applied for assessing quality of both small learning units and entire courseware. They fall within four categories concerned with the quality of the content, of the instructional design, of the technology-related aspects, and with the assessment of the courseware, as a whole. These criteria correspond to the quality characteristics of quality in use, internal and external product quality according to ISO/IEC 25000 SQuaRE standard, and they cover the next user needs: effectiveness, efficiency, satisfaction, reliability, security, context coverage, learnability, and accessibility [2-4]. A very concise presentation of these quality criteria is included in Table 1, which works as a rubric for our quality model (where the scoring meaning is as follows: 0=absence, 1=poor, 2=satisfactory, 3=good, 4=very good and 5=excellent). For the time being the evaluation of OCW/OERs is subjective, being based on many decades of evaluators' experience in Higher Education. However, this seems to be the tendency in other works in this area [4, 22-25].

Table 1. Quality Rubric for Quality Assurance of OCW and OER

Content related criteria	*To what degree an educational resource allows learners to have* ***engaging learning experiences*** *that provide for* ***mastery of the content.***	
	CR1: readability	0-5
	CR2: uniformity of language, terminology, and notations	0-5
	CR3: availability of the course syllabus	0-5
	CR4: comprehensiveness of the lecture notes	0-5
	CR5: modularity of the course content	0-5
	CR6: possibility to select the most suitable learning unit	0-5
	CR7: opportunity to choose the most appropriate learning path	0-5
	CR8: top-down, bottom-up or combined approach	0-5
	CR9: availability of assignments (with or without solutions)	0-5
	CR10: *resource related*: accuracy[1], reasonableness[2], self-containedness[3], context[4], relevance[5], multimedia inserts[6], interactive elements[7], correlation with the entire course[8], links to related readings[9], links to other resources (audio, video etc.)[10]	0-5 x 10

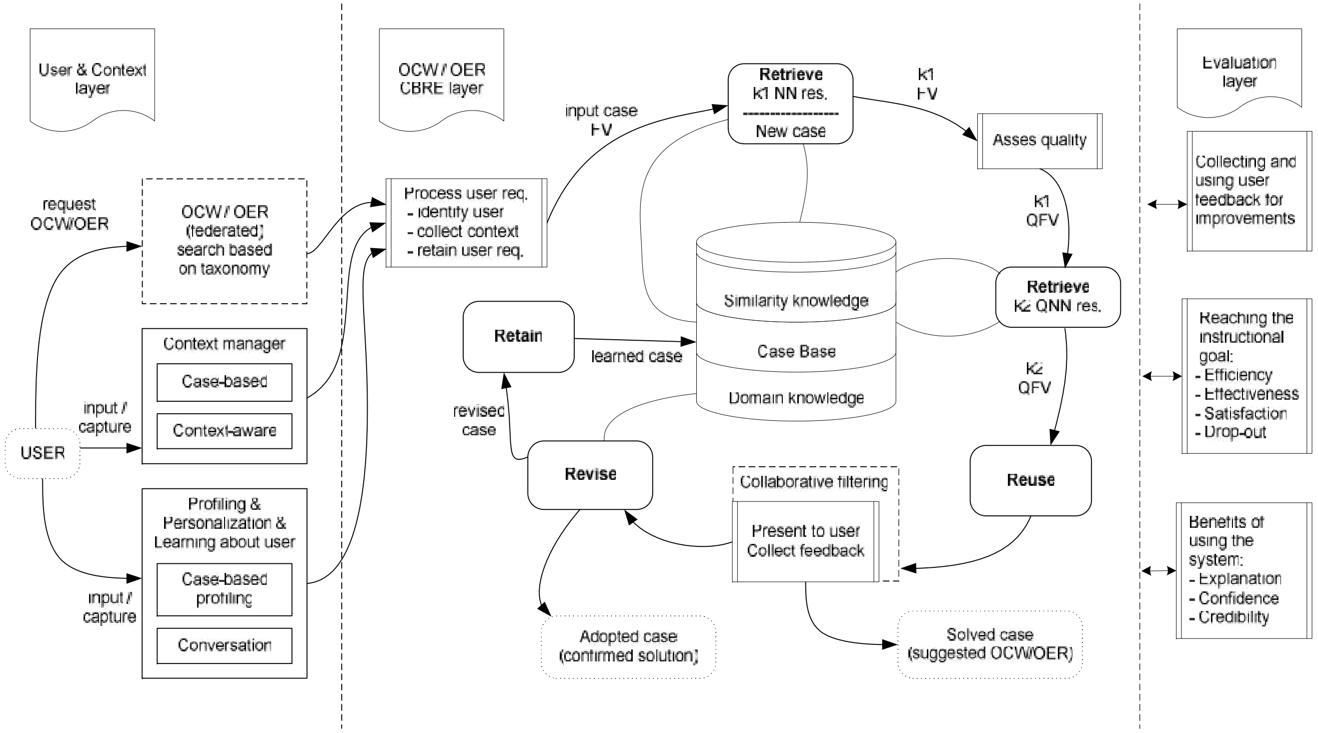
Table 1. (*Continued*)

Instructional design criteria	*Address **instructional design** and other resource's **pedagogical aspects***	
	ID1: goal and learning objectives (outline the material)	0-5
	ID2: learning outcomes (knowledge, skills, abilities, attitudes)	0-5
	ID3: appropriate instructional activities	0-5
	ID4: availability of the evaluation and auto-evaluation means	0-5
	ID5: learning theory	0-5
	ID6: instructional design model	0-5
	ID7: *reflective learning opportunities:* desired outcome of education becomes the construction of coherent functional knowledge structures adaptable to further lifelong learning	0-5
Technology related criteria	*OCW/OERs are expected to **benefit fully from ICT technologies**, to have **user-friendly interfaces**, and to **comply with various standards**.*	
	TR1: conformity with standards for interoperability	0-5
	TR2: compliance with standards for accessibility	0-5
	TR3: extensibility wrt to adding content, activities, and assessments, from a technological viewpoint(developers and learners)	0-5 0-5
	TR4: user interface's basic technological aspects	0-5
	TR5: supporting technology requirements at user's end	0-5
	TR6: prerequisite skills to use the supporting technology	0-5
	TR7: multi-platform capability	0-5
	TR8: supporting tools	0-5
Courseware evaluation criteria	*All major OCW initiatives have become lately more **involved with their learners,** and therefore regular assessment of **OCW effectiveness** and using the results for **further improvements** is essential.*	
	CW1: *courseware overview*: content scope[1] and sequence[2], intended audience[3], grade level[4], periodicity[5] of content updating, author's credentials[6], source credibility[7], multiple-languages[8], instructor facilitation[9] or semi-automated support[10], suitableness for self-study[11], classroom-based[12] study, and/or peer collaborative[13] study, time requirements[14], grading policy[15], instructions on using[16] the courseware, reliability[17], links to other[18] educational resources (readings, OCW, OERs etc.)	0-5 x 18
	CW2: availability of prerequisite knowledge	0-5
	CW3: availability of required competencies	0-5
	CW4:matching the course schedule with learner's own pace	0-5
	CW5: *terms of use/service:* availability of repository policies wrt copyright&licensing issues, security for primary, secondary and indirect users, anonymity, updating and deleting personally identifiable information, age restrictions, netiquette, etc.	0-5
	CW6: freeness of bias and advertising	0-5
	CW7: suitable design and presentation of educational content	0-5
	CW8: *user interface richness (style):* navigational consistency[1], friendliness[2], multimedia[3], interactivity[4], adaptability[5](both to user's needs and context) etc.	0-5 x5
	CW9: providing a formal degree or a certificate of completion	0-5
	CW10: *participatory culture and Web 2.0 facets*: contribution to the content[1], collection of users' feedback[2], collaboration with fellows[3], sharing the development[4]/using experience[5]	0-5 x5

2.2 The QORECT Architecture

The architecture we propose for recommending open courseware and OERs based on a quality model is three layered, as it can be seen in Fig. 1, and it is called QORECT (Quality driven Open educational resource/courseware case-based RECommending Tenet). First layer is *User and Context Layer*, which is dedicated to user's request and to collecting information about the user and about his context. Thus, a user may address a request for specific OCW/OERs that includes her instructional goal, the subject, the material level (graduate, undergraduate, K12 etc.), an indication about her preference for OCW or OERs, an expectation with respect to the resource quality (i. e. more than good), and so on. For the time being the resources are manually collected and inserted within a local pool of resources, but for future versions of the system we intend to include an automatic OCW/OER federated search engine, based on the taxonomy introduced in a previous work [26]. Additionally, she is expected to provide information about the context in which the resource will be used (for example, within a classroom setting, for self-study, either independent or in a learning network etc.) to be processed by the Context Manager. We also foresee for future versions the capability of automatic capturing of context information within a context-aware architecture. Finally, this layer is responsible with processing information about the user (input and capture) for (case-based) profiling, personalization, and for creating opportunities for learning about the user. A conversation module is included here. Automatic capturing of information about the user behavior is also envisaged for future versions.

The second layer, called *OCW-OER Case-Based Recommender Engine*, is dedicated to the recommendation of OCW/OERs, which results from a CBR process. First, the user request is re-constructed from its parts in form of an *input case FV (Feature Vector)*, by including user's descriptive information and user's context, and by retaining user requirements. Then, the first *Retrieve* step in the CBR cycle is activated, and the best k1-NN resources are selected (k1 FVs that describe those resources) based on a simple similarity measure (Eq. 1 and Eq. 2). Further on, the quality of these k1 resources is assessed based on our quality model, and k1 quality-enhanced FVs are obtained. Based on the user's initial quality expectation, k2-NN (k2<=k1) resources are retained, in the second *Retrieve* step, for further processing within the CBR cycle (described by their *Quality Feature Vectors - QFV*). Next, the *Reuse* phase is on, and these resources are presented to the user. If she is happy with the results, then we have "solved the case" for her. Still, we need her feedback for future collaborative resource filtering based on her appreciations of both quality and usefulness of the resource for her instructional goal in a particular context. Her feedback may be collected also in the case she is not content with the recommendation, and further *Revise*-ing is needed, resulting an adapted case that is again presented to her. Finally, the system is able to *Retain* as learned cases the new cases within the revised ones. Within our framework, a case consists of information (and learned knowledge) about the user, about the desired resource (quality included), about the context of use, and on user's feedback.



-

Fig. 1. QORECT – a Case-Based Architecture for Collaborative Recommending OCW and OERs based on a Quality Model

The *Evaluation layer* is concerned with collecting user feedback, and with the evaluation of the system within educational settings from a dual point of view: reaching the instructional goal, and benefiting from using the system. With respect to the instructional aspects, there are several issues to be considered: effectiveness, efficiency, satisfaction, the drop-out rate etc. [16]. Effectiveness may be measured as the total number of completed, visited or studied OCW or OERs during a learning session, while efficiency may be quantified as the total amount of time needed for reaching the instructional goal. Satisfaction reflects the user's contentment with regard to the accomplishment of his instructional goal. If this goal is learning, then the drop-out rate, which shows the number of learners that abandon during the learning phase, is an important indicator of learning process as it is shown in a literature survey [16]. Satisfaction with the recommendations made, explanations supporting these recommendations (quality rubrics to motivate the score, collaborative filtering results), confidence and credibility are specific benefits from using the recommendation system.

2.3 The Case-Based Recommending Process

In this sub-section we present briefly the recommendation process:

```
Step0: process OCW/OER pool → Feature Vectors (FVs)
Step1: process user request:
  - identify (profile of) the user
  - collect the context data (input + automatic capture)
  - retain user requirements (goal, subject, quality expectation, some relevant quality criteria, and so on)
  - construct the input (target) case (as a FV)
Step2: retrieve the k1 most similar resources or create a new case (new case not similar to stored ones)and go to 8
Step3: assess quality for the k1 resources → k1QFV
Step4: retrieve the k2 most similar cases wrt quality expectation and relevant criteria and collaborative rating
Step5: present the k2 resources to the user (reuse); if s/he is content with the suggested OCW/OER → solved case
Step6: collect user feedback (reaching instructional goal, resource usefulness etc.) and collaborative rating
Step7: revise the case (adapted case - confirmed)
Step8: retain the new or adapted case as a learned case
```

3 Related Work

In the TEL domain there is a variety of RSs that propose learning resources to users using several recommending approaches, and the most related to our work –in some particular aspects - will be presented further here. A well-known system that has been proposed for the recommendation of learning resources is RACOFI (Rule-Applying Collaborative Filtering), which combines recommendation based on users' ratings with results of an inference rules based engine that determines associations between the learning resources and their use. However, no assessment of the pedagogical value of the resources or of the recommending engine is available yet [7]. QSIA (Questions Sharing and Interactive Assignments) is a system for sharing, assessing, and recommending learning resources that is used by online communities, which has a rather atypical approach, different from mainstream RSs, by putting users in control of the recommendation process. Thus, s/he may choose on whom to advise or whether to use a collaborative filtering service or not [6, 27]. In [28], recommendations based on NN collaborative filtering of learning objects are performed, but the novel twist regards the multi-attribute algorithms that provide for multi-dimensional user ratings of the learning resources. It is interesting to notice that the same algorithms perform differently depending on the context where the testing takes place. RecoSearch, an engine that combines content, collaboration, collaborative filtering and searching techniques provides for a collaborative infrastructure for authoring, searching, recommending, and presenting Java source code learning objects [29]. A web-based learning system that can both find relevant content on the open Web and adapt it to learners and their situated learning characteristics, based on system's observation on learners and their ratings, is presented in [30]. Pro-active recommending of learning objects that combine content and social aspects, and that is able to adapt to learner's profile and his navigation history, relying on an ontology of topics from programming (as an index of the learning objects) is presented in [31]. An interesting idea is put to practice in [32], where authors propose that the users with greater knowledge have greater weight when computing recommendations than the ones with less knowledge. A totally different approach of obtaining the quality rating of learning objects is proposed in [33]. It contains a hybrid approach that includes content-based and collaborative filtering, and that implements a Markov model to verify the quality evaluation of the learning objects. Their system uses Bayesian Belief Networks to overcome the incompleteness and lack of learning object quality reviews, as well as the differences between evaluations of different reviewers. A work that uses CBR to make personalized recommendation in online learning object repositories is [34]. The authors present their combination of content-based filtering techniques with collaborative filtering mechanisms to be applied to a repository with more than 200 programming examples written in different programming languages. Students are expected to include ratings for the learning objects, both existing and new.

One of these works takes into account very few quality aspects when make recommendations [33], while other considers briefly the pedagogical value and other educational aspects of the resources [30]. The majority of them approach closed repositories that contain very specific instructional resources. Therefore, we hope that our approach will contribute to better recommendations of diverse educational resources (particularly OCW and OERs) for a large variety of users.

4 Conclusions and Future Work

We introduced here our hybrid approach of recommending framework for open OCW and OERs that combines case based recommending based on a quality model with (collaborative) feedback from users. The current stage of the project is as follows: we are developing the first prototype, which collects in a common pool of resources several OCW and OERs (around 10 resources per subject) that are necessary to graduate majoring in Computer Science. Additionally, we are evaluating the quality of these resources using the rubrics presented briefly here. Our first goal is to use this prototype for our Computer Science students both in formal and informal environments and to evaluate and, hopefully, validate the viability of our approach. Further on, we consider automating some activities of our framework: the federated search of resources based on the taxonomy, some quality evaluations, capturing context and knowledge about user etc. to obtain a true adaptive recommender system.

References

1. Kortemeyer, G.: Ten Years Later: Why Open Educational Resources Have Not Noticeably Affected Higher Education, and Why We Should Care, Educase Review online, http://www.educause.edu/ero/article/ten-years-later-why-open-educational-resources-have-not-noticeably-affected-higher-education-and-why-we-should-ca
2. Vladoiu, M.: Quality Criteria for Open Courseware and Open Educational Resources. In: 11th ICWL 2012 Workshops. LNCS, vol. 7697. Springer, Heidelberg (2012)
3. Vladoiu, M., Constantinescu, Z.: Evaluation and Comparison of Three Open Courseware Based on Quality Criteria. In: Grossniklaus, M., Wimmer, M. (eds.) ICWE Workshops 2012. LNCS, vol. 7703, pp. 204–215. Springer, Heidelberg (2012)
4. Moise, G., Vladoiu, M., Constantinescu, Z.: MASECO - Multi-Agent System for Evaluation and Classification of OERs and OCW based on Quality Criteria (in press, 2013)
5. Nicoara, E.S.: The Impact of Massive Online Open Courses in Academic Environments. In: 9th Int. Conf. eLearning and Software for Education. Ed. Universitara, Bucharest (2013)
6. Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H.G.K., Koper, R.: Recommender Systems in Technology Enhanced Learning. In: Kantor, P.B., Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender System Handbook, pp. 387–415. Springer, Berlin (2011)
7. Lemire, D., Boley, H., McGrath, S., Ball, M.: Collaborative Filtering and Inference Rules for Context-Aware Learning Object Recommendation. International Journal of Interactive Technology and Smart Education 2(3), 179–188 (2005)
8. Cechinel, C., Sicilia, M.-A., Sánchez Alonso, S., García Barriocanal, E.: Evaluating Collaborative Filtering Recommendations Inside Large Learning Object Repositories. Information Processing and Management 49(1), 34–50 (2013)
9. Zapata, A., Menéndez, V.H., Prieto, M.E., Romero, C.: A Framework for Recommendation in Learning Object Repositories: An Example of Application in Civil Engineering. Advances in Engineering Software 56, 1–14 (2013)
10. Resnick, P., Varian, H.R.: Recommender Systems. Commun. ACM 40(3), 56–58 (1997)
11. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
12. Burke, R.: Hybrid Web Recommender Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 377–408. Springer, Heidelberg (2007)

13. Burke, R.: Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12(4), 331–370 (2002)
14. Burke, R.: Knowledge-based Recommender Systems. In: Kent, A. (ed.) Encyclopedia of Library and Information Systems, vol. 69(32). Marcel Dekker, New York (2000)
15. Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating Contextual Information in Recommender Systems using a Multidimensional Approach. ACM Trans. Inf. Syst. 23(1), 103–145 (2005)
16. Manouselis, N., Drachsler, H., Verbert, K.: TEL as a Recommendation Context, Recommender Systems for Learning, pp. 21–36. Springer, New York (2013)
17. Buder, J., Schwind, C.: Learning with Personalized Recommender Systems: A Psychological View. Computers in Human Behavior 28, 207–216 (2012)
18. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 7(1), 39–59 (1994)
19. Smyth, B.: Case-Based Recommendation. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 342–376. Springer, Heidelberg (2007)
20. Lee, J.S., Lee, J.C.: Context Awareness by Case-Based Reasoning in a Music Recommendation System. In: Ichikawa, H., Cho, W.-D., Satoh, I., Youn, H.Y. (eds.) UCS 2007. LNCS, vol. 4836, pp. 45–58. Springer, Heidelberg (2007)
21. Văduva, I., Albeanu, G.: Introduction to fuzzy modelling. Ed. of Univ. of Bucharest (2004)
22. Nesbit, J.C., Li, J.Z., Leacock, T.L.: Web-Based Tools for Collaborative Evaluation of Learning Resources. J. of Systemics, Cybernetics and Informatics 3(5), 102–112 (2005)
23. Burgos Aguilar, J.V.: Rubrics to evaluate OERs (2011), `http://www.temoa.info/sites/default/files/OER_Rubrics_0.pdf`
24. ACHIEVE, `http://www.achieve.org`
25. OER Commons, `http://www.oercommons.org`
26. Vladoiu, M., Constantinescu, Z.: A Taxonomy of Opportunities for Searching, Browsing, and Retrieving OCW and OERs (submitted for publication, 2013)
27. Rafaeli, S., Barak, M., Dan-Gur, Y., Toch, E.: QSIA: a Web-based Environment for Learning, Assessing and Knowledge Sharing in Communities. Computers & Education 43(3), 273–289 (2004)
28. Manouselis, N., Vuorikari, R., Van Assche, F.: Simulated Analysis of MAUT Collaborative Filtering for Learning Object Recommendation. In: Proc. of the Workshop on Social Information Retrieval in Technology Enhanced Learning (SIRTEL 2007), pp. 17–20 (2007)
29. Fiaidhi, J.: RecoSearch: a Model for Collaboratively Filtering Java Learning Objects. Int. J. Instruct. Technol. Distance Learning 1(7), 35–50 (2004)
30. Tang, T.Y., McCalla, G.I.: Smart Recommendation for an Evolving e-Learning System: Architecture and Experiment. Int. J. E-Learning 4, 105–129 (2005)
31. Ruiz-Iniesta, A., Jimenez-Diaz, G., Gómez-Albarrán, M.: Recommendation in Repositories of Learning Objects. In: The 9th IEEE International Conference on Advanced Learning Technologies (ICALT 2009), pp. 543–545 (2009)
32. Bobadilla, J., Serradilla, F., Hernando, A.: Collaborative Filtering Adapted to Recommender Systems of E-learning. Knowl.-Based Syst. 22(4), 261–265 (2009)
33. Kumar, V., Nesbit, J., Winne, P., Hadwin, A., Jamieson-Noel, D., Han, K.: Quality Rating and Recommendation of Learning Objects. In: Pierre, S. (ed.) E-learning Networked Environments and Architectures, pp. 337–373. Springer, London (2007)
34. Gomez-Albarran, M., Jimenez-Diaz, G.: Recommendation and Students'Authoring in Repositories of Learning Objects: A Case-Based Reasoning Approach. International Journal of Emerging Technologies in Learning (iJET) 4(1), 35–40 (2009)

The Differential Evolution with the Entropy Based Population Size Adjustment for the Nash Equilibria Problem

Przemyslaw Juszczuk and Urszula Boryczka

Institute of Computer Science, University of Silesia,
ul.Bedzinska 39, Sosnowiec, Poland
{przemyslaw.juszczuk,urszula.boryczka}@us.edu.pl

Abstract. The Differential Evolution (DE) is a simple and powerful optimization method, which is mainly applied to numerical optimization. Many modifications are inclined to use adaptive or dynamic parameters values. One of the most important parameters in the algorithm is the population size. Many existing methods focus only on the decreasing of population size over the time. In this article we propose a new approach based on the entropy of the population. In successive iterations, entropy measure is based on the phenotype of every individual in the population. This new approach is adapted to the problem of finding the approximate Nash equilibrium in n–person games in the strategic form. Finding the Nash equilibrium may be classified as continuous problem, where two probability distributions over the set of pure strategies of both players should be found.

Keywords: Differential Evolution, population size, entropy, ϵ–Nash equilibrium.

1 Introduction

The performance of the Differential Evolution is affected by its control parameters which in turn are depended on the landscape characteristics of the objective function. From these parameters population size seems to be an element, which is given the least attention. Other parameters: mutation and the crossover where deeply analysed in the literature. The adaptive or self–adaptive Differential Evolution algorithms have shown faster and more reliable convergence performance than the classical approach. Main idea of this paper is to use simple, entropy based measure to calculate differences between individuals in the population. In a situation where the differences are significant, population size should be increased. On the other hand, while approaching global optimum, population diversity is decreasing, so the population size should be reduced. For these two situations we use the formula, which allows to calculate new population size. The whole operation is not time consuming, and moreover we are not using any additional operators.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 691–700, 2013.

We use the modified Differential Evolution to solve very difficult problem known as the Nash equilibrium problem. It is one of the main concepts in the game theory. Game theory is a mathematical theory that deals with the features of competitive situations in a formal and abstract way. It emphasises the decision–making processes of the players. Game theory has influenced many fields, including economics (its initial focus) [1], political science [13], and many others. In recent years, its presence in computer science has become very important. Game theory is an integral part of artificial intelligence (AI) [20], e–commerce [7], and other areas of computer science.

Our motivation is to show simple and effective approach that allows dynamic population size adjustment over the time. The vast mojority of existing solutions assume, that population size should be only reduced. The whole optimization process based on the Differential Evolution is in some way random, and this assumption is not consistent with the observations. In order to confirm the effectiveness of the proposed solution, we use the modified Differential Evolution to find multiple approximate Nash equilibria in the n–person strategic game. We also introduce a slight modification, which allows to generate a set of solutions in one single algorithm run.

This paper is organised as follows: first we give short information about related works, next we describe the basic Differential Evolution algorithm and the Nash equilibrium problem. In the section five we explain details of the solution. We end with experiments and some conslusions.

2 Related Works

There are many different modifications of the original DE algorithm. Many of them focus on optimal values for the crossover and mutation parameters. In the article [22] optimal values for the classical set of functions were given. One of the first approaches based on the dynamic parameters was [5], where mutation factor F is calculated on the basis of the formula:

$$F_{t+1} = F_t + (0.5 + F_0)/g, \tag{1}$$

where F_{t+1} is the F value in subsequent iterations, g is the number of the iterations and F_0 is the starting value for the F parameter. D. Zaharie proposed different approach for the F parameter based on the normal distribution. Such modification allows substantially increase the diversity of the population, and prevents the stagnation [10]. Stagnation occurs in the problems with small number of dimensions. The reason for introducing above modification was work on convergence analysis for the evolutionary algorithms presented in the algorithm [6].

One of the first articles on dynamic population size was [21]. DESAP algorithm (Differential Evolution with Self–Adapting Populations) has several modifications including dynamic mutation parameter, crossover parameter and population size. In the article [2] J. Breast used simple approach in which F and CR values were modified randomly. Moreover, population size was set as constant value. In his next work [3], the population size was decreased in time. It is worth mentioning, that the reduction process was rarely used. Modifications based on frequent mutation schema changes were proposed in the article [4]. The vast majority of above modifications and parameter values changes cencentrated on randomly selecting new values. There are no assumptions concerning impact of individual parameters on each other. Only D. Zaharie in the article [23] proposed formula, in which all parameters are linked:

$$2 \cdot F^2 - \frac{2}{NP} + \frac{CR}{NP} = 0, \tag{2}$$

where F is the mutation coefficient, CR is the crossover coefficient and NP is the population size. An approach associated with the random selection of an active mutation strategy has been proposed in the article [8]. Authors assumed, that in each iteration there is equal probability of choosing one of the mutation schemas. This approach seems to be modification of the algorithm with two different mutation schemas described in the article [15]. The similar method assuming the use of more mutation schemas was also proposed in the article [19].

One of the newest and the most efficient approaches related to the parameter adaptation was the JADE algorithm [17]. Values of the CR and the F parameters were selected on the basis of the Cauchy distribution. As it may be seen, population size problem seems to be omitted very often. Some empirical studies on the above subject were given in the article [16] and refers to the impact of population size on the quality of solutions.

3 The Classical Differential Evolution

Differential Evolution (DE) is a stochastic, population–based search strategy developed by Storn and Price in 1995, and deeply studied by Jouni Lampinen, Ivan Zelinka [14,11], and others. Differential Evolution was successfully used in many practical applications: neural network train [12], filter design [18] or image analysis [9]. During the selection stage, three parents are chosen and they generate a single offspring which competes with a parent to determine who passes to the next generation. The DE algorithm begins with the initialization of population P(0) which consist of NP individuals. Over time, as the search progresses, the distances between individuals become smaller, with all individuals converging to the same solution. For the each individual (vector), firstly, its fitness is evaluated. Then the mutation process follows. The pseudocode of the general DE algorithm is presented on alg. 1.

The DE mutation operator produces a trial vector for each individual of the current population by mutating a target vector with a weighted differential. This vector (also called individual) is treated like a standard individual in the basic DE. In DE, mutation step sizes are influenced by differences between individuals of the current population:

$$\forall_i \; \forall_j \; v_{i,j} = s_{r_1,j} + F \cdot (s_{r_2,j} - s_{r_3,j})$$

Algorithm 1. Basic DE algorithm

```
Create the initial population of genotypes $P_0 = \{\boldsymbol{X}_{1,0}, \boldsymbol{X}_{2,0}, ..., \boldsymbol{X}_{n,0}\}$;
Set the generation number $g = 0$;
while stop criterion is not met do
    Compute the fitness function for every genotype in the population
    $\{f(\boldsymbol{X}_{1,g}), f(\boldsymbol{X}_{2,g}), ..., f(\boldsymbol{X}_{n,g})\}$ ;
    Create the population of trial genotypes $V_g$ based on $P_g$;
    Make crossover of genotypes from the population $P_g$ and $V_g$ to create
    population $U_g$;
    Choose the genotypes with the highest fitness function from population $U_g$
    and $P_g$ for the next population;
    $generation = generation + 1$, go to step 4;
```

where $s_{r_1,j}$ is the j-th element of the individual r_1 in the parent population S. The parameter F specifies the strength of impact of the difference vector (between the two genotypes from the population). r_1, r_2, and r_3 are three randomly selected indexes of individuals.

The crossover process consists of the creation of a new individual (offspring) u_i. Some of the elements of the vector u_i is derived from individual p_i and the others from the trial vector v_i. The crossover operation operates both the genotype from the population S and the trial genotype (population V). In the process of crossing over, the parameter $CR\langle 0, 1\rangle$ and the randomly chosen number are used.

$$\forall_i \forall_j u_{i,j} = \begin{cases} v_{i,j} \text{ when } RAND[0,1) < CR, \\ p_{i,j} \text{ in other case.} \end{cases}$$

The fitness function is computed for each individual from population U. The genotype with the lower fitness function value is transferred to the next population. Fitness function equal to 0 is identified as global optimum–Nash equilibrium. To construct the population for the next generation, deterministic selection is used: the offspring replaces the parent if the fitness of the offspring is better than its parent; otherwise the parent survives to the next generation. This ensures that the average fitness of the population does not deteriorate.

4 The Nash Equilibrium Problem

A non–cooperative game in a strategic form consists of a set of players, and, for each player, a set of strategies available to him. There is also the payoff function mapping each strategy profile (i.e. each combination of strategies, one for each player) to a real number that captures the preferences of the player over the possible outcomes of the game:

$$\Gamma = \langle N, \{A_i\}, M\rangle, i = 1, 2, ..., n$$

where:

- $N = \{1, 2, ..., n\}$ is the players set;
- $\{A_i\}$ is the finite set of strategies for the i-th player with m-strategies;
- $M = \{\mu_1, \mu_2, ..., \mu_n\}$ is the finite set of the payoff functions.

Next we define the strategy profile $a = (\boldsymbol{a_1}, ..., \boldsymbol{a_n})$ for all players. Moreover:

$$a_{-i} = (\boldsymbol{a_1}, ..., \boldsymbol{a_{i-1}}, \boldsymbol{a_{i+1}}, ..., \boldsymbol{a_n}),$$

will be the strategy profile excluding i-th player. Mixed strategy for the i-th player will be denoted as:

$$\boldsymbol{a_i} = (P(a_{i_1}), P(a_{i_2}), ..., P(a_{i_m})),$$

where $P(a_{i_1})$ will be probability of chosing strategy 1 by the player i.

Nash equilibrium is a strategy profile such that no deviating player could achieve the payoff higher than the one that the specific profile gives him:

$$\forall_i, \forall_j \ \mu_i(a) \geq \mu_i(a_{i_j}, a_{-i}),$$

where i is the i-th player, j is the number of the strategies for given player, $\mu_i(a)$ is the payoff for the i-th player for the strategy profile a and $\mu_i(a_{i_j}, a_{-i})$ is the payoff for the i-th player using strategy j agaist profile a_{-i}.

5 Proposed Approach

Recently it was shown, that any random game has Nash equilibrium with support equal $\ln n$, where support is the number of the active strategies for the player. Modification proposed in this article uses above asumption to sample the player strategy set. Subset of strategies for every player is called the candidate solution. This candidate solution is optimized on the basis of the proposed algorithm ADE (Adaptive Differential Evolution). Pseudocode of this algorithm is presented in the alg. 2.

The first stage of the algorithm (rows 2 and 3) allow for preliminary estimate quality of the solutions for both players. Solution is considered as the approximate Nash equilibrium if estimated value is lower than the fixed ϵ value. Second stage are rows 5 to 13. It is the main part of the algorithm. The population size update is performed in the row 14. After the first few hundreds of the iterations (the first stage of the algorithm) it is possible to reject poor solutions, in which the ϵ value is greater than the value determined at the beginning of the algorithm.

Complexity of the proposed algorithm is equal to $O(c{\cdot}\, G{\cdot}\, D^2)$ where G is the number of iterations, D is the problem dimension and c is a number of the considered strategy subsets (($c \ll G$).

Every considered strategy subset, in which ϵ value after g iterations is greater than the fixed value may be rejected and described as unacceptable solution. On the other hand, ϵ value lower than the fixed value is solved by the adaptive Differential Evolution. The fitness function for the algorithm consists of two parts, where first part is given as follows:

Algorithm 2. The ADE algorithm

```
for Number of the considered subsets of the strategies do
    test current subset $a$;
    if estimated value $< \epsilon$ then
        Create initial population $S$ with $NP$ individuals;
        for Number of the ADE iterations do
            for Population size do
                Calculate fitness function;
                Create trial individual $\boldsymbol{U_i}$ ;
                Create child individual $\boldsymbol{V_i}$ ;
                if Fitness function value $\boldsymbol{V}_i \geq \boldsymbol{S}_i$ then
                    $\boldsymbol{S}_i^{t+1} = \boldsymbol{V}_i^t$;
                else
                    $\boldsymbol{S}_i^{t+1} = \boldsymbol{S}_i^t$;
            if gen mod $\frac{gen}{10} = 0$ then
                Update $NP$
    Select new subset of strategies $a$;
Solutions = the set of $\epsilon$–Nash equilibria;
```

$$f_1 = \max\{\max\{u_1(a_{11}, a_{-1}), ..., u_1(a_{1j}, a_{-1})\} - \\ u_1(a), ..., \max\{u_i(a_{i1}, a_{-i}), ..., u_i(a_{ij}, a_{-i})\} - u_i(a)\}$$

where $u_i(a_{ij}, a_{-i})$ is the payoff for the i-th player using strategy j. Second necessary condition is to check, if sum of probabilities for every player is equal to one. The second condition may be described as follows:

$$f_2 = \sum_{i=1}^{n} |1 - \sum_{j=1}^{m} P(a_{ij})|$$

where f_1 means maximal deviation from the optimal strategy–denoted as the worst solution from all players. This value is also denoted as the ϵ value (ϵ Nash equilibrium means the approximate Nash equilibrium). Sum of this two above functions gives the fitness function, which is represented by the formula:

$$f = f_1 + c\cdot f_2,$$

where c is the constant value. Only after the fulfillment of the condition f_2, the f_1 may be minimized.

Entropy used in the algorithm is so called probability distribution entropy and it is denoted as $H(\cdot)$. It concerns the diversity of the phenotype values in the whole population. The entropy measure is given as follows:

$$H(g) = ln(\sigma\cdot \sqrt{2\cdot \Pi\cdot e}), \tag{3}$$

where:

$H(g)$ - entropy value in the iteration g,

σ - the standard deviation for the phenotype values.

Initial position for all individuals in the search space is calculated on the term of the uniform distribution. In successive iterations, phenotype values for the whole population begins to resemble the normal distribution. In this case, small part of the population will be characterized by a very good adaptation. At the same time, the value for some individuals will be very small. The remaining individuals phenotype will be between these two values. This indicates the possibility of using the normal distribution entropy. The above assumption provides variability of the population size based on the actual phenotype values. Moreover, the population size is limited by the maximal value NP_{max} and the minimal value NP_{min} equal to the $\frac{NP_{max}}{3}$.

6 Experiments

The aim of this research work is to determine if the algorithm with the dynamic population size based on the entropy measure is capable to solve problem of finding the approximate Nash equilibrium in the n–person games. Unfortunately because of space limitation, we are not able to present the complete parameter tuning. In our experiments for the basic DE we used the following values:

- population size equal to the product of the players and strategies;
- basic mutation schema with $F = 0.7$;
- binomial crossover with $CR = 0.5$;

All parameters for the ADE are calculated on the term of the current population. Initial parameters where the same as above. Minimal ϵ value was set to the 0.4. For the ADE algorithm, number of active strategies per player were equal to 2, 3 and 4. In some experiments, implementation of the GNM algorithm from the GAMBIT package was used.

In the table 1 we can see minimal, maximal, average, median ane standard deviation for the two algorithms. Table was divided into two parts. First part contains results for the basic Differential Evolution (DE), and the second part contains results for the adaptive DE with three active strategies. All presented results apply only to the games, where solution was found in less than 300 seconds (3–players games) or 600 seconds (4 and 5–players games). Moreover, only time of the first solution was calculated. On the fig. 1 we can see box plots for the GNM algorithm. DE and ADE algorithms give similar results in every execution, so the standard deviation value is very small.

Each presented algorithm for every game was run 30 times. It should be mentioned, that GNM algorithm wasn't capable to find solution in every run. Moreover, with the increase of the size of the problem, number of satisfactory solutions was decreasing. In the fig. 2 we can see modified dispersion chart which allows to specify a range of generated results. The accumulation of the points in the bottom of the chart means that solution was found in the allowed time.

Table 1. Computing time–in seconds. DE and ADE algorithms (min–minimum, max–maximum, avg–average, med–median and std–standard deviation)

game	DE					ADE				
	min	max	avg	med	std	min	max	avg	med	std
3p10s	2,38	2,4	2,38	2,38	0,01	1,25	1,42	1,34	1,34	0,05
3p12s	5,85	5,88	5,87	5,87	0,01	2,36	2,63	2,45	2,41	0,09
3p14s	12,84	12,87	12,86	12,86	0,01	4,45	5,16	4,61	4,52	0,21
3p15s	18,35	18,57	18,4	18,37	0,07	5,99	6,82	6,26	6,19	0,28
3p17s	24,86	24,96	24,89	24,89	0,02	10,33	12,12	10,79	10,63	0,54
3p19s	54,38	55,43	54,5	54,42	0,26	15,49	16,68	15,85	15,92	0,38
3p20s	157,8	158,91	158,09	158,02	0,27	20,25	22,88	21,17	20,76	1,06
4p5s	0,73	0,94	0,83	0,82	0,04	1,12	1,85	1,37	1,25	0,28
4p7s	4,46	4,56	4,51	4,52	0,03	6,13	6,65	6,29	6,24	0,2
4p9s	20,06	20,37	20,21	20,23	0,1	22,57	23,73	23,01	22,98	0,41
4p10s	39,14	40,84	39,72	39,57	0,39	39,56	40,94	40,11	39,98	0,53
4p12s	124,83	125,46	125,08	125,08	0,16	102,75	107,46	104,81	104,8	1,54
5p5s	28,07	29,05	29,03	29,03	0,01	22,66	25,13	23,69	23,62	0,7
5p6s	106,84	107,51	107,08	107,04	0,18	57,87	68,24	62,47	62,28	3,69
5p7s	201,9	204,75	203,38	203,43	0,86	170,84	190,62	177,83	176,3	6,57

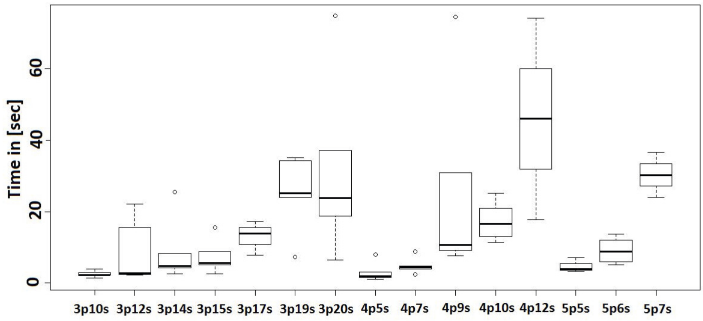


Fig. 1. Boxplot for the solution generating time–random games (the GNM algorithm)

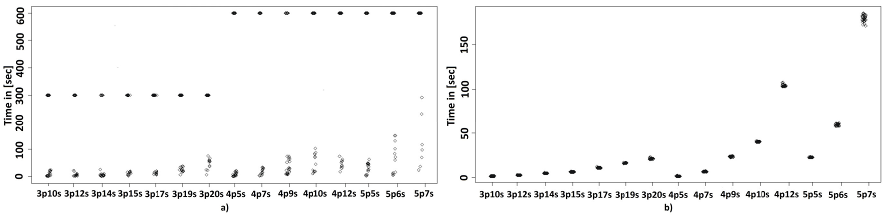


Fig. 2. Dispersion chart for the solution generating time, a) GNM algorithm, b) ADE algorithm withm 3 active strategies

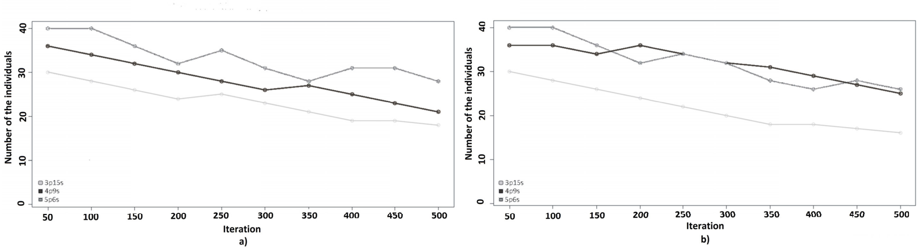


Fig. 3. Dynamic population size, a) ADE with 2 active strategies; b) ADE with 3 active strategies

Dynamic population size adjustment based on entropy allows for a significant reduction of the function evaluations and leads to decreasing computation time. Over the time, population diversity decreases. Above observation allows the reduction of the population withoud losing its diversity. The following study shows games with: 3, 4 and 5 players. On fig. 3 we can see the population size for the ADE algorithm (three example games). The chart presents only first 500 iterations, where the most visible changes are presented. Population size is modified every 50 iterations, and it is clear that the number of individuals is gradually reduced until it reaches a certain established minimum. The effectiveness of this approach is easily noticed in the case of games with 4 and 5 players.

7 Conclusions

After the analysis of the test results of experiments with the new entropy based population size the following conclusions may be reached. Proposed approach allows not only to decrease the population size over the time, but also reacts in case, where diversity population is very high, and there is need of extra individuals. The Adaptive Differential Evolution gives very good and repeatable results in the problem of finding Nash equilibria in n–person games, and moreover it is able to find sets of solutions in one algorithm run. This approach also shows a good scalability–we tested games with 5 players and 7 strategies per players and results indicate that algorithm seems to be also effective for more complicated problem. ADE seems to be good alternative for existing algorithms like GNM.

References

1. Aubin, J.: Mathematical Methods of Game and Economic Theory. North-Holland Publ. Co., New York (1979)
2. Boskowic, B., Brest, J., Greiner, S.: Selfadapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems. IEEE Transactions on Evolutionary Computation 10(6), 646–657 (2006)
3. Boskowic, B., Brest, J., Greiner, S., Maucec, M.S.: Performance Comparison of Self-adaptive and Adaptive Differential Evolution Algorithms. Soft Computation: A Fusion of Foundations, Methodologies and Applications 11(7), 617–629 (2007)

4. Brest, J., Maucec, M.S.: Self-adaptive Differential Evolution Algorithm Using Population Size Reduction and Three Strategies. Special Issue on Scalability of Evolutionary Algorithms and Other Metaheuristics for Large-Scale Continuous Optimization Problems 15, 2157–2174 (2011)
5. Chang, C.S., Du, D.: Differential Evolution Based Tuning of Fuzzy Automatic Train Operation for Mass Rapid Transit System. IEE Proceedings of Electric Power Applications 147, 206–212 (2000)
6. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Professional (1989)
7. Griss, M., Letsinger, R.: Games at Work-Agent-Mediated E-Commerce Simulation. In: Proceedings of the Fourth International Conference on Autonomous Agents (2000)
8. Huang, Q.A.K., Suganthan, V.L.: Differential evolution algorithm with strategy adaptation for global numerical optimization. IEEE Transactions on Evolutionary Computation 13, 398–417 (2009)
9. Kasemir, K.U.: Detecting ellipses of limited eccentricity in images with high noise levels. Image and Vision Computing 21(7), 221–227 (2003)
10. Lampinen, J., Zelinka, I.: On stagnation of the differential evolution algorithm. In: International Conference on Soft Computing, pp. 76–83 (2000)
11. Lampinen, J., Zelinka, I.: Mixed variable non-linear optimization by differential evolution. In: Proceedings of Nostradamus (1999)
12. Magoulas, G.D., Vrahatis, M.N., Androulakis, G.S.: Effective backpropagation training with variable stepsize. Neural Netw. 10, 69–82 (1997)
13. Ordeshook, P.: Game theory and political theory. Cambridge University Press (1986)
14. Price, K., Storn, R., Lampinen, J.: Differential evolution: A practical approach to global optimization. Springer (2005)
15. Qin, A.K., Suganthan, P.N.: Self-adaptive differential evolution algorithm for numerical optimization. In: Proceedings of IEEE Congress on Evolutionary Computation, vol. 2, pp. 1785–1791 (2005)
16. Rammohan, M.: Empirical study on the effect of population size on Differential evolution Algorithm, Evolutionary Computation. In: CEC 2008, pp. 3663–3670 (2008)
17. Sanderson, A.C., Zhang, J., Meng-Hiot Lim, Y.-S.O.: Adaptive Differential Evolution - A Robust Approach to Multimodal Problem Optimization. Springer (2009)
18. Storn, R.: Differential evolution design of an iir-Filter. In: IEEE International Conference on Evolutionary Computation, ICEC 1996, pp. 268–273 (1996)
19. Tang, K., Yang, Z., Yao, X.: Self-adaptive differential evolution with neighborhood search. In: Proceedings of IEEE Congress on Evolutionary Computation, pp. 1110–1116 (2008)
20. Tennenholtz, M.: Game Theory and Artificial Intelligence. In: d'Inverno, M., Luck, M., Fisher, M., Preist, C. (eds.) UKMAS 1996-2000. LNCS (LNAI), vol. 2403, pp. 49–58. Springer, Heidelberg (2002)
21. Teo, J.: Exploring dynamic self-adaptive populations in differential evolution. Soft Computation: A Fusion of Foundations, Methodologies and Applications 10, 673–686 (2006)
22. Zaharie, D.: Critical Values for the Control Parameters of Differential Evolution Algorithms. In: International Mendel Conference on Soft Computing (2002)
23. Zaharie, D.: Critical Values for the Control Parameters of Differential Evolution Algorithms. In: International Mendel Conference on Soft Computing (2002)

Dynamic Version of the ACDT/ACDF Algorithm for H-Bond Data Set Analysis

Jan Kozak and Urszula Boryczka

Institute of Computer Science, University of Silesia, Będzińska 39, 41–200 Sosnowiec, Poland
{jan.kozak,urszula.boryczka}@us.edu.pl

Abstract. This article is devoted to the new application of the ACDT/ACDF algorithms. In this work we distinguish ant colony optimization and join it with decision tree construction algorithms, the proposed approach builds more stable decision forests. Additionally, we would like to mention that it is possible to analyze the overloaded data sets. Several methods are proposed in this study, each considered different pseudo-samples from training data sets. We combine ideas from ACO, Boosting and Random Forests. We show that our algorithms perform comparable to common approaches. Moreover, we demonstrate the suitability of our method to H-bonds detections in protein structures.

Keywords: Ant Colony Optimization, Ant Colony Decision Trees, Ant Colony Decision Forest, Decision Forest, Molecular Dynamics, Hydrogen Bond.

1 Introduction

Our goal is to train ants in ACDT/ACDF approaches to predict the stability of H-bonds in every protein. This is the reason to design an algorithm that analyzes difficult data sets characterized by irregularly numerous decision classes. As an effect of algorithm performance an improvement of classification accuracy concerning objects, belonging to this specific decision class, should be obtained, this improvement ought not to influence the precision of the classification.

Data mining and machine learning have been the subject of increasing attention over the past 30 years. Ensemble methods, popular in machine learning and pattern recognition, are learning algorithms that construct a set of many individual classifiers, called base learners, and combine them to classify new data points or samples by taking a weighted or unweighted vote of their predictions. It is now well-known that ensembles are often much more accurate than the individual classifiers that make them up. The success of ensemble approaches on many benchmark data sets has raised considerable interest in understanding why such methods succeed and identifying circumstances in which they can be expected to produce good results. This article provide a summary of widely used heuristic methods and their modifications used in different benchmark problems and to identify its development of the random forests in reference to Ant Colony

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 701–710, 2013.

Decision Tree algorithm. Our goal is to design a new algorithm for construction of decision forests, where we obtain better accuracy of classification, especially in case of difficult data sets making explicit reference to Ant Colony Optimization.

This paper is organized as follows: section 2 reviews the basic concept of decision trees. Section 3 is devoted to decision forests. Section 4 and 5 propose the ant colony optimization to tackle the decision trees and forest construction. In section 6, we present our experiment schemes and show the results. We also give a simple analysis of our experimental outcomes. Finally, section 7 presents the conclusion.

2 Decision Trees

One of the most efficient and widely applied learning algorithms search the hypothesis (solution) space consisting of decision trees [11,13]. The term hypothesis is understood as a combination of attribute values which determine the way to undertake a specific decision. A decision tree learning algorithm searches the space of such trees by first considering trees that test only one attribute and making an immediate classification. Then they consider expanding the tree by replacing one of the leaves by a test of the second attribute. Various heuristics are applied to choose which test to include in each iteration and when to stop growing the tree [5]. The evaluation function for decision trees will be calculated according to the following formula:

$$Q(T) = \phi \cdot w(T) + \psi \cdot a(T, P) \tag{1}$$

where: $w(T)$ – the size (number of nodes) of the decision tree T; $a(T, P)$ – the accuracy of the classification samples from a test set P by the tree T; ϕ and ψ – constants determining the relative importance of $w(T)$ and $a(T, P)$.

Constructing optimal binary decision trees is an NP–complete problem, where an optimal tree is one which minimizes the expected number of tests required for identification of the unknown samples, as shown by Hyafil and Rivest in [10]. Classification And Regression Tree (CART) approach was developed by Breiman et al. in 1984 [6] .

Twoing criterion, firstly proposed in CART, will search for two classes that will make up together more then 50% of the data. Twoing splitting rule maximizes the following change-of-impurity measure which implies the following maximization problem for nodes m_l, m_r:

$$\underset{a_j \le a_j^R, j=1,\ldots,M}{\arg\max} \left(\frac{P_l P_r}{4} \left[\sum_{k=1}^{K} |p(k|m_l) - p(k|m_r)| \right]^2 \right), \tag{2}$$

where: $p(k|m_l)$, $p(k|m_r)$ – the conditional probability of the class k provided in node m_l, m_r; P_l, P_r – the probability of transition samples into the left or right node m_l, m_r; K – number of decision classes; a_j – j–th variable, a_j^R is the best splitting value of variable a_j.

3 Decision Forests

A decision forest is a collection of decision trees [5,7]. We defined the decision forest by following formula:

$$DF = \{d_j : X \to \{1, 2, ..., g\}\}_{j=1,2,...,J}, \tag{3}$$

where J is a number of decision trees j $(J \geqslant 2)$.

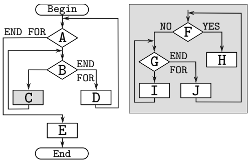
In decision forests, predictions of decision trees are combined to make the overall prediction for the forest. Classification is done by a simple voting. Each decision tree votes on the decision for the sample and the decision with the highest number of votes is chosen. The classifier created by a decision forest *DF*, denoted as $dDF : X \to 1, 2, ..., g$, uses the following voting rule:

$$dDF(x) := \arg\max_k N_k(x), \tag{4}$$

where k a decision class, such that $k \in \{1, 2, \ldots, g\}$; $N_k(x)$ is the number of votes for the sample $x \in X$ classification in to class k, such that $N_k(x) := \#\{j : d_j(x) = k\}$.

4 Ant Colony Decision Trees Algorithm

Ant Colony Optimization (ACO) approach has been successfully applied to many difficult combinatorial problems. Ant Colony Decision Trees (ACDT) algorithm is the first ACO adaptation to the task of rule induction and constructing decision trees, but also rule induction approach – Ant-Miner [2,12].



A- for each colony of ants
B- for each angent-ant
C- build tree
D- update pheromone trail
E- the decision tree
F- the completed tree
G- for all samples and attributes
H- calculate the accuracy rate
I- select the division
J- calculate the probability of choosing the appropriate split

Fig. 1. Construction the tree by ACDT algorithm

In each ACDT step an ant chooses an attribute and its value for splitting the samples in the current node of the constructed decision tree. The choice is made according to a heuristic function and pheromone values. The heuristic function is based on the Twoing criterion (equ. (2)), which helps ants select an attribute-value pair which well divides the samples into two disjoint sets, i.e. with the intention that samples belonging to the same decision class should be put in the same subset. The best splitting is observed when similar number of samples exists in the left subtree and in the right subtree, and samples belonging to the same decision class are in the same subtree. Pheromone values indicate the best way (connection) from the superior to the subordinate nodes – all possible combinations are taken into account.

The diagram of the proposed algorithm is presented in Fig. 1. As mentioned before, the value of the heuristic function is determined according to the splitting rule employed in CART approach (see formula (2)). The probability of choosing the appropriate split in the node is calculated according to a classical probability used in ACO [9]:

$$p_{i,j} = \frac{\tau_{m,m_{L(i,j)}}(t)^{\alpha} \cdot \eta_{i,j}^{\beta}}{\sum_{i}^{a} \sum_{j}^{b_i} \tau_{m,m_{L(i,j)}}(t)^{\alpha} \cdot \eta_{i,j}^{\beta}}, \tag{5}$$

where:

$\eta_{i,j}$ – a heuristic value for the split using the attribute i and value j,
$\tau_{m,m_{L(i,j)}}$ – an amount of pheromone currently available at time t on the connection between nodes m and m_L, (it concerns the attribute i and value j),
α, β – the relative importance with experimentally determined values 1 and 3, respectively.

The initial value of the pheromone trail is determined similarly to the Ant–Miner approach and depends on the number of attribute values. The pheromone trail is updated (6) by increasing pheromone levels on the edges connecting each tree node with its parent node:

$$\tau_{m,m_L}(t+1) = (1-\gamma) \cdot \tau_{m,m_L}(t) + Q(T), \tag{6}$$

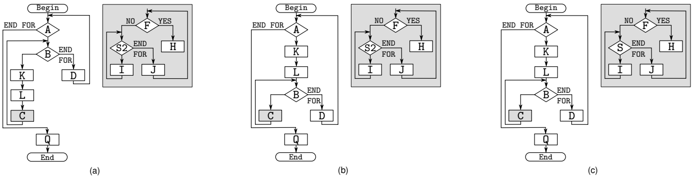
where $Q(T)$ is a quality of the decision tree (see formula (1)), and γ is a parameter representing the evaporation rate, equal to 0.1.

5 Ant Colony Decision Forest

A computational problem arises when the proposed algorithm cannot guarantee to find the best hypothesis within hypotheses space. In ACO and Random Forest approaches, the task of finding the suitable hypothesis that best fits the training data is computationally intractable, so more sophisticated method should be employed in this situation. An algorithm ACDF is based on two approaches: Random Forest and ACDT. The ACDF algorithm can be applied for difficult data sets analysis by adding randomness to the process of choosing which set of features or attributes may be distinguish during the construction decision trees [3].

In case of the ACDF, agent-ants create a collection of hypotheses in random manner complying the threshold or rule to split on. The challenge is to introduce a new random subspace method for growing collections of decision trees – it means that agents-ants can create the collection of hypotheses from the hypothesis space using random-proportional rules. At each node of the tree agent-ant choose from the random subset (random pseudo-samples) of attributes and then constrain the tree-growing hypothesis to choose its splitting rule from among this subset. Because of the re-labeled randomness proposed in our approach we resign from the different subsets of attributes chosen for each agent-ant or colony to favor of greater stability of the undertaken hypotheses. It is a consequence of the proposition firstly used in Random Forest.

The ACDT approach that suffer from the representational problem is said to have a good diversity in (random pseudo-samples) training and testing samples and balance in decision making. ACDF is characterized as an algorithm with high diversity, because agents-ants make a cascade of choices consist of attribute and value choosing at each internal node in the decision tree (for creating a special hypothesis). Consequently, ensembles of decision tree classifiers perform better than individual decision trees. It is due to the independently performed exploration/exploitation the subspace of hypotheses.



A- for each colony of ants B- for each angent-ant C- build tree
D- update pheromone trail F- the completed tree H- calculate the accuracy rate
I- select the division J- calculate the probability of choosing the appropriate split
K- create the pseudo-samples L- create the subset of attributes Q- the decision forest
S- for pseudo-samples and subset of attributes S2- for pseudo-samples and all attributes

Fig. 2. Diagram of the construction ACDF forest

6 Experiments

Evaluation of the performance behavior of ACDT has been presented in Tab. 2 with 100 executions for 15 generations consisted of 10 ants. Appropriate results obtained in ACDF performance are presented in Tab. 3 and 4 with 30 executions for 30 generations consisted of 25 ants (with arithmetical average of presented values). For both algorithms (ACDT and ACDF), values of the parameters (including shown above) have been determined experimentally. The best values for each of the approaches are established independently.

Both situations are analyzed in the context of standard observations, but the small values of the analyzed characteristic cause the exception those values in presented tables. In the presented approaches the twoing criterion as well as Error-Based Pruning procedure have been used. The parameters values employed in ACO are established in the way firstly presented in [1,3].

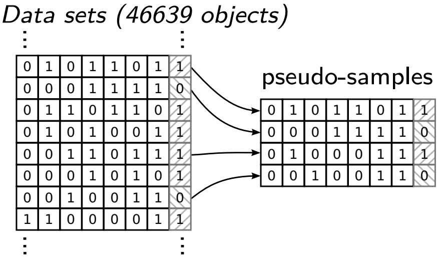
6.1 Data Set – Characteristic of Dynamic, Testable Environment

Experiments have been conducted on the real data sets concerning H-bonds [8]. After discretization data sets consist of objects described by 35 conditional attributes (each of 4 values) and one decision attributes (two values) three data sets: training, test and clear are created for further examination (the presentation of that classification is obtained in Tab. 1).

The new approach with specific pseudo-samples creation is presented. The goal is to obtain better classification accuracy of objects belonging to the "0" decision class, with minimal cost function. This relation of high classification accuracy

Table 1. Data set used in experimental study

Training set	46 639	objects
	4,5%	objects represents the decision class 0
	95,5%	objects represents the decision class 1
Test set	126 151	objects
	5,0%	objects represents the decision class 0
	95,0%	objects represents the decision class 1
Clean set	111 143	objects
	6,1%	objects represents the decision class 0
	93,9%	objects represents the decision class 1



0	1	0	1	1	0	1	1
0	0	0	1	1	1	1	0
0	1	1	0	1	1	0	1
0	1	0	1	0	0	1	1
0	0	1	1	0	1	1	1
0	0	0	1	0	1	0	1
0	0	1	0	0	1	1	0
1	1	0	0	0	0	1	1

0	1	0	1	1	0	1	1
0	0	0	1	1	1	1	0
0	1	0	0	0	1	1	1
0	0	1	0	0	1	1	0

Fig. 3. Way of construction pseudo-samples

of objects from the "1" class and the small deterioration of the precision in the second "0" class, is used in cost function.

In order to tackle the problem of such data, we divide the data set into pseudo-samples (see Fig. 3). We explain the process of division with more details as follows. We choose objects to the pseudo-samples in accordance to the number of objects belonging to the decision classes which must be the same (in two cases). Finally, the prepared data sets have been used in the following stages of analysis in accordance to rules presented below:

- Construction decision trees: pseudo-samples.
- Testing of the trees/forest: test set.
- Result: clean set.

6.2 Dynamic Version of ACDT Algorithm

In order to find the relationship between the number of object in the training set and the quality of the result, we conduct the consecutive analysis the construction of the decision trees can be demonstrate by the changeable pseudo-samples employment. To improve the performance of our algorithm, we use the following approaches:

- **Learn 46600** – all objects from the training set (46639 objects) are used, where 4,5% belongs to the decision class "0", and 95,5% to class "1".
- **Learn 3300** – one pseudo-samples consisting of randomly chosen (draw without replacement) about 3300 objects, where 60% belongs to the decision class "0", a 40% to class "1".
- **Learn 4150** – one pseudo-samples consisting of randomly chosen (draw without replacement) about 4150 objects, where 60% belongs to the decision class "0", a 40% to class "1".

- **Learn 8700** – one pseudo-samples consisting of randomly chosen (draw without replacement) about 8700 objects, where 22% belongs to the decision class ”0”, a 78% to class ”1”.
- **Learn 3300, 4150 and 8700** – dynamic algorithm ACDT in with randomly chosen pseudo-samples are employed (one by one: **Learn 3300**, **Learn 4150** and **Learn 8700**).
- **Learn 3300, 4150 and 8700 V2** – dynamic algorithm ACDT with three pseudo-samples sequential used with greater impact of pruning.
- **Learn 15 Sets** – dynamic algorithm ACDT in which 15 pseudo-samples with 3200 randomly chosen pseudo-samples are employed (draw without replacement), where 50% belongs to the decision class ”0”, a 50% to class ”1”.
- **Learn 15 Sets V2** – dynamic algorithm ACDT in which 15 pseudo-samples with 3200 randomly chosen pseudo-samples are employed (draw without replacement). It contains different number of objects representing decision classes ”0” and ”1”, respectively: 60–40%, 55–45%, 50–50%, 45–55%, 30–60%.

6.3 Pseudo Boosting Version of ACDF Algorithm

New version of ACDF algorithm in the context of the performance behavior is similar to the idea of Boosting [14]. Each created pseudo-sample is based on the weighted draw so that objects from the smaller number of participants have the greater probability of being chosen to the pseudo-sample. In this way, decision trees are created on the basis of such pseudo-samples containing about 3200 objects with the same number of participants for class ”0” and ”1”.

We proposed three version concerning the pseudo-samples construction and the moment of the creation (this problem had previously been analyzed in [3]):

ant – all – a pseudo-sample is created for each agent-ant and all attributes occur (similar to the Bagging [4], see Fig. 2 (a)).

col – all – a pseudo-sample is created for each colony of ants, and the same situation of occurrence of all attributes is analyzed (see Fig. 2 (b)).

col – rand – a pseudo-sample is created for each colony of ants, separately. In this case the $\sqrt{p}$ number of attributes is analyzed. We examine 6 attributes (similar to the Random Forest [5], Fig. 2 (c)).

6.4 Results of Experiments

Results of experiments point out the possibility of improvement concerning the classification accuracy for objects belonging to the class ”0”. The chart - receiver operating characteristic (ROC) for decision trees (Fig. 4) confirm the improvement of the mentioned above parameter, while the classification accuracy concerning objects belonging to class ”1” relatively deteriorates. Particular in case of the approach denoted as ”Learn 3300, 4150 and 8700 V2”; ”Learn 3300, 4150 and 8700” and ”Learn 8700”. Especially good results are obtained

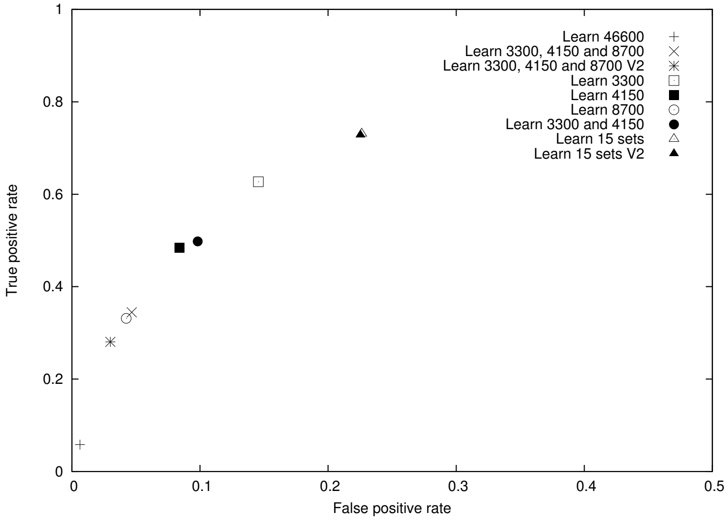


Fig. 4. The ROC curve for the results of dynamic ACDT algorithm

Table 2. ACDT algorithm

Learning	Accuracy for class 0	Accuracy for class 1	N. of nodes	Precision	N. of obj. from 0 class
Learn 46600	0.0580	0.9937	180.4	0.3748	1046
Learn 3300, 4150 and 8700	0.3444	0.9533	310.1	0.3234	7205
Learn 3300, 4150 and 8700 V2	0.2806	0.9700	39.3	0.3774	5029
Learn 3300	0.6271	0.8545	257.1	0.2184	19429
Learn 4150	0.4847	0.9160	212.3	0.2721	12052
Learn 8700	0.3314	0.9576	254.1	0.3362	6669
Learn 3300 and 4150	0.4980	0.9018	271.3	0.2473	13624
Learn 15 sets	0.7309	0.7738	241.6	0.1732	28558
Learn 15 sets V2	0.7283	0.7748	267.2	0.1733	28431

using pseudo-samples ”Learn 3300, 4150 and 8700 V2” simultaneously considered number of decision tree nodes. At the same time, we reinforced pruning decision trees. Number of nodes for decision trees is presented in Tab. 2.

Significantly better results are obtained using new version of ACDT, which is illustrated in the Fig. 5. In this case we observe the break through Pareto front achieved by ACDF algorithm performance.

Particularly interesting is the approach ”col–rand” in which individual decision tree (treated as a result of the forest) obtains very poor result (Tab. 3). However, the set of such weak and small decision trees inflicts better results, which is confirmed in [3]. Also good classifier embodied as decision forest could have been built applying these trees (Tab. 4).

Table 3. ACDF algorithm - one tree

Method	Accuracy for class 0	Accuracy for class 1	N. of nodes	Precision	N. of obj. from 0 class
ant – all	0.5747	0.9009	302	0.2612	3754
col – all	0.5945	0.8954	281	0.2734	3763
col – rand	0.0470	0.9722	12	0.0475	286

Table 4. ACDF algorithm - forest

Method	Accuracy for class 0	Accuracy for class 1	N. of nodes	Precision	N. of obj. from 0 class
ant – all	0.6296	0.9042	9068	0.2905	4211
col – all	0.6671	0.8984	8488	0.2985	4315
col – rand	0.6783	0.8745	1987	0.2750	4494

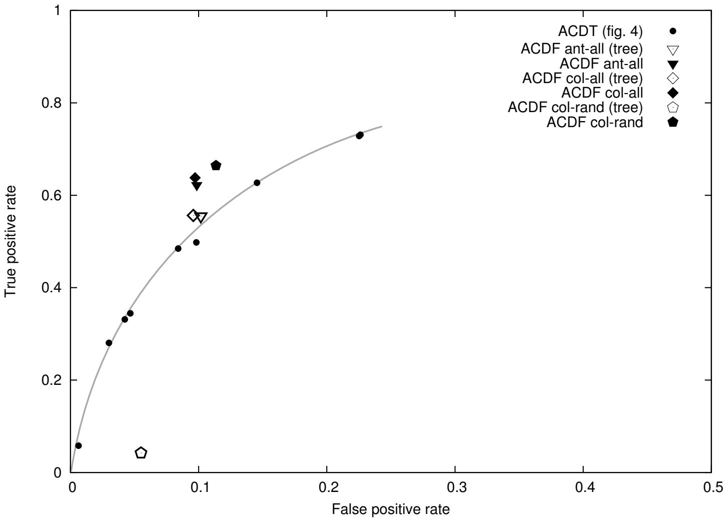


Fig. 5. The ROC curve for the results of ACDF algorithm

7 Conclusions

Applying pseudo-samples during the construction of single decision tree by ACDT algorithm affects the improvement of classification accuracy for objects belonging to the decision class "0" while good precision for this class is preserved. Especially good results are obtained in case of dynamic version of ACDT, where the training set undergoes change during the performance of the algorithms.

After all the ACDF algorithm, in which pseudo-samples with weighted draw are employed, allow to get more satisfied outcome. All proposed version of ACDF algorithm outperform those obtained by ACDT algorithm. However, particularly interestingly presents the approach concerning limited number of attributes.

Despite substantial reduction of decision tree growth, decision forest built as a result of algorithm performance, represents very good outcome.

Decision trees belonging to the analyzed decision forest represent differential local optima in solution space. Applying dynamically changing set of solutions (during the algorithm performance) that allows to construct stable classifiers. It can be applied to analyze difficult data sets.

In the future it should be consider to refine the manner of creation pseudo-samples. It is possible to build such sets according weighted observations, in which subsequent pseudo-samples will be created considering results from previously constructed decision trees. This method will bring the algorithm ACDF closer to the Boosting idea.

Acknowledgement. We thank Mikhail Ju. Moshkov and Jean-Claude Latombe for access to data set and for helpful discussions.

References

1. Boryczka, U., Kozak, J.: Ant colony decision trees – A new method for constructing decision trees based on ant colony optimization. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part I. LNCS, vol. 6421, pp. 373–382. Springer, Heidelberg (2010)
2. Boryczka, U., Kozak, J.: New Algorithms for Generation Decision Trees – Ant–Miner and Its Modifications. In: Abraham, A., Hassanien, A.-E., de Leon F. de Carvalho, A.P., Snášel, V. (eds.) Foundations of Computational Intelligence 6. SCI, vol. 206, pp. 229–264. Springer, Heidelberg (2009)
3. Boryczka, U., Kozak, J.: Ant colony decision forest meta-ensemble. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part II. LNCS, vol. 7654, pp. 473–482. Springer, Heidelberg (2012)
4. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
5. Breiman, L.: Random forests. Mach. Learn. 45, 5–32 (2001)
6. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York (1984)
7. Bühlmann, P., Hothorn, T.: Boosting algorithms: Regularization, prediction and model fitting. Statistical Science 22(4), 477–505 (2007)
8. Chikalov, I., Yao, P., Moshkov, M., Latombe, J.C.: Learning probabilistic models of hydrogen bond stability from molecular dynamics simulation trajectories. BMC Bioinformatics 12(S-1), S34 (2011)
9. Dorigo, M., Birattari, M., Blum, C., Clerc, M., Stützle, T., Winfield, A.F.T. (eds.): ANTS 2008. LNCS, vol. 5217. Springer, Heidelberg (2008)
10. Hyafil, L., Rivest, R.: Constructing optimal binary decision trees is NP–complete. Inf. Process. Lett. 5(1), 15–17 (1976)
11. Murphy, O., McCraw, R.: Designing Storage Efficient Decision Trees. IEEE Transactions on Computers 40, 315–320 (1991)
12. Otero, F.E.B., Freitas, A.A., Johnson, C.G.: Handling continuous attributes in ant colony classification algorithms. In: CIDM, pp. 225–231 (2009)
13. Rokach, L., Maimon, O.: Data Mining With Decision Trees: Theory and Applications. World Scientific Publishing (2008)
14. Schapire, R.E.: The strength of weak learnability. Machine Learning 5, 197–227 (1990)

Ant Colony System with Selective Pheromone Memory for SOP

Rafał Skinderowicz

Institute of Computer Science, Silesia University, Sosnowiec, Poland
rafal.skinderowicz@us.edu.pl

Abstract. Ant Colony System (ACS) is a well known metaheuristic algorithm for solving difficult optimization problems inspired by the foraging behaviour of social insects (ants). Artificial ants in the ACS cooperate indirectly through deposition of pheromone trails on the edges of the problem representation graph. All trails comprise a *pheromone memory*. In this paper we extend the previous work on a novel *selective pheromone memory model* for the ACS in which pheromone values are stored only for the selected *subset* of trails. Results of the experiments conducted on several Sequential Ordering Problem (SOP) instances show that it is possible to significantly reduce the ACS memory requirements without impairing the quality of the solutions.

Keywords: ant colony system, selective pheromone memory, Sequential Ordering Problem.

1 Introduction

The *Sequential Ordering Problem* (SOP) can be used to model important real-world problems such as vehicle routing problem with pick-up and delivery constraints or production planning problem [5]. The SOP is a NP-hard problem and solving it to optimality is impractical in terms of computation time required. If one can accept non-optimal but good quality solutions, *metaheuristic* algorithms should be considered. In recent years a new class of nature inspired metaheuristic algorithms proved to be successful in finding good quality solutions to many NP-hard problems. Among them are Ant Colony Algorithms introduced by Dorigo in the context of solving Travelling Salesman Problem (TSP) [3].

Dorigo and Gambardella have combined Ant Colony System algorithm (ACS) with efficient Local Search (LS) heuristic for the SOP and proposed Hybrid Ant Colony Optimization System for the Sequential Ordering Problems (HAS-SOP) [5]. Montemanni et al. presented a Heuristic Manipulation Technique (HMT) based on the HAS-SOP algorithm [7]. The HAS-SOP local search method was also used in an efficient Discrete Particle Swarm Optimization (DPSO) algorithm for the SOP [1]. Recent work by Gambardella et al. resulted in Enhanced ACS algorithm which can be considered the current *state-of-the-art* algorithm for the SOP [6].

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 711–720, 2013.

In the previous our work a modification of the classic ACS pheromone memory model was introduced [9]. The modification called *selective pheromone memory* (SPM) allowed to significantly (to about 2%) reduce the size of the pheromone memory while preserving the same quality of the solutions to the TSP as the classic ACS algorithm. The goal of the present work is to check if similar results can be obtained for a problem with a more complicated solution search space, namely the Sequential Ordering Problem. Although some ideas from the HAS-SOP algorithm will be used our goal is not to improve the results obtained by this algorithm.

The rest of the paper is organized as follows. In Section 2 a brief description of the SOP problem is given. Section 3 describes the ACS for SOP. Section 4 contains description of the proposed selective pheromone memory for the ACS and the Section 5 shows results of experimental evaluation of the proposed algorithms.

2 Sequential Ordering Problem

Sequential ordering problem was first formulated by Escudero in the context of a production planning system but it is often formulated as a generalization of an Asymmetric Travelling Salesman Problem (ATSP) [4]. We stick to the latter formulation in which SOP can be modeled as follows. A complete directed graph $G = (V, A)$ is given, where $V = \{0, 1, \ldots, n-1\}$ is a set of nodes corresponding to cities and $A = \{(i, j) | i, j \in V, i \neq j\}$ is a set of directed arcs corresponding to the roads between each pair of cities. A cost d_{ij} is associated with each arc $(i, j) \in A$. Without loss of generality it can be assumed that a fixed starting node, $0 \in V$, and a fixed final node, $n-1 \in V$, are given and are connected with all other nodes.

Also an additional precedence digraph $P = (V, R)$ is given, defined on the same set of nodes V as the graph G. An arc $(i, j) \in R$ denotes a precedence constraint between the nodes i and j, i.e. i has to precede j in every *feasible* solution. By definition $(0, i) \in R \;\; \forall i \in V \backslash \{0\}$ and $(i, n-1) \in R \;\; \forall j \in V \backslash \{n-1\}$. The relation of precedence is transitive, i.e. if i precedes j and j precedes k, than i precedes k. It can be seen that the precedence graph P has to be acyclic in order for a feasible solution to exist.

A solution to the SOP is a Hamiltonian path in which the sum of the edge weights is *minimal* and the precedence constraints are satisfied. The presence of the precedence constraints makes the SOP somewhat more difficult than the ATSP even if the size of the feasible solutions search space is smaller.

3 ACS for SOP

The ACS algorithm belongs to the Ant Colony Optimization algorithms family consisting of algorithms inspired by the behaviour (mainly foraging) of real ant colonies [2]. In the algorithm a set of simple computational agents called *ants* iteratively construct solutions to the problem tackled. At each step an agent appends a new component to its partial solution. The selection is made using both *a priori* knowledge about the problem and artificial *pheromone trails* associated with available solution components. The pheromone trails comprise a *pheromone memory* which is shared by the agents. The higher the value (level) of a pheromone trail the more likely the associated solution component will be selected by an agent. Each agent can modify the pheromone trails values which correspond to the selected solution components. The pheromone levels are increased on the components which belong to the high quality solutions found. Modification of the pheromone levels can be seen as a form of *indicrect* communication between agents often referred to as a *stigmergy.*

3.1 Pheromone Memory Model

A set of n decision variables X, $X = \{X_0, X_1, \ldots, X_{n-1}\}$, with domains $D_i = \{v_i^j : (i,j) \in A\}$ is given, where A is a set of edges as defined earlier. In other words, there is one decision variable for each city. A variable assignment $X_i = v_i^j$ means that edge (i,j) is a part of the solution. A solution to the SOP is a complete assignment of the variables X_i to values taken from their domains, that satisfies the precedence constraints and constitutes a Hamiltonian path over graph G. The set $\mathcal{C}$ contains solution components c_i^j for every possible assignment $X_i = v_i^j$, where $i \in 0, 1, \ldots, n-1$ is a node index such that $(i,j) \in A$. For each component c_i^j a *pheromone trail* parameter $\mathcal{T}_i^j$ is defined. The value of a $\mathcal{T}_i^j$ is denoted by τ_i^j ($\tau_i^j > 0$) and represents pheromone trail value (level). The vector of all pheromone trails values, or shortly the *pheromone memory*, is denoted by $\mathcal{T}$. From the definition it follows that $|\mathcal{T}| = \sum_{i=0}^{n-1} |D_i|$. For the SOP the $|\mathcal{T}|$ is the order of n^2 where n is the number of nodes (cities).

In the context of SOP, an agent begins being placed at the *starting node* and in the consecutive steps moves to one of the unvisited adjacent nodes. The selected edge is appended to the current partial solution s^p. The next solution component c_i^j for the ant k placed at node i is selected according to the formula:

$$j = \begin{cases} \arg\max_{l \in J_k^i} [\tau_i^l] \cdot [\eta(c_i^l)]^\beta, & \text{if } q \le q_0 \\ J, & \text{if } q > q_0, \end{cases} \tag{1}$$

where $\eta(c_i^l)$ is a heuristic knowledge associated with component c_i^l, J_k^i is a list of unvisited (candidate) cities of the ant k, q_0 is a parameter, $q \in [0,1]$ is a random number and J is a city selected with probability:

$$P(J|i) = \frac{[\tau_i^J] \cdot [\eta(c_i^J)]^\beta}{\sum_{l \in J_k^i} [\tau_i^l] \cdot [\eta(c_i^l)]^\beta}. \tag{2}$$

Parameter q is an enhancement introduced in the ACS in order to drive algorithm's search process towards the areas of the problem solution space containing the solutions of high quality. If the value of q is high, the choice given by the (1) is mostly deterministic and focused on the *exploitation* driven by the possessed knowledge – both heuristic and accumulated in the pheromone memory.

There are two kinds of pheromone trails updates performed in the ACS. The first is called *local pheromone update* and is performed after an ant has selected a new solution component, represented by an edge in the graph $G(V, A)$. The second is called *global pheromone update* and is performed after all ants have completed construction of their solutions. It consists in deposition of additional pheromone on the trails associated with the components of the best quality solution found in an iteration of the ACS.

3.2 Local Search

The ACS and other ant colony related algorithms are usually used combined with a local search (LS) algorithm applied after an agent has completed the construction of a solution. Gambardella et al. adapted the well known 3-Opt algorithm designed for the TSP [5]. The main problem was to preserve all the precedence constraints while improving the solution quality. A naive implementation would have $O(n^5)$ time complexity because of the necessity to check precedence constraints between all the pairs of exchanged solution's segments. However the clever *labelling procedure* proposed reduced the complexity to only $O(n^3)$ [5]. Nevertheless, the cost of the LS is still higher than the cost of the solution construction phase which has the complexity $O(n^2)$. To reduce running time in the EACS algorithm the LS is applied only to solutions of quality no more than 20% worse than the best solution found to date [6].

4 Selective Pheromone Memory

In the previous our work a *selective pheromone memory* (SPM) model for the ACS was introduced [9]. The idea can be summarized as follows. The pheromone memory vector $\mathcal{T}$ was replaced with a pair $(\widehat{\mathcal{T}}, U)$ where the vector $\widehat{\mathcal{T}}$ contains pheromone values but only for the *selected subset* of the pheromone trails, i.e. trails for which there is a corresponding element in the set U defined as:

$$U = \{u_i | i \in \{1, \ldots, m\} \wedge u_i \in A\} , \tag{3}$$

where A is a set of edges, and m is a positive number ($m \leq n$) denoting the size of the SPM vector $\widehat{\mathcal{T}}$. The pheromone trails for which there is no entry in $\widehat{\mathcal{T}}$ are assumed to have value of τ_0, i.e. minimal or initial pheromone level. It is worth realizing that there is no need to actually store any data for such trails in the computer's memory. What this means is that if $m \ll n$ then the SPM memory storage requirements can be much lower compared to the standard pheromone memory in the ACS which we will refer to as the *classic pheromone memory* (CPM). Of course, if $m = n$, then the set $U = A$ and hence $\widehat{\mathcal{T}} = \mathcal{T}$.

By storing only some of the values the ACS algorithm loses a potentially important part of the knowledge gathered in the form of pheromone trails. It would be desirable for this change not to impair the quality of the results obtained. Obviously, the loss of information depends both on the size, m, of the SPM and a *selection criterion* used to indicate the least important pheromone trail when updating the SPM. Additional *a priori* knowledge about the solutions search space would be required to make an optimal choice. When no such knowledge is available the choice becomes difficult and thus a *dynamic* selection criteria were proposed in the previous work. Specifically, the set of pheromone trails for which the values were stored in the SPM could be changed as the result of the ACS *local* and *global* pheromone update operations.

```
1  Input: $(\widehat{\mathcal{T}}, U)$ – selective pheromone memory
2  Input: $\widehat{\tau}_i^j$ – a new value for the pheromone parameter $\widehat{\mathcal{T}}_i^j$
3  Input: $SC$ – a criterion for selecting the least important parameter from $(\widehat{\mathcal{T}}, U)$

4  if $(i,j) \in U$ then {Entry for $\widehat{\mathcal{T}}_i^j$ exists in the memory}
5      $\widehat{\mathcal{T}}_i^j := \widehat{\tau}_i^j$ {Update pheromone parameter's value}
6  else{No entry for $\widehat{\mathcal{T}}_i^j$ exists in the memory, create one}
7      if $|U| \geq m$ then {Is the memory full?}
8          select_pheromone_parameter($SC$, $\widehat{\mathcal{T}}_k^l$)
9          $\widehat{\mathcal{T}}_k^l := \tau_0$ {In the actual implementation an entry for $\widehat{\mathcal{T}}_k^l$ may be
            removed from the computer's memory}
10         $U := U - \{(k,l)\}$
11     end
12     $U := U \cup \{(i,j)\}$
13     $\widehat{\mathcal{T}}_i^j := \widehat{\tau}_i^j$ {In the actual implementation an entry for $\widehat{\mathcal{T}}_i^j$ is created
        in the computer's memory}
14 end
```

Fig. 1. Procedure for updating the *selective pheromone memory* $(\widehat{\mathcal{T}}, U)$ using the given selection criterion *SC*

Figure 1 shows a pseudo-code of the pheromone update procedure. The algorithm starts with empty memory, i.e. all pheromone trails with values set to τ_0. When local or global pheromone update is performed the contents of the SPM is updated. If the pheromone trail is already in the SPM its value is updated (line 5), otherwise if the size of the memory is lower than the limit m, the updated pheromone value is appended to the vector $\widehat{\mathcal{T}}$ and its edge to the set U (lines 12–13). Otherwise the least important trail, chosen according to a specified selection criterion (SC), has to be removed before (lines 8–10).

Two selection criteria were proposed in the previous work, namely *minimum pheromone value* (MPV) criterion and *least recently used* (LRU) criterion [9]. The MPV criterion results in removal of the pheromone trail with the lowest value, i.e. it is based on the assumption that the higher the pheromone value is, the more important the trail is for the algorithm's performance. The LRU criterion selects for removal the least recently used trail. For both criteria the quality of the results obtained was similar and at the same level as for the standard ACS even when the size of the SPM was equal to only 2% of the corresponding CPM size. In the present work, we will investigate also an additional *random* (RND)

selection criterion which selects a pheromone trail to remove *randomly* from all the trails in the SPM.
There are two kinds of operations performed on the pheromone memory:

- *read operations* – when an agent (ant) selects a new solution component as stated by (1) and (2),
- *write (update) operations* – when an agent performs local or global pheromone update.

Not surprisingly, the reduction of the pheromone memory size results in a *time-memory trade-off*. An amortized constant time read operation complexity can be achieved with a *hash table*, but complexity of the update operation depends on the selection criterion used. The MPV criterion requires fast access to a pheromone trail with the lowest value what can be achieved in $O(\log m)$ time (m is the size of the SPM) if one uses a *binary heap* combined with a hash table. An amortized constant time complexity of the update operation for the LRU criterion can be achieved with a *doubly linked* list combined with a hash table. Similarly, an amortized constant time update operation complexity for the RND criterion can be achieved. Obviously, maintaining additional data structures results in higher memory per element (pheromone trail) overhead therefore *actual* memory savings can be observed if the size of the SPM is at least a few times smaller than the size (denoted by n) of the CPM, preferably $m \ll n$.

5 Experiments

In order to investigate how the SPM affects the ACS performance for the SOP a number of experiments were conducted using a few tests from the TSPLIB repository [8]. Both the algorithms without and with the local search were tested. Various versions of the SPM were compared with the classic pheromone memory (CPM) and a version with *constant pheromone memory* (CONST). In the CONST version all pheromone trails had a value of τ_0 which was not changed during the computations. In other words, the pheromone had no influence on the solution construction process (see Eq. 1 and Eq. 2). This comparison was made to verify the pheromone memory influence on the ACS performance.

The ACS has several parameters important for its performance. For all the algorithms the following *common* values were used: number of ants equal to $\lceil 0.1n \rceil$, $q_0 = 0.9$ (see Eq. 1), $\beta = 2$ (see Eq. 2), $\rho_l = 0.01$ (local pheromone update evaporation rate), $\rho_g = 0.2$ (global pheromone update evaporation rate). The values for the rest of the parameters depended on the LS application.

5.1 Algorithms without Local Search

The first series of experiments considered algorithms without the local search applied. Few different values of the SPM size, m, were investigated: $1 \cdot n$, $2 \cdot n$, $4 \cdot n$, $8 \cdot n$, $16 \cdot n$ and $0.5 \cdot n^2$. Of course, the smaller the value of m the more "forgetful" the ACS becomes, i.e. the smaller number of pheromone trails can have non-default (τ_0) value. In the extreme case of $m = 1 \cdot n$ the size of the SPM was equal to the number of nodes (cities) and exactly n times smaller than the size of the classic pheromone memory. For the rest of the parameters the values were as follows: the number of iterations equaled to 5000, the number of repetitions for each test and parameters values combination was set to 50.

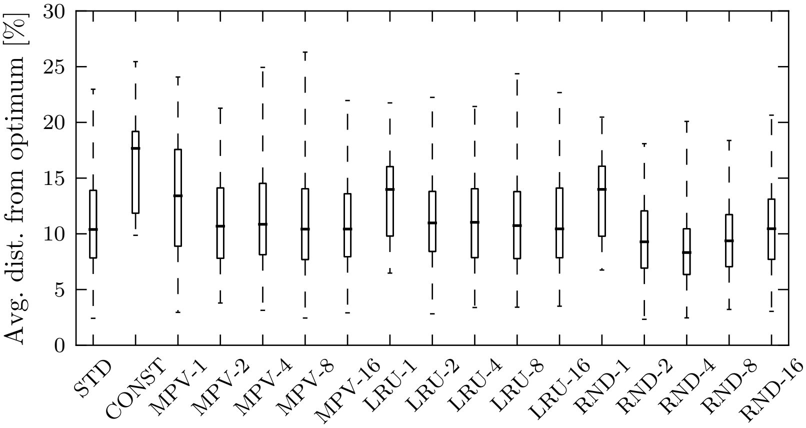


Fig. 2. Box plot of the mean solution quality obtained for all the tests for the algorithms investigated *without* the local search applied (STD – standard ACS; CONST – ACS with the constant pheromone memory; MPV-r / LRU-r / RND-r – the SPM version with the respective selection criterion, where r points to the SPM size, e.g. LRU-2 means that the LRU criterion was used and the SPM size was equal to $m = 2 \cdot n$).

Figure 2 shows a box plot of the average distance of the solutions obtained from the optimum[1] for the tests: *ft53.1*, *ft53.4*, *ft70.1*, *ft70.4*, *kro124p.1*, *kro124p.4*. As can be seen the worst results were obtained for the ACS with CONST pheromone memory in which all the pheromone trails had the same fixed and unmodifiable value. This confirms that pheromone memory is essential for the performance of the ACS. The second worst quality was observed for the ACS with SPM for all selection criteria when the size of the memory was set to $1 \cdot n$. It shows that if the size of the memory is too small too much *knowledge* accumulated in the pheromone trails is being lost. For all the bigger sizes of the SPM, $2n, 4n, 8n, 16n$, the solutions were of higher quality. Surprisingly, even for the SPM with size of $2n$ the quality of the results was similar to that of the standard ACS.

Statistical comparison of the investigated algorithms was also conducted using *non-parametric two-sample Wilcoxon rank-sum test.* The results are summarized in Tab. 1. As stated earlier when the size of the SPM was equal to n for almost all tests the results were significantly worse than for the standard ACS, irrespective of the selection criterion used. For the cases with $m > n$ the performance of the

[1] For the tests for which the optimum was not available the best known estimate was taken [8].

algorithm with MPV and LRU criteria was (except for the LRU with $m = 4n$) not significantly different than the performance of the standard ACS. Surprisingly, the algorithm with RND criterion achieved significantly better results in several cases. Especially, the with the size $m = 4n$ the results were better in 5 out of 6 cases.

Overall, the quality of the solutions obtained was not very good but one has to remember that the number of iterations was relatively small (5000). If larger number was used the quality would definitely improve but even more important is the use of an efficient local search method.

Table 1. The values in the table represent the number of times the selection criterion in the column resulted in the ACS with SPM performance significantly better or worse than the standard ACS algorithm according to *two-sided pairwise Wilcoxon rank sum* (the confidence level was 99%). First column shows the size of the SPM.

	MPV		LRU		RND	
m	Better	Worse	Better	Worse	Better	Worse
n	0	4	1	3	0	4
$2n$	0	0	0	0	3	0
$4n$	0	0	0	1	5	0
$8n$	0	0	0	0	3	0
$16n$	0	0	0	0	1	0
$0.5n^2$	0	0	0	0	0	0

5.2 Algorithms with Local Search

Ant colony algorithms are usually applied in combination with an efficient local search method [2]. It was important to check how the selective pheromone memory influences the quality of the results in such case. We used the LS method for the SOP proposed in [5]. Because of the increased computation time required by the ACS with LS the computations were conducted for a smaller set of parameters values selected on the basis of the previous experiments. Namely, the number of iterations was set to 1000 and the number of repetitions to 30. Two values of the SPM size, m, were investigated – $2 \cdot n$ and $4 \cdot n$. The six tests from the first part of the experiments were used plus two larger ones – *rbg253a* and *rbg378a*.

Figure 3 shows a box plot of the average distance of the solutions from the optimum (over all tests). The quality of the results was much better than for the ACS without LS, with the median value below 1%. As can be seen, the quality for the algorithm with MPV criterion was at the similar level as for the standard ACS. Results for the LRU version were slightly worse. Results for the RND selection criterion were better than for the other algorithms.

Statistical tests were conducted, as before, to confirm the differences noticed. Table 2 shows the number of times the SPM version of the ACS with respective selection criteria achieved significantly better or worse results than the standard

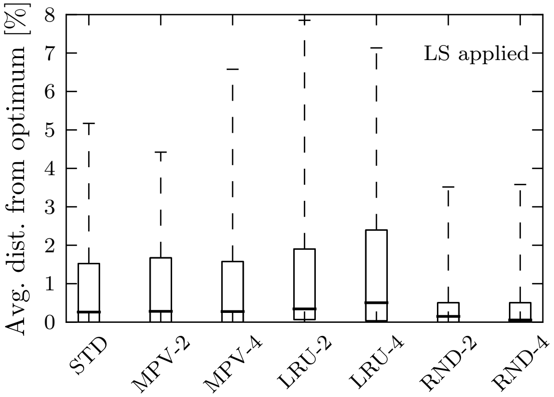


Fig. 3. Box plot of the mean solution quality obtained for all the tests for the algorithms investigated *with* the local search applied (meaning of the symbols is the same as before)

ACS algorithm. The algorithm with MPV selection criterion achieved statistically the same quality as the STD version. Both the LRU and RND criteria resulted in significantly worse results for the two largest tests: *rbg253a* and *rbg378a*. Surprisingly, the RND criterion occurred to be significantly better for a few smaller tests. It can be explained if one considers how the pheromone memory works. Pheromone accumulates in the trails for which the corresponding solution components belong to high quality solutions found. As the result the search process becomes more and more focused on those components, i.e. the *exploitation to exploration* ratio raises. But in the SPM with RND criterion *randomly* selected pheromone trails are removed from the memory what can result in decreased exploitation in favour of broader exploration. For the smaller tests the number of iterations was large enough for the algorithm to benefit from the broader exploration, but for the bigger tests it was too small and the decreased exploitation resulted in lower quality of the solutions.

Table 2. The values in the table represent the number of times the selection criterion in the column resulted in the ACS with SPM and with *local search* performance significantly better or worse than the standard ACS algorithm according to *two-sided pairwise Wilcoxon rank sum* (the confidence level was 99%)

	MPV		LRU		RND	
m	Better	Worse	Better	Worse	Better	Worse
$2n$	0	0	0	2	4	2
$4n$	0	0	0	1	3	2

6 Summary

Pheromone memory is essential for the performance of the Ant Colony System algorithm, but its size can be significantly reduced if the classic pheromone memory (CPM) model is replaced with the proposed selective pheromone memory model (SPM). Three versions of the SPM differing in the selection criterion used were compared with the CPM for the SOP. Computational experiments were performed for the algorithms with and without local search applied for several tests from the TSPLIB repository. The results showed that it is possible to reduce the size of the pheromone memory by order of magnitude (from $O(n^2)$ to $O(n)$), mostly without sacrificing the quality of the solutions and, in some cases, even improving the results. The SPM read and write operations come with additional computational overhead compared with the CPM, being an example of time-memory trade-off. Nevertheless, the SPM may be useful when the available memory capacity is very small. The forgetfulness of the SPM could also be beneficial when applied to dynamic problems such as Dynamic TSP.

Acknowledgments. This research was supported in part by PL-Grid Infrastructure.

References

1. Anghinolfi, D., Montemanni, R., Paolucci, M., Gambardella, L.M.: A hybrid particle swarm optimization approach for the sequential ordering problem. Computers & Operations Research 38(7), 1076–1085 (2011)
2. Dorigo, M., Blumb, C.: Ant colony optimization theory: A survey. Theoretical Computer Science 344, 243–278 (2005)
3. Dorigo, M., Gambardella, L.M.: Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation 1(1), 53–66 (1997)
4. Escudero, L.F.: An inexact algorithm for the sequential ordering problem. European Journal of Operational Research 37(2), 236–249 (1988)
5. Gambardella, L.M., Dorigo, M.: An ant colony system hybridized with a new local search for the sequential ordering problem. INFORMS Journal on Computing 12(3), 237–255 (2000)
6. Gambardella, L.M., Montemanni, R.: An enhanced ant colony system for the sequential ordering problem. Operations Research for Complex Decision Making, 80 (2010)
7. Montemanni, R., Smith, D.H., Gambardella, L.M.: A heuristic manipulation technique for the sequential ordering problem. Computers & Operations Research 35(12), 3931–3944 (2008)
8. Reinelt, G.: Tsplib95, `http://www.iwr.uni-heidelberg.de/groups/comopt/-software/tsplib95/index.html`
9. Skinderowicz, R.: Ant colony system with selective pheromone memory for TSP. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part II. LNCS, vol. 7654, pp. 483–492. Springer, Heidelberg (2012)

Efficient DPSO Neighbourhood for Dynamic Traveling Salesman Problem

Urszula Boryczka and Łukasz Strąk

University of Silesia, Intitute of Computer Science,
Będzińska 39, 41-205 Sosnowiec, Poland
{urszula.boryczka,lukasz.strak}@us.edu.pl

Abstract. This paper introduces a new Discrete Particle Swarm Optimization algorithm for solving Dynamic Traveling Salesman Problem (DTSP). An experimental environment is stochastic and dynamic. Changeability requires quick adaptation ability from the algorithm. The introduced technique presents a set of advantages that fulfill this requirement. The proposed solution is based on the virtual pheromone first applied in Ant Colony Optimization. The pheromone serves as a communication topology and information about the landscape of global discrete space. To improve a time bound, the α-measure proposed by Helsgaun's have been used for the neighborhood. Experimental results demonstrate the effectiveness and efficiency of the algorithm.

Keywords: dynamic traveling salesman problem, pheromone, particle swarm optimization, α-measure, neighborhood, held-karp lower bound.

1 Introduction

There has been a growing interest in studying evolutionary algorithms in dynamic environments in recent years due to its importance in real applications [4]. A problem where input data are changeable depending on time is called Dynamic Optimization Problem (DOP). The purpose of the optimization for DOPs is to continuously track and adapt the changes through time and to find quickly the currently best solution [16]. Metaheuristics that proved their effectiveness for static problems are being modified by different adaptation strategies for use in dynamic environments.

Particle Swarm Optimization is a technique based on swarm population created by Russell Eberhart and James Kennedy in 1995 [13]. This technique is inspired by the social behavior of a bird flocking or fish schooling. The algorithm was created primarily to optimize the function of a continuous space exploration. The PSO algorithm quickly became popular due to the fact that it has small number of parameters and it is easy to implement [13]. The growing interest has caused that this technique was adapted to solve different problems: static and dynamic [22,20].

In contrast with the classic Traveling Salesman Problem in the DTSP distances between the cities which are subject to changes and cause that it increases the computational complexity of the algorithm. The PSO algorithm in

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 721–730, 2013.

the problem of dynamic TSP is unexplored, but there are many publications of the PSO in a dynamic environment. The PSO in the dynamic environment was presented by [2,21]. First it was adapted for the Moving Peaks Benchmark, secondly for automatically tracking various changes in a dynamic system. The PSO with the virtual pheromone had first appeared in [12]. Kalivarapu [12] modified the parameters so that the pheromone was treated as additional information on the global landscape of optimized functions.

In this paper we introduce the algorithm for solving Dynamic Traveling Salesman Problem based on digital pheromone trail known in the ACO. The main idea is to use various proven techniques rather than creating a new one from scratch. Our research focus also on it as a new topology for communication between particles. To reduce the problem size we used Helsgaun [11] neighborhood. We also propose values for the parameters of the most successful tests.
This paper is structured as follows: in section 2 we present the basic concepts: Dynamic Traveling Salesman Problem and miscellaneous implementations of Discrete Particle Swarm Optimization. Section 3 presents ours DPSO algorithm proposals. Research results are shown in Section 4. Section 5 contains summary and conclusions.

2 Background

The Traveling Salesman Problem is a classic discrete combinatorial optimization problem. The objective of TSP is to find a shortest Hamilton cycles in an undirected graph $G = (V, E)$, where V is a vertex set (cities) and E is the edge set (routes). This problem plays an important role in the discrete optimization. This follows from the fact that many problems can be transformed to the problem of TSP by adjusting the encoding of the problem [1]. An interesting example is the DNA Computing [18] where cities represent gene sequences. It has been proved that the TSP is the $NP-complete$ problem [8]. For this reason, exact algorithms can not be used in practice.

There are several approaches to discrete optimization for the TSP with the PSO. Kennedy and Eberhart [14,22] defined first discrete binary version of PSO. All particles were coded as binary strings. The predefined velocity was interpreted as the probability of a bit state transition from zero to one and from one to zero. In a velocity formula we can use a sigmoid function restricting value to 0 or 1. Zhong, Zhang and Chen [17,9] proposed a new algorithm for TSP in which the position of the particle was not a permutation of numbers but a set of edges. It used the parameter c_3 named by the authors a mutation factor, which allowed control (balance) between exploration and exploration in discrete space. Most current PSO adaptations to the problem of static TSP were developed and compared in [9]. The study [17,9] has shown that this is the best adaptation of the PSO to the problem of static TSP.

The dynamic TSP is expressed through changes in both the number of vertices and a cost matrix. Every change in the input data may imply the optimum change. To track the algorithm results (for testing purpose) the optimum should be estimated. Younes et al. proposed a tool that unifies these changes - the Benchmark Generator (BGM) [23]. Basically, algorithms works as follows: in the first phase (half iterations), modifications are added sequentially. In the second phase changes are withdrawn. The disadvantage of this solution is only the first and the last iteration has known optimum and only those results can be compared. Some researchers use the static TSP problem combined with geostationary points satellite, whose orbit is well-defined pattern. However, the dynamics of such an environment is very small. For problem CHN146 + 3 - have 146 fixed points, and only 3 variables. Held-Karp created lower bound algorithm to fast optimum approximation, based on 1-*tree* (relaxation of *tree*) and subgradient optimization [10,11]. The result is an estimated value $T \leq T_{opt}$. Studies have shown that the algorithm very frequently gives lower bound less than 1% optimum [11]. However, so far, this technique has not been used in DTSP.

To reduce a TSP instance's size, the TSP solvers often used the neighborhood. It seems reasonable to do the same for the DTSP. Not every neighborhood can be used for dynamic TSP. In a static problem time for generating neighbors is not so important, because usually it is done only once at the beginning. In dynamic version, after every change, neighbors have to be recreated. It is required that the time needed to create the neighborhood was relatively smaller than the time spent on optimization. Helsgaun has introduced the α-measure and used it to construct the neighborhood [11]. He based his measure on lower tolerances and used for speed up Lin-Kernighan Heuristic (LKH) [19]. Helsgaun proved that the algorithm can be implemented in $O(n^2)$ [11]. The experiments showed that the optimum edges occur more frequently in neighborhood created by α-measure then nearest neighbor heuristics. For example, in a 532-city problem one of the links in the optimal solution is the 22nd nearest neighbor city for one of its end points. Using α-measure just 14th on candidates list [11]. Use of this measure in the DTSP, literature is relatively poor. Li et al. use α-measure with Gene Pool and Inver-Over operator [15]. In our solution we also used the α-measure to create a neighborhood. A more detailed description of this algorithm can be found in [11].

3 DPSO with Pheromone

The common feature of the PSO implementations is the general symbolic equation describing steps of the algorithm, which is inspired by the social behavior of bird flocking or fish schooling (equation 1) [13].

$$v_i^{k+1} = w \cdot v_i^k + c_1 rand() \cdot (pBest - x_i^k) + c_2 rand() \cdot (gBest - x_i^k) \tag{1}$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \tag{2}$$

where i denotes the number of particles, k - iteration number, $rand()$ is a random value from $[0, 1]$. Variables $pBest$ and $gBest$ denote particle personal best position and global best position respectively. Variable w is called an inertia weight and decides how much the pre-velocity will affected the new position. Parameters c_1 and c_2 are cognitive and social parameters for scaling *pBest* and *gBest* respectively. At the beginning, the algorithm initializes the distributed population and randomly distributed particles. At the second step this algorithm iteratively assigns velocity (which means assign the search direction) using equation 1 and it assigns position from previously calculated velocity using equation 2. After each iteration, variables *pbest* and *gbest* are updated. Depending on the problem being solved particle, position and velocity can be a vector, a set or a number.

The algorithm proposed by Zhong et al. [17] is based on the concept of a set of edges. It requires the introduction of many new concepts to adapt PSO and to operate with the edges. Most important concepts are described later in this article. The concepts are also described in [17].

Each edge can be represented by $A(x, y)$, where A denotes the probability of choosing edge, x and y are the endpoints. A is constrained to a number between 0 and 1 and $A(x, y) = A(y, x)$. When the algorithm is on velocity updating phase, a random number R between 0 and 1 is given, and if $R \leq X$ this edge is chosen. Velocity v is a set of elements as $A(x, y)$ like $\{A(a, b), B(a, c), C(b, d)\}$. To reduce the size of the edges set, vertex may be found at most 4 times. Therefore, the velocity can create sub tours (1-2-3-1) and some vertices can be omitted. The same edge but with a different probability is also acceptable. A subtraction between two positions is defined as a set of edges which exist in x_1 but not in x_2. To adapt this result to edge representation form, the probability 1 is added to these edges and make it uniform as velocity. A multiplication between a real number and a set of edges is defined like multiplication number and probability. Sum of two velocities v_1,v_2 is a set of $v_1 \cup v_2$, but vertex can't exist more than 4 times in edges. The new parameter c_3 is introduced. Equation 1 and 2 take a new form given by equations 3 and 4. New position is calculated in three steps. There are also more restrictive rules that obtained edge must fulfill like added edge can not create sub-cycle, or vertex can not occur more than 2 times.

$$v_i^{k+1} = c_2 rand() \cdot (gBest - x_i^k) + c_1 rand() \cdot (pBest - x_i^k) + w \cdot v_i^k \tag{3}$$

$$x_i^{k+1} = v_i^{k+1} \oplus c_3 rand() \cdot x_i^k \tag{4}$$

In the first step we choose edges depending on their corresponding probability in v_i to construct temporary position x_i'. If edge is chosen but construct an incorrect tour it will discard this edge. In second step, the algorithm selects edges from x_i according to the probability to complete the x_i'. Like before, an illegal edge is discarded. Third, if the first two steps not make a whole Hamiltonian cycle, it will add the absent with nearest principle heuristic. First and second steps have more restricted rules because third step must create feasible tour. Changing the order of performed operations in equation 3 is caused due to increasing importance of *gBest* rather than *pBest*. The full description of the algorithm may be found in [17].

When the problem size is growing linearly, solution space is growing exponentially. Exact algorithm (Concorde) required 85 CPU and year to prove a known tour's optimality for the problem sw24978 [5]. To reduce a TSP instance's size, our solution used α-measure. This allows to solve larger instances of the problem.

The concept of α-nearness is based on lower tolerance [11] and *min* 1-*tree* (*tree* relaxation concept which contains exactly one cycle) [11,10]. The $\alpha(i,j)$ can be defined as subtraction between the input *min* 1-*tree* cost and the cost of that *tree* after the addition of the edge (i,j). Formal algorithm is given by equation 5.

$$\alpha(i,j) = L(T^{+}(i,j)) - L(T) \tag{5}$$

where T^{+} is *min* 1-*tree* T with (i,j) edge. The value of $\alpha(i,j)$ specifies importance of edge (i,j) in 1-*tree*-ness. The neighborhood for vertex i, based on this measure is created by calculating α-value for every vertex incident with the vertex i. Then formed values are sorted in ascending order and limited to the parameter specifying the size of the neighborhood. Due to the *min* 1-*tree* consists of the shortest edges (created from the Minimum Spanning Tree), $\alpha(i,j) \geq 0$ and $\alpha(i,j) = 0$ if edge $e \subset T$ [11]. Helsgaun show that time complexity of this algorithm is $O(n^2)$.

The higher number of optimal edges contains *min* 1-*tree*, the measure gives a better neighborhood. For this reason, Held-Karp algorithm is used for generating the neighborhood, and estimating the optimal tour. Helsgaun proposed its own Held-Karp theory implementation (*Ascent* method) [11]. The Held-Karp lower bound is used twice - in the optimum tour estimation and with creating neighborhood. The first case is used for testing and is not part of the computation time. Both cases are physically performed twice (independently).

The initial population is created using the algorithm 1. The results shown that the quality of the final solution does not depend strongly on the initial tour. However, running time reduction can be archived by choosing initial tour that are close to optional.

Algorithm 1. Initialize population

```
choose random $i$
while not tour do
    choose $j$ if $(i,j)$ are neighbors
    otherwise choose $j$ among those node not already chosen
    $i = j$
end while
```

We have applied similar algorithm to the edge complete (fulfill edge set to tour). The search space is limited by the neighborhood definition. For this reason, if the vertex has in its neighborhood vertices previously used, the vertex is added to the missing list. There is a chance that this vertex will be a part of a different neighborhood (in further iterations). But if not, then the nearest neighbor heuristics is used for each node from the list.

The above algorithm is designed for searching of the solution space from the random position. In dynamic environment the DPSO algorithm starts from some partial solutions, which are the solutions obtained from previous calculations. To adapt this algorithm and to exploit the partial solutions, virtual pheromone was used as a form of attraction to the edges, which are frequently visited in best positions. So the algorithm starts running with some information about search space. This use of the virtual pheromone works like backbones concept. The backbone of a TSP instance consists of all edges, which are contained in each optimum tour of the instance [6]. In addition, the global pheromone stores information about the best solutions better than *gBest* because the information is more complete. Edge diversity depends on pheromone distribution. When just some edges will accumulate much pheromone value, the diversity is decreasing. The pheromone value depends on a frequency visit matrix and function that calculates value stored in matrix to the real amount of pheromone. More details about the use of pheromone can be found in [3]. The only difference is the use of pheromone reset. After each change, the pheromone is inherited in percent of a number (p_{shake}) like in [7]. We have also introduced visit frequency matrix scale factor (p_{scale}). Previously, the matrix has been passed with constant value equal to 1. In the DPSO$^{\alpha}$ algorithm the value is variable depending on the problem being solved. Real value is calculated using formula: $1/(p_{scale} \cdot n)$, where n denotes number of cities. This is related to larger changes in the data according to our previous DTSP framework (each change modify one edge).

4 Experimental Results

The first experiment investigates the property of Held-Karp theory (lower bound) for the DTSP testing framework. We implement Helsgaun improvements for this purpose [11]. Experiments are based on the TSPLIB instances which optimal solutions are known. An important parameter is the computing time. Table 1 presents the test results. Each instance is repeated 10 times.

Almost all instances provides a lower bound less than 1% optimum. In the berlin52 problem instance the algorithm has found an optimum. Solutions quality is not dependent on the size of the instance (small and large instance have a similar gap). The computation time is also satisfactory.

Each change modified coordinates and required recalculation of distance matrix. Changes takes two parameters fixed for the entire test. First parameter specifies how many points has to be changed and parameter determines the influence of the change to current position (equation 6):

$$\begin{aligned} x &= (1 - p_{inf}) \cdot x_{old} + p_{inf} \cdot rand() \\ y &= (1 - p_{inf}) \cdot y_{old} + p_{inf} \cdot rand() \end{aligned} \tag{6}$$

where $rand()$ is random value, x_{old}, y_{old} is previous x, y position and $p_{inf} = [0, 1]$. In all experiments p_{inf} is set to 0.3. This better reflect the real-world scenario. Figure 1 shows the changes in the berlin52 problem using presented algorithm.

Table 1. The results of the Held-Karp algorithm, columns: running time, obtained estimated optimum (HK), TSPLIB optimum (a priori), Gap between these values

Problem	Avg. time (ms)	HK	Optimum	Gap in %
berlin52	56	7542	7542	0
eil76	58	536,9188	538	0,2
eil101	165	627,4619	629	0,2
gr120	411	6886,966	6942	0,8
kroA200	1183	29016,27	29368	1,2
gr202	8987	39984,04	40160	0,4
a280	2100	2566	2579	0,5
pcb442	7946	50485,7	50778	0,6
ali535	120589	200296,2	202310	1
gr666	65651	290042,8	294358	1,5
pr1002	113212	256692,2	259045	0,9

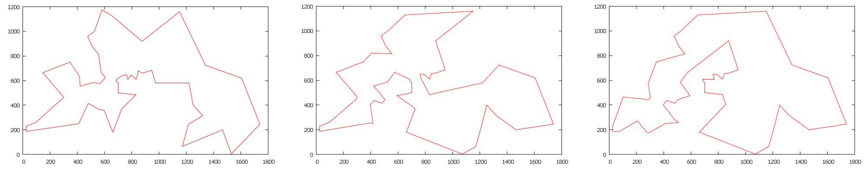


Fig. 1. Berlin52 problem in iteration 0 (original), 5, 10 (5% changes per iteration)

From the figure 1 it can be concluded that the rapid changes may modified input data so much, that any inference from past data becomes meaningless.

Next experiment, where DPSO$^{\alpha}$ algorithm ability has been investigated to solve the Dynamic Traveling Salesman Problem. Configuration is set to $c_1 = 1.5$, $c_2 = 2$, $c_3 = 2$, $w = 0.6$, n_s - swarm size $= 30$. This configuration was originally proposed by Zhong et al. in [17]. Experiments has been running on computer with Intel i7 processor 3.2 GHz and 12 GB of RAM memory. Operating system is Microsoft Windows Server 2008 R2. All tests run on a single core. The results are shown in table 2. The configuration is presented in table 3. It should also be noted that the gap is calculated based on the lower bound, which the tolerance is about 1% (see table 1).

In all experiments (table 2), the algorithm based on α-measure returns better results. This applied to the execution time and results quality. However, the best results are achieved with *Ascent* method (with and without the pheromone). Unfortunately, this method has the additional time overhead. Therefore, the variant of the algorithm depends on the assumptions - the accurate solution or faster optimization. It can be concluded that with further changes in the input data, the complexity of the problem grows (see the run-time). This is due to the fact that the growing complexity of the problem, for example some points move closer together. This trend can be seen in both versions - with and without pheromone. We also observe that the declining importance of pheromone in

Table 2. The results of our algorithms: Iter. - data change step, DPSO$^{\alpha}$ with *min* 1-*tree* (a), *Ascent* method (b), equivalent to a, b but without pheromone (c, d), DPSO (nearest neighbor) - with pheromone (e), without pheromone (f)

Prob.	Iter.	Avg. time [s]						Avg. gap [%]						Lower
		a	b	c	d	e	f	a	b	c	d	e	f	bound
berlin52	0	0,4	0,46	0,4	0,45	1,04	1,01	0,52	0	2,87	0	7,18	8,18	7542
	1	0,7	0,73	0,39	0,47	1,36	1	0,33	0,05	3,12	0,07	4,47	6,77	7491
	2	0,69	1,11	0,4	0,82	1,4	1,01	0,13	0,2	3,4	0,43	3,88	4,97	7489
	3	0,7	0,82	0,41	0,53	1,36	1	0,1	0,28	3,58	0,33	2,07	5,45	7502
	4	0,69	2,48	0,41	2,16	1,37	1,02	0,55	0,45	3,17	0,88	2,68	4,45	7481
	5	0,7	0,92	0,41	0,69	1,38	1	0,93	0,78	2,67	1,22	2,58	3,23	7476
	6	0,72	0,78	0,41	0,5	1,39	1,01	1,32	0,7	2,63	1,07	2,6	3,17	7498
	7	0,7	0,88	0,42	0,6	1,37	1,01	1,18	0,37	2,47	1,15	2,48	2,47	7479
	8	0,72	0,93	0,42	0,66	1,37	1,02	2,38	1,4	2,75	1,63	3	2,95	7434
	9	0,71	1,57	0,41	1,29	1,38	1,04	1,95	1,77	2,65	1,52	2,85	2,38	7458
	10	0,7	0,88	0,41	0,58	1,41	1	1,52	1,28	3,05	1,75	2,98	2,3	7445
Avg. summary		0,68	1,05	0,41	0,79	1,35	1,01	0,99	**0,66**	2,94	0,91	3,34	4,21	n/a
eil101	0	1,34	1,29	1,3	1,26	5,19	5,17	1,33	0,42	2,63	0,6	15,75	14,2	627
	1	2,2	2,24	1,28	1,39	6,23	5,15	2,27	1,57	3,97	2,87	8,63	8,55	612
	2	2,2	2,25	1,29	1,39	6,25	5,14	2,47	1,28	4,08	1,95	7,85	7,88	602
	3	2,18	2,18	1,27	1,34	6,18	5,16	2,53	1,2	5,05	1,98	7,33	6,85	595
	4	2,17	2,2	1,28	1,35	6,2	5,11	3,5	1,35	4,97	2,67	7,1	8,58	587
	5	2,19	2,2	1,27	1,34	6,23	5,11	3,53	1,53	5,08	2,42	6,05	7,82	577
	6	2,24	2,17	1,26	1,33	6,2	5,15	2,67	2,07	6,2	2,37	6,67	7,38	562
	7	2,21	2,18	1,28	1,31	6,19	5,06	2,67	1,57	5,83	2,12	6,48	6,32	558
	8	2,21	2,13	1,28	1,31	6,2	5,08	2,8	1,72	5,82	2,4	6,13	7,12	552
	9	2,23	2,28	1,29	1,42	6,22	5,1	3,07	1,43	6,3	3,13	6,68	8,82	546
	10	2,22	2,2	1,27	1,33	6,2	5,07	2,78	1,07	5,65	3,1	5,95	7,33	546
Avg. summary		2,13	2,12	1,28	1,34	6,12	5,12	2,69	**1,38**	5,05	2,33	7,69	8,26	n/a
gr202	0	7,08	13,06	7,09	13,12	60,9	61,48	3,23	1,4	4,47	1,87	9,5	9,07	39984
	1	11,67	29,61	6,96	26,09	64,63	58,88	4,7	1,93	5,6	2,87	9,63	9,17	44580
	2	11,7	21,83	7,04	18	63,67	58,3	4,23	2,13	4,97	2,8	9,3	9,47	46134
	3	11,57	21,83	7,07	18,04	63,51	57,71	4,97	2,53	6,23	3,07	9,87	9,77	47466
	4	11,75	19,22	7,18	15,24	62,65	57,02	4,97	2,8	6,67	2,93	10,53	9,3	47939
	5	11,65	17,38	7,15	13,19	62,62	57,11	4,93	1,53	7,03	2,03	11,53	11,23	48507
	6	11,53	21,07	7,08	17,08	62,32	56,85	6,07	3,03	6,63	3,9	10,63	10,77	48943
	7	11,54	19,47	7,02	15,04	61,99	56,37	5,4	2,97	5,83	2,87	11,23	9,5	50041
	8	11,58	19,2	7,17	15,18	61,44	55,74	5,8	2,7	6,9	3,23	9,67	9,53	49604
	9	11,69	21,18	7,12	17,45	61,16	55,42	5,3	1,87	6,77	2,5	9,7	9	49628
	10	11,72	19,69	7,19	16,3	63,86	61,73	4,43	2,2	6,67	2,93	6,87	7,83	50089
Avg. summary		11,23	20,32	7,1	16,8	62,62	57,87	4,91	**2,28**	6,16	2,82	9,86	9,51	n/a
pcb442	0	54,77	61,55	56,01	60,74	940,11	935,75	6,4	3,05	6,65	4,5	31,3	28,6	50486
	1	88,07	90,3	55,35	60,3	963,7	910,58	7,55	4,2	7,55	5,2	18,25	18,6	50004
	2	89,92	91,71	57,02	60,83	954,73	922,07	6,35	5,65	8,25	5,2	15,85	17,55	50645
	3	91,34	92,57	56,82	62,3	939,34	889,05	7,2	3,8	9,4	5,15	16,2	18,35	50771
	4	91,24	91,85	56,9	60,06	928,3	880,51	8,15	5,85	9,85	6	15,85	17,25	50233
	5	89,84	92,86	57,52	62,16	925,18	893,74	7,2	6,4	8,35	5,9	14,3	14,3	49966
	6	90,04	92,64	56,9	64,12	913,51	873,22	7,8	4,35	9	4,95	15,15	14,5	50018
	7	89,52	92,06	56,99	66,64	911,73	866,96	8,75	3,85	8,9	5,15	16,15	15,05	49713
	8	92,45	95,28	58,31	64,04	913,46	855,8	7,8	3,95	9,65	4,7	13,9	15,05	49357
	9	91,21	101,93	57,52	75,59	892,32	856,63	6,65	3,6	9,4	3,7	14,05	14,55	49752
	10	93,5	92,62	57,64	65,3	897,27	849,63	7,3	4,6	10,6	4,95	15,95	13,7	49643
Avg. summary		87,44	90,49	57	63,83	925,42	884,9	7,38	**4,48**	8,87	5,04	17	17,05	n/a

Table 3. Parameters for DPSO test optimization

Problem	Parameters						
	Solution				Testing environment		
	Iterations	p_{scale}	p_{shake}	Neighborhood size	Problem repeats	Iteration repeats	Changes [%]
berlin52	52 · 18	10	20	30	3	10	5
eil101	101 · 15	10	10	60	3	10	5
gr202	202 · 20	10	10	100	3	5	5
pcb442	442 · 30	5	10	75	2	5	5

relation to the increasing size of the instance. This approach applied both - the quality of the obtained solution and the execution time. For instance, more than 101 cities (eil101) time overhead is too costly and does not increase the results quality. It should also be noted that the version without the pheromone does not inherit the population from the previous solution - for objective neighborhoods comparison.

5 Conclusions

In this paper, we proposed a set-based DPSO with pheromone and neighborhood for the DTSP. Particle swarm optimization algorithm is responsible for the convergence to the optimum and filtering promising edge. The α-values neighborhood is responsible for providing the best possible edges. Both in terms of execution time and results quality, this combination outperforms our previous DPSO algorithm with the nearest neighbor heuristic. Application of the Held-Karp theory to DTSP proved to be good testing framework. The algorithm was sufficiently ”sensitive” to estimated optimum distance even for the small changes. The future work will focus on improving the pheromone updating to faster convergence in larger instances and reduce time overhead.

References

1. Applegate, D.L., Bixby, R.E., Chvátal, V., Cook, W.J.: The traveling salesman problem: A computational study. Princeton University (2006)
2. Blackwell, T.: Particle swarm optimization in dynamic environments. In: Yang, S., Ong, Y.-S., Jin, Y. (eds.) Evolutionary Computation in Dynamic and Uncertain Environments. SCI, vol. 51, pp. 29–49. Springer, Heidelberg (2007)
3. Boryczka, U., Strąk, Ł.: A hybrid discrete particle swarm optimization with pheromone for dynamic traveling salesman problem. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part II. LNCS, vol. 7654, pp. 503–512. Springer, Heidelberg (2012)
4. Branke, J.: Evolutionary approaches to dynamic environments. In: In GECCO Workshop on Evolutionary Algorithms for Dynamics Optimization Problems (2001)
5. Cook, W.: Sweden computation log (June 2013)

6. Dong, C., Ernst, C., Jäger, G., Richter, D., Molitor, P.: Effective heuristics for large euclidean tsp instances based on pseudo backbones
7. Eyckelhof, C.J., Snoek, M.: Ant systems for a dynamic tsp: Ants caught in a traffic jam. In: Dorigo, M., Di Caro, G.A., Sampels, M. (eds.) Ant Algorithms 2002. LNCS, vol. 2463, pp. 88–99. Springer, Heidelberg (2002)
8. Garey, M.R., Johnson, D.S.: Computers and intractability: A guide to the theory of NP-completeness. W.H. Freeman (1979)
9. Goldbarg, E.F.G., Goldbarg, M.C., de Souza, G.R.: Particle swarm optimization algorithm for the traveling salesman problem (2008)
10. Held, M., Karp, R.M.: The traveling-salesman problem and minimum spanning trees. Operations Research (18), 1138–1162 (1970)
11. Helsgaun, K.: An effective implementation of the lin-kernighan traveling salesman heuristic. European Journal of Operational Research 126, 106–130 (2000)
12. Kalivarapu, V., Foo, J.-L., Winer, E.: Improving solution characteristics of particle swarm optimization using digital pheromones. Structural and Multidisciplinary Optimization (2009)
13. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, pp. 1942–1948 (1995)
14. Kennedy, J., Eberhart, R.: A discrete binary version of the particle swarm algorithm. In: Systems, Man, and Cybernetics. Computational Cybernetics and Simulation. IEEE (1997)
15. Li, C., Yang, M., Kang, L.: A new approach to solving dynamic traveling salesman problems. In: Wang, T.-D., Li, X., Chen, S.-H., Wang, X., Abbass, H.A., Iba, H., Chen, G.-L., Yao, X. (eds.) SEAL 2006. LNCS, vol. 4247, pp. 236–243. Springer, Heidelberg (2006)
16. Li, W.: A parallel multi-start search algorithm for dynamic traveling salesman problem. In: Pardalos, P.M., Rebennack, S. (eds.) SEA 2011. LNCS, vol. 6630, pp. 65–75. Springer, Heidelberg (2011)
17. Zhong, W.L., Zhang, J., Chen, W.N.: A novel set-based particle swarm optimization method for discrete optimization problems. In: Evolutionary Computation, CEC 2007, vol. 14, pp. 3283–3287. IEEE (1997)
18. Azimi, P., Daneshvar, P.: An efficient heuristic algorithm for the traveling salesman problem. In: Dangelmaier, W., Blecken, A., Delius, R., Klöpfer, S. (eds.) IHNS 2010. LNBIP, vol. 46, pp. 384–395. Springer, Heidelberg (2010)
19. Richter, D., Goldengorin, B., Jäger, G., Molitor, P.: Improving the efficiency of helsgaun's lin-kernighan heuristic for the symmetric tsp. In: Janssen, J., Prałat, P. (eds.) CAAN 2007. LNCS, vol. 4852, pp. 99–111. Springer, Heidelberg (2007)
20. Schutte, J.F., Groenwold, A.A.: A study of global optimization using particle swarms. Journal of Global Optimization 31, 93–108 (2005)
21. Xiaohui, H., Eberhart, R.: Adaptive particle swarm optimisation: detection and response to dynamic systems. In: Proceedings of the 2002 Congress on Evolutionary Computation, CEC 2002 (2002)
22. Hu, X., Shi, Y., Eberhart, R.: Recent advances in particle swarm (2004)
23. Younes, A., Basir, O., Calamai, P.: A benchmark generator for dynamic optimization. In: Digest of the Proceedings of the Wseas Conferences (2003)

Breaking LFSR Using Genetic Algorithm

Iwona Polak and Mariusz Boryczka

University of Silesia, Institute of Computer Science,
Będzińska 39, 41-200 Sosnowiec
ipol@vp.pl,
mariusz.boryczka@us.edu.pl

Abstract. In this paper it is shown how to find LFSR using genetic algorithm. LSFRs are part of many cryptographic structures and pseudorandom number generators. Applying genetic algorithms to Linear Feedback Shift Registers (LFSR) cryptanalysis is not quite obvious. Genetic algorithms – being one of heuristic techniques – give approximate solution. The solution could be very good, but not necessarily the best, whereas cryptographic problems require one exact answer, every other being not good enough. But as it will be shown, even if it is not intuitive, breaking LFSRs using genetic algorithms can give some interesting and promising results.

Keywords: LFSR, genetic algorithm, cryptanalysis.

1 Introduction

Various optimisation heuristics were applied in the field of cryptanalysis. Very good results were achieved in classical cipher's analysis [3,11], but these ciphers are not used in practice any more. There were also some research in cryptanalysis of ciphers based on knapsack problem [4,15], nevertheless these ciphers are also of no practical application. Modern ciphers as DES or AES were broken only on weakened versions – with smaller number of rounds, shorter keys or without initial or final phase [5,8,13,14]. The main problem here is that heuristic techniques often give solutions close to optimal, however not necessarily the optimal one – this could be sufficient in many problems, but in cryptography it could be not enough.

This work focuses on applying genetic algorithms to LFSR cryptanalysis. Many cryptographic structures and pseudorandom number generators contain LFSRs. Breaking LFSRs using genetic algorithms can give some interesting and promising results.

There are some papers devoted to the similar topic, but they discuss a little bit different issues, for instance the goal of [1] is to find shortest LFSR that approximate given key stream, in other words the paper presents equivalent linear system. Authors study registers of length 5 to 8. Whereas this paper focuses on finding the exact register that was used in analysed system and examine registers of length up to 32.

C. Bădică, N.T. Nguyen, and M. Brezovan (eds.): ICCCI 2013, LNAI 8083, pp. 731–738, 2013.

In section 2 genetic algorithm is presented and basic version of this method is listed. Section 3 is the introduction to LFSR shift registers. Section 4 gives a detailed description of experimental results, preceded by defining chosen parameters. The last section concludes and discusses future evolution of this approach.

2 Genetic Algorithms

Genetic algorithms (GA) were first introduced by J. H. Holland [6]. Their behaviour is based on living organisms evolution and they use common natural mechanisms. GAs are methods of searching best solutions in the space of all possible solutions. They belong to the set of probabilistic methods, which means that for every run of the algorithm different results could be obtained.

In the presented work the classic version of the genetic algorithm was used. This algorithm acts as follows:

1. Randomly generate initial population.
2. Repeat until termination condition has been reached (here: number of iterations):
   (a) Evaluate the fitness function for every individual.
   (b) Apply reproduction for the best fit individuals (parents):
      i. Crossover;
      ii. Mutation;
      iii. New population replaces old one.
3. Output the best solution found.

When applying GA to any problem it is very important to properly define all components of it, e.g. parameters (probability of crossover and mutation), representation of individuals, the exact way that crossover and mutation work, how the best individuals are evaluated and chosen for the next population, and so on. All listed elements of tuning GA are more or less crucial and they have very significant influence whether obtained results are satisfying or totally irrelevant. These components always have to be defined individually for considered problem.

3 Linear Feedback Shift Register

Linear Feedback Shift Register (LFSR) is a structure that appears in many cryptographic constructions and pseudorandom number generators, e.g. in encryption schemes A5/1 and A5/2 that are used in mobile communication in GSM networks [7,9]. More examples are listed in [2].

LFSRs are part of FSRs – Feedback Shift Registers. Scheme of FSR is shown in fig. 1. LFSR's input bit for the next state is a result of some function applied to chosen bits, the most common function is XOR. When shifting, this bit is inserted at the beginning of the sequence, while the last significant bit is added to the output stream; all other bits are just shifted one position to the right.

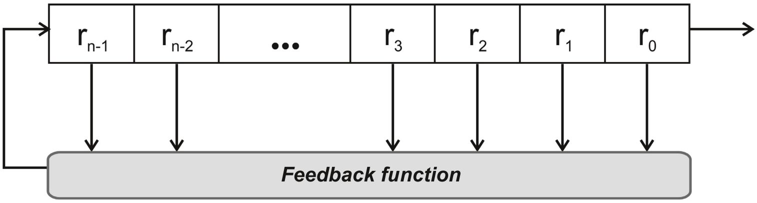


Fig. 1. Feedback Shift Register – FSR

Shift register of that kind can be described by a polynomial in which powers show bits where the taps appear.

In A5/1 scheme among others there is LFSR described by equation (1) and corresponding LFSR register is shown in fig. 2.

$$T(x) = x^{22} + x^{21} + 1 \tag{1}$$

In order to create a register which for given length cycles through the maximum number of states, equation describing this register has to be a primitive polynomial modulo 2 (primitive polynomial k degree is an irreducible polynomial, which divides polynomial $x^{2k-1}+1$, but doesn't divide polynomials x^d+1, for every d dividing 2^k-1). Such register of length n will output a string of maximum length 2^{n-1}; after this period sequence will start to repeat. (One state is not included, as string consisted of only 0s leads to infinite 0s output string, which definitely is not a good stream for cryptographic or random generator purposes).

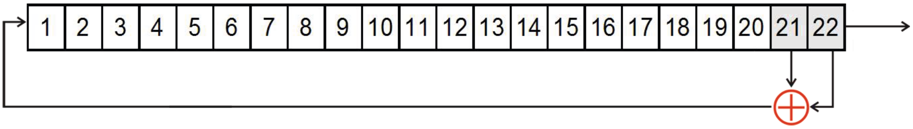


Fig. 2. LFSR scheme from equation $x^{22} + x^{21} + 1$

4 Experimental Results and Discussion

In this section results of experiments are shown, preceded by description of parameter's values that were used.

The goal of the research is to find primitive polynomial of the attacked LFSR register, without knowledge neither about its length nor placement and number of taps. Revealing construction of this LFSR could lead to decryption of message encrypted with stream generated by this register.

4.1 Parameters

Parameters of the genetic algorithm were set as follows:

- number of individuals: 40,
- number of generations: 50,
- crossover probability: 0.5,
- mutation probability: 0.02.

The fitness function was Hamming's distance between output string of attacked LFSR and compared individual. Obtained value was divided by output length in order to get values from interval [0,1], independently of register's length. Starting bits were not considered in this calculation, as the starting bits of output are the register's initial state and will always be correct – amount of subtracted bits was equal to the length of the individual generated by GA.

For the experiments there were five LFSRs chosen, which differ in length and number of taps – in order to get general result. The studied set consisted of four registers of maximum-length, used in practice [2,7,9]:

- $x^{22} + x^{21} + 1$ (one register from A5/1 GSM encryption),
- $x^{16} + x^5 + x^4 + x^3 + 1$ (USB 3.0 scrambler),
- $x^7 + x^3 + 1$ (CRC-7 / telecom systems, MMC),
- $x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x^1 + 1$ (CRC-32-MPEG2, CRC-32-IEEE 802.3)

and one generated randomly, not of maximum-length:

- $x^{23} + x^{21} + x^{20} + x^{18} + x^{17} + x^{16} + x^{11} + x^{10} + x^9 + x^7 + x^3 + 1$.

Every LFSR was tested with five initial states – four of following sequences:

- all 1s (1111...),
- alternating 1s and 0s (1010...),
- sequence of 0s beginning with 1 (1000...),
- sequence of 0s ending with 1 (...0001),

and one generated randomly, separately for every register.

For LFSRs there were also additional parameters to be specified:

- output length,
- minimum and maximum register length,
- minimum and maximum taps amount.

In every case an output length was set to 100, whereas register length and taps amount scopes were selected individually for every examined register, as shown in table 1.

4.2 Results

For every pair of LFSR and initial state there were 10 runs of genetic algorithm and the best results for each register are shown in tables from 2 to 6.

Table 1. Parameters for registers

LFSR	Length		Taps	
	min	max	min	max
$x^{22} + x^{21} + 1$	15	40	2	6
$x^{16} + x^5 + x^4 + x^3 + 1$	10	30	2	6
$x^7 + x^3 + 1$	5	15	2	6
$x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10}$ $+x^8 + x^7 + x^5 + x^4 + x^2 + x^1 + 1$	25	50	10	25
$x^{23} + x^{21} + x^{20} + x^{18} + x^{17} + x^{16}$ $+x^{11} + x^{10} + x^9 + x^7 + x^3 + 1$	15	31	6	16

Table 2. Results for $x^{22} + x^{21} + 1$ (one register from A5/1 GSM encryption)

Initial state	Maximum fit	Polynomial found	Generation
1111...	1	$x^{23} + x^{21} + x^1 + 1$	22
1010...	0.911	$x^{21} + x^{20} + 1$	0
1000...	1	$x^{22} + x^{21} + 1$	2
...0001	1	$x^{22} + x^{21} + 1$	5
0010110010100000111010 (random)	1	$x^{22} + x^{21} + 1$	16

Table 3. Results for $x^{16} + x^5 + x^4 + x^3 + 1$ (USB 3.0 scrambler)

Initial state	Maximum fit	Polynomial found	Generation
1111...	0.720	$x^{25} + x^{10} + 1$	1
1010...	0.713	$x^{20} + x^{17} + x^8 + x^7 + 1$	13
1000...	0.682	$x^{18} + x^{17} + x^{16} + x^{15} + x^{14} + x^{12}$ $+x^{11} + x^9 + x^8 + x^7 + x^5 + x^2 + 1$	15
...0001	0.690	$x^{29} + x^{23} + x^{19} + x^{18} + x^{12} + x^1 + 1$	0
0111101111100110 (random)	0.679	$x^{22} + x^7 + 1$	13

Table 4. Results for $x^7 + x^3 + 1$ (CRC-7 / telecom systems, MMC)

Initial state	Maximum fit	Polynomial found	Generation
1111...	1	$x^{10} + x^7 + x^6 + 1$	20
1010...	1	$x^7 + x^3 + 1$	7
1000...	1	$x^7 + x^3 + 1$	8
...0001	1	$x^9 + x^7 + x^5 + x^3 + x^2 + 1$	0
0001100 (random)	1	$x^7 + x^3 + 1$	5

Table 5. Results for $x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^{8} + x^{7} + x^{5} + x^{4} + x^{2} + x^{1} + 1$ (CRC-32-MPEG2, CRC-32-IEEE 802.3)

Initial state	Maximum fit	Polynomial found	Generation
1111...	0.678	$x^{41} + x^{38} + x^{30} + x^{19} + x^{18} + x^{17} + x^{12}$ $+x^{10} + x^{9} + x^{7} + x^{6} + x^{3} + x^{2} + x^{1}$ $+1$	22
1010...	0.790	$x^{38} + x^{37} + x^{32} + x^{30} + x^{29} + x^{28} + x^{27}$ $+x^{26} + x^{25} + x^{24} + x^{23} + x^{22} + x^{21}$ $+x^{20} + x^{18} + x^{17} + x^{12} + x^{9} + x^{6} + x^{5}$ $+x^{1} + 1$	0
1000...	0.742	$x^{38} + x^{37} + x^{36} + x^{29} + x^{27} + x^{26} + x^{22}$ $+x^{21} + x^{19} + x^{16} + x^{15} + x^{14} + x^{11}$ $+x^{10} + x^{9} + x^{5} + x^{2} + 1$	2
...0001	0.769	$x^{48} + x^{45} + x^{44} + x^{43} + x^{42} + x^{40} + x^{39}$ $+x^{38} + x^{34} + x^{33} + x^{32} + x^{30} + x^{28}$ $+x^{27} + x^{25} + x^{23} + x^{20} + x^{17} + x^{16}$ $+x^{15} + x^{14} + x^{11} + x^{10} + x^{6} + x^{5}$ $+x^{3} + x^{2} + 1$	5
1110111010101110 0000000000000000 (random)	0.730	$x^{26} + x^{25} + x^{24} + x^{23} + x^{22} + x^{21} + x^{20}$ $+x^{18} + x^{17} + x^{16} + x^{13} + x^{12} + x^{7} + x^{6}$ $+x^{4} + x^{3} + x^{2} + x^{1} + 1$	16

Table 6. Results for $x^{23} + x^{21} + x^{20} + x^{18} + x^{17} + x^{16} + x^{11} + x^{10} + x^{9} + x^{7} + x^{3} + 1$ (generated randomly, not of maximum-length)

Initial state	Maximum fit	Polynomial found	Generation
1111...	1	$x^{29} + x^{28} + x^{27} + x^{23} + x^{21} + x^{17} + x^{13}$ $+x^{7} + x^{5} + x^{4} + x^{3} + 1$	0
1010...	0.704	$x^{29} + x^{27} + x^{22} + x^{20} + x^{17} + x^{16} + x^{10}$ $+x^{8} + x^{6} + x^{3} + x^{2} + 1$	3
1000...	0.714	$x^{30} + x^{29} + x^{25} + x^{23} + x^{18} + x^{11} + x^{10}$ $+x^{9} + x^{7} + x^{4} + x^{3} + x^{2} + x^{1} + 1$	25
...0001	0.746	$x^{29} + x^{27} + x^{26} + x^{25} + x^{23} + x^{22} + x^{19}$ $+x^{16} + x^{11} + x^{5} + x^{4} + 1$	1
110010010110 11100000000 (random)	0.710	$x^{31} + x^{28} + x^{23} + x^{20} + x^{18} + x^{16} + x^{14}$ $+x^{9} + x^{3} + x^{2} + x^{1} + 1$	29

Achieved results are very promising – between 68% to 100% of output bits match the original LFSR, with average value of fitness function 0.84 (that is to say that 84% output bits match). This means that for a given message encrypted with a stream generated by this register there can be significant amount of message's bits decrypted properly. The best fitted individual was found no later than in 29th generation.

The best results were obtained for registers with small amount of taps (91-100%, with average 99%) – primitive polynomials which describe them have a small number of coefficients and therefore they are called low density polynomials. But also for longer LFSRs with more taps the results were satisfying (68-79%, with average 72%).

For every initial state of the LFSR the results were similar, which shows that these results are independent of initial sequence of LFSR. There was one limitation: initial state of 0s sequence beginning with value 1 (1000...) was not very helpful – when genetic algorithm generated a register shorter than attacked LFSR, then this generated register's initial state was all 0s, producing infinite sequence of just 0s and giving results of no value, even if fitness function was high.

5 Conclusion

In this paper the new approach to breaking LFSR was presented – cryptanalysis of LFSR shift register using genetic algorithms inspired by evolution mechanisms. This method occurred to be quite effective and promising. Output bits stream of the best individual matched the attacked register's output bits stream in 68% to 100%, with average at 84% – this indicates that the same percent of encrypted bits will be decrypted properly. It is planned to test this method on longer registers. Future work will focus also on further development of applying nature-inspired techniques in modern cryptography.

References

1. Abd, A.A., Younis, H.A., Awad, W.S.: Attacking of stream Cipher Systems Using a Genetic Algorithm. Journal of the University of Thi Qar 6, 1–6 (2011)
2. Ahmad, A., Hayat, L.: Selection of Polynomials for Cyclic Redundancy Check for the use of High Speed Embedded – An Algorithmic Procedure. Wseas Transactions on Computers 10(1) (January 2011)
3. Bergmann, K.P.: Cryptanalysis Using Nature-Inspired Optimization Algorithms, master's thesis (2007)
4. Garg, P., Shastri, A.: An Improved Cryptanalytic Attack on Knapsack Cipher using Genetic Algorithm. World Academy of Science, Engineering and Technology, 553–560 (2007)
5. Garg, P.: A Comparison between Memetic algorithm and Genetic algorithm for the cryptanalysis of Simplified Data Encryption Standard algorithm. International Journal of Network Security & Its Applications (IJNSA) 1(1), 34–42 (2009)

6. Goldberg, D.E.: Genetic algorithms in search, optimization, and machine learning, 3rd edn. Wydawnictwa Naukowo-Techniczne, Warszawa (2003) (in Polish)
7. Hołubowicz, W., Płóciennik, P.: GSM cyfrowy system telefonii komórkowej, 2nd edn. Przedsiębiorstwo Wielobranżowe "Rokon", Poznań (1997)
8. Hospodar, G., et al.: Machine learning in side-channel analysis: a first study. J. Cryptogr. Eng., 293–302 (2011)
9. Mehrotra, A.: GSM System Engineering. Artech House, London (1997)
10. Robling-Denning, D.E.: Cryptography and Data Security. Wydawnictwa Naukowo-Techniczne, Warszawa (1995) (in Polish)
11. Russell, M.D., Clark, J.A., Stepney, S.: Making the Most of Two Heuristics: Breaking Transposition Ciphers with Ants (2003)
12. Schneier, B.: Applied Cryptography, 2nd edn. Wydawnictwa Naukowo-Techniczne, Warszawa (2002) (in Polish)
13. Selvi, G., Purusothaman, T.: Cryptanalysis of Simple Block Ciphers using Extensive Heuristic Attacks. European Journal of Scientific Research 78(2), 198–221 (2012)
14. Song, J., et al.: Cryptanalysis of Four-Round DES Based on Genetic Algorithm (2007)
15. Yaseen, I.: Breaking multiplicative knapsack ciphers using agenetic algorithm (1998)

Author Index