Advances in Intelligent Systems and Computing 290

Sigeru Omatu · Hugues Bersini
Juan M. Corchado · Sara Rodríguez
Paweł Pawlewski · Edgardo Bucciarelli *Editors*

Distributed Computing and Artificial Intelligence, 11th International Conference

Springer

Advances in Intelligent Systems and Computing

Volume 290

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/11156

About this Series

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within “Advances in Intelligent Systems and Computing” are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Sigeru Omatu · Hugues Bersini
Juan M. Corchado · Sara Rodríguez
Paweł Pawlewski · Edgardo Bucciarelli
Editors

Distributed Computing and Artificial Intelligence, 11th International Conference

Springer

Editors
Sigeru Omatu
Faculty of Engineering
Osaka Institute of Technology
Osaka
Japan

Hugues Bersini
Université Libre de Bruxelles
Brussels
Belgium

Juan M. Corchado
Faculty of Science
Department of Computing Science and Control
University of Salamanca
Salamanca
Spain

Sara Rodríguez
Faculty of Science
Department of Computing Science and Control
University of Salamanca
Salamanca
Spain

Paweł Pawlewski
Faculty of Engineering Management
Poznan University of Technology
Poznan
Poland

Edgardo Bucciarelli
Dep. PPEQS, Section of Economics and Quantitative Methods
University of Chieti-Pescara
Pescara
Italy

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-07592-1 ISBN 978-3-319-07593-8 (eBook)
DOI 10.1007/978-3-319-07593-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014939945

Printed on acid-free paper

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

Preface

Artificial intelligence is changing our society. Its application in distributed environments, such as internet, electronic commerce, environment monitoring, mobile communications, wireless devices, distributed computing, to mention only a few, is continuously increasing, becoming an element of high added value with social and economic potential, in industry, quality of life and research. These technologies are changing constantly as a result of the large research and technical effort being undertaken in both universities and businesses.

The 11^{th} International Symposium on Distributed Computing and Artificial Intelligence 2014 (DCAI 2014) is a forum to present applications of innovative techniques for solving complex problems in these areas. The exchange of ideas between scientists and technicians from both the academic and industrial sector is essential to facilitate the development of systems that can meet the ever-increasing demands of today's society. The present edition brings together past experience, current work and promising future trends associated with distributed computing, artificial intelligence and their application in order to provide efficient solutions to real problems. This conference is a stimulating and productive forum where the scientific community can work towards future cooperation in Distributed Computing and Artificial Intelligence areas.

Nowadays it is continuing to grow and prosper in its role as one of the premier conferences devoted to the quickly changing landscape of distributed computing, artificial intelligence and the application of AI to distributed systems. This year's technical program will present both high quality and diversity, with contributions in well-established and evolving areas of research. Specifically, 75 papers were submitted from over 21 different countries (Algeria, Brazil, China, Croatia, Czech Republic, Denmark, France, Germany, Ireland, Italy, Japan, Malaysia, Mexico, Poland, Portugal, Republic of Korea, Spain, Taiwan, Tunisia, Ukraine, United Kingdom), representing a truly "wide area network" of research activity. The DCAI'14 technical program has selected 64 papers and, as in past editions, it will be special issues in journals such as Neurocomputing, Journal of Artificial Intelligence (IJAI), the International Journal of Imaging and Robotics (IJIR) and the International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI). These special issues will cover extended versions of the most highly regarded works. Moreover, DCAI'14 Special Sessions have been a very useful tool in order to complement the regular program with new or emerging topics of particular interest to the participating community. Special Sessions that emphasize on multi-disciplinary and transversal aspects, such as *AI-driven methods for Multimodal Networks and Processes Modeling* and *Multi-Agents Macroeconomics* have been especially encouraged and welcome.

This symposium has been organized by the Bioinformatics, Intelligent System and Educational Technology Research Group (http://bisite.usal.es/) of the University of Salamanca. The present edition was held in Salamanca, Spain, from 4th to 6th June 2014.

We thank the sponsors (Indra, INSA - Ingeniería de Software Avanzado S.A., IBM, JCyL, IEEE Systems Man and Cybernetics Society Spain, AEPIA *Asociación Española para la Inteligencia Artificial*, APPIA *Associação Portuguesa Para a Inteligência Artificial*, CNRS *Centre national de la recherche scientifique*), the Ministerio de Economía y Competitividad (Spain) with the project *Sociedades Humano-Agente: Inmersión, Adaptación y Simulación* (TIN2012-36586-C03-03 - Project co-financed with FEDER funds), and finally, the Local Organization members and the Program Committee members for their hard work, which was essential for the success of DCAI'14.

Salamanca
June 2014

Sigeru Omatu
Hugues Bersini
Juan M. Corchado
Sara Rodríguez
Paweł Pawlewski
Edgardo Bucciarelli (Eds.)

Organization

General Chairs

José M. Molina	Universidad Carlos III de Madrid, Spain
James Llinas	State University of New York, USA
Andre Ponce de Leon F. de Carvalho	University of Sao Paulo at Sao Carlos, Brazil
Ajith Abraham	Norwegian University of Science and Technology

Honorary Chairman

Masataka Inoue	President of Osaka Institute of Technology, Japan

Scientific Committee

Sigeru Omatu (Chairman)	Osaka Institute of Technology, Japan
Hugues Bersini (Co-Chairman)	Université Libre de Bruxelles, Belgium
Adriana Giret	Politechnich University of Valencia, Spain
Alberto Fernández	University Rey Juan Carlos, Spain
Álvaro Herrero	University of Burgos, Spain
Ana Carolina Lorena	Federal University of ABC, Brazil
Ângelo Costa	University of Minho, Portugal
Antonio Moreno	University Rovira y Virgili, Spain
Antonio Manuel de Jesús Pereira	Leiria Polytechnic Institute, Portugal
Araceli Sanchís	University Carlos III of Madrid, Spain
B. Cristina Pelayo García-Bustelo	University of Oviedo, Spain
Bianca Innocenti	University of Girona, Spain
Bogdan Gabrys	Bournemouth University, UK

Bruno Baruque	University of Burgos, Spain
Carina González	University of La Laguna, Spain
Carlos Carrascosa	Politechnich University of Valencia, Spain
Carmen Benavides	University of Leon, Spain
Daniel Glez-Peña	University of Vigo, Spain
David Griol Barres	University Carlos III of Madrid, Spain
Davide Carneiro	University of Minho, Portugal
Dídac Busquets	University of Girona, Spain
Dongshik Kang	Ryukyu University, Japan
Eladio Sanz	Unversity of Salamanca, Spain
Eleni Mangina	University College Dublin, Ireland
Emilio Corchado	University of Burgos, Spain
Eugenio Aguirre	University of Granada, Spain
Eugénio Oliveira	University of Porto, Portugal
Evelio J. González	University of La Laguna, Spain
Faraón Llorens Largo	University of Alicante, Spain
Fernando Díaz	Univesity of Valladolid, Spain
Fidel Aznar Gregori	University of Alicante, Spain
Florentino Fdez-Riverola	University of de Vigo, Spain
Francisco Pujol López	Polytechnic University of Alicante, Spain
Fumiaki Takeda	Kochi Institute of Technology, Japan
Germán Gutiérrez	University Carlos III, Spain
Grzegorz Bocewicz	Koszalin University of technology, Poland
Helder Coelho	University of Lisbon, Portugal
Gustavo Santos	University of Salamanca, Spain
Ivan López Arévalo	Lab. of Information T. Cinvestav, Mexico
Jamal Dargham	University of Malaysia, Saba, Malaysia
Javier Carbó	University Carlos III of Madrid, Spain
Javier Martínez Elicegui	Telefónica I+D, Spain
Jesús García Herrero	University Carlos III of Madrid, Spain
Joao Gama	University of Porto, Portugal
Johan Lilius	Åbo Akademi University, Finland
José R. Villar	University of Oviedo, Spain
Juan A. Botía	University of Murcia, Spain
Juan Pavón	Complutense University of Madrid, Spain
José M. Molina	University Carlos III of Madrid, Spain
José R. Méndez	University of Vigo, Spain
José V. Álvarez-Bravo	University of Valladolid, Spain
Joseph Giampapa	Carnegie Mellon, USA
Juan Manuel Cueva Lovelle	University of Oviedo, Spain
Juan Gómez Romero	University Carlos III of Madrid, Spain
Kazutoshi Fujikawa	Nara Institute of Science and Technology, Japan

Lourdes Borrajo	University of Vigo, Spain
Luis Alonso	University of Salamanca, Spain
Luis Correia	University of Libon, Portugal
Luis F. Castillo	Autonomous University of Manizales, Colombia
Luís Lima	Polytechnic of Porto, Portugal
Manuel González-Bedia	University of Zaragoza, Spain
Manuel Pegalajar Cuéllar	University of Granada, Spain
Manuel Resinas	University of Sevilla, Spain
Marcilio Souto	Federal University of Rio Grande do Norte, Brazil
Margarida Cardoso	ISCTE, Portugal
Maria del Mar Pujol López	University of Alicante, Spain
Masanori Akiyoshi	Osaka University, Japan
Masaru Teranishi	Hiroshima Institute of Technology, Japan
Michifumi Yoshioka	Osaka Prefecture University, Japan
Miguel Delgado	University of Granada, Spain
Miguel Ángel Patricio	University Carlos III of Madrid, Spain
Miguel Rebollo	University of Vigo, Spain
Miguel A. Vega-Rodriguez	University of Extremadura, Spain
Mohd Saberi Mohamad	University of Technology Malaysia, Malaysia
Michael Zaki	University Rostok, Germany
Noki Mori	Osaka Prefecture University, Japan
Norihisa Komoda	Osaka University, Japan
Paulo Novais	Polytechnic University of Minho, Portugal
Pawel Pawlewski	Poznan University of Technology, Poland
Peter Fobrig	University Rostok, Germany
Ichiro Satoh	National Institute of Informatics, Japan
Rafael Corchuelo	Catholic University of Sevilla, Spain
Ramón Rizo	University of Alicante, Spain
René Leistikow	University Rostok, Germany
Ricardo Campello	University of Sao Paulo, Brazil
Ricardo Costa	Polytechnic of Porto, Portugal
Rubén Fuentes	Complutense University of Madrid, Spain
Rui Camacho	University of Porto, Portugal
Sebastian Ahrndt	Technische Universität Berlin, Germany
Sebastian Bader	University of Rostock, Germany
Seiichi Ozawa	Kobe University, Japan
Seiji Ito	Hiroshima Institute of Technology, Japan
Seyedsaeid Mirkamali	University of Mysore, India
Shanmugasundaram Hariharan	B.S.Abdur Rahman University Vandalur, India
Silvana Aciar	University of Girona, Spain

Xinyu Que	IBM Research, USA
Shiro Tamaki	University of the Ryukyus, Japan
Sofya V. Zhukova	St. Petersburg State University, Russia
Toru Yamamoto	Hiroshima University, Japan
Vicente Botti	Politechnich University of Valencia, Spain
Vicente Julián	Politechnich University of Valencia, Spain
Victor J. Sosa-Sosa	Laboratory of Information Technology - Cinvestav, México
Viet-Hoang VU	Hanoi University of Technology, Vietnam
Waldo Fajardo	University of Granada, Spain
Worawan Diaz Carballo	Thammasat University, Thailand
Yi Fang	Purdue University, USA
Yusuke Nojima	Osaka Prefecture University, Japan
Yutaka Maeda	Kansai University, Japan
Zbigniew Pasek	IMSE/University of Windsor, Canada

AI–Driven Methods for Multimodal Networks and Processes Modeling Special Session Committee

Grzegorz Bocewicz (Chairman)	Koszalin University of Technology, Poland
PawełPawlewski (Chairman)	Poznan University of Technology, Poland
Irena Bach-Dábrowska (Co-Chairman)	Gdańsk University of Technology, Poland
Izabela E. Nielse (Co-Chairman)	Aalborg University, Denmark
Peter Nielsen (Co-Chairman)	Aalborg University, Denmark
Robert Wójcik (Co-Chairman)	Wrocław University of Technology, Poland
Zbigniew Banaszak (Co-Chairman)	Warsaw University of Technology, Poland

Multi-Agents Macroeconomics Special Session Committee

Edgardo Bucciarelli (Chairman)	University of Chieti-Pescara, Italy
Gianfranco Giulioni (Chairman)	University of Chieti-Pescara, Italy
Alan Kirman	Aix-Marseille Université, France
Anwar Shaikh	The New School for Social Research, New York, United States
Barry Cooper	University of Leeds, United Kingdom
Enrico Rubaltelli	University of Padua, Italy

Felix Freitag	Universitat Politècnica de Catalunya, Spain
José Carlos R. Alcantud	University of Salamanca, Spain
Katsunori Yamada	Osaka University, Japan
Leandro Navarro	Universitat Politècnica de Catalunya, Spain
Nicola Mattoscio	University of Chieti-Pescara, Italy
Shu-Heng Chen	National Chengchi University, Taipei, Taiwan
Stefano Zambelli	University of Trento, Italy

Organizing Committee

Juan M. Corchado (Chairman)	University of Salamanca, Spain
Sara Rodríguez (Co-Chairman)	University of Salamanca, Spain
Juan F. De Paz (Co-Chairman)	University of Salamanca, Spain
Javier Bajo	Polytechnic University of Madrid, Spain
Fernando de la Prieta Pintado	University of Salamanca, Spain
Davinia Carolina Zato Domínguez	University of Salamanca, Spain
Gabriel Villarrubia González	University of Salamanca, Spain
Antonio Juan Sánchez Martín	University of Salamanca, Spain
Dante I. Tapia	University of Salamanca, Spain
Emilio S. Corchado	University of Salamanca, Spain
Belén Pérez Lancho	University of Salamanca, Spain
Angélica González Arrieta	University of Salamanca, Spain
Vivian F. López	University of Salamanca, Spain
Ana de Luis	University of Salamanca, Spain
Ana B. Gil	University of Salamanca, Spain

Contents

Artificial Intelligence Applications

Multi-Agent Systems

Distributed, Grid, Cloud and Mobile Computing

Bioinformatics, Biomedical Systems, E-health

Data Mining, Information Extraction, Semantic, Knowledge Representation

Image Processing, Tracking, Robotic, Control and Industrial Systems

Reachability Modeling for Multimodal Networks Prototyping

Grzegorz Bocewicz[1], Robert Wójcik[2], and Zbigniew Banaszak[3]

[1] Department of Electronics and Computer Science, Koszalin University of Technology, Sniadeckich 2, 75-453 Koszalin, Poland
bocewicz@ie.tu.koszalin.pl

[2] Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wroclaw, Poland
robert.wojcik@pwr.wroc.pl

[3] Department of Business Informatics, Warsaw University of Technology, Narbutta 85, 02-524 Warsaw, Poland
Z.Banaszak@wz.pw.edu.pl

Abstract. A declarative model aimed at reachability-driven refinement of the multimodal networks (MNs) cyclic steady state space is proposed. The concept of multimodal processes executed in goods/passengers transportation or data transmission networks where several closed loop structure subnetworks interact each other via distinguished subsets of common shared hubs as to provide a variety of demand-responsive goods or data transportation/handling services is employed. Multimodal processes throughput depends on their cycle time that is on cycle time reachable in considered MN. Therefore, searching for the MN's cyclic steady state behavior the following question is considered: Is the cyclic steady state space reachable in the given network structure? The declarative approach employed makes it possible to evaluate the reachability of cyclic behaviors on a scale that reflects real practice.

Keywords: multimodal network, initial states reduction-free method, cyclic scheduling, constraint programming.

1 Introduction

Multimodal route planning that aims to find an optimal route between the source and the target of a trip while utilizing several transportation modes including different passenger/cargo transportation systems, e.g. ship, airline, AGV systems, train and subway networks, are of significant importance [1, 2, 3, 6]. Multimodal processes (MP) executed in the multimodal networks (MN), depending on the networks' nature can be seen either as passengers and/or goods flow transferred between different modes to reach their destination [3] or as stream of packets transferred between the source and destination nodes.

MPs planning problems, i.e. taking into account MPs routing and scheduling can be found in different application domains (such as manufacturing, intercity fright transportation, multimodal passenger transport network combining several unimodal

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_1,

networks as well as distance learning environment and data and supply media flows, e.g., cloud computing, oil pipeline, computer networks) [1, 3, 6, 8].

The local transportation/transmission processes serviced by different modes, while executed along unimodal networks (lines, channels), are usually cyclic. Hence, MPs supported by them also have the periodic character. That means that the periodicity of MPs depends on periodicity of unimodal (local) processes executed in MN. Of course, the MN throughput is maximized by the minimization of its cycle time.

Apart from such typically used, above mentioned, large scale networks let us focus on smaller scale objects responsible for MP synchronization. In computer networks such a role plays a router [8]. Assuming the given packet routings linking assumed source-destination nodes have to be dedicated, e.g. for emergency cases. Since the routings should be deadlock-free while of guaranteed capacity the capability of routers employed play of primary role. In this context the paper discusses the importance of transmission channels structure in assuring routers robustness. In other words, a router's channels structure design, enabling assumed output of packet routings seen as MPs, is of our main interest.

The declarative models employing the constraint programming techniques implemented in modern platforms such as OzMozart, ILOG, [2, 3] seems to be well suited to cope with MPs planning problems. The existing approach to the concurrently flowing cyclic processes scheduling problem is based upon the simulation models, e.g. the Petri nets [13], the algebraic models [14] upon the (max,+) algebra or the artificial intelligent methods [7, 15-17]. The constraints programming driven models of multimodal networks composed of a set of local cyclic processes have been studied in [2]. Therefore, this work can be seen as a continuation of our former investigations conducted in [2, 3]. In that context, our paper provides contribution to reachability modeling of multimodal networks as well as to multimodal processes prototyping.

Section 2 provides a description of a multimodal network while modeling its structural and behavioral features. In Section 3 we formulate the problem of cyclic states space generation for the given MN structure. The method of solving this problem is shown in Section 4. Then, Sections 5 and 6 describe the conducted experiments and provide conclusions resulting from them.

2 Modelling of Multimodal Networks

2.1 Structure

Fig. 1 shows an example of a **four-port router** for Video on Demand [7] where four local packets transmission processes (P_1, P_3, P_2 and P_4) support two MP servicing a computer network dedicated routings. MP are transmitted along two routings: distinguished by blue – mP_1 and red – mP_2, lines. MP of this kind can be modeled as the SCMCP shown in Fig. 1b). The class SCMCP is assumed to include two types of processes:

- *local processes* (representing modes of transport – P_1, P_2, P_3, P_4), whose operations are cyclically repeated along the set routes (sequences of successively used resources). For the system from Fig. 1b), the routes of local processes are defined as follows:

  $p_1 = (R_6, R_2, R_3)$, $p_2 = (R_5, R_1, R_2)$, $p_3 = (R_4, R_1, R_8)$, $p_4 = (R_3, R_4, R_7)$,

- *multimodal processes* $(mP_1,\ mP_2)$ representing streams of packets Operations of the multimodal processes are implemented cyclically along routes being compositions of fragments of routes of local processes representing resources used for transporting materials along a given route. For the system from Fig. 1b), the routes of multimodal processes are defined as follows:

$$mP_1 = \big((R_5, R_1, R_2), (R_2, R_3), (R_3, R_4, R_7)\big),\ mP_2 = \big((R_6, R_2, R_3), (R_3, R_4), (R_4, R_1, R_8)\big).$$

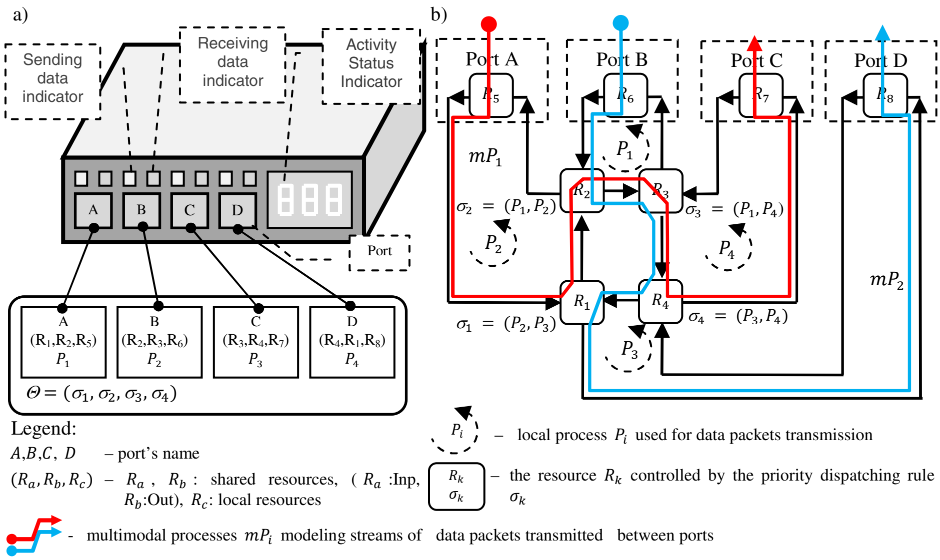


Fig. 1. Exemplary four-port router for Video on Demand [7] a), and its SCMCP model b)

Process operation are implemented on two kinds of resources: own resources (each of them is used by only one process of a given kind – R_5, R_6, R_7, R_8) and shared resources (each of them is used by more than one process of a given kind – R_1, R_2, R_3, R_4). Process uses resources that are shared in the mode of mutual exclusion, i.e. in a given moment only one process operation of a given kind can be implemented on a resource (yet, process operations of different kinds, local and multimodal, can be implemented simultaneously).

The access to shared resources is given in the sequence determined by the dispatching rules $\Theta = \{\Theta^0, \Theta^1\}$ [2, 4]. It is assumed that $\Theta^l = \{\sigma_1^l, \ldots, \sigma_k^l, \ldots, \sigma_{lk}^l\}$, where σ_k^l – is the sequence whose elements determine the order in which the processes (local $l = 0$ / multimodal $l = 1$) are provided with access to the resource R_k. In case of the system from Fig. 1b), the access to shared resources is determined by the following rules:

$$\sigma_1^0 = (P_2, P_3),\ \sigma_2^0 = (P_1, P_2),\ \sigma_3^0 = (P_2, P_3),\ \sigma_4^0 = (P_3, P_4),$$
$$\sigma_1^1 = (mP_2, mP_1),\ \sigma_2^1 = (mP_1, mP_2),\ \sigma_3^1 = (mP_1, mP_2),\ \sigma_4^1 = (mP_2, mP_2).$$

The subsequent operation starts right after the current operation is completed, providing that the resource indispensible to its implementation is available. While waiting for the busy resource, the process does not release the resource which was assigned for implementing the previous operation. Moreover, an assumption is made that processes are of no-expropriation nature, and the times and sequence of operations performed by the processes do not depend on external interferences.

The parameters described above constitute the structure of SCMCP that determines its behavior. Formally, the structure of SCMCP is defined as the following tuple [2]:

$$SC = \big((R, SL), SM\big)\ , \tag{1}$$

where: $R = \{R_k \mid k = 1, \ldots, lk\}$ – set of resources,

$SL = (P, U, O, T, \Theta^0)$ – structure of local cyclic processes, where:

P – set of local processes; U – set of routes of local cyclic processes; O – set of operations; T – set of operation times; Θ^0 – set of dispatching rules;

$SM = (mP, mU, mO, mT, \Theta^1)$ – structure of multimodal processes, where:

mP – set of multimodal processes mP_i; mU – set of routes of multimodal processes, mO – set of operations; mT – set of operation times; Θ^1 – set of dispatching rules for multimodal processes.

2.2 Behavior

In the systems of concurrent cyclic processes, the behavior is usually presented [2], as schedules determining the moments of initiating all the operations implemented within them. Fig. 2a) provides an example of such a schedule that determines the way of implementing the processes of SC structure from Fig. 1b).

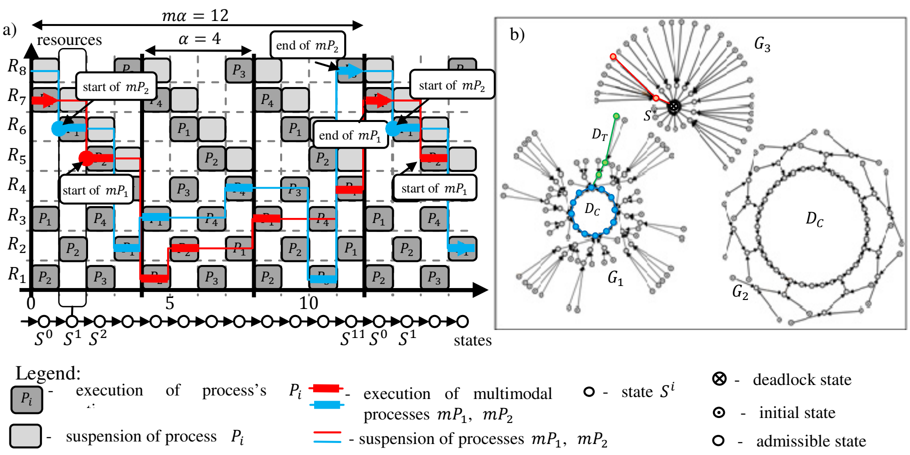


Fig. 2. Cyclic schedule for structure from Fig. 1b) a), and the corresponding states space $\mathcal{P}$ b)

The presented schedule is an example of cyclic behavior, i.e. the successive states of the processes are reachable with the constant period. The operations of local processes are repeated with the period $\alpha = 4$ t.u. (time units) and the multimodal processes with $m\alpha = 12$ t.u. Of course, the periodicity of multimodal processes $m\alpha$ depends on SCMCP periodicity [2].

In this approach, each behavior can be comprehended as a sequence of successive states (subsequent allocations of processes, as well successively changing, according to the rules Θ of access rights). In case of the schedule from Fig. 2a), it is a sequence of the following 12 states $S^0, S^1, S^2, \ldots, S^{11}$.

Formally the SCMCP state is defined as follows [2]:

$$S^r = (Sl^r, mS^r), \tag{2}$$

where Sl^r means the r^{th} state of local processes:

$$Sl^r = (A^r, Z^r, Q^r),$$

$A^r = (a_1{}^r, a_2{}^r, \ldots, a_k{}^r, \ldots, a_{lk}{}^r)$ – **allocation** of local processes In the r-th state, $a_k{}^r \in P \cup \{\Delta\}$.

$Z^r = (z_1{}^r, z_2{}^r, \ldots, z_k{}^r, \ldots, z_m{}^r)$ – **sequence of semaphores** of the r-th state, $z_k{}^r \in P$ – **semaphore** determining the process (an element of rule σ_k^0), which has an access to the resource R_k next in the sequence.

$Q^r = (q_1{}^r, q_2{}^r, \ldots, q_k{}^r, \ldots, q_m{}^r)$ – **sequence of semaphore indexes** of the r-th state, $q_k{}^r$ – **index** determining the position of the semaphore value $z_k{}^r$ in the dispatching rule σ_k^0, $z_k{}^r = s_{k,(q_k{}^r)}$, $q_k{}^r \in \mathbb{N}$..

mS^r means the r^{th} state of multimodal processes:

$$mS^r = (mA^r, mZ^r, mQ^r),$$

$mA^r = (ma_1{}^r, ma_2{}^r, \ldots, ma_k{}^r, \ldots, ma_m{}^r)$ – sequence of multimodal processes allocation in the r^{th} state, $ma_k{}^r \in mP \cup \{\Delta\}$,

$mZ^r = (mz_1{}^r, mz_2{}^r, \ldots, mz_k{}^r, \ldots, mz_m{}^r)$ – sequence of semaphores of the r-th state, $mz_k{}^r \in mP$ – determines the process (an element of the rule σ_k^1, ascribed to R_k), which has the access right to the resource R_k,

$mQ^r = (mq_1{}^r, mq_2{}^r, \ldots, mq_k{}^r, \ldots, mq_m{}^r)$ – sequence of semaphore indexes of the r^{th} state, $mq_k{}^r$ determines the position of the semaphore value $mz_k{}^r$ in the dispatching rule σ_k^1, $mz_k{}^r = s_{k,(mq_k{}^r)}$, $mq_k{}^r \in Z$.

Behaviors of the system characterized by various sequences of subsequently reachable states S^r (2) can be illustrated in a graphical form as a state space $\mathcal{P}$. Fig. 2b) shows an example illustrating this possibility for the SCMCP model from Fig. 1b). If we take the graph-theoretical interpretation of the space $\mathcal{P}$ the diagraph corresponding to it is represented by the pair $\mathcal{P} = (\mathbb{S}, \mathbb{E})$, where $\mathbb{S}$ means a set of admissible SCMCP states [2], $\mathbb{E} \subseteq \mathbb{S} \times \mathbb{S}$ means a set of arcs representing transitions between SCMCP states (transitions take place according to the function $S^f = \delta(S^e)$ described in [2].

Cyclic behaviors appearing in the space $\mathcal{P}$ form cyclic subgraphs, e.g. digraph G_1, see Fig. 2b). The initiation of process implementation from an arbitrary state of the digraph G_1 results in attaining (e.g. by means of transition states) the states being part of a cycle (states marked with blue). A set of such sates depicted by a cyclic route is called as a **cyclic steady state**, and defined [2] as the sequence $D_C = (S^{d_1}, \ldots, S^{d_i}, S^{d_{i+1}}, \ldots, S^{d_{ld}})$, of admissible states: S^{d_i}, $S^{d_{i+1}} \in \mathbb{S}$, following conditions: $S^{d_{i+1}} = \delta(S^{d_i})$, $i = 1 \ldots (ld-1)$, $S^{d_1} = \delta^{lp}(S^{d_{ld}})$. In the case, the all operations have the same execution times ($mt_{i,j} = t_{i,j} = 1$ [t.u]), a period of SCMPC is equal to the number of states contained by cyclic steady state: $m\alpha = ld$.

Besides of cyclic steady states D_C the so called transient states leading to above mentioned cyclic steady states (see the sequence of states marked with green on Fig. 2b): $D_T = (S^{t_1}, \ldots, S^{t_i}, S^{t_{i+1}}, \ldots, S^{t_{lt}})$, where: $S^{t_i}, S^{t_{i+1}} \in \mathbb{S}$, $S^{t_{i+1}} = \delta(S^{t_i})$,

$i = 1 \dots (lt-1)$ and $\delta(S^{t_{lt}})$ belongs to the cyclic steady state Dc, are distinguished.

It should be emphasized that not all of the state space $\mathcal{P}$ result in such a cyclic steady states. Most states lead to deadlock states (marked with the symbol ⊗), which in practice mean an interrupt of the system resulting from the occurrence of a closed chain of requests.

In general case the space $\mathcal{P}$ may not include cyclic steady states. Such situations are quite common in the systems with high density of processes (e.g. in transportation systems). In the context of such systems, the question of evaluating the cyclic behavior attainability becomes crucial.

3 Problem Formulation

Problems of admissible states space $\mathcal{P}$ generation and than reachability of cyclic steady states have been studied in our previous work [2, 3]. In this paper a special case of cyclic steady state reachability problem is considered. The cyclic steady states periodicity $m\alpha$ of which do not exceeds arbitrarily assumed value $m\alpha_H$ are sought. This kind of problem is defined in following way.

There is SCMCP with the structure SC (1) including:

- set of resources R and set of dispatching rules Θ,
- local processes P described by routes p_i, operation sequences O_i and their duration times T_i,
- multimodal processes mP dscribed by the routes mp_i, operation sequences mO_i and their duration times mT_i,

An answer is sought to the question: *Does the state space $\mathcal{P}$ generated by SCMCP structure SC, consist cyclic steady states D_C periodicity of which $m\alpha$ do not exceeds an arbitrarily given value $m\alpha_H$?*

4 Cyclic Steady State Generation

In order to answer the question posed above, we have developed an approach that uses the following properties of the states space $\mathcal{P}$:

(i) initial state (see Fig. 1b) marked with the symbol ⊙) is a state without input arcs, $SB \subseteq \mathbb{S}$ – means a set of all initial states of the space $\mathcal{P}$ [2],

(ii) deadlock state (marked with the symbol ⊗) is a state without output arcs [2],

(iii) the length lt of any transient state (i.e., the number of states belonging to transient state) leading to the cyclic steady state D_C composed of ld states follows the inequality $lt \leq ld$. This property results from assumptions imposed on SCMPC [2].

The proposed approach assuming "reduction" of possible initial states leading to the cyclic steady states follows from the method of exhaustive initial states SB elimination [3]. The property (iii) is used in the course of cyclic steady state of assumed periodicity $m\alpha = ld \leq m\alpha_H$ searching. That means the transient states containing more than $2m\alpha_H$ can be eliminated from the searching process.

Therefore, an idea standing behind of the proposed method of initial states reduction (ISRM) assumes successive generation of states of the set SB successors till obtaining again the state that has been already generated (the obtained states sequence determines the admissible cyclic steady state). The searching ends in case the number of successive generated states exceeds $2m\alpha_H$. The following algorithm is implemented in ISRM:

Algorithm 1

function $\textsc{CyclicSteadyStateAlphaGeneration}(D_S, C_N, \delta, \alpha_H)$
  $PS_{SB} \leftarrow \left((S^r, D_S), C_N \cup \{S^r \neq \delta(S^e), \forall S^e \in \mathbb{S}^{lp}\}\right)$
  $SB \leftarrow \textsc{SearchAll}(PS_{SB})$; $V_{DC} \leftarrow \emptyset$
  for $i = 1$ **to** $|SB|$
    $S^{i,1} \leftarrow SB(1)$; $SB \leftarrow SB \backslash S^{i,1}$; $V^i \leftarrow \{S^{i,1}\}$
    for $j \leftarrow 2$ **to** $2m\alpha_H$
      $S^{i,j} \leftarrow \delta(S^{i,j-1})$
      if $S^{i,j} \notin V^i$ **then** $V^i = V^i \cup \{S^{i,j}\}$
    end
    $V_{DC,i} \leftarrow \textsc{CheckCyclic}(V^i)$; $V_{DC} \leftarrow V_{DC} \cup V_{DC,i}$
  end
return V_{DC}

where: $PS_{SB} = \left((S^r, D_S), C_N \cup \{S^r \neq \delta(S^e), \forall S^e \in \mathbb{S}^{lp}\}\right)$ – constraint satisfaction problem, which solution is an initial state S^r, see property (i); D_S – domain determining admissible values of variables (allocation, semaphores, indexes) characterizing the state S^r;C_N – set of constraints of admissibility of states S^r; $\{S^r \neq \delta(S^e), \forall S^e \in \mathbb{S}\}$ – set of constraint that guarantees that S^r is an initial state (there is no state S^e leading to the state S^r); α_H – assumed, admissible value of the cyclic steady state period; SB – set of initial states, $V_{DC,i}$ – set of states contained in the cyclic steady state $D_{C,i}$ reached from the initial state $S^{i,1}$; V_{DC} – set of states contained in cyclic steady states periodicity of which do not increase α_H;$\textsc{SearchAll}(PS)$ – function returning the set of all solutions to the problem PS (in case of no solution the function returns an empty set), $\textsc{CheckCyclic}(V^i)$ – function returning the subset $V_{DC,i} \subseteq V^i$.

First of all the set of initial states SB is calculated (i.e., solution to the problem PS_{SB} is determined) with help of $\textsc{SearchAll}$ procedure implemented in constraint programming environment (OzMozart [2, 3], ECL^{i}PS^e [11, 12], ILOG). Then the successive successors $S^{i,j}$ of initial states $S^{i,1}$ are generated. Due to the property (iii) the number of non-repeating successors of each initial state cannon succeed value equal to $2m\alpha_H$. In case the condition of the property (iii) does not hold the algorithm stops that means the rest of potentially reachable states lead either to a cyclic steady state periodicity of which is greater than $m\alpha_H$ or just to a deadlock state.

5 Computational Experiment

Consider a system from Fig. 1b) that represents the network of four packets transmission processes of router from Fig. 1a). In the network two MP servicing a computer network dedicated routings take place, see multimodal processes: mP_1, mP_2. All operation times of operations are equal to 1 [t.u.] : $mt_{i,j} = t_{i,j} = 1$.

For thus defined system an answer is sought to the question: whether exits the cyclic steady state periodicity of which do not exceed $m\alpha_H = 12$. In order to calculate an admissible cyclic steady state an Algorithm 1 (it is assumed the Algorithm 1 stops after first admissible cyclic steady state is found) has been used. The total time of calculations was no longer than 2 seconds (Oz Mozart, Intel Core Duo2 3.00 GHz).

The state space generated for the initial states SB of SCMCP model from Fig. 1b) is shown in Fig. 3a). In order to calculate the cyclic steady state $D_{C,1}$ periodicity of which do not exceeds $m\alpha_H = 12$, generation of 117 states have been required. That means twice less than in an algorithm from [3].

In order to evaluate an efficiency of the method proposed several experiments for different $m\alpha_H$ values have been executed. Some results are show in Fig. 3b), where the efficiency of an initial states reduction-free method [3] is underlined by the green line. Presented comparison shows that for some values α_H (i.e. for $ld < 15$ and $ld < 30$) the ISRM has a better efficiency than a methods implementing exhausting search algorithm.

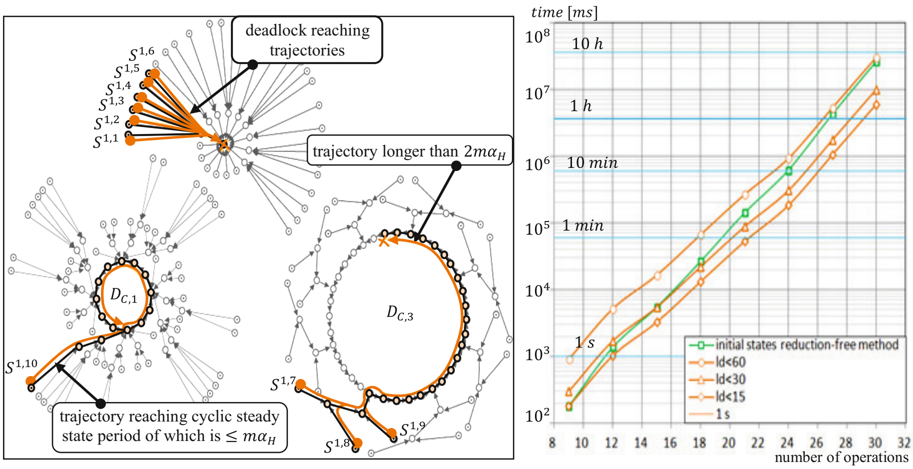


Fig. 3. Trajectories obtained by Algorithm 1 for the system from Fig 1b), comparison of ISRM and initial states reduction-free method [3] efficiency

6 Conclusions

The presented method of prompt prototyping of cyclic steady state space $\mathcal{P}$ (i.e. the states reachable in the given structure of the system), is one of the computer-assisted methods of creating different variants of multimodal networks. The use of the declarative approach makes it possible to evaluate the reachability of cyclic behaviors on a scale that reflects real practice (in the considered case the calculation time was no longer than 2 seconds).

The real-life cases of multimodal networks are strongly affected by the imprecise character of available information (such as time duration, moment of initiation, etc.). The presented expectations specify the development directions for the presented model and method, providing a possibility of taking into consideration the imprecise character of decision variables, e.g. within the Fuzzy Constraints Satisfaction Problem [9], framework.

References

1. Abara, J.: Applying integer linear programming to the fleet assignment problem. Interfaces 19, 4–20 (1989)
2. Bocewicz, G., Banaszak, Z.: Declarative approach to cyclic steady state space refinement: periodic process scheduling. The International Journal of Advanced Manufacturing Technology 67(1-4), 137–155 (2013)
3. Grzegorz, B., Zbigniew, B., Paweł, P.: Cyclic Steady State Space Refinement. In: Szewczyk, R., Zieliński, C., Kaliczyńska, M. (eds.) Recent Advances in Automation, Robotics and Measuring Techniques. AISC, vol. 267, pp. 11–20. Springer, Heidelberg (2014)
4. Korytkowski, P., Rymaszewski, S., Wiśniewski, T.: Ant Colony Optimization for job shop scheduling using multi-attribute dispatching rules. International Journal of Advanced Manufacturing Technology 67, 231–241 (2013)
5. Levner, E., Kats, V., Alcaide, D., Pablo, L., Cheng, T.C.E.: Complexity of cyclic scheduling problems: A state-of-the-art survey. Computers & Industrial Engineering 59(2), 352–361 (2010)
6. Pawlewski, P.: Multimodal approach to model and design supply chain. In: Proceedings of 7th IFAC Conference on Manufacturing Modelling, Management, and Control, St. Petersburg, pp. 2110–2115 (2013)
7. Polak, M., Majdzik, P., Banaszak, Z., Wójcik, R.: The performance evaluation tool for automated prototyping of concurrent cyclic processes. Fundamenta Informaticae 60(1-4), 269–289 (2004)
8. Puente, V., Gregorio, J.A., Beivide, R., Vallejo, F.: A New Routing Mechanism for Networks with Irregular Topology. In: Proceedings of Conference of Supercomputing, ACM/IEEE (2001)
9. Relich, M.: A declarative approach to new product development in the automotive industry. In: Environmental Issues in Automotive Industry, EcoProduction, pp. 23–45 (2014)
10. Różewski, P., Małachowski, B.: System for creative distance learning environment development based on competence management. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part IV. LNCS, vol. 6279, pp. 180–189. Springer, Heidelberg (2010)
11. Sitek, P., Wikarek, J.: A hybrid approach to supply chain modeling and optimization. In: Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1223–1230 (2013)
12. Sitek, P., Wikarek, J.: A hybrid method for modeling and solving constrained search problems. In: Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 385–392 (2013)
13. Song, J.-S., Lee, T.E.: Petri net modeling and scheduling for cyclic job shops with blocking. Computers & Industrial Engineering 34(2), 281–295 (1998)
14. Von Kampmeyer, T.: Cyclic scheduling problems. Ph.D. Dissertation, Fachbereich Mathematik/Informatik, Universität Osnabrück (2006)
15. Yu, H., Lu, F.: A multi-modal route planning approach with an improved genetic algorithm. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 38, 343–348 (2010)
16. Zhang, X., He, Z., Pan, Y.: Study on Multimodal Transport Network Model Base on Genetic Algorithm Method. In: ICLEM, pp. 3514–3520 (2010)
17. Zidi, S., Maouche, S.: Ant Colony Optimization for the rescheduling of multimodal transport networks. In: IMACS Proceedings of Multiconference on Computational, Engineering in Systems Applications, vol. 1, pp. 965–971 (2006)

Hybrid Solution Framework for Supply Chain Problems

Paweł Sitek and Jarosław Wikarek

Institute of Management and Control Systems, University of Technology, Kielce, Poland
{sitek,j.wikarek}@tu.kielce.pl

Abstract. The paper presents application and implementation aspects of a hybrid approach to modeling and optimization for supply chain problems. Two environments of mathematical programming (MP) and constraint programming (CP) were integrated into Hybrid Solution Framework (HSF). The strengths of MP and CP, in which constraints are treated in a different way and different methods are implemented, were combined to use the strengths of both. The proposed approach is particularly important for the decision models in logistic and supply chain management, where an objective function and many discrete decision variables added up in multiple constraints.

Keywords: constraint logic programming, mathematical programming, optimization, supply chain management, hybrid methods.

1 Introduction

The issue of the supply chain management (SCM) is the area of science and practice that has been strongly developing since the '80s of the last century. The aim of supply chain management is to increase sales, reduce costs and take full advantage of business assets by improving interaction and communication between all the actors forming the supply chain. The supply chain management is a decision process that not only integrates all of its participants but also helps to coordinate the basic flows: products/services, information and funds. Changes in the global economy and the increasing globalization lead to the widespread use of IT tools, which enables continuous, real-time communication between the supply chain links. The above mentioned problems are characterized by a very large number of integer decision variables, and different types of constraints (time, resource, capacity, transportation, financial, etc).

The vast majority of the works reviewed [3] have formulated models for supply chain management as linear programming (LP), integer programming (IP) and mixed integer linear programming (MILP) problems and solved them using the Operations Research methods. Despite this, constraint programming (CP) seems to offer the best environments for modeling supply chain management problems. This is due to the declarative nature of CP–based environments and the support for various types of constraints.

Most problems that the constraint programming concerns belong to the group which conventional programming techniques find hardest. Time needed to solve such

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_2,

problems using unconstrained search increases exponentially with the problem size. Unfortunately, CP–based environments poorly deal with the optimization problems.

The main contribution of this paper is the concept and description of the Hybrid Solution Framework (HSE). HSE is based on a combination of constraint programming (CP) and mathematical programming (MP).

2 Constraint Logic Programming (CLP)

CLP is a form of CP, in which LP (logic programming) is extended to include concepts from constraint satisfaction. A constraint logic program is a program that contains constraints in the body of clauses. Constraints can also be present in the goal. These environments are declarative. CLP is also a tool for solving constraint satisfaction problems (CSP) [1]. For the important combinatorial case CSP is characterized by following features:

- a finite set S of integer variables $X_1,..,X_n$, with values from finite domains $D_1,..,D_n$;
- a set of constraints between variables. The i–th constraint $C_i(X_{i1},..,X_{ik})$ between k variables from S is given by a relation defined as subset of the Cartesian product $D_{i1} \times,..,\times D_{ik}$ that determines variable values corresponding to each other in a sense defined by the problem considered;
- a CSP solution is given by any assignment of domain values to variables that satisfies all constraints.

The semantics of constraint logic programs can be defined in terms of a virtual interpreter that maintains a pair <G,S> during execution. The first element of this pair is called the current goal; the second element is called the constraint store. The current goal contains the literals the interpreter is trying to prove and may also contain some constraints it is trying to satisfy; the constraint store contains all constraints the interpreter has assumed satisfiable so far. At the beginning, the current goal is the goal and the constraint store is empty. The interpreter proceeds by removing the first element from the current goal and analyzing it. In the end this analysis should produce a successful termination or a failure. This analysis could involve recursive calls and addition of new literals to the current goal and new constraint to the constraint store. The interpreter backtracks if a failure is generated. A successful termination is generated when the current goal is empty and the constraint store is satisfiable. CLP can use Artificial Intelligence (AI) techniques to improve the search: constraint propagation, data–driven computation, “forward checking” and “lookahead” [1].

3 Motivation and Contribution

The declarative approach and the use of logic programming provide incomparably greater possibilities for decision problems modeling than the pervasive approach based on mathematical programming. Unfortunately, these environments deal with optimization to a lesser degree.

Based on [1,2,3],[6],[9,10,11] and our previous work [7,8] we observed and checked some advantages and disadvantages of CP–based and MP–based environments. An integrated approach of constraint logic programming (CLP) and mixed integer programming / mixed integer linear programming (MIP/MILP) can help to solve decision and optimization problems that are intractable with either of the two methods alone [4,5].

Both MIP/MILP and finite domain CP/CLP involve variables and constraints. However, the types of the variables and constraints that are used, and the way the constraints are solved, are different in the two approaches. Thus, some types of constraints are easier to solve in the CLP and the other in the MP, and vice versa [5].

The motivation and contribution behind this work was to create a hybrid framework for constrained decision and optimization problems modeling, solving and optimization instead of using CP–based and MP–based environments separately.

In presented approach to modeling, solving and optimization supply chain management problems we proposed the hybrid solution environment, where:

- knowledge related to the problem can be expressed as linear and logical constraints (implementing all types of constraints of the previous MILP/MIP/IP models [8] and introducing new types of constraints (logical, nonlinear, symbolic etc.));
- the decision models solved using the proposed framework can be formulated as a mathematical programming model or constraint programming model, or it can also be a hybrid model;
- the problem is modeled in CP–based environment, which is far more flexible than MP–based environment;
- the novel method of constraint propagation is introduced (obtained by transforming the decision model to explore its structure);
- constrained domains of decision variables, new constraints and values for some variables are transferred from CP–based environment into MP–based environment.

4 Hybrid Solution Environment (HSE)

Both environments have advantages and disadvantages. CP–based environments are declarative and ensure a very simple modeling of decision problems, even those with poor structures if any. The problem is described by a set of logical predicates. The constraints can be of different types (linear, non–linear, logical, binary, etc.). The CLP does not require any search algorithms. This feature is characteristic of all declarative backgrounds, in which modeling of the problem is also a solution, just as it is in Prolog, SQL, etc. The CLP seems perfect for modeling any decision problem.

Constantly improved methods and mathematical programming algorithms, such as the simplex algorithm, branch and bound, branch–and–cost [2] etc., have become classics now. Traditional methods when used alone to solve complex problems provide unsatisfactory results. This is related directly to different treatment of variables and constraints in those approaches [4,5]. The proposed hybrid approach, a composition of methods, offers the optimal system for specific contexts. The concept of this framework with its phases (P1..P5, G1..G3) is presented in Fig. 1. The names and descriptions of the phases and the implementation environment is shown in Tab. 1.

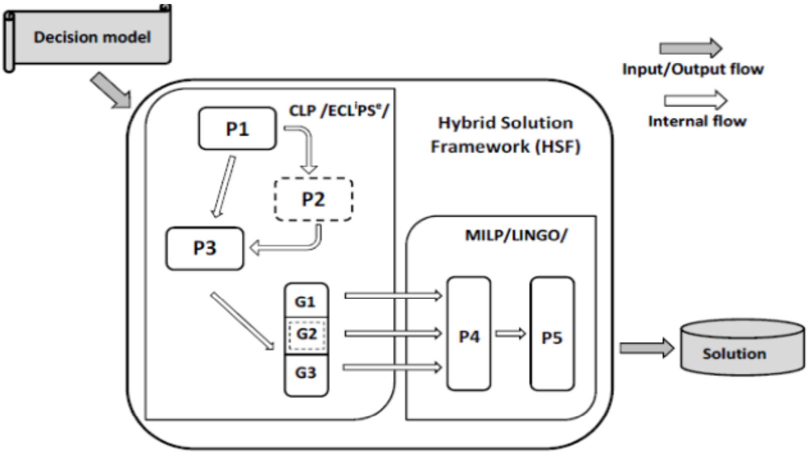


Fig. 1. Scheme of the hybrid solution framework (HSF)

Table 1. Description of Phases

Name	Description
P1-Implementation of decision model – [environment: CLP].	The implementation of the model in CLP, the term representation of the problem in the form of predicates.
P2-Transformation of implemented model for better constraint propagation (optional) – [environment: CLP].	The transformation of the original problem aimed at extending the scope of constraint propagation. The transformation uses the structure of the problem. The most common effect is a change in the representation of the problem by reducing the number of decision variables, and the introduction of additional constraints and variables, changing the nature of the variables, etc.
P3-Constraint propagation – [environment: CLP]	Constraint propagation for the model. Constraint propagation is one of the basic methods of CLP. As a result, the variable domains are narrowed, and in some cases, the values of variables are set, or even the solution can be found.
G1-Generation of MILP/MIP/IP model – [environment: CLP].	Generation of the model for mathematical programming. Generation performed automatically using CLP predicate. The resulting model is in a format accepted by the MILP system.
G2-Generation of additional constraints (optional) – [environment: CLP].	Generation of additional constraints on the basis of the results obtained in step PH3.
G3- Generation domains of decision variables and other values - [environment: CLP]	Generation of domains for different decision variables and other parameters based on the propagation of constraints. Transmission of this information in the form of fixed value of certain variables and/or additional constraints to the MP.
P4-Merging MILP/MIP/IP model – [environment: MILP].	Merging files generated during the phases G1-G3 into one file. It is a model file format in MILP system.
P5-Solving MILP/MIP/IP model – [environment: MILP].	The solution of the model from the previous stage by MILP solver.

From a variety of tools for the implementation of the CP–based environment in HSF, ECL^{i}PS^{e} software [12] was selected. ECL^{i}PS^{e} is an open-source software system for the cost–effective development and deployment of constraint programming applications. MP-based environment in HSF was LINGO by LINDO Systems [13]. ECL^{i}PS^{e} software is the environmental leader in HSE. ECL^{i}PS^{e} was used to implement the following phases of the framework: P1, P2, P3, G1, G2, G3 (Fig. 1, Table 1.). The transformed files of the model were transferred from ECL^iPS^e to LINGO where they were merged (P4). Then the complete model was solved using LINGO efficient solvers (P5). Constraint propagation (phase–P3) greatly affected the efficiency of the solution. Therefore phase P2 was introduced. During this phase, the transformation was performed using the structure and properties of the model. This is an optional phase that depends on the modeled problem. The details of this phase will be presented in one of the illustrative examples in Chapter 5 (cost optimization of supply chain).

5 Illustrative Example-Implementation of the Model

The proposed framework was verified and tested for illustrative example. The example is the authors' original model of cost optimization of supply chain with multimodal transport. The mathematical model and its discussion are presented in [8]. The model was formulated as a MILP problem [2],[8] in order to test the HSF (Fig. 1) against the classical integer–programming environment [2],[13]. The simplified structure of the supply chain network for this model, composed of factories, distributors and customers is presented in Fig. 2.

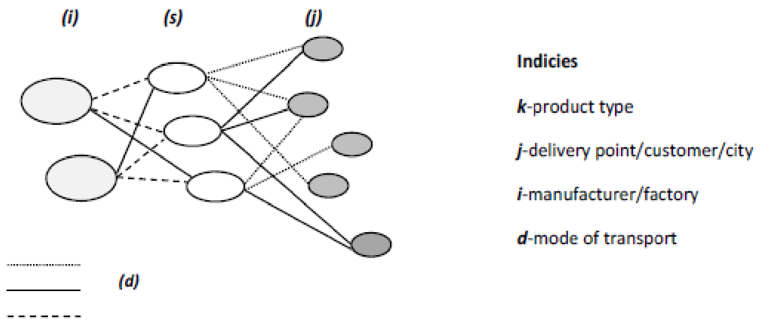


Fig. 2. The simplified structure of the supply chain network for illustrative example

In the construction of this model following assumptions are valid:

- the shared information process in the supply chain consists of resources (capacity, versatility, costs), inventory, production, product, transport, demand, etc;
- the transport is multimodal. (several modes of transport. A limited number of means of transport for each mode);
- the environmental aspects of use of transport modes;
- different products are combined in one batch of transport;
- the cost of supplies is presented in the form of a function;
- can occur constraints on the common distribution of certain products.

Another is also in this approach, the representation of the problem. Because it contains only one value is not set while in the classical approach, there are five. Each

CLP predicate has a corresponding multi-dimensional vector representation. While modeling both problems, quantities *i, s, k, d* and decision variable (deliver quantity) $X_{i,s,k,d}$ were vector parameters (Fig. 3a). As shown in Fig. 3b, for each vector instance there are 5 values to be determined. They define the size of the delivery, factories, distributors involved in the delivery and the mode of transport. Symbols necessary to understand both the representation of the problem and their descriptions are presented in Table 2.

[O_n,K,J,D,F,Tu,Tu,Oq,X,T]

Fig. 3a. Representation of the problem in the classical approach-vector definition

[[o_1,k1,j1,_,_,_,_,10,_,8], [o_2,k1,k2,_,_,_,_,20,_,6],…]

Fig. 3b. Representation of the problem in the classical approach- a set of vector instances in the process of finding a solution

[[name_1,f1,k1,c1,j1,s1,s1,5,12,100,_],[name_2,f1,k1,c1,j1,s1,s2,6,14,100,_],
[name_3,f1,k1,c1,j1,s2,s1,6,22,100,_],...]

Fig. 4. Representation of the problem in the hybrid approach-route model

Table 2. Symbols used in the representation of the problem

Symbol	Description	Symbol	
O_n	order number	Tu	transport unit, Tu $\in \{d_1,d_2, \ldots, d_l\}$
K	products, K $\in \{k_1,k_2, ..., k_o\}$	T	delivery time/period
J	customers J $\in \{j_1,j_2, \ldots, j_m\}$	Oq	order quantity
D	distributors, S $\in \{s_1,s_2, \ldots, s_e\}$	X	delivery quantity
F	factories, F $\in \{f_1,f_2, \ldots, f_n\}$	Name_	routes name-number

The process of finding the solution may consist in using the constraint propagation methods, variable labeling and the backtracking mechanism. The numbers of parameters that must be specified/labeled in the given predicate/vector critically affect the quality of constraint propagation and the number of backtracks. In both models presented above, the classical problem representation included five parameters*: i, s, k, d* and $X_{i,s,k,d.}$ Considering the domain size of each parameter, the process was complex and time–consuming. In addition, the above representation (Fig. 3a, Fig. 3b) arising from the structure of the problem is the cause of many backtracks. Our idea involved the transformation of the problem by changing its representation without changing the very problem. All permissible routes were first generated based on the fixed data and the set of customer's orders, then the specific values of parameters *i, s, k, d* were assigned to each of the routes. In this way, only decision variables $X_{i,s,k,d}$ (deliveries) had to be specified (Fig. 4). This transformation fundamentally improved the efficiency of the constraint propagation and reduced the number of backtracks. A route model is a name adopted for the models that underwent the transformation.

6 Numerical Experiments

In order to verify and evaluate the proposed approach, many numerical experiments were performed for both illustrative examples. All the examples relate to the supply chain with seven manufacturers *(i=1..7)*, three distributors *(s=1..3)*, ten customers *(j=1..10)*, three modes of transport *(d=1..3)*, and twenty types of products *(k=1..20)*.

The series of experiments was designed to show the advantages of the hybrid approach and framework used. The experiments started with eight examples: E1..E8 for the problem formulated in MILP [8]. Two approaches were used to implement the proposed model: mathematical programming (LINGO) and the hybrid framework HSF). The examples E1..E8 varied in terms of the number of orders (*No*).

Table 3. Results of numerical experiments

E(No)	***MILP-LINGO***				***MILP-HSF***			
	Fc	***T***	***V (int)***	***C***	***Fc***	***T***	***V(int)***	***C***
E1(5)	806	6	6881(6429)	5326	806	1	223(207)	536
E2(10)	1756*	600**	6881(6429)	6271	1756	2	299(277)	541
E3(20)	2960*	600**	6881(6429)	8161	2878	15	487(455)	552
E4(30)	3609*	600**	6881(6429)	10051	2488	34	645(603)	562
E5(50)	6186*	600**	6881(6429)	12831	5596	370	896(853)	563
E6(60)	6853*	600**	6881(6429)	15721	6309	298	1027(993)	564
E7(80)	8854*	600**	6881(6429)	19501	8379	343	1348(1301)	566
E8(100)	15459*	600**	6881(6429)	23659	10791	423	1703(1657)	566

Fc	the optimal value of the objective function
T	solution finding time
V(int)/C	the number of variables (integer variables) /constraints
*	the feasible value of the objective function after the time T=600s
**	calculation was stopped after T=600s

The analysis of the outcome indicates that the hybrid framework provided better results in terms of the time needed to find the solution in each case, and to obtain the optimal solution in some cases, which was impossible to do within the acceptable time limits using the traditional approaches. The results confirmed the superiority of the applied framework over the MP–based approach. The application of HSF allows solving larger examples than those of the classic approach (order of magnitude), as was checked on the same decision-making model [8].

7 Conclusion and Discussion on Possible Extension

The efficiency of the proposed approach is based on the reduction of the combinatorial problem and using the best properties of both environments. The hybrid approach (HSF) makes it possible to find better solutions in the shorter time (Table 3).

In addition to solving larger problems faster, the proposed approach provides virtually unlimited modelling options. Therefore, the proposed solution is recommended for decision–making problems that have a structure similar to the presented models [8]. This structure is characterized by the constraints and objective function in which the decision variables are added together. Further work will focus on running the optimization models with non-linear and logical constraints, multi-objective, uncertainty etc. in the hybrid optimization framework. It is also planned to use the proposed approach to other areas related to the multimodality [14].

References

1. Apt, K., Wallace, M.: Constraint Logic Programming using Eclipse. Cambridge University Press (2006)
2. Schrijver, A.: Theory of Linear and Integer Programming. John Wiley & Sons (1998) ISBN 0-471-98232-6
3. Mula, J., Peidro, D., Diaz-Madronero, M., Vicens, E.: Mathematical programming models for supply chain production and transport planning. European Journal of Operational Research 204, 377–390 (2010)
4. Williams, H.P.: Logic and Integer Programming. Springer, Berlin (2009)
5. Bockmayr, A., Kasper, T.: Branch-and-Infer: A Framework for Combining CP and IP. In: Constraint and Integer Programming. Operations Research/Computer Science Interfaces Series, vol. 27, pp. 59–87 (2004)
6. Dang, Q.V., Nielsen, I.: Simultaneous scheduling of machines and mobile robots. In: Corchado, J.M., et al. (eds.) PAAMS 2013. CCIS, vol. 365, pp. 118–128. Springer, Heidelberg (2013)
7. Sitek, P., Wikarek, J.: A Declarative Framework for Constrained Search Problems. In: Nguyen, N.T., Borzemski, L., Grzech, A., Ali, M. (eds.) IEA/AIE 2008. LNCS (LNAI), vol. 5027, pp. 728–737. Springer, Heidelberg (2008)
8. Sitek, P., Wikarek, J.: Cost optimization of supply chain with multimodal transport. In: Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1111–1118 (2012)
9. Pawlewski, P.: Multimodal Approach to Model and Design Supply Chain. In: Proceedings of IFAC MIM Conference, St. Petersburg (2013)
10. Bocewicz, G., Wójcik, R., Banaszak, Z.: Cyclic Scheduling for Supply Chain Network. In: Rodríguez, J.M.C., Pérez, J.B., Golinska, P., Giroux, S., Corchuelo, R. (eds.) Trends in PAAMS. AISC, vol. 157, pp. 39–48. Springer, Heidelberg (2012)
11. Bocewicz, G., Banaszak, Z.: Declarative approach to cyclic steady states space refinement: periodic processes scheduling. International Journal of Advanced Manufacturing Technology 67(1-4), 137–155 (2013)
12. Eclipse – The Eclipse Foundation open source community website, `http://www.eclipse.org`
13. LINDO Systems – Optimization Software: Integer Programming, Linear Programming, Nonlinear, `http://www.lindo.com`
14. Różewski, P., Małachowski, B.: System for Creative Distance Learning Environment Development Based on Competence Management. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part IV. LNCS, vol. 6279, pp. 180–189. Springer, Heidelberg (2010)

Scheduling of Mobile Robots with Preemptive Tasks

Izabela Nielsen[1], Quang-Vinh Dang[1], Peter Nielsen[1], and Pawel Pawlewski[2]

[1] Dept. of Mechanical and Manufacturing Engineering, Aalborg University, Aalborg, Denmark
{izabela,vinhise,peter}@m-tech.aau.dk
[2] Faculty of Engineering Management, Poznań University of Technology, Poznań, Poland
pawel.pawlewski@put.poznan.pl

Abstract. This paper deals with the problem of scheduling of mobile robots taking into account preemption cases in a flexible manufacturing system (FMS). In addition to capability of transporting materials between some machines, mobile robots are able to perform manufacturing tasks at other machines by using their manipulation arms. These manufacturing tasks can be preempted to allow mobile robots to transport materials when needed. The performance criterion is to minimize time required to complete all tasks, i.e. makespan. A mixed-integer programming (MIP) model is formulated to find the optimal solutions for the problem. Numerical experiments are investigated to demonstrate results of the proposed approach.

Keywords: preemptive scheduling, mobile robots, mixed-integer programming.

1 Introduction

The automation technology in combination with advances in production management has dramatically changed the equipment used by manufacturing companies as well as the issues in planning and control. With these changes, mobile robots have become more popular in several industrial areas, e.g., automotive, chemical and pump manufacturing. Mobile robot is a term used to refer to robotic systems consisting of a robot arm mounted on a mobile platform which allows performance of tasks requiring both locomotion and manipulation abilities [10]. Some mobile robots are capable of transporting a variety of part types from one location to another location similar to material handling devices, e.g. automated guided vehicles (AGVs). Moreover, the mobile robots have the capabilities to perform manufacturing tasks at machines, workstations, or production lines. These tasks consist of such processes as: machine tending, pre-assembly and quality inspection [10]. These mobile robots can have enough flexibility to switch from one type of tasks (e.g. pre-assembly tasks) to another (e.g. transporting tasks). Furthermore, using mobile robots can lead to production efficiency gains, e.g. less energy usage or lower tool-changing costs than conventional industrial robots fixed to one location [8]. The superior capabilities of the mobile robots can pave the way for meeting the growing needs of production systems. Within the scope of this study, a given problem is particularly considered in a flexible manufacturing system (FMS) for a fleet of mobile robots which will not

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_3,

only transport materials of some tasks from one machine to another but also process some manufacturing tasks at the other machines. However, to utilize the mobile robots in an efficient manner require the ability to properly schedule transporting tasks and manufacturing tasks with respect to the needs of manufacturing factories. Therefore, it is important for scheduling to determine in which sequences the mobile robots should process those tasks such that they could effectively work while satisfying a number of practical constraints.

The problem of simultaneous scheduling of machines and mobile robots in an FMS has been modeled in several respects comparable to the joint scheduling problems of machines and AGVs. However, it is different from the problems concerning AGVs in the sense that besides transporting tasks, mobile robots are capable of carrying out manufacturing tasks on the shop floor level. Furthermore, they are flexible enough to switch between manufacturing tasks and transporting tasks. Several approaches and models for exact algorithms have been proposed to address problems of this type ([2], [3], [4], [6], [13]). Heuristic methods (e.g., neural network, genetic algorithm, iterated local search, hybrid method, etc.) are well adapted to study most of the FMS. On the one hand, some works are dedicated to simplified forms of the material handling system of the FMS considering only one transport device ([7], [14]). On the other hand, many works are undertaken on the FMS scheduling with multiple AGVs ([1], [2], [5], [9], [11], [12], [15]).

There have been carried out a number of researches related to the class of mobile robot scheduling problem. However, little attention is paid to the problem of scheduling of multiple mobile robots processing manufacturing tasks that can be preempted in an FMS. In other words, the mobile robots are allowed to interrupt their processing tasks to carry transporting tasks when needed. This constitutes the main novelty of the problem. The remainder of this paper is organized as follows: in Section 2, problem description is discussed while the mathematical model is formulated in Section 3 to find the optimal solutions for the problem. Numerical experiments are conducted to demonstrate results of the proposed model in Section 4. Finally, conclusions and future work are drawn in Section 5.

2 Problem Description

The work is developed for an FMS which products parts or components for the pump manufacturing industry at a factory. Fig. 1 below shows a typical layout of the FMS with mobile robots.

The FMS consists of a number of tasks processed on a set of machines and a set of mobile robots. These tasks are classified into two types which are non-preemptive and preemptive. On the one hand, each non-preemptive task consists of a set of operations that cannot be interrupted, i.e. each operation in this type must be executed without interruption from its starting time to its ending time. On the other hand, each preemptive task considered in this study has only one operation that can be interrupted at any time to let some other operations be executed. There is no restriction on the number of interruptions or on the duration of an interruption. During operation, non-preemptive tasks require mobile robots to transport materials or parts between some machines while preemptive tasks need the participation of the mobile robots in the processing on the other machines. Each mobile robot may carry out the transportation of different non-preemptive tasks, but it is assigned to process only one preemptive task on a specific machine.

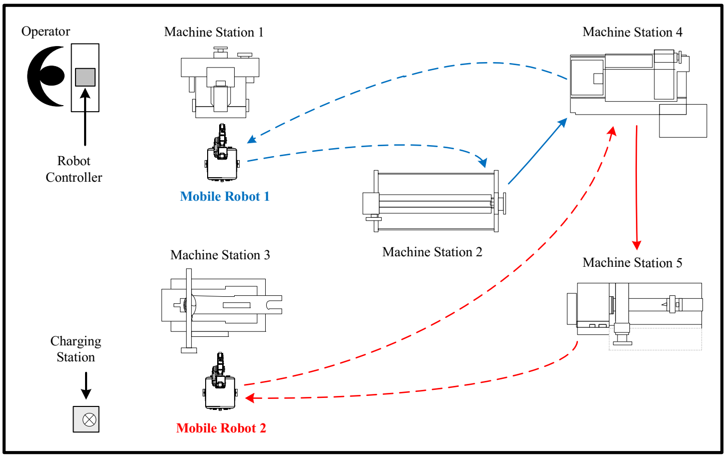


Fig. 1. Typical layout of the FMS with mobile robots

As being occupied by a preemptive task, a mobile robot may be invoked for transportation of a non-preemptive task at some points in the scheduling period. This mobile robot will pause to process the preemptive task, carry out transportation of the non-preemptive task and go back to processing the preemptive task if the preemptive task has not been finished. In practice, e.g. in a pump part manufacturing factory, production operators may set the maximum number of operations of non-preemptive tasks which mobile robots can transport each time being away from their preemptive tasks. This prevents the mobile robots from leaving their preemptive tasks for a long period of time, which may lead to the cancellation of these tasks due to some practical issues on the shop floor. To some extent, this also helps to increase the utilization of these mobile robots. Within the scope of this study, any mobile robot is set to come back to its unfinished preemptive task after each achieved transportation. An example illustrating such preemption case is depicted in Fig. 2. The objective is to find an optimal schedule that minimizes the time required to complete all tasks, i.e. makespan.

To enable the construction of a schedule of mobile robots and machines, the following assumptions are made.

- Each task is available at the beginning of the scheduling period.
- The first operation of each task is available at a machine at the beginning of the scheduling period.
- Each operation sequence of each task (the route of each part type) is available before making scheduling decisions.

- Each mobile robot can transport only one kind of parts at a time.
- There is sufficient input and output buffer space at each machine.
- Traveling time is only machine-dependent and deterministic.
- Loading and unloading time are included in the traveling time of loaded trips.
- Processing time is deterministic.

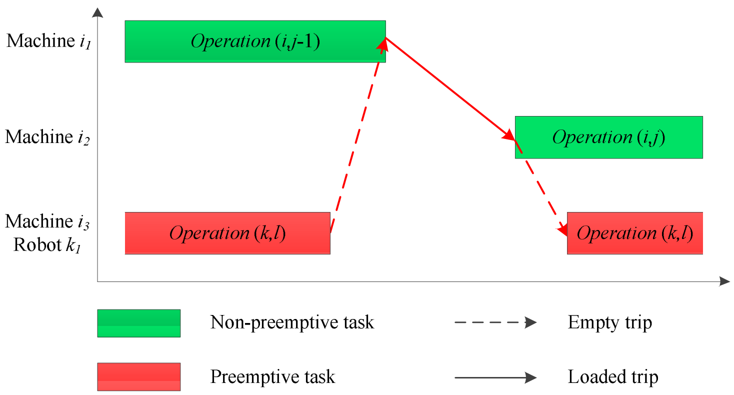


Fig. 2. Illustration of a preemption case

To describe and find optimal solutions for the presented problem, it is necessary to formulate a mathematical model. Solutions found by the mathematical model can be used as reference points as well as for supporting decision making about schedules of mobile robots and machines.

3 Mathematical Model

In this section, a MIP model is formulated to determine an optimal schedule for the problem of scheduling of mobile robots considering preemptive tasks. The notations and formulation of the MIP model are given in the following.

3.1 Notations

i, i', i^* : index of tasks
j, j', j^* : index of operations
k : index of mobile robots
N : set of non-preemptive tasks
P : set of preemptive tasks
M : set of machines
R : set of mobile robots
O_m : set of operations of non-preemptive tasks to be performed on machine m
o_{ij} : the j-th operation of task i
T_{ij} : transportation of operation o_{ij}
n_i : number of operations for task i

p_{ij} : processing time of operation o_{ij}
$t_{ij,i'j'}$: traveling time from machine of operation o_{ij} to machine of operation $o_{i'j'}$

Decision variables

T : time required to complete all tasks
s_{ij}^{p} : starting time of operation o_{ij}
s_{ij}^{t} : starting time of transportation T_{ij}

$$y_{iji'j'} = \begin{cases} 1 & \text{if operation } o_{ij} \text{ precedes operation } o_{i'j'} \\ 0 & \text{otherwise} \end{cases}$$

$$z_{iji'j'} = \begin{cases} 1 & \text{if transportation } T_{ij} \text{ precedes transportation } T_{i'j'} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ijk} = \begin{cases} 1 & \text{if robot } k \text{ carries out transportation } T_{ij} \text{ for operation } o_{ij} \\ 0 & \text{otherwise} \end{cases}$$

3.2 Mixed-Integer Programming Model

The formulation of the MIP model for the presented problem is described as follows. The objective function (1) minimizes makespan *T*. Constraints (2) and (3) represent the operation non-overlapping constraints which imply that a machine cannot process more than one operation at a time, where *L* is a very large number. Constraints (4) and (5) represent the mobile robot non-overlapping constraints which imply that a mobile robot cannot carry out more than one transportation task at a time. Constraint (6) ensures that the transportation of an operation could only start after the predecessor of that operation completes. Constraint (7) ensures that an operation could only start after the transportation of that operation finishes. Constraint (8) ensures that the transportation of an operation is carried by only one mobile robot. Constraints (9) and (10) ensure that completion times of operations of non-preemptive tasks and preemptive tasks respectively cannot exceed the makespan *T*. Constraints (11) and (12) imply the types of variables. The MIP model is given as follows.

Objective function: minimize T (1)

Subject to:

$$s_{ij}^{p} - s_{i'j'}^{p} + Ly_{iji'j'} \geq p_{i'j'} \qquad \forall (i,j),(i',j') \in O_m, m \in M \tag{2}$$

$$s_{i'j'}^{p} - s_{ij}^{p} + L\left(1 - y_{iji'j'}\right) \geq p_{ij} \qquad \forall (i,j),(i',j') \in O_m, m \in M \tag{3}$$

$$s_{i'j'}^{t} - s_{ij}^{t} + L\left(1 - z_{iji'j'}\right) \geq t_{ij-1,ij} + t_{ij,i^*j^*} + t_{i^*j^*,i'j'} \tag{4}$$

$$\forall i,i' \in N, j = 2,\ldots,n_i, j' = 2,\ldots,n_{i'}, \forall i^* \in P, j^* = 1$$

$$z_{iji'j'} + z_{i'j'ij} = x_{ijk}x_{i'j'k} \qquad \forall i,i' \in N, j = 2,\ldots,n_i, j' = 2,\ldots,n_{i'} \tag{5}$$

$$s_{ij}^{p}+p_{ij}\le s_{i,j+1}^{t} \qquad \forall i\in N, j=1,...,n_i-1 \qquad (6)$$

$$s_{ij}^{t}+t_{ij-1,ij}\le s_{ij}^{p} \qquad \forall i\in N, j=2,...,n_i \qquad (7)$$

$$\sum_{k\in R}x_{ijk}=1 \qquad \forall i\in N, j=1,...,n_i \qquad (8)$$

$$s_{ij}^{p}+p_{ij}\le T \qquad \forall i\in N, j=1,...,n_i \qquad (9)$$

$$s_{i^*j^*}^{p}+p_{i^*j^*}+\sum_{i\in N}\sum_{j=2}^{n_i}x_{ijk}\left(t_{i^*j^*,ij-1}+t_{ij-1,ij}+t_{ij,i^*j^*}\right)\le T \qquad \forall i^*\in P, j^*=1, k\in R \qquad (10)$$

$$x_{ijk}, y_{iji'j'}, z_{iji'j'}\in\{0,1\} \qquad \forall i,i'\in N, j=1,...,n_i, j'=1,...,n_{i'}, k\in R \qquad (11)$$

$$s_{ij}^{p}, s_{ij}^{t}\ge 0 \qquad \forall i\in N, j=1,...,n_i \qquad (12)$$

As Constraint (5) is nonlinear, it is replaced by the three following constraints:

$$z_{iji'j'}+z_{i'j'ij}\le x_{ijk} \qquad \forall i,i'\in N, j=2,...,n_i, j'=2,...,n_{i'} \qquad (13)$$

$$z_{iji'j'}+z_{i'j'ij}\le x_{i'j'k} \qquad \forall i,i'\in N, j=2,...,n_i, j'=2,...,n_{i'} \qquad (14)$$

$$z_{iji'j'}+z_{i'j'ij}\ge x_{ijk}+x_{i'j'k}-1 \qquad \forall i,i'\in N, j=2,...,n_i, j'=2,...,n_{i'} \qquad (15)$$

4 Numerical Experiments

In this section, the performance of the MIP model will be tested on a number of problem instances. 10 problem instances are generated with different number of tasks, operations, machines, mobile robots and other parameters. The number of all tasks and number of operations in each non-preemptive task are randomly generated in the ranges of [5, 20] and [2, 6], respectively. The number of machines and mobile robots are respectively distributed within the ranges of [5, 10] and [2, 4]. The processing time of non-preemptive and preemptive operations in time unit are respectively distributed within the ranges of [25, 50] and [100, 200] while the traveling time of mobile robots in time unit are generated in the interval [8, 18]. Note that the time/cost matrices of the generated traveling time should satisfy the triangle inequality. For real time use in production environments, the MIP model is solved considering the maximum computation time of 30 seconds. In the experiments, the MIP model has been programmed and solved by the mathematical modeling language ILOP CPLEX. All the experiments run on a PC having an Intel® Core i5 2.67 GHz processor and 4 GB RAM. The problem sizes and results of the MIP model are shown in Table 1.

It can be observed from Table 1 that the MIP model found the optimal solutions for 5 instances (problem 1 to 5) and feasible solutions for 2 instances (problems 6 and 7). For the other problems, the MIP cannot find any feasible solution within 30 seconds. These results show that the MIP model can be used in practice for small and medium problem sizes (up to around 60 operations, 8 machines and 3 mobile robots). Fig. 3 below depicts the optimal solution of problem instance 1 on a Gantt chart. It can be seen from Fig. 3 that each mobile robot has to interrupt its preemptive task two times to carry out transportation of the non-preemptive tasks. For instance, mobile robot 2 interrupts task 5 the first time to transport task 1 from machine 1 to machine 3 and the second time to transport task 2 also from machine 1 to machine 3. These interruptions consequently divide the duration of each preemptive task (operation) into three separate parts as shown in the Gantt chart. In general, the experiments have illustrated the results of the MIP model.

Table 1. Problem sizes for 10 randomly generated problems

No.	Problem size						MIP	
	Number of tasks	Number of operations	Number of machines	Number of robots	Total variables	Number of constraints	Objective (time unit)	Com. Time (second)
1	5	9	5	2	57	75	164	1.76
2	6	16	5	2	174	296	306	5.68
3	7	25	6	2	460	839	458	13.17
4	9	33	6	2	783	1457	576	21.25
5	11	44	7	3	1985	3748	687	27.95
6	13	53	8	3	2838	5405	1131	30.00
7	14	62	7	3	4124	7930	1595	30.00
8	16	76	9	4	8026	15524	-	-
9	18	85	9	4	9969	19343	-	-
10	20	100	10	4	14017	27330	-	-

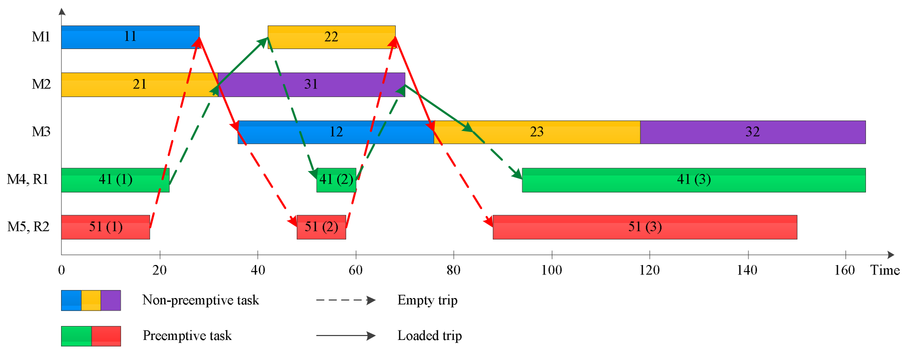


Fig. 3. Gantt chart for the optimal solution of problem instance 1

5 Conclusions

In this paper, a problem of scheduling of mobile robots in an FMS is studied. Two types of tasks which are preemptive and non-preemptive are taken into account. The mobile robots in this study have capabilities to not only transport materials or parts of non-preemptive tasks similar to material handling devices but also process preemptive tasks by using their manipulation arms. The main novelty of this study lies in the fact that the mobile robots must interrupt their preemptive tasks to do transportation of non-preemptive tasks when needed. The objective is to minimize the makespan while satisfying a number of practical constraints. A MIP model is formulated to find optimal solutions for the problem. The results from the numerical experiments show that the MIP model can be applied for small and medium-scale problems in practice. For further research, the complexity of the problem will rapidly grow as taking account of larger number of operations and mobile robots. A meta-heuristic method, namely a genetic algorithm can be considered for solving large-scale mobile robot scheduling problems. Furthermore, re-scheduling mechanism based on obtained schedules and feedback from the shop floor will be developed to deal with real-time disturbances.

Acknowledgments. This work has partly been supported by the European Commission under grant agreement number FP7-260026-TAPAS.

References

1. Abdelmaguid, T.F., Nassef, A.O., Kamal, B.A., Hassan, M.F.: A Hybrid GA/Heuristic Approach to the Simultaneous Scheduling of Machines and Automated Guided Vehicles. Int. J. Prod. Res. 42, 267–281 (2004)
2. Bilge, Ü., Ulusoy, G.: A Time Window Approach to Simultaneous Scheduling of Machines and Material Handling System in an FMS. Oper. Res. 43, 1058–1070 (1995)
3. Blazewicz, J., Eiselt, H.A., Finke, G., Laporte, G., Weglartz, J.: Scheduling Tasks and Vehicles in a Flexible Manufacturing System. Int. J. Flex. Manuf. Syst. 4, 5–16 (1991)
4. Bocewicz, G.: Robustness of Multimodal Transportation Networks. Eksploatacja i Niezawodnosc – Maintenance and Reliability 16, 259–269 (2014)
5. Bocewicz, G., Banaszak, Z.: Declarative Approach to Cyclic Steady States Space Refinement: Periodic Processes Scheduling. Int. J. Adv. Manuf. Tech. 67, 137–155 (2013)
6. Caumond, A., Lacomme, P., Moukrim, A., Tchernev, N.: A MILP for Scheduling Problems in an FMS with One Vehicle. Eur. J. Oper. Res. 199, 706–722 (2009)
7. Dang, Q.-V., Nielsen, I.E., Bocewicz, G.: A Genetic Algorithm-based Heuristic for Part-Feeding Mobile Robot Scheduling Problem. In: Rodríguez, J.M.C., Pérez, J.B., Golinska, P., Giroux, S., Corchuelo, R. (eds.) Trends in PAAMS. AISC, vol. 157, pp. 85–92. Springer, Heidelberg (2012)
8. Dang, Q.V., Nielsen, I., Steger-Jensen, K.: Scheduling a Single Mobile Robot Incorporated into Production Environment. In: Golinska, P. (ed.) EcoProduction and Logistics. EcoProduction, pp. 185–201. Springer, Heidelberg (2013)
9. Deroussi, L., Gourgand, M., Tchernev, N.: A Simple Metaheuristic Approach to the Simultaneous Scheduling of Machines and Automated Guided Vehicles. Int. J. Prod. Res. 46, 2143–2164 (2008)
10. Hvilshøj, M., Bøgh, S., Nielsen, O.S., Madsen, O.: Multiple Part Feeding – Real-world Application for Mobile Manipulators. Assemb. Autom. 32, 62–71 (2012)
11. Lacomme, P., Larabi, M., Tchernev, N.: Job-shop Based Framework for Simultaneous Scheduling of Machines and Automated Guided Vehicles. Int. J. Prod. Econ., 24–34 (2013)
12. Lin, L., Shinn, S.W., Gen, M., Hwang, H.: Network Model and Effective Evolutionary Approach for AGV Dispatching in Manufacturing System. J. Intell. Manuf. 17, 465–477 (2006)

13. Sitek, P., Wikarek, J.: A Hybrid Method for Modeling and Solving Constrained Search Problems. In: Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 385–392. IEEE Press, Kraków (2013)
14. Soylu, M., Özdemirel, N.E., Kayaligil, S.: A Self-organizing Neural Network Approach for the Single AGV Routing Problem. Eur. J. Oper. Res. 121, 124–137 (2000)
15. Ulusoy, G., Sivrikaya-Şerifoğlu, F., Bilge, Ü.: A Genetic Algorithm Approach to the Simultaneous Scheduling of Machines and Automated Guided Vehicles. Comput. Oper. Res. 24, 335–351 (1997)

Multimodal Processes Approach to Supply Chain Modeling

Patrycja Hoffa[1], Pawel Pawlewski[2], and Izabela Nielsen[3]

[1] Poznan University of Technology, ul.Strzelecka 11, 60-965 Poznan, Poland
patrycja.hoffa@doctorate.put.poznan.pl
[2] Faculty of Engineering Management, Poznan University of Technology, Poznan, Poland
pawel.pawlewski@put.poznan.pl
[3] Dept. of Mechanical and Manufacturing Engineering, Aalborg University, Aalborg, Denmark
izabela@m-tech.aau.dk

Abstract. The paper presents the results of conducted research works which concern one of the reference models for transport operations in supply chains. The research focuses mainly on defining the multimodal approach to supply chain simulation. In the paper, five basic models for a transport organization are defined. One of them, the relay model, is the base for the case study. Authors propose multimodal approach to prepare a simulation model. According to the defined approach, the procedure of analyzing and building a simulation model is described in detail.

Keywords: supply chain, transport, multimodal process, modeling, simulation.

1 Introduction

A supply chain is defined by the Council of Supply Chain Management Professionals [14] as follows: "Supply Chain Management encompasses the planning and management of all activities involved in sourcing and procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third-party service providers, and customers. In essence, supply chain management integrates supply and demand management within and across companies. Supply Chain Management is an integrating function with primary responsibility for linking major business functions and business processes within and across companies into a cohesive and high-performing business model. It includes all of the logistics management activities noted above, as well as manufacturing operations, and it drives coordination of processes and activities with and across marketing, sales, product design, finance and information technology."

Transport enables cooperation between companies that constitute a supply chain. It is one of the most significant factors affecting lives of the whole communities. Since 2011 the European Commission has been promoting a new transport policy, published in the new White Paper on Transport. The goals of this new policy are set into four main directions [6]:

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_4,

- reduction of the greenhouse gases emission by 60% by 2050 comparing to the 1990 level;
- efficient core network for multimodal intercity travel and transport;
- clear urban transport and commuting;
- creating global hubs in the European Union for long-distance travel and intercontinental flights.

Achieving these objectives requires joint (common) action of both producers of means of transport (eg. environmental aspect [8]) and their users, but especially companies providing transport services.

The organization of transport is a very complex issue which includes many aspects, such as route selection, modes of transport, problems with coordination. According to customers' requirements concerning the time and price, companies find new solutions. The analysis of activities performed in a supply chain enables evaluation and improvement of some aspects. Taking into consideration the solutions of organizing transport tasks we can define 5 basic models [11]: swinging model, swinging-continuous model, radial model, circular model, relay model. These models are the foundation of building and creating appropriate types of a supply chain. The conducted investigations concern the simulation models of the defined models of organizing transport tasks. Based on these models and using modern simulation technologies, we can perform experiments with different rules of management and coordination. We introduce the multimodal process approach to building simulation models of organizing transport tasks. This approach is used in other application like distance learning netowrks, e-learning [7]

The present paper concentrates mainly on:

- the new approach to analyzing transportation tasks,
- the definition of multimodal processes,
- using multimodal processes as the main idea of building a simulation model.

The article is divided into 5 sections. Section 2 refers to multimodal transport in a supply chain. In this section multimodal networks and multimodal processes are defined. Section 3 defines five models of organizing transport tasks and section 4 focuses on the case study based on a more complicated model, the relay model, which is most commonly used in supply chains. Finally, section 5 provides conclusions and suggestion for the further stages of the project.

2 Multimodal Processes in Supply Chain

This section discusses issues related to the analysis of supply chains using multimodal approach. A supply chain is actually a complex and dynamic supply and demand network [13],[10]. This network is formed by suppliers and companies which are connected with use of different kinds of transport. The identification of ineffective processes in both internal and external flow chains requires proper tools for the analysis and decision making.

Both logistic practice and literature on the subject provide various interpretations of intermodal and multimodal networks and transport [3]. These notions are similar in their scope and meaning, yet there are two significant differences between them. According to the Convention on International Multimodal Transport of Goods [12], the multimodal system of transport is defined as an internally integrated system of carrying goods along with accompanying services provided with use of at least two modes of transport on the basis of a multimodal transport contract. The multimodal transport contract is concluded by a multimodal transport operator who takes the responsibility for its performance. In case of intermodal transport, also at least two modes of transport are involved, however its specific feature is the fact that in the whole freight lane only one unit load is used.

The advantages of limiting the use of intermodal transport are as follows [9]:

- the possibility of offering combined freight services so that benefits can be derived from various modes of transport,
- the possibility of reducing the cost of moving goods without deteriorating the quality of freight services,
- the possibility of reducing damages and losses as well as handling and storage of goods by using pallets and containers,
- increasing the elasticity of deliveries by providing customers with better availability of services over time and space.

In simple words, we can assume that an intermodal chain of deliveries is a specific kind of multimodal logistics, which is characterized by a unified load unit, constant in the whole freight line. In the logistic practice, the most commonly used multimodal solutions are sea-air transport and rail-air transport. The multimodality in transport results from the development of containerization: various modes of transport have become more closely related due to fact that the modes of freight, storage and loading of unified load units had to become similar[1].

In the context of transportation networks, a multimodal network is the one in which two or more types of transportation modes (such as walking, riding a train, or driving a car) are modeled. Alternatively, with utility networks, a multimodal network may consist of the differing transmission and distribution systems [4].

The considerations presented above lead to the definition of multimodal processes, i.e. processes executed along the routes consisting of parts of the routes of local processes [2],[5].

3 Models Organizing Transport Tasks

A model, in which a means of transport runs regularly between two points (loading and unloading) is a swinging model. An example of such a model is the daily delivery of fresh meat to the company's butcher shop and then returning an empty vehicle to the base [11].

The swinging-continuous model of transport is a model, where a means of transport is not expected at a loading and unloading place. When it comes to the point, it takes a loaded or empty semi-trailer, trailer, container, body interchangeable and goes to another place. An example of such a model is the delivery of auto parts from subcontractors to a factory. These parts, for better protection, are often shipped in special boxes and containers. A subcontractor provides a full container of specified elements to the factory (the place of unloading), and takes empty containers that are transported to the loading place [11].

When goods are delivered by a means of transport from one place of loading to a lot of places of unloading, it means that we are dealing with the radial model. After unloading at a given point, a means of transport goes back to the place of loading, where it is re-loaded and products are transported to other collection points. An example of such a model is the delivery of products from a central warehouse to regional warehouses, which handle further distribution tasks [11].

Loading at one point and then delivering to the next unloading points is called circular model. It is characteristic for the courier and distribution activities [11].

Another model is the relay model (Fig. 1). It is also called cross-docking. In this model, transport is organized from a loading point to the final point of unloading with use of transshipment points. Goods are usually delivered to the reloading point in a large batch (by vehicles 33-trucks, trains or ships), and then reloaded to the lower means of transport and delivered to the customer [11].

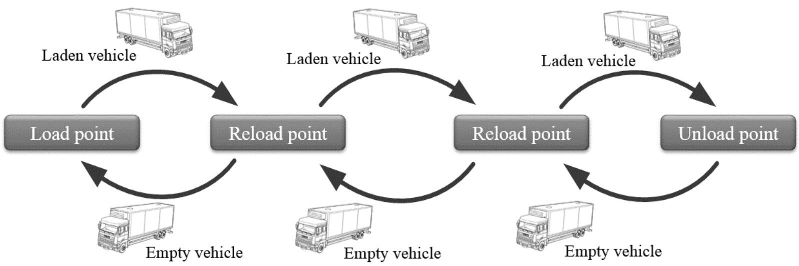


Fig. 1. Relay model. Source: [11].

4 Case Study – Relay Models

According to the definition, multimodal processes are processes consisting of local processes. The implementation of local processes determines the execution of multimodal processes.

In case of the relay model, we can distinguish one multimodal process (PM), whose goal is the delivery of goods from the point of loading to the point of unloading, taking into account the intermediate points – points of reloading – Fig.2. This process consists of three local processes:

- transport of goods from the load point to the reload point no. 1 (PL1),
- transport of goods from the reload point no. 1 to the reload point no. 2 (PL2),
- transport of goods from the reload point no. 2 to the unload point (PL3).

No	Resource
RA	Load point
RB	Reload point no 1
RC	Reload point no 2
RD	Unload point

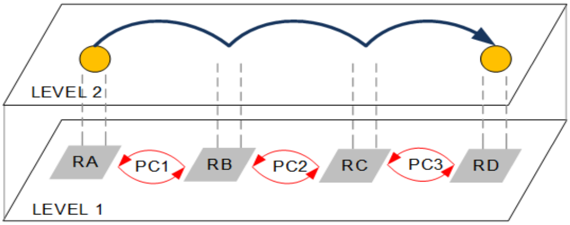


Fig. 2. Relay model – multimodal and local processes

It should be noted that Figure 2 does not show all the processes carried out within the framework of the planned supply chain. When discussing the issue of delivering goods from point A to point B, it is necessary to take into account the activities in the various points of loading, handling and unloading. For example, while analyzing the time of transport service, the time of loading and unloading at specified locations must also be taken into account. Therefore, for each of the highlighted points, it was crucial to present the processes that take place there and affect the transport tasks.

At the loading point, we take into account processes occurring between the part of the warehouse, where the goods are stored as part of the loading. In order to send goods to a customer it is necessary to download them from the storage, deliver them to the loading ramps, and then load them onto a means of transport. Goods can be transported by an employee with no use of equipment or with use of equipment such as a forklift. All these activities have an impact on the whole organized transport task, which directly affects the time of departure from the starting point. Here, we can distinguish two cyclic processes – Fig.3:

- transport of goods from the storage to the temporary storage in the loading parts of a warehouse (PC1),
- transport of goods from the temporary storage in the loading parts of a warehouse to the trailer vehicle.

No	Resource
RA	Warehouse – storage place
RB	Warehouse – place of temporary storage
RC	Load ramps – vehicle of trailer

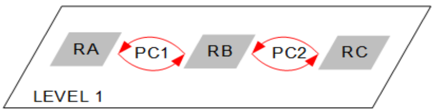


Fig. 3. Loading point – cyclic processes

Other places for consideration are reloading points. Depending on the role of a reloading point, we can distinguish different patterns of action. For the purpose of the present work, the following patterns of action, depending on their functions, can be defined:

1. unloading goods in unloading points and loading goods onto a means of transport, without storage operations (case 1),
2. unloading goods, temporary storage of them and then loading them onto means of transport (case 2),

3. unloading goods, reloading them, and then loading onto a means of transport, without storage operations (case 3),
4. unloading goods, reloading them, temporary storage of them and then loading onto a means of transport (case 4) – Fig.4.

No	Resource
RA	Unload ramps – vehicle of trailer
RB	Warehouse – place of temporary storage
RC	Warehouse – repackaging part
RD	Warehouse – storage part
RE	Warehouse – place of temporary storage
RF	Load ramps – vehicle of trailer

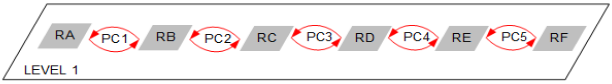


Fig. 4. Cyclic processes in a loading point – case 4

At the point of unloading, we should take into account the processes occurring between the unloading part (vehicle-trailer on the unloading ramp) and the storage area where goods are stored – Fig 5.

No	Resource
RA	Unload ramps – vehicle of trailer
RB	Warehouse – place of temporary storage
RC	Warehouse – storage part

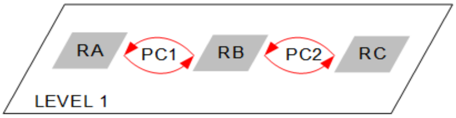


Fig. 5. Cyclic processes in unloading point

For this relay model, we decided to consider the largest number of activities (reloading points are responsible for repackaging and storage). Therefore, the processes are presented with use of the 4-layer approach [5]; Using this approach four levels are considered: business processes (layer 3), multimodal processes (layer 2), local cyclic processes (layer 1), resources (layer 0).

The presented analysis of the processes shows the influence of the operation at one level upon the activity located at a higher level. Therefore, it is possible to map operations that do not create any added value from the standpoint of the object of the process (in this case goods delivered from a company to a customer) yet, they are necessary for the functioning of the process. An example of such a situation is the return of a means of transport from the reloading point to the company or selected rides forklift.

The purpose of a supply chain is to transport goods from a supplier to the end customer (business process level).

The analyzed transport consists of one multimodal process. This multimodal process consists of a number of cyclic process (local), for example:

- load point - get the goods from the storage and transport them to a temporary storage (with a forklift or only an employee),

- the transport of goods from the load point to the reload point no. 1 (using means of transport)
- reload point no. 1 - the transport of goods from a trailer to TMS (with a forklift or only an employee),
- etc.

The lowest level represents resources that are involved in the next stages of the operations throughout the analysis process. Resources are as follows: place of storage in a warehouse (place of temporary storage/rack); place of temporary storage in a warehouse; place of temporary storage in a warehouse; unloading ramps, vehicle trailer; reloading point; loading ramps, vehicle trailer; transport routes at the point of loading; public transport routes etc.

Figure 6 shows the supply chain model as a multilayered model of the system behaviors of concurrent cyclic processes

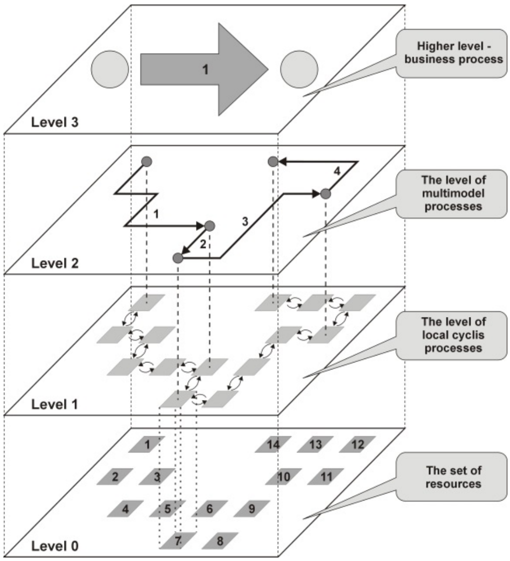


Fig. 6. Four levels approach to process analysis. Source: [5].

5 Conclusions and Further Investigations

The article presents the results of current investigations on developing multimodal approach to building simulation models of organizing transport tasks. A special emphasis was put on deploying the idea of multimodal processes. Based on this approach we analyze and build a simulation model of a logistics center where all transportation tasks are designed with use of multimodal approach. Figure 7 presents the whole simulation model built with use of FlexSim simulation program. Blue arrows present local cyclic processes, red arrows present multimodal processes.

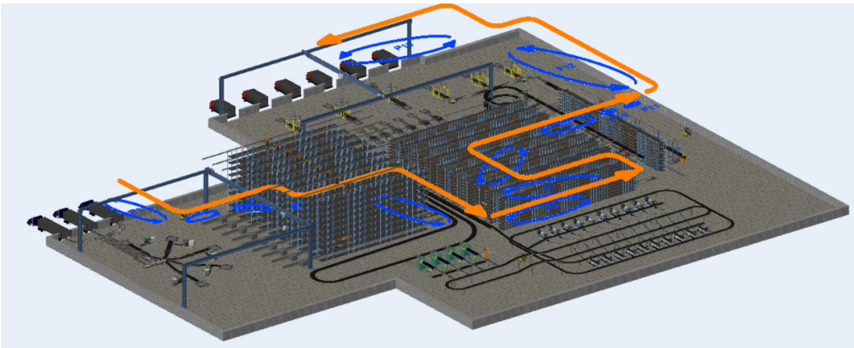


Fig. 7. Logistics Center simulation model constructed on multimodal 4 level approach

The main conclusions are:

- Specification of local processes and multimodal processes (there can be many) makes it possible to analyze every of them independently. It means that we can observe their behaviors independently with no necessity to observe other levels. We can specify the production tact time for every multimodal level and cycle time for a local process.
- It is easy to specify relations between the levels: to evaluate how behaviors of multimodal processes depend on behaviors of local processes and, on the other hand, how behaviors of local processes depend on multimodal processes.

Further stages of investigations will include extending this approach in order to build a complete supply chain simulation model based on the data from practice.

Acknowledgements. Presented research are carried out under the LOGOS project (Model of coordination of virtual supply chains meeting the requirements of corporate social responsibility) under grant agreement number PBS1/B9/17/2013.

References

1. Bocewicz, G., Banaszak, Z.: Declarative modeling of multimodal cyclic processes. In: Golinska, P., Fertsch, M., Marx-Gómez, J. (eds.) Information Technologies in Environmental Engineering. Environmental Science and Engineering – Environmental Engineering, vol. 3, pp. 551–568. Springer, Heidelberg (2011)
2. Bocewicz, G., Nielsen, P., Banaszak, Z.A., Dang, V.Q.: Cyclic Steady State Refinement: Multimodal Processes Perspective. In: Frick, J., Laugen, B.T. (eds.) Advances in Production Management Systems. IFIP AICT, vol. 384, pp. 18–26. Springer, Heidelberg (2012)
3. Golinska, P., Hajdul, M.: Agent-Based System for Planning and Coordination of Intermodal Transport. In: Corchado, J.M., Pérez, J.B., Hallenborg, K., Golinska, P., Corchuelo, R. (eds.) Trends in Practical Applications of Agents and Multiagent Systems. AISC, vol. 90, pp. 99–107. Springer, Heidelberg (2011)

4. Hoel, E.G., Heng, W.-L., Honeycutt, D.: High Performance Multimodal Networks. In: Medeiros, C.B., Egenhofer, M., Bertino, E. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 308–327. Springer, Heidelberg (2005)
5. Pawlewski, P.: Multimodal Approach to Model and Design Supply Chain. In: Proceedings of IFAC MIM Conference, St. Petersburg (2013)
6. Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system, COM, Brussels (March 28, 2011)
7. Różewski, P.: A Method of Social Collaboration and Knowledge Sharing Acceleration for e-learning System: the Distance Learning Network Scenario. In: Bi, Y., Williams, M.-A. (eds.) KSEM 2010. LNCS, vol. 6291, pp. 148–159. Springer, Heidelberg (2010)
8. Rychlik, A., Jasiulewicz-Kaczmarek, M.: Diagnostics systems as a tool to reduce and monitor gas emissions from combustion engines. In: Golinska, P. (ed.) Environmental Issues in Automotive Industry, pp. 95–128. Springer, Heidelberg (2014)
9. Seidl, M., Šimák, L.: Integracja transportu w systemach logistycznych. Zeszyty Naukowe. Logistyka i Transport / Międzynarodowa Wyższa Szkoła Logistyki i Transportu we Wrocławiu 1(8), 159–167 (2009)
10. Sitek, P., Wikarek, J.: A hybrid approach to supply chain modeling and optimization. In: Federated Conference on Computer Science and Information Systems, pp. 1223–1230 (2013)
11. Stajniak, M.I.: Transport and Spedition. ILiM, Poznań (2008)
12. UNC - United Nations Convention on International Multimodal Transport of Goods, Geneva (May 24, 1980)
13. Wieland, A., Wallenburg, C.M.: Supply-Chain-Management in stürmischen. Zeiten, Berlin (2011)
14. http://cscmp.org/about-us/supply-chain-management-definitions (2012)

Multimodal Perspective on Ontologies Combining Problem in Production Management Systems

Przemysław Różewski and Justyna Bednarz

West Pomeranian University of Technology in Szczecin,
Faculty of Computer Science and Information Systems,
ul. Żołnierska 49, 71-210 Szczecin, Poland
{prozewski,jbednarz}@wi.zut.edu.pl

Abstract. The issue of the integration of ontologies in production management systems has not been solved yet. There are a number of methods, each of which is adapted to some special cases. In our opinion the multimodal approach can reduce cost of the ontologies combining problem. We recognize two multimodal-based approaches: Multimodality Theory and Multimodal Network. The first allows to find other instances (modalities) for current ontology. The second creates mechanism to include new domain in ontologies. The paper presents the conceptual approach to the problem of integration of ontologies based on multimodal theory.

Keywords: multimodal network, multimodality theory, ontology integration.

1 Introduction

Knowledge is the base of modern enterprises. Most of manufacturing organizations are trying to meet the constantly increasing demands of the market, directing its activities to improve its efficiency and competitiveness. The development of enterprises is currently based primarily on the knowledge management skills [13]. The problem of many organizations is the complicated access to the knowledge resources and the difficulty of locating their sources. Main objective of ontology is to enable the sharing and reuse of knowledge. Ontologies provide a description of the domain to which they relate, define and classify objects, define the relationships between them and create a common vocabulary [10]. Dissemination of this solution improves communication and knowledge exchange, which allows avoiding misunderstandings between users. Technological advances, globalization, increased customer demands and shorter product life cycle force the implementation of the knowledge management system in production enterprises [9]. Knowledge properly prepared and presented to the outside, helps to attract and keep customers, potential partners, valuable employees and stakeholders. It also helps to build customer satisfaction, create and strengthen brand image, so is the main source of competitive advantage.

The multimodal approach is applied to the problems of the multimodal transportation network analyzing [4] or supply chain modelling and optimization [12,14]. In this

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_5, © Springer International Publishing Switzerland 2014

area multimodal system is using two or more transport modes for a trip between which a transfer is necessary [4]. A mode might be defined by vehicle type or by transport function. According to [5], the content of multimodal processes depending on the networks' nature can be seen either as passengers and/or goods flow transferred between different modes to reach their destination or as stream of packets transferred between the source and destination nodes.

Other view on multimodal approach is related to the human-computer interaction design [1]. In reality person communicated with other person in fully multimodal way. The research issue is to reflect such communication channel in the graphic user interface. The multimodal interface has to covers in some way following six physical media [2]:

- light / vision / graphics;
- sound waves / hearing / acoustics;
- mechanical touch sensor contact / touch / haptics;
- molecular smell sensor contact / smell / olfaction;
- molecular taste sensor contact / taste / gustation;
- proprioceptor stimulation (as when you sense that you are being turned upside down).

In this paper we want to discuss how multimodal perspective helps to solve ontologies combining problem in the production management systems. This is possible because each ontology can be treated as mode characterized by a certain modality. The paper provides conceptual work for the problem and integrated both views on multimodal issue.

2 Ontologies Combining Problem

Ontology mapping, the search for relationships between ontologies and defining the rules of correspondence between their elements, is required to connect dispersed and heterogeneous models [15]. Generally there are classified three types [6]: (1) mapping between integrated global ontology and local ontologies, (2) mapping between local ontologies and (3) mapping in ontology merging and alignment. One of the significant differences between the categories is the way they are built and maintained. There are several basic methods of ontologies combining. Main of them are [6]:

- Ontology alignment - involves creating links between classes and relationships of merged models by their lexical and semantic analysis. It is used in complementary domains and when the sources of ontologies are distinct but mutually consistent.
- Ontology integration - creating new resulting ontology from existing ontologies. Characteristic for this method is that the source ontology domain is not equal to resulting ontology domain, but there is a specific relationship between them.
- Ontology merging - specific case of ontologies integration. Consists of creating a separate, consistent ontology, on the basis of two or more existing and different from each other ontologies of the same domain.

The terms: mapping, matching, alignment, merging and integration are used to describe ontology connections and mappings. C. Keet [8] summarizes the differences in the use of these terms:

- ontology mapping - find equivalent items in the ontologies with different sources;
- ontology matching - find similar parts in the source ontologies or sets translation rules between them;
- ontology alignment - identify equivalents in the source ontologies;
- ontology merging - creates new ontology based on the combination of two ontologies in the same domain;
- ontology integration - creates new ontology based on the combination of two ontologies in various domains.

There may be different models of the same part of the real world. The diversity of views causes heterogeneous solutions. Many of problems with connection of independently developed ontologies were classified. Chalupsky [8] defined the differences in the syntax and expressiveness of the languages, but now there are four types of inconsistencies, which often overlap to some extent.

Heterogeneity of syntax is caused by variety of ontology building languages. Languages differ in their expression and the way they describe reality. Syntax incompatibility refers to both specification of ontology and metamodel and is mainly related definition of elements. This is the most obvious inconsistency and easiest to be observed.

Heterogeneity in the logical representation [8] results from the way the logical concepts were defined. It is a metamodel-level variance. An example is the relationship that in some languages can unequivocally state that two classes are disjoint (eg Disjoint AB), whereas in other languages it is necessary to use negation in statements subclasses (eg A subclass-of (NOT B) , B subclass-of (NOT A)).

Lexical (semantic) heterogeneity [8] refers to the natural language used for naming of items that make the model up. Incompatibility in the terminology occurs typically in the case of terms for which have/are synonyms and homonyms. Sometimes, the same name is used in the syntax of two languages, but the semantics of this two components may vary.

The fourth form of heterogeneity is an incompatibility connected with interpretation of the domain. The differences result from the way that creator look at the surrounding reality, the ground of ontology. Covering problem usually occurs when the models differ in the description of reality, detail, the level of abstraction and perspective.

The quality of ontology integration depends largely on the quality of their creation process [7]. The basic condition for models merging is to provide accuracy and consistency of the source ontology. It is achievable by avoiding common mistakes and leading to certain guidelines. So far does not exist an officially established methodology to create an ontology, but many sources claim various tips on how to build ontologies clearly, consistent and complete. The W3C has taken up the challenge of standardization of ontology creating rules.

For both ontology aligning and merging the first step is manual annotation of mappings and bridges of tested ontologies. Unfortunately, the process turns into increasing collection of links, creating problem that is ultimately impossible to solve. It is also problematic that any change in any base ontology requires modification of all mappings made before on this ontology. Still unresolved problem is the integration of ontologies in case of necessity to find the common point of reference for the different patterns of the domain. Automating this process is unfortunately, so far unattainable.

Over the time, there are more and more proposed solutions in the area of merging ontologies, but they are still inadequate, especially with regard to their level of automation. Most of them require human participation, especially in the process of making final decisions about the shape of the final ontology version. Many methods focus on the lexical similarities ignoring semantic or vice versa. Major disadvantage of most solutions is their complexity. Simple and efficient algorithms would be much easier to implement to computer systems.

Extremely difficult to solve, is the problem caused by the complexity and diversity of natural language, as well as the differences between languages. A great assistance in ontologies combining is WordNet. However, it does not allow the development of ontologies created in another language than English or integration of ontologies created in two or more languages.

3 Multimodality Theory

The Multimodality Theory has gender in humane-computer interaction. Different media are combining in order to create meaning. In this case the multimodal system is a system which somehow involves several modalities [2]. Each modalities is a unimodal system and combination of unimodal systems make that something multimodal. The components of the multimodal system exchange information and modality of information representation is a way of representing information in some physical medium [2].The same physical medium perceived by the same set of sensors system can placed many different modalities. According to [2] the reasons for that are following: modalities differ in expressiveness and representing different kinds of information, the recipient of the information is characterised by the unique set of cognitive skills. The term multimodal representation designates combinations of two or more pure or unimodal representational modalities [3]. This modaliteies have to be external to the human cognitive system and encoded in computer-processable language [3]. Based on the modality theory following cases can occur between two uimodal systems:

- Standard meaning: the meaning is standard and the uimodal systems have already shared meaning of part of knowledge in way that communicative interaction is highly possible.
- Ad hoc assignment of meaning: some part of the other system become familiar to because external information input.

- Non-standard meaning function: Defined through some function operating over Metaphor: representing target domain through analogy with source domain, or Metonym: representing a complex entity through part of it.

In [1], the purpose of every modality is decomposed into set of parameters: application type, user, user group (user population profile), user task or other activity, application domain, use environment, interaction type, interaction devices. Moreover, the multimodality theory can help to answer for following question [2]:

- is modality M(a) useful or not useful for development purpose P?
- is modality M(a) more or less useful for purpose P than an alternative modality M(b)?
- is modality M(a) in combination with modalities M(c, c+1, … c+n) the best multimodal choice given purpose P?

4 Multimodal Approach for Ontologies Combining Problem

Multimodal approach has a different connotation depending on the area in which it is applied. The discussion contained in the article focuses on the Multimodality Theory describing creating interfaces to optimize human-computer cooperation and Mulitmodal Network permits to integrate and synchronize different modes from different areas. Looking for generic multimodal concept in these approaches, we can come to conclusions shown in Fig. 1 and Fig. 2. In the Multimodality Theory (Fig. 1) same object A is represented in different modalities. We cannot go to the assumption that $A \neq A_1 \neq A_2 \neq A_3$. However, it is possible to create a mapping that convert A_1, A_2 or A_3into A. An example is the description of the accident, using a variety of media, or a description of the process by different experts.

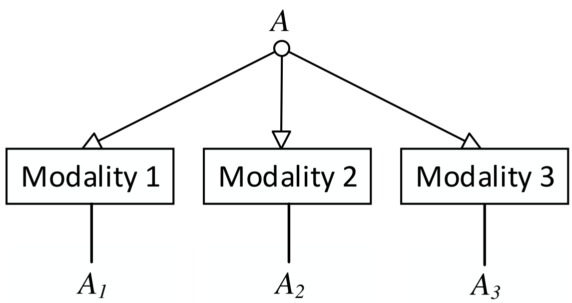


Fig. 1. The idea of multimodal approach in the Multimodality Theory

In the mulitmodal network (Fig. 2) emphasis is placed on the transfer of the object through different modes (environments) assuming object's integrity ($B \cong B' \cong B''$) [11]. The object can be processed, however, his distinguished features must be preserved. In network of modes the (intelligent) transformation operation is needed in order to transfer complex object to other modes. An example of this approach is a transport networks, where goods are transmitted by various means of transport. Another example is the transfer of knowledge to student where the final object is built/supplemented by subsequent set of courses and trainings (modes).

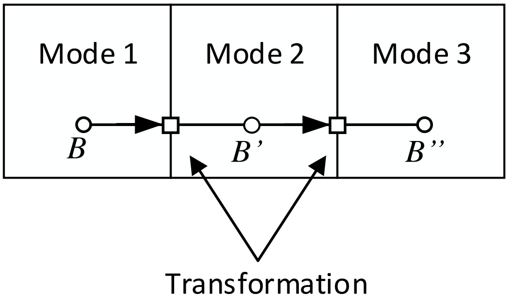


Fig. 2. The idea of multimodal approach in the multimodal network

While analyzing the issue of ontologies integration or alignment we can formulate following restriction:

$$cost\ (\Lambda, \Theta) > \beta \tag{1}$$

where: Λ - ontology alignment or integration operation procedure code, Θ - set of ontologies $\Theta = \{O_i^d\}$, where i is a ontologies index and d is a domains index, β - maximal cost accepted by user.

Since the methods of ontology alignment/integration/merging, in the vast majority, are not automatically the main cost is the work of operator/expert associated with system. The cost can be estimated knowing the number of objects in the ontology, and calculating the number of conflict between objects. In the case of automated method the cost is located in method's preparation phase when system must be supplemented with number of data, information or ontologies in order to collect initial knowledge. When operation cost is bigger, then β we can use following multimodal based procedures in order to reduce the cost:

- New modality: based on the ontology merging methods expert creates new ontology by combination of discussed ontology with his/her ontology in the same domain. The initial ontology is changing as a result it will be easier to integrate with other ontologies.
- New mode: expert mapping the ontology and find equivalent items in the ontologies in other domain. As a result the ontology background is extended into new domain and is easier to find similarities with other ontologies.

The costs of such operations are relatively small because we only work with our ontology, which we know well and our expert is able to process it efficiently.

5 Conclusion

This paper presents concept of multimodal procedures in terms of mode-based and modality-based approach. The work is only the beginning and require a special attention for each method of ontology alignment/integration/merging separately. However, it seems that the proposed approach is interesting because of the possibility of costs reducing.

Ontology-based computer system can efficiently cooperate with other systems only if the ontologies that describe these systems are correctly integrated. The interoperability of these systems provides a combination of source ontologies into one result model. There is no perfect ontology combining method, but every attempt to improve the quality of the resulting ontology, brings researchers to its invention. As long as the inconsistencies will appear in the source ontologies, so long there will be errors in the output ontologies. The algorithms simplify the elimination of these errors on the level of ontology description and structure, but to be able to analyze the relationships between concepts, they must be based on well-developed semantic dictionary, such as WordNet or human expert.

References

1. Bernsen, N.O., Dybkjær, L.: Multimodal Usability. Human-Computer Interaction Series. Springer (2010)
2. Bernsen, N.O.: Multimodality Theory. In: Multimodal User Interfaces. Signals and Communication Technology. Springer (2008)
3. Bernsen, N.O.: Foundations of multimodal representations: a taxonomy of representational modalities. Interacting with Computers 6(4), 347–371 (1994)
4. Bocewicz, G., Banaszak, Z.: Declarative approach to cyclic steady state space refinement: periodic process scheduling. The International Journal of Advanced Manufacturing Technology 67(1-4), 137–155 (2013)
5. Grzegorz, B., Zbigniew, B., Paweł, P.: Cyclic Steady State Space Refinement. In: Szewczyk, R., Zieliński, C., Kaliczyńska, M. (eds.) Recent Advances in Automation, Robotics and Measuring Techniques. AISC, vol. 267, pp. 11–20. Springer, Heidelberg (2014)
6. Choi, N., Song, I.-Y., Han, H.: A Survey on Ontology Mapping. ACM SIGMOD Record 35, 34–41 (2006)
7. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering: With Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web. Springer (2004)
8. Keet, C.M.: Aspects of Ontology Integration. Technical Report, Napier University, Edinburgh, Scotland (2004)
9. Konys, A., Wątróbski, J., Różewski, P.: Approach to Practical Ontology Design for Supporting COTS Component Selection Processes. In: Selamat, A., Nguyen, N.T., Haron, H. (eds.) ACIIDS 2013, Part II. LNCS, vol. 7803, pp. 245–255. Springer, Heidelberg (2013)
10. Kushtina, E., Różewski, P., Zaikin, O.: Extended ontological model for distance learning purpose. In: Reimer, U., Karagiannis, D. (eds.) PAKM 2006. LNCS (LNAI), vol. 4333, pp. 155–165. Springer, Heidelberg (2006)

11. Heath, L.S., Sioson, A.A.: Multimodal Networks: Structure and Operations. IEEE/ACM Trans. Comput. Biol. Bioinformatics 6(2), 321–332 (2009)
12. Pawlewski, P.: Multimodal approach to model and design supply chain. In: Proceedings of 7th IFAC Conference on Manufacturing Modelling, Management, and Control, St. Petersburg, pp. 2110–2115 (2013)
13. Różewski, P., Małachowski, B., Jankowski, J., Prys, M., Dańczura, P.: Preliminaries for Dynamic Competence Management System building. In: Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1327–1333 (2013)
14. Sitek, P., Wikarek, J.: A hybrid approach to supply chain modeling and optimization. In: Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1223–1230 (2013)
15. Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEEE Transactions on Knowledge and Data Engineering 25(1), 158–176 (2013)

Behavioral Macroeconomics and Agent-Based Macroeconomics

Shu-Heng Chen* and Umberto Gostoli

AI-ECON Research Center, Department of Economics,
National Chengchi University, Taipei 11605
chen.shuheng@gmail.com

Abstract. In this article, we review the recent development of the agent-based macroeconomic models (ABMMs) using the skeleton of behavioral macroeconomics as proposed by George Akerlof. Based on the 15 models surveyed in this paper, we find that most behavioral elements addressed by Akerlof have not been well incorporated into the ABMMs. The only element which has been successfully well applied is the decision heuristics (rule-of-thumb behavior). We discuss the fundamental difficulties which cause the current deviations and highlight the expected efforts to meet the gap.

Keywords: Agent-Based Macroeconomic Models, Behavioral Macroeconomics, Asymmetric Information, Social Preferences, Rule of Thumbs, Wage Rigidity, Hyperbolic Discounting.

1 Introduction

The paper is motivated by our on-going research project which examines critically the growing literature on agent-based macroeconomics and, in particular, presents more than 15 relatively recent agent-based macroeconomic models with the aim to describe the state of art of this field and, at the same time, to highlight some possible (and, in our view, desirable) directions for future developments. In order to perform this critical survey, we assess the extent to which the behavioral micro-foundations of the agent-based models considered in this project include the behavioral elements addressed by [1] and which have been shown to affect, generally speaking, the macroeconomic dynamics.

With years of efforts, agent-based macroeconomic modelers have claimed some real progresses. First, the agent-based macroeconomic models can generate the dynamic behavior, related to economic fluctuation and growth, of almost all macroeconomic variables that are conventionally studied by the equation-based mainstream models [20,19,17,8,6,4,15]. These variables are, for example, GDP, consumption, investment, wage, unemployment rate, interest rate, and inflation rate. Therefore, in this regard, the agent-based macroeconomic models have gradually made themselves the alternatives to the conventional equation-based

* Corresponding author.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_6,

macroeconomic models. However, agent-based macroeconomic models can do more than that. The additional strengths and advantages of agent-based macroeconomic models come from the removing of the representative agent and the Walrasian auctioneer. These two removals lead to a number of features that distinguish the agent-based macroeconomic models.

First, households and firms in the agent-based macroeconomic models are no longer simplified into a single representative agent; instead, they are each individually modeled. Their complex interaction processes may further contribute to their evolving heterogeneities. This feature enables us to observe simultaneously multiple kinds of heterogeneities that interest economists, such as income distribution, wealth distribution, firm size distribution, firm's growth rate distribution, price distribution, profit-rate distribution, wage distribution, and the indexes used to characterize these distributions, including the Herfindahl-Hirschman index and the Gini Index [20,19,15,14,7,4]. The three aspects of macroeconomics–growth, fluctuation and distribution–can now be studied in one unified framework. Over the last half of the 20th century, great efforts have been made to model the phenomenon of growth with cycles; however, incorporating the distribution behavior of households and firms into the growth-with-cycles model is definitely a challenge for the equation-based macroeconomic models. Hence, this integration in the agent-based models is definitely a milestone for macroeconomics.

Second, the labor markets in the agent-based macroeconomic models are constructed in a decentralized manner, which enables us to further see the details of labor markets, such as the dynamics (time series) of vacancy rate, labor contract duration, and the unemployment duration [19,3]. Needless to say, these labor market details are generally hard to obtain in the equation-based models.

Third, since firms are modeled individually, the agent-based macroeconomic models can demonstrate the industrial dynamics and the macroeconomic dynamics simultaneously; that is, the macro-dynamics is generated with the underpinning competitive dynamics of firms. The emergent growth-with-cycles is now accompanied by various characteristics of firms, such as size, growth rate, bankruptcy rate, duration, and markups [20,19,5,16,3,7,4,9]. This kind of two-in-one feature is again hard to be observed in the equation-based macroeconomic models.

Fourth, in the agent-based models, the relationships among all economic variables are emergent and not assumed or imposed. These include the famous Philips Curve, Beverage Curve and Okun's Law. Unlike the conventional equation-based models which take these relationships as given, the agent-based models allow these relationships to be studied as endogenously emergent properties [19,15,7].

Last, in addition to the substitute and complementary relationships with the conventional equation-based models, the agent-based macroeconomic models have far-reaching implications for economic models because the agent-based macroeconomic models can serve as a platform to integrate various agent-based markets currently studied in an isolated manner, such as agent-based financial markets, housing markets, electricity markets, educational markets, and innovation networks. Needless to say, these markets all play important roles in the operation of macroeconomy, but studying them in an isolated manner may limit our understanding of their possible interactions with other markets. Currently, a number of agent-based macroeconomic models have been developed to include a stock market [17,4], although these built-in ones are much simpler than the general agent-based financial markets.

2 Behavioral Foundations

In his 2001 Nobel Price lecture, George A. Akerlof described how, in the late 1960's, some macroeconomists started adopting behavioral assumptions grounded in psychological and sociological observations in order to develop models that could account for the macroeconomic phenomena, such as the existence of involuntary unemployment, the prevalence of under-saving for retirement and the excessive volatility of stock prices relative to their fundamental, that the competitive, general-equilibrium models developed by the New Classical macroeconomists could not satisfactorily explain. He called this new field of research *Behavioral Macroeconomics* [1].

In this section, we will assess to what extent the behavioral assumptions listed in [1] can constitute the micro foundations of the ABMMs.

2.1 Asymmetric Information

In all agent-based models, information sets are intrinsically heterogeneous as agents use private and local information to make their decisions. Therefore, in these models, adverse selection and moral hazard problems may naturally arise through the agents' adaptive and/or learning process. However, the problem of asymmetric information is not considered in most of the models mentioned in this survey as the agents are generally supposed to have perfect knowledge regarding the goods and resources exchanged or the behavior of the agents with whom they interact. In most models, in fact, the very structure of the behavioral models adopted rules out the possibility of shirking or other kinds of deceitful behavior.

In [13], while the workers have the possibility to exert different levels of effort, a variable which enters both in their utility function and the firms' production function, the firms are assumed to have perfect knowledge of the effort supplied by their workers and, therefore, shirking cannot emerge. Among the models known to us, the only exception is [14] where the firms cannot observe the workers' ability. This is a typical situation of asymmetric information, although the effect of this asymmetric information may be mitigated by the workers' educational degree, which serves as a signal of the workers' ability to potential employers. Even in this case, however, no analysis of the effect of asymmetric information on the aggregate outcomes was performed by the authors.

Given the influence of asymmetric information on the functioning of an economic system, particularly the labor and credit markets, its inclusion in the next generation of ABMMs would certainly represent an important step towards the development of more realistic and relevant economic models.

2.2 Social Preferences

A growing body of empirical literature shows the crucial effects of social preferences on people's behavior in many economic contexts. [10] showed how social preferences, particularly reciprocity, affect market competition, cooperation and, in general, the way incentive works. With reference to the last point, the experiments in [11] showed that reciprocity has a powerful economic effect on the efficiency of complete and incomplete contracts. Notwithstanding the pervasiveness and the economic importance of social preferences, almost none of the existing ABMMs adopted behavioral models with social preferences.[1] The only exception is the model presented in [4] where one determinant of the households' consumption decision is the consumption level of their neighbors.

One of the reasons social preferences have been generally excluded form agent-based macroeconomic models is that the inclusion of social preference would require the introduction of behavioral models with more complex structures so to allow for a wider behavioral repertoire. Such a modeling choice would certainly make the model more realistic but, on the other hand, it would make the number of parameters increase, with the consequence of making the analysis of the simulations' results more complex. So, as for asymmetric information, the inclusion of social preferences into the agents' behavioral models represents a line of development which, in large part, is still to pursue.

2.3 Rule-of-Thumb Pricing Behavior

Among all of the aspects of real-world agents' behavior highlighted by empirical research, the rule-of-thumb pricing is the only one which appears in virtually every agent-based model. In fact, agent-based models, by their very nature, contain behavioral models based on some form of rule-of-thumb behavior, or heuristics, as the complexity of the environment does not allow the agents to perform the computations needed to adopt a maximizing behavior. The agents are forced to explore their strategy space in the attempt to improve their welfare.

By and large, three rule-of-thumb pricing mechanisms have been adopted in agent-based models: (a) those based on mark-up pricing (with and without mark-up evolution) [16,3,6,4]; (b) those based on some firms' internal variable (typically, inventory levels) [19,7,18]; and (c) those based on imitation.[13,7,12]

Examples of the first pricing method are given by [4] and by the Eurace model [6]. Variants of this pricing behavior are proposed by [16], where the mark-up is not fixed, but evolves through imitation and mutation processes, and by [9], where the mark-up variation depends on the evolution of the firms' market share. [3] adopt a hybrid price adjustment mechanism combining mark-up pricing with

[1] Apart from the inclusion of imitation, which may be also considered a form of social behavior.

an adaptive adjustment by means of which the price is decreased (increased) by a fixed percentage if the difference between the actual and the desired shop's inventory levels is above (below) a fixed threshold.

Examples of the second price mechanism are provided by [19] and by [18], where firms increase their prices by a stochastically determined percentage if they have not sold all of their inventories and decrease it otherwise. A variant of this mechanism is adopted by [7] where the price adjustment mechanism involves additional information represented by the economy's average price. In particular, in this model, firms increase their price by a stochastically determined percentage if it is below the average price and they have sold all their production (i.e. their inventory level is equal to zero), while they decrease it if it is above the average price and they have a positive inventory level.

A 'pure' imitation mechanism is introduced by [12], who compare this mechanism with the pricing behavior proposed by [7]. In particular, firms copy (with a fixed probability) the price of the best performing firms among a randomly chosen subset of the firms' population, price which is varied through a random mutation. A hybrid pricing behavior combining the two latter mechanisms is proposed by [13], where firms adjust their price according to their inventory levels and through a combined imitation/mutation process which is activated with a certain probability.

As we can see from this brief review, the pricing behaviors adopted are the most diverse and so far very little effort has been spent on (a) the study of how different pricing behaviors affect the macroeconomic dynamics; (b) the development of a unified model based on the available empirical evidence.

2.4 Downward Wage Rigidity (Reference Dependence)

Many empirical studies have shown downward wage rigidity to be a pervasive feature of industrialized economies. Downward wage rigidity has been explained by referring to asymmetric information, social preferences (such as reciprocity) and the presence of utility functions characterized by reference points and loss aversion. Agent-based macroeconomic models have, in general, abstracted away from these behavioral assumptions and, therefore, do not allow for the endogenous emergence of this feature of real economic systems. A few studies have taken wage rigidity as given, exogenously introducing wage adjustment mechanisms where the wage can be changed only upwards [17,7].

For example, in [17] the wage is centrally determined by a trade union which keep increasing it (according to an inflation rate equal to a planned inflation rate determined by the central bank) as long as the previous increases determined an increase of the aggregate real labor income, while keeps the wage unchanged otherwise. Another example is provided by [7], where firms keep the last wage if they have no vacancies while they increase their wage offer by a stochastically drawn percentage if they have vacancies. In other models, although downward changes are possible, mechanisms favoring upward over downward changes are adopted [13]. In [13], for example, although downward changes can be the result of the combined imitation/mutation process by means of which firms revise, with a certain probability, their operating characteristics, a 'preference' for upward changes is introduced by allowing firms, in certain conditions (an inventory equal to zero and a workforce below their target employment level), to revise their wage offer up by a firm-specific percentage.

In general, downward wage rigidity is among the 'behavioral' phenomena which have been to a large extent neglected by agent-based macroeconomic literature.

2.5 Hyperbolic Discount Function (Under-Saving)

A vast and growing body of experimental economics literature shows that individuals have preference for present payoff which is not consistent with exponential discounting and may be better approximated by hyperbolic functions. This feature of individuals' behavior may help explaining some economic puzzles such as under-saving and the need for compulsory pension schemes.

In many of the ABMMs reviewed in this paper, the households' saving decision is based on simple rules of thumb (typically, a 'buffer stock' saving rule). Therefore, discounting, either exponential or hyperbolic, has no role in individual's decision-making process. The reason for the absence of discounting in agent-based models is due to the behavioral assumptions and the structural features characterizing these models. As regards to the former, the bounded rationality assumed in ABMMs does not allow the agents to compute the intertemporal payoff accruing from their decisions. This assumption is often justified by the complexity of the environment. Indeed, in complex environments characterized by fast and often unpredictable changes, agents with limited working memory and cognitive capabilities are not in the condition to determine the stream of payoffs associated with a certain strategy. Often, the best way they can do is trying to increase their immediate payoff. In some way, this behavior could be considered as a case of 'extreme' hyperbolic discounting as they will always choose the action leading to the highest immediate payoff, no matter its long-term consequences.

The second reason is related, somewhat paradoxically, to the relative simplicity of the current agent-based macroeconomic models' temporal structure. In fact, discounting mechanism will affect the economic system to the extent that the agents' behavior is characterized by path-dependence, i.e. to the extent that the agent's actions will determine (by constraining or expanding) their future actions' space. A typical example is a specific, and therefore irreversible, investment. However, the agent-based macroeconomic models which have been developed so far, have a rather simple productive structure, in many cases characterized by the absence of physical capital. In such models, the temporal horizon over which the agent's actions exert their effect is very short, making the discounting mechanism less relevant. In other words, if agents are allowed to make a choice in every period and this choice will generate a payoff only in that period, without influencing the agent's future actions' set, their decisions will not be affected by their particular discounting behavior. Therefore, agent-based macroeconomic models with more sophisticated, forward-looking behavioral models and with a structure allowing for path-dependent processes must be developed before hyperbolic discounting can be meaningfully introduced and its effects on the aggregate dynamics can be assessed.

2.6 Irrationality of Stock Prices (Bubbles and Crashes)

Irrationality, which seems to characterize the financial markets in some periods, represents one of the main challenges to the neoclassical economic paradigm and motivated the development of many behaviorally founded agent-based models in the last 15 years, following the seminal Santa Fe stock market model [2]. Yet, only a few of the ABMMs have included financial markets. Moreover, of these models, only [4] generates phenomena such as booms in stock prices, as shown in the previous section. Analogously, while many models include the credit market, few generate phenomena such as credit crunches and bank runs. Given the importance the financial and credit crisis have on aggregate consumption and investment dynamics, it is crucial that the current ABMMs are extended to include models of credit and financial markets which can account for the periodic occurrence of bank runs, bubbles and crashes and other 'irrational' phenomena observed in these markets.

3 Conclusion

To sum up, we can say that, in general, the ABMMs developed so far have only very partially drawn the insights from the field of behavioral economics. Therefore, there is significant room for developing behaviorally micro-founded macroeconomic models.

Acknowledgements. The authors are grateful to two anonymous referees for their comments and suggestions. The NSC grant NSC 101-2410-H-004-010-MY2 is also gratefully acknowledged.

References

1. Akerlof, G.: Behavioral macroeconomics and macroeconomic behavior. American Economic Review 92(3), 411–433 (2002)
2. Arthur, W., Holland, J., LeBaron, B., Palmer, R., Tayler, P.: Asset pricing under endogenous expectations in an artificial stock market. In: Arthur, B., Durlauf, S., Lane, D. (eds.) The Economy as an Evolving Complex System II, pp. 15–44. Addison-Wesley, Reading (1997)
3. Ashraf, Q., Gershman, B., Howitt, P.: Banks, market organization, and macroeconomic performance: An agent-based computational analysis. Working paper, Brown University (2011)
4. Bruun, C.: Rediscovering the economics of Keynes in an agent-based computational setting. New Mathematics and Natural Computation (2013) (Forthcoming)
5. Chan, C., Steiglitz, K.: An agent-based model of a minimal economy. Working Paper of the Department of Computer Science, Princeton University (2008)

6. Cincotti, S., Raberto, M., Teglio, A.: The Eurace macroeconomic model and simulator. In: Aoki, M., Binmore, K., Deakin, S., Gintis, H. (eds.) Complexity and Institution: Markets, Norms and Corporations, Palgrave, pp. 81–104 (2012)
7. Delli Gatti, D., Desiderio, S., Gaffeo, E., Cirillo, P., Gallegati: Macroeconomics from the Bottom-up. Springer (2011)
8. Dosi, G., Fagiolo, G., Roventini, A.: Schumpeter Meeting Keynes: A Policy-Friendly Model of Endogenous Growth and Business Cycles. Journal of Economics Dynamics and Control 34(9), 1748–1767 (2010)
9. Dosi, G., Fagiolo, G., Napoletano, M., Roventini, A.: Income distribution, credit and fiscal policies in an agent-based keynesian model. Journal of Economic Dynamics and Control 37(8), 1598–1625 (2013)
10. Fehr, E., Fischbacher, U.: Why social preferences matter: The impact of non-selfish motives on competition, cooperation and incentives. Economic Journal 112(478), C1–C33 (2002)
11. Fehr, E., Klein, A., Schmidt, K.: Fairness, incentives, and contractual incompleteness. Mimeo, University of Munich (2001)
12. Gaffeo, E., Gallegati, M., Gostoli, U.: An agent-based "proof of principle" for Walrasian macroeconomic theory. CEEL Working Papers 1202, Cognitive and Experimental Economics Laboratory, Department of Economics, University of Trento, Italy (2012)
13. Gintis, H.: The dynamics of general equilibrium. Economic Journal 117(523), 1280–1309 (2007)
14. Kinsella, S., Greiff, M., Nell, E.: Income distribution in a stock-flow consistent model with education and technological change. Eastern Economic Journal 37, 134–149 (2011)
15. Lengnick, M.: Agent-based macroeconomics: A baseline model. Journal of Economic Behavior & Organization 86, 102–120 (2013)
16. Mandel, A., Jaeger, C., Fuerst, S., Lass, W., Lincke, D., Meissner, F., Pablo-Marti, F., Wolf, S.: Agent-based dynamics in disaggregated growth models. CES Working Paper 2010. 77 (2010)
17. Raberto, M., Teglio, A., Cincotti, S.: Integrating real and financial markets in an agent-based economic model: An application to monetary policy design. Computational Economics 32, 147–162 (2008)
18. Riccetti, L., Russo, A., Gallegati, M.: Unemployment benefits and financial leverage in an agent based macroeconomic model. Economics 7, 2013–2042 (2013), `http://dx.doi.org/10.5018/economics-ejournal.ja.2013-42`
19. Russo, A., Catalano, M., Gaffeo, E., Gallegati, M., Napoletano, M.: Industrial dynamics, fiscal policy and RD: Evidence from a computational experiment. Journal of Economic Behavior and Organization 64, 426–447 (2007)
20. Wright, I.: The social architecture of capitalism. Physica A 346, 589–620 (2005)

Heterogeneous Households: Monopolistic Capitalists, Entrepreneurs and Employees

Jonathan Swarbrick*

University of Surrey, Guildford, GU2 7XH, UK
j.m.swarbrick@surrey.ac.uk

Abstract. What are the implications of an uneven distribution of welfare on optimal stabilisation policy? I build a dynamic stochastic general equilibrium model with household heterogeneity in income and consumption with which to solve optimal fiscal and monetary policy over the business cycle. I include three types of household; capitalists, entrepreneurs and workers, and endogenise the selection process between the latter two.

Keywords: limited asset market participation, stabilisation policy, income inequality, business cycles.

1 Introduction

What are the implications of an uneven distribution of welfare on optimal stabilisation policy? This paper presents a framework with which to solve optimal monetary and fiscal policy in the presence of uneven distributions of wealth, income and consumption. To achieve this, I introduce household heterogeneity into the model with idiosyncratic productivity and the possibility of default. There are *ex post* three types of households in the economy including a fixed proportion of capitalists, that is, households whom own all the productive capital, with the remaining households either behaving as entrepreneurs or workers. The entrepreneurs borrow from the capitalists and combine labour and capital to produce output for consumption and investment. The entrepreneurs cannot be forced to commit to the loan contract and with idiosyncratic risk some will renege on the loan. If they do so they can no longer deal with the capitalists and must provide labour as workers. Monopolistic capitalists formulate a contract given that it is possible for the entrepreneurs to default.

We choose to differentiate between capitalists that are able to hold wealth and smooth consumption, and other households in the economy that are not able to do so. The motivation for this set-up begins with Mankiw in [8] who highlights that many households live hand-to-mouth holding virtually zero net worth. This

* The author would like to thank Tom Holden and Paul Levine for their suggestions and advice. He also acknowledges the support by the Economic and Social Research Council grant numbers ES/J500148/1 and ES/K005154/1 (Agent-Based and DSGE Macroeconomic Models: A comparative Study).

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_7,

has a significant effect on aggregate consumption smoothing and the transmission mechanism of stabilisation policy. Indeed, further research (e.g. Amato & Laubach in [1], Bilbiie in [2], and Bilbiie & Straub in [3]) has emphasised this, showing that the presence of this behaviour leads to endogenous persistence in output and inflation, and alters both the transmission of monetary policy and the welfare objective. We then endogenise the selection of households to act as either entrepreneurs or workers. This follows a literature beginning with Quadrini in [9], followed by Cagetti & De Nardi in [5]. Doing so enriches the analysis of distributional variables, and by capturing firm exit and employment, models labour market dynamics in a novel way. The model then distinguishes between the optimising behaviour of three types of household; the capitalist, the entrepreneur, and the worker.

2 The Model

The economy is comprised of a large number of identical families, each of which contains *ex post* three types of household; capital owners, entrepreneurs and workers.

2.1 Firms and Production

Firms use capital and labour to produce output according to a standard Cobb-Douglas production function

$$Y_{f,i,t} = z_t \hat{H}^{\alpha}_{f,i,t} K^{1-\alpha}_{f,t-1}$$

for firm i where subscript f identifies a firm-level variable. Capital is invariant across firms as the contract is agreed prior to realisation of the idiosyncratic productivity. The variable z_t represents total factor productivity experienced by the whole economy. In addition to the economy-wide productivity, each firm experiences an independent across time and population productivity shock such that at time t, firm i is defined by idiosyncratic labour productivity status $\varepsilon_{i,t}$. We assume that ε is uniformly distributed across time and space with mean 1 and with $\varepsilon_l \leq \varepsilon \leq \varepsilon_h$, and enters the production function by introducing efficiency hours $\hat{H}_{f,i,t}$ where

$$\hat{H}_{f,i,t} = \varepsilon_{i,t} H_{f,i,t}$$

Within each firm there is an employer and some workers. Indexing the total population to 1, the proportion of households that are the capitalists is exogenously determined and given by $1-\phi$. The remaining ϕ of households are then made up of the entrepreneurs and workers. The proportion of households that have defaulted and supply labour as workers is given by $\phi\theta_{t-1}$. It follows then that the proportion of households in the economy that borrow capital and hire workers is given by $\phi(1-\theta_{t-1})$. The firm's labour input is given by

$$H_{f,i,t} = H_{e,i,t} + H_{w,i,t}.$$

which is simply the sum of the labour input by the employer $H_{e,i,t}$ and that by all workers at the firm $H_{w,i,t}$.

2.2 Households

All households seek to maximise expected, discounted lifetime utility under the standard assumptions of rational expectations. We assume all households derive utility according the standard log utility

$$U_t = \ln(C_t) + \chi \ln(1 - H_t).$$

where C_t is household consumption, and H_t household labour supply. Other than access to the asset market, the capitalists differ from the other households in the economy only in that they do not provide labour. Just below we will describe the optimal default decision faced by the entrepreneur. We will then look at the labour market before solving the capitalist's problem.

Idiosyncratic Productivity and Default. If each firm is identical other than the idiosyncratic labour productivity ε then we can attempt to find some critical values of productivity at which point an entrepreneur will be indifferent between defaulting and not. The following equation describes the evolution of θ (or the firm entry and exit in the economy).

$$\theta_t = (1 - \zeta)\theta_{t-1} + \xi_t(1 - \theta_{t-1})$$

A worker faces a fixed probability ζ or being forgiven and so ξ_t is then the proportion of employers that default, or in other words, the proportion of firms that exit. Recalling that $\varepsilon \sim \mathcal{U}(a, b)$, if all employers that experience $\varepsilon > \varepsilon_c$ will decide to default, then we can say

$$\xi_t = (b - \varepsilon_{c,t})(b - a)^{-1}$$

Finding the critical productivity status, $\varepsilon_{c,t}$ will imply ξ_t and given θ_{t-1} we can solve θ_t. This, with the capital stock K_t, total factor productivity z_t, and any new exogenous shocks, determine the state of the economy in period $t + 1$. The value of $\varepsilon_{c,t}$ is solved by finding the firm that is indifferent between defaulting and not.

We let $V_{e,i,t}$ be the value to entrepreneur i of keeping to the contract at time t. This is given as the utility gained from doing so in period t plus the discounted expected future utilities of defaulting or not multiplied by their probabilities. Similarly, $V_{d,i,t}$ is the value to the household if they have defaulted. These are given by

$$V_{e,i,t} = U_{e,i,t} + \beta \mathbb{E}_t \left[\xi_{t+1} \left(V_{d,i,t+1} + \omega_{i,t+1}\right) + \left(1 - \xi_{t+1}\right) V_{e,i,t+1}\right] \tag{1}$$

$$V_{d,i,t} = U_{d,i,t} + \beta \mathbb{E}_t \left[\zeta V_{e,i,t+1} + (1 - \zeta) V_{d,i,t+1}\right] \tag{2}$$

where $\omega_{i,t}$ represents the additional utility from stealing the output, and $U_{e,i,t}$ and $U_{d,i,t}$ the period utility gained by an employer sticking to a contract, and by a worker respectively. For a firm, j, that realises the critical ε_c and is indifferent between defaulting or not, the condition $V_{e,j,t} = V_{d,j,t} + \omega_{j,t}$ must hold. Solving this leads to the indifference condition for the marginal entrepreneur

$$U_{e,j,t} = U_{s,j,t} + \beta \mathbb{E}_t \begin{bmatrix} (\zeta + \xi_{t+1} - 1)\left(V_{e,j,t+1} - V_{d,j,t+1}\right) \\ -\xi_{t+1}\left(U_{s,j,t+1} - U_{d,j,t+1}\right) \end{bmatrix}, \tag{3}$$

where $V_{e,t}$ and $V_{d,t}$ are expressed recursively as in equations 1 and 2 and $U_{s,j,t}$ is the utility gained in default. This condition determines the critical value of ε_t and so the proportion of employees that will default in period t, ξ_t. This then determines the value of θ_t which is the proportion of households that are workers in the following period $t+1$.

Capital Investment and the Contract Form. Before looking at the labour market, we state our assumption of the evolution of capital and the functional form of the contract. Recent research (such as Bose, Pal & Sappington in [4]) has shown linear contracts to be close approximations of fully optimal contracts and so we assume that contract offered by the capitalist will be of linear form. The contract is agreed between capitalist and entrepreneur and offers the payout

$$\mathcal{C}_{f,i,t} = \gamma_{1,t} + \gamma_{2,t} Y_{f,i,t}$$

where $\mathcal{C}_{f,i,t}$ is the payout to a firm based on the firm output and will be consumed by the entrepreneurs and workers. The capitalists own the capital and are responsible for the investment decisions given the capital accumulation process

$$K_t = I_t + (1-\delta)K_{t-1}.$$

with K_t and δ as the end of period capital stock and the depreciation rate respectively. The contracts are agreed prior to idiosyncratic productivity and so all firms are identical. We can then state, for all firms,

$$K_{f,t-1} = (\phi(1-\theta_{t-1}))^{-1} K_{t-1}.$$

Timing of Events. The timing of events in each period are important to clarify.

1. The proportion of workers that have defaulted θ_{t-1} and the capital stock K_{t-1} is inherited from the end of the previous period.
2. Any exogenous aggregate shocks occurring in the period are realised.
3. The monopolistic capitalists offer contracts to maximise their lifetime utility.
4. The firms realise idiosyncratic productivity status ε and choose labour demand, entrepreneurs and workers choose labour supply. Firms yield output Y_t.
5. The proportion of entrepreneurs ξ_t default and ζ workers are forgiven, realising θ_t.
6. The capitalists choose to consume or invest the income from the contracts which will realise the capital stock K_t to be used in the next period.

We can now derive the labour market conditions.

Labour Demand. The firm owner seeks to maximise their consumption and will choose labour demanded to do so. Let $C_{e,i,t}$ be the consumption of firm owner i with productivity status ε_i at time t.

$$C_{e,i,t} = \gamma_{1,t} + \gamma_{2,t} Y_{f,i,t} - w_t H_{w,i,t} \tag{4}$$

$H_{w,i,t}$ is the worker labour input and is allowed to take negative values in which case it represents the entrepreneur selling their own labour back to the labour market at the market wage rate. With many firms and many workers it is assumed that the labour market is perfect and so all agents are price takers. Maximising consumption and solving for firm i labour demand $H_{w,i,t}$ gives

$$H_{w,i,t} = \left(\alpha \gamma_{2,t} z_t w_t^{-1} \varepsilon_i^{\alpha}\right)^{\frac{1}{1-\alpha}} K_{f,t-1} - H_{e,i,t}, \tag{5}$$

Workers Labour Supply. Given that all workers receive the same wage rate, and with no wealth accumulation, every worker supplies the same labour. Whilst the labour demanded by each firm varies, the workers can be seen to provide labour to a central pool which is then distributed to the firms unevenly according to the firm's demand at the market wage. The labour supply condition for the representative worker is found by maximising their utility subject to the budget constraint

$$C_{d,t} = w_t H_{d,t}.$$

This leads to

$$H_d = (1 + \chi)^{-1}$$

for all workers. The time subscript is dropped as this is time-invariant.

Entrepreneur Labour Supply. The entrepreneurs have the same objective function as the workers but with the budget constraint as in equation 4. This leads to the labour supply condition

$$\chi \frac{1}{2} \frac{C_{e,i,t}}{1 - H_{e,i,t}} = w_t \tag{6}$$

Labour Market Equilibrium. We assume the labour market clears and so the total labour demanded must equate that supplied as in

$$\theta_{t-1} H_d = (1 - \theta_{t-1}) \frac{1}{\varepsilon_h - \varepsilon_l} \int_{\varepsilon_l}^{\varepsilon_h} H_{w,i,t} d\varepsilon$$

Substituting in the labour demand equation 5, we can solve for the market wage

$$w_t = A^{1-\alpha} \alpha \gamma_{2,t} z_t G_t^{1-\alpha}$$

where

$$A = \left(\varepsilon_h{}^{\frac{1}{1-\alpha}} - \varepsilon_l{}^{\frac{1}{1-\alpha}} \right) (\varepsilon_h - \varepsilon_l)^{-1}(1-\alpha)$$

and $G_t = K_{t-1}/H_t$ is the capital-labour ratio over the whole economy with H_t as measure of aggregate labour given by

$$H_t = \phi(\theta_{t-1}H_d + (1-\theta_{t-1})H_{e,t}) \tag{7}$$

with the representative entrepreneur labour supply as $H_{e,t} = \frac{1}{\varepsilon_h - \varepsilon_l} \int_{\varepsilon_l}^{\varepsilon_h} H_{e,i,t} d\varepsilon$. Taking the total labour demanded by firm i, $H_{f,i,t}$ and substituting in the expression for the market wage, we can find

$$H_{f,i,t} = \varepsilon_i^{\frac{\alpha}{1-\alpha}} A^{-1} H_t(\phi(1-\theta_{t-1}))^{-1}.$$

From here and 6 we can express the entrepreneur labour supply as a closed form expression

$$H_{e,i,t} = (2+\chi)^{-1} \left(2 - \chi\gamma_{1,t}w_t^{-1} - \chi A^{-1}\alpha H_t(\phi(1-\theta_{t-1}))^{-1}\varepsilon_i^{\frac{\alpha}{1-\alpha}} \right)$$

So we now conveniently have all firm-level variables as closed form functions of aggregate variables and the firm productivity status $\varepsilon_{i,t}$.

2.3 Aggregations

The aggregate variables can be expressed by integrating over all productivity status realisations. This leads to expressions for aggregate output, output from non-defaulting firms which will appear in the capitalist budget constraint and the representative employee labour supply

$$Y_t = A^{1-\alpha} z_t H_t^{\alpha} K_{t-1}^{1-\alpha}$$

$$\bar{Y}_t = \left(\varepsilon_{c,t}^{\frac{1}{1-\alpha}} - \varepsilon_l{}^{\frac{1}{1-\alpha}} \right) (\varepsilon_{c,t} - \varepsilon_l)^{-1}(1-\alpha)A^{-\alpha} z_t H_t^{\alpha} K_{t-1}^{1-\alpha}$$

$$H_{e,t} = (2+\chi)^{-1} \left(2 - \chi\gamma_{1,t}w_t^{-1} - \chi\alpha(1-\alpha)H_t(\phi(1-\theta_{t-1}))^{-1} \right)$$

2.4 Optimal Contract

To ensure saddle-path stability, we introduce a small adjustment cost to the contract. We can think of these as administrative costs that enter the capitalist's budget constraint as $X_{\gamma_1,t} + X_{\gamma_2,t}$. We want the adjustment costs to disappear in the long-run and so specify the functional forms of these costs as

$$X_{\gamma_j,t} = \varrho_1 \left(1 - \gamma_{j,t}/\gamma_{j,t-1}\right)^2$$

so that the costs are zero in steady state. This is then adequate to ensure saddle-path stability as required.

With the assumption of the availability of a commitment technology, the capitalists offer the contracts to the entrepreneur with full knowledge of the household behaviour and internalising their impact on aggregate conditions. The representative capitalist maximises expected lifetime utility subject to the budget constraint

$$(1-\phi)C_{c,t} + I_t = (1-\gamma_{2,t})\bar{Y}_t - (1-\xi_t)\,\phi\,(1-\theta_{t-1})\,\gamma_{1,t} - (X_{\gamma_1,t} + X_{\gamma_2,t})\,.$$

They achieve this through the appropriate choice of policy instruments $\gamma_{1,t}$ and $\gamma_{2,t}$, and investment I_t. The economy is described in full by the state variables K_{t-1} and θ_{t-1}, the policy choice, $\gamma_{1,t}$ and $\gamma_{2,t}$, and the exogenous shock z_t. The optimization problem for the capital owner is then to choose a path for capital stock, $\{K_{t+s}\}$, and the worker population, $\{\theta_{c,t+s}\}$, to maximise lifetime utility subject to the implementability constraints given by equations 1, 2, 3, and 7. All other variables can then be expressed recursively. We can express these four constraints as

$$\begin{aligned}
0 &= \Theta_t \begin{pmatrix} E_t\left[V_{e,t+1}\right], E_t\left[V_{d,t+1}\right], K_{t-1}, \theta_{t-1}, H_t, \gamma_{1,t}, \gamma_{2,t}, K_t, \theta_t, \\ E_t\left[\theta_{t+1}\right], E_t\left[\gamma_{1,t+1}\right], E_t\left[\gamma_{2,t+1}\right], E_t\left[H_{t+1}\right], z_t \end{pmatrix} \\
0 &= \mathcal{H}_t\left(K_{t-1}, \theta_{t-1}, H_t, \gamma_{1,t}, \gamma_{2,t}, z_t\right) \\
0 &= \mathcal{V}_{e,t} \begin{pmatrix} V_{e,t}, E_t\left[V_{e,t+1}\right], E_t\left[V_{d,t+1}\right], K_{t-1}, \theta_{t-1}, H_t, \gamma_{1,t}, \gamma_{2,t}, K_t, \theta_t, \\ E_t\left[\theta_{t+1}\right], E_t\left[\gamma_{1,t+1}\right], E_t\left[\gamma_{2,t+1}\right], E_t\left[H_{t+1}\right], z_t \end{pmatrix} \\
0 &= \mathcal{V}_{d,t}\left(V_{d,t}, E_t\left[V_{e,t+1}\right], E_t\left[V_{d,t+1}\right], K_{t-1}, H_t, \gamma_{2,t}, z_t\right)
\end{aligned}$$

The representative capitalist then maximises their objective by solving the Lagrangian leading to the *timeless perspective* optimal solution.

$$\mathcal{L} = E_0 \sum_{t=0}^{\infty} \beta^t \left\{U_{c,t}(\cdot) + \lambda_{1,t}\Theta_t(\cdot) + \lambda_{2,t}\mathcal{H}_t(\cdot) + \lambda_{3,t}\mathcal{V}_{e,t}(\cdot) + \lambda_{4,t}\mathcal{V}_{d,t}(\cdot)\right\}$$

3 Results and Analysis

We solve the zero-growth steady-state around which we log-linearize to second order. Standard methods are employed to simulate, and find the moments and auto-correlations of the model. Impulse response functions to the productivity shock are also computed. This information is then compared to a benchmark real business cycle model and a set of stylized facts regarding business cycle dynamics.

Table 1 includes the correlations of key variables with output, and the mean of the variables and standard deviation relative to that of output. The last column contains stylized business cycle facts from Canova in [6]. In comparison to the data, the behaviour of the aggregate variables over the business cycle are similar to those in a benchmark real business cycle model. Investment is found to be more volatile and hours less so in comparison to the benchmark model and the data. The latter is partly consequence of the inelastic labour supply of the workers but it also finds that entrepreneur labour drops following a technology shock, a feature supported by the data (e.g. see Canova in [7]) and typically requiring nominal rigidities to replicate. We consider the proposed model a useful framework to begin to answer the question posed at the start.

Table 1. Correlation with output, mean and standard deviation relative to output, and some stylised facts

	Heterogeneous			RBC Baseline			Canova
	Corr w/ Y	Mean	rel. σ	Corr w/ Y	Mean	rel. σ	rel. σ
Y	1	0.894	1	1	1.168	1	1
I	0.91	0.106	7.3036	0.929	0.186	5.184	2.82
K	0.60	4.598	0.7538	0.577	8.07	0.764	0.61
w	0.99	0.155	0.7281	0.964	1.603	0.670	0.70
H	0.48	0.443	0.0937	0.892	0.510	0.403	1.06
$H_{e,a}$	0.01	0.467	0.1843	-	-	-	-
$H_{e,b}$	-0.21	0.396	0.2810	-	-	-	-
θ	0.81	0.219	1.7719	-	-	-	-
C	0.737	0.652	0.555	0.610	0.982	0.453	0.49
C_c	0.71	5.813	0.5076	-	-	-	-
$C_{e,a}$	0.99	0.207	0.7311	-	-	-	-
$C_{e,b}$	0.98	0.234	0.7734	-	-	-	-
C_d	0.99	0.086	0.7296	-	-	-	-

References

1. Amato, J.D., Laubach, T.: Rule-of-thumb behaviour and monetary policy. European Economic Review 47(5), 791–831 (2003)
2. Bilbiie, F.O.: Limited asset markets participation, monetary policy and (inverted) aggregate demand logic. Journal of Economic Theory 140(1), 162–196 (2008)
3. Bilbiie, F.O., Straub, R.: Asset market participation, monetary policy rules, and the great inflation. The Review of Economics and Statistics 95(2), 377–392 (2013)
4. Bose, A., Pal, D., Sappington, D.E.M.: On the performance of linear contracts. Journal of Economics & Management Strategy 20(1), 159–193 (2011)
5. Cagetti, M., Nardi, M.D.: Entrepreneurship, frictions, and wealth. Journal of Political Economy 114(5), 835–870 (2006)
6. Canova, F.: Detrending and business cycle facts. Journal of Monetary Economics 41(3), 475–512 (1998)
7. Canova, F., Lopez-Salido, D., Michelacci, C.: The effects of technology shocks on hours and output: a robustness analysis. Journal of Applied Econometrics 25(5), 755–773 (2010)
8. Mankiw, N.G.: The Savers-Spenders Theory of Fiscal Policy N. Gregory Mankiw Harvard University. American Economic Review 90(2), 120–125 (2000)
9. Quadrini, V.: Entrepreneurship, saving and social mobility. Review of Economic Dynamics 3(1), 1–40 (2000)

When Can Cognitive Agents Be Modeled Analytically versus Computationally?*

Leigh Caldwell

Inon Cognitive Economics, Inon Company, United Kingdom
leigh@inon.com
http://www.knowingandmaking.com/

Abstract. When agents are modelled with psychologically realistic decision making rules, the assumptions made about their decisions and behaviour are usually more complex than those made in traditional economics. This has made it harder for behavioural and cognitive economists to derive mathematical results analogous to those of the Arrow-Debreu theorem and similar economic findings. It has also made it difficult to generalise behavioural economic results – for example Fudenberg [Fud06] criticises the approach of modifying individual assumptions instead of considering them as a group.

As a result of this additional complexity, many modellers in behavioural economics are turning to agent-based computational methods instead of attempting to find analytic, closed-form solutions to economic problems (e.g. [Boq11], [Tes02]).

Computational methods have many advantages over traditional analytic methods, but also some disadvantages (for example, they make it harder to make very general findings, or to place an economic interpretation on some results).

This paper proposes a new method of modelling agent decision making and behaviour, based on information processing rather than utility and preferences. Models can be built whose agents follow such rules; these produce different micro and macroeconomic predictions to conventional economic models. By shifting the basis of the model to information and how it is transformed by agents, it becomes possible to develop new kinds of economic models which can be understood analytically, not just by computer simulation.

The agents in the model have goals which stochastically become salient at different times ([DTM08]). They learn and use strategies to achieve those goals, such as adaptive heuristics ([HMC03]) and fast-and-frugal rules ([GG02]). They process information using the approach of Payne, Bettman and Johnson 1993, and make choices based on heuristics such as attribute substitution (for example [KF02]).

This approach can complement computational methods and provides some of the generality and elegance which is often thought to be missing

* Thanks to David Laibson, Liam Delaney and others for encouragement and discussions on this topic; to Elina Halonen for debate and development of the ideas, and suggestions of useful sources. Especial thanks to four anonymous reviewers for useful suggestions and comments on an earlier version of this manuscript, which helped to move it forward to the current version.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_8,

from behavioural approaches. It can also replicate some of the standard economic results as well as being compatible with a number of empirically discovered "anomalies" from the judgement and decision-making (JDM) and behavioural economics literatures. It may also offer a way to consider aspects of certain phenomena at the cognitive level which are outside of the scope of traditional choice-based economics, but which are clearly important to real individuals and have real-world consequences: motivation, happiness, deliberate ignorance, and the learning of new preferences.

The paper is currently a work in progress and represents one step towards a possible closed-form macroeconomic model based on cognitive and behavioural microfoundations.

1 Introduction

Under close inspection, the general equilibrium (and to a large extent, other microeconomic) theorems at the heart of modern economics rely on a perhaps surprising number of strict assumptions; assumptions which are required to constrain the mathematical form of their elements to allow tractable analysis. Most of these assumptions can be fulfilled by agents who conform to a narrow standard of utility-based economic rationality; a few others (such as convex preferences) are harder to justify purely in rationality terms, but are needed to make the sums add up.

Over time, as psychological experiments have provided more counterexamples to these rationality assumptions, behavioural and cognitive economists have wanted to relax the assumptions and find out the effects on the findings of economic theory. Unfortunately, most attempts to do this have rendered the mathematical superstructure of microeconomics unworkable – throwing out most standard economic results rather than replacing them with weaker (but more empirically realistic) versions.

A notable exception is Xavier Gabaix's "sparse bounded rationality" [Gab11], which is mathematically tractable (under certain assumptions about utility functions) but only allows relaxation of a subset of the assumptions.

As a result, modellers have turned to computational methods to generate economic predictions. Computation allows agents to be specified with any kind of behavior or decision-making strategy, and for the outcomes of populations of agents over time to be predicted. This approach, however, lacks generality and risks creating "black boxes", where the behavior of the system depends on the implementation of the software environment as well as on the psychological or behavioural assumptions made about the agents.

This paper attempts to show that an alternative approach, based on information processing, can provide both mathematical tractability and psychological realism. As a result, theorems can both be highly general and fit empirical data.

2 A New Approach

Building models of agents based on (modified) utility maximization seems to be difficult. The game theory proofs that allow a group of agents to collectively achieve a Pareto optimal price for a good are not robust to most psychological modifications; for example, when agents learn from each other at finite rather than infinite speed, the theory makes no prediction about how a population will converge on a price or what that price will be ([Cam11]).

However, by changing the key variables, often a system can be looked at in a new way. Instead of focusing on the allocations of goods to each agent, and predicting their behavior based on expected utility with equilibrium constraints, we model information.

Agents are considered to possess certain information, about themselves and about the world. From this information they infer goals and ascertain what strategies can be followed to achieve those goals.

As agents pursue strategies and interact with other agents, they gain new information and forget old information. We look at the properties of the whole system in terms of total information, and consider the welfare of agents using properties such as their ability to achieve their goals.

How might we model the propagation and processing of information in such a system? The theory of decision making gives us some guidelines. A descriptive model of how individuals process information is given in Payne, Bettman and Johnson 1993. This approach suggests that agents break down problems into a finite number of smaller steps which consist of simple operations such as matching two pieces of in-formation or comparing two values. The resulting processes must necessarily obey mathematical constraints on information processing, such as those given by Turing [Tur37], but are likely to obey tighter constraints still when psychological constraints such as short-term memory limits and attentional resources are incorporated.

Agents' choices can be described by heuristics such as attribute substitution (for example [KF02]), and fast-and-frugal rules ([GG02]).

It will be possible to replace the specific assumptions made about heuristics while retaining the more general model of the constraints on how information can be processed. This should insulate economic discoveries from the exact assumptions made about microfoundations.

3 A Possible Model

The following model is one possible way to describe an economy in terms of information processing agents, information stocks and flows, and agents' actions in response. This paper does not attempt to work out the full consequences of this model, but uses it to illustrate some possible methods for analytical work.

Let an economy be composed of a finite but large number n of agents $\{A_1, A_2, A_3, ..., A_n\}$. Each agent A_i possesses beliefs: a set of m(i) strings $s_{i1}, s_{i2}, s_{i3}, ..., s_{im(i)}$ of data. Agents are capable of interpreting the strings as statements about the world (we might imagine statements as diverse as "I have \$2400 in my bank account", "elephants cannot jump" or "I am hungry", or meta-statements about other agents or about other statements, such as "most people believe that the President can be trusted"). Agents also possess endowments of a set of g goods or real resources $\{r_{i1}, r_{i2}, r_{i3}, ..., r_{ig}\}$.

Agents evaluate these statements about the world in an order given by some salience ranking, and will in some cases judge that the state of the world described by the statement can be improved. (Without such judgements, agents will never be motivated to take any action.) This provides the agent with a goal – to improve the world in this respect. An agent then acts on one of these judgements using a strategy S from a set of strategies $\{S_{i1}, S_{i2}, ...\}$ Strategies are processes that agents can execute, each composed of a sequence of steps which use real resources in the world, change the agent's own set of beliefs, or send messages to other agents.

The agent then receives messages from the world in turn, which it interprets and compares to its existing beliefs, which it updates if appropriate.

To specify this model fully, various properties and capabilities of the agents would need to be considered:

- How agents interpret strings to determine the state of the world
- The semantic structure of their mental model of the world
- The process by which they judge that the state of the world could be improved
- What kinds of strategies they possess to change the state of the world, and how they learn or design new strategies
- How they update their beliefs based on messages received

In addition, the state of the world itself is subject to physical constraints arising from the relationships between the real resources it contains, which also need to be described in detail for a fully specified model.

Applying the principle of the universal Turing machine, the beliefs of an agent, its strategies, its mental model of the world, and the internal mechanism it uses to interpret beliefs as statements about the world, can be seen interchangeably as "program" and "data" within the same mental state. This provides a mechanism for new strate-gies to be learned as agents update their beliefs based on the outcomes of previous strategies.

Treating the agent in this way as a computational device we can apply results on such concepts as Bennett's logical depth (as discussed in [ZD13]) and on information flow in the sense of Shannon [Sha49].

4 Results

How agents process new information and accumulate knowledge is yet to be rigorously determined, but some fruitful directions are suggested by the proposed model.

Established results on algorithmic complexity impose constraints on the kinds of calculation that can be carried out by agents within realistic time limits. This implies that agents' judgements about the state of the world cannot be based on numerical utility calculations. Instead, we can infer that special-purpose hardware in the brain allows agents to recognize similarities between current and remembered or hypothe-sized states, and make (inexact) comparisons between them. Of course, special-purpose hardware is not immune to the effects of complexity theory, but the massive parallelism of the brain can allow certain limited types of processing to be carried out quickly in time, at the cost of using larger spatial resources and a loss of generality.

It seems likely that we can derive further results about the total size of the belief set of agents (individually and collectively): intuitively, it should increase over time as messages are communicated between agents. Over time, beliefs should also evolve to more accurately reflect the relationships between real resources in the external world.

Most microeconomic propositions – right down to the claim that people buy more of something when the price goes down – can be re-expressed in terms of this model of information, goals and strategies. The model provides tools to explore not just whether these propositions are correct, but the conditions under which they are a more or less close approximation to the truth, and how they evolve over time.

Further work will mathematically model how economic phenomena take place in a world of information processing agents following fast and frugal heuristics. A theory of how agents determine and prioritise goals will be developed based on the same model. Two key differences between this approach and traditional economic models are, first, that agents do not necessarily know the allocation of resources that will best help them to meet their goals – they will learn more over time about which resources to use and how - and second, that it takes time to apply strategies which reallocate resources. As a result, agents with multiple goals can only make progress towards those goals over time. They are likely to need to make compromises between goals and to leave some goals unsatisfied, even if they could in principle acquire the resources to achieve them.

Different microfoundations might suggest different market-level or macroeconom-ic outcomes; one objective of the model is to ascertain boundary conditions so that system modellers can understand individual behavioural thresholds, in order to predict when systems will cross those thresholds and change phase.

The model also offers a new way to ask important questions about economic welfare – such as how successfully will particular economic structures allow agents to achieve their goals; and how will their goals change in the face of different beliefs or different resource constraints. The answers to those questions may be some way off, but the absence of a way to even formalize the questions has been a problem for neo-classical economics since the emergence of the behavioural challenge, and one not yet satisfactorily answered by empirical behavioural research.

The model therefore offers a way to reconcile some of the conflicting results of judgement and decision making theory, equilibrium and welfare economics, complex systems theory and empirical behavioural psychology.

5 Limitations

Beckage et al [BKG+13] point out that results on computational irreducibility make it impossible to fully predict the behaviours or outcomes of complex systems in any faster or simpler way than by observing the system itself. Therefore, any analytical results achieved from the above model (or any similar one) will only be an approximation to the true behaviour of the world.

However, this does not mean the model is not useful. Velupillai [Vel13] says:

"The mischief indulged in by economists, particularly those advocating blind agent-based modelling in economics and finance, claiming that their practice makes the case against formal mathematics in its deductive underpinnings, enhancing the case for a 'new mathematics' that is inductively based, shunts research towards pointless ends."

Velupillai shows that boundedly rational agents are equivalent to universal Turing machines, and thus are constrained by results on incompleteness and the Halting Problem. However, despite these constraints we can still make analytical inferences about the behaviour of such machines. For example, Ruivo and Oliveira [Rd13] use spectral analysis of cellular automata to draw conclusions about (some of) their behaviour without the need to simulate them entirely. And Rucker [Ruc13] shows that the natural world itself is subject to the complexity and reducibility constraints of Turing machines, without invalidating the many powerful analytical results achieved in the physical sciences.

The point, then, is to use the right tools for the right job. Computational methods will still be needed to predict certain specific outcomes – just as we can better predict next week's weather with measurements and simulations than with general principles. Analytic models, for their part, will allow relationships between certain aggregate variables to be understood and general principles of agent behaviour to be inferred.

The answer to the question posed by the title of this paper is: agents can be modeled analytically when we find the right simplifications to make their aggregate behaviour tractable. Traditional economics uses one set of simplifications; behavioural economics challenges these, and the model proposed here provides an alternative basis for simplifying agents that lets us study them in a different way.

6 Conclusions

The model suggests that it will be possible to infer some aggregate facts about economic behavior from individual-level axioms of information processing. These facts will have the generality of an analytic theory and will illuminate general properties of the economic world, while having the psychological realism of behavioural economics.

This modeling approach will not replace computational methods but will complement them in domains where analytic proofs are necessary, and where a simplified model of agent behavior is acceptable in return for this additional generality.

References

[BKG+13] Beckage, B., Kauffman, S., Zia, A., Koliba, C., Gross, L.J.: More complex complexity: Exploring the nature of computational irreducibility across physical, biological and human social systems. In: Zenil, H. (ed.) Irreducibility and Computational Equivalence. ECC, vol. 2, pp. 79–88. Springer, Heidelberg (2013)

[Boq11] Boqvist, S.: A neurocomputational perspective on behavioural economics: a study of emotional processes. Master's thesis, KTH Royal Institute of Technology, Stockholm (2011)

[Cam11] Camerer, C.F.: Behavioural Game Theory: Experiments in Strategic Interaction. Princeton University Press (2011)

[DTM08] Dastani, M.M., Tinnemeier, N.A.M., Meyer, J.-J.C.: A programming language for normative multi-agent systems. In: Dignum, V. (ed.) Multi-Agent Systems: Semantics and Dynamics of Organizational Models. IGI Global (2008)

[Fud06] Fudenberg, D.: Advancing beyond advances in behavioural economics. Journal of Economic Literature 44(3), 694–711 (2006)

[Gab11] Gabaix, X.: A sparsity-based model of bounded rationality. NBER Working Papers (16911) (2011)

[GG02] Goldstein, D., Gigerenzer, G.: Models of ecological rationality. Psych. Review 109(1), 75–90 (2002)

[HMC03] Hart, S., Mas-Colell, A.: Regret-based continuous-time dynamics. Games and Economic Behavior 45(2), 375–394 (2003)

[KF02] Kahneman, D., Frederick, S.: Representativeness revisited: Attribtue substitution in intuitive judgment. In: Gilovich, T., Griffin, D., Kahneman, D. (eds.) Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press (2002)

[Rd13] Ruivo, E.L.P., de Oliveira, P.P.B.: A spectral portrait of the elementary cellular automata rule space. In: Zenil, H. (ed.) Irreducibility and Computational Equivalence. ECC, vol. 2, pp. 211–235. Springer, Heidelberg (2013)

[Ruc13] Rucker, R.: An incompleteness theorem for the natural world. In: Zenil, H. (ed.) Irreducibility and Computational Equivalence. ECC, vol. 2, pp. 185–198. Springer, Heidelberg (2013)

[Sha49] Shannon, C.E.: The Mathematical Theory of Communication. University of Illinois Press (1949)

[Tes02] Tesfatsion, L.S.: Agent-based computational economics: Growing economies from the bottom up. Staff General Research Papers (5075) (2002)

[Tur37] Turing, A.M.: On computable numbers, with an application to the entscheidungsproblem. Proceedings of the London Mathematical Society 42(2), 230–265 (1937)

[Vel13] Velupillai, K.V.: The relevance of computation irreducibility as computation universality in economics. In: Zenil, H. (ed.) Irreducibility and Computational Equivalence. ECC, vol. 2, pp. 101–111. Springer, Heidelberg (2013)

[ZD13] Zwirn, H., Delahaye, J.-P.: Unpredictability and computational irreducibility. In: Zenil, H. (ed.) Irreducibility and Computational Equivalence. ECC, vol. 2, pp. 273–295. Springer, Heidelberg (2013)

Designing a Homo Psychologicus More Psychologicus: Empirical Results on Value Perception in Support to a New Theoretical Organizational-Economic Agent Based Model

Andrea Ceschi[1], Enrico Rubaltelli[2], and Riccardo Sartori[1]

[1] University of Verona, Verona, Italy
{andrea.ceschi,riccardo.sartori}@univr.it
[2] University of Padua, Padua, Italy
enrico.rubaltelli@unipd.it

Abstract. The study presents a new approach of modelling human behavior based on empirical evidence on individual differences in cognitive science and behavioral economics fields. Compared to classical studies of economics, empirical research makes use of the descriptive approach to analyze human behavior and to create models able to explain the behavior of investors and organizational traders in a more realistic way. Consistently, an economic assumption that has been strongly disputed by scientists is the concept of Homo Economicus, which is currently considered unable to capture all the details and variability that characterize human behavior (which we define, in opposition to the economic label, Homo Psychologicus). Thanks to recent empirical studies and the development of such advanced techniques as agent based models, new simulation studies are now capable of investigating a higher number of psychological variables. However, models which implement heuristics or fallacies often distribute these characteristics among all agents without distinction. The present study shows how it is possible to design multiple agents considering individual differences, which can have a different impact on organizational and economic behavior. Starting from several empirical studies, which show a negative relation between optimism and loss aversion, coefficients of the Value function of the Prospect theory have been reviewed to create agents characterized by different psychological strategies used to manage costs and risks.

Keywords: Agent Based Model, Homo Economicus, Loss aversion, Optimism, Prospect Theory, Value function.

1 An Introduction to the Contribution of Cognitive Science and Behavioral Economics Studies

In economics many models are based on such assumptions as the Homo Economicus conception which can never explain and predict in its integrity human behavior (Thaler, 2000). Considering past work on bounded rationality, it has been proved that humans do not adhere to the principles that drive the actions of such a rational agent

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_9,

as the Homo Economicus (Simon, 1955, 1991). This depends on several psychological processes like heuristics, the satisfaction seeking orientation, and so on.

Since the Seventies, these concepts have been further explored by psychologists like Amos Tversky and Daniel Kahneman and economists like Richard Thaler, Hersh Shefrin and Shlomo Benartzi who have uncovered automatic cognitive strategies like Sunk cost, Extra cost and Loss aversion. These strategies have been proved to have an important role in human decision-making and can also influence the management of financial assets. Starting from these assumptions, it was possible to elaborate new financial and economic theories and models useful to understand such phenomena as Money illusion (Shafir, Diamond, & Tversky, 1997) or to provide alternative explanations to economic problems like the Equity premium puzzle (Benartzi & Thaler, 1995). These authors demonstrated that empirical studies are extremely important to understand how people make decisions in financial markets and organizations. These studies are also of great value to elaborate new theories of human behavior and decision-making that consider psychological factors and individual differences (Sartori & Ceschi, 2012).

1.1 Cost Biases versus Optimistic Fallacies

Such fallacies as the Sunk cost bias, the Extra cost bias, the Loss aversion, the Zero risk bias, the Regret aversion and the Endowment effect have in common a pessimistic approach toward losses, a disposition to give a high value to the possibility of risking to lose, or a tendency of subjects to give a lower value to future gains. These effects are caused by similar processes characterized by a desire to avoid risks and losses, plus a pessimistic disposition to manage these events. The idea that these fallacies are related has been put forward in several studies (Ceschi, Sartori, & Weller, 2013). Soman and Cheema (2001) proposed the Loss aversion as a potential explanation for the Sunk cost fallacy. Consistent with this explanation, Frisch (1993) found that young adults reported the loss aversion as the main reason they were prone to the Sunk cost effect. Also pessimism is strictly connected to loss aversion, cost aversion, and risk aversion as showed in several studies (Chapman & Polkovnichenko, 2009; de Palma et al., 2008).

On the other side, fallacies related to the Optimistic disposition present an opposite profile. Fallacies belonging to this dimension are featured by a lack of experience and by the presence of a sort of optimism for the future. In fact, if experience is a factor that over time induces people to evaluate losses as more probable, the lack of experience can affect people's judgments inducing an under-evaluation of costs or making them more optimistic and willing to accept risks. These fallacies are characterized by the influence of hope in future events like in the Optimistic bias, the Planning fallacy and the Hindsight bias. More in general, these biases are found in situations in which people desire positive outcomes from decisions that they have to take, they have already taken or they are going to take.

1.2 The Importance of Individual Differences in Value Perception

Recent research has suggested that setting a value or a price is a more deliberative thinking process than reporting attractiveness ratings, that is to say, cost and price issues are likely to depend on rather deliberative, conscious cognitive processes (Rubaltelli, Dickert, & Slovic, 2012). This conclusion is also supported by a study conducted by Iyer, Lindner, Kagan, and Andersen (2010). Iyer and colleagues used a magnetic resonance to monitor brain areas in subjects who were asked to perform a task in which different amounts of money could be gained or lost. Two types of subjects were identified. Those who reported they were "good" at the task were defined optimists. They showed a higher brain activity when they expected gains. Instead, those who were defined pessimists showed higher brain activity when trying to avoid losses. The Sunk costs fallacy and, in general, all the fallacies driven by the Losses-pessimism dimension show that people could learn across the life the "do not waste" heuristic in order to be more efficient in cost management (Strough, Karns, & Schlosnagle, 2011).

1.3 The Difficulty in Capturing Cost Dynamics and Individual Changes over Time

Basically the "do not waste" heuristic is the disposition to give a higher importance to the costs incurred and theory suggests that people can develop over time an aversion to wasting resources (Arkes & Blumer, 1985). Evidence coming from research on adult decision-making abilities show that the Sunk Cost depends on people's experience (Bruine de Bruin, Parker, & Fischhoff, 2007). This aspect is difficult to analyze because most of the studies and models based on the descriptive approach lack time series data. Most of the experiments carried out are static and based on different situations, but, for what concerns individual differences, they are often unable to capture dynamics over time.

Consequently a question arises: How is it possible to create studies and models in decision-making that are fit to consider individual differences over time? For instance, what happens in a market when a huge number of interactions among different investors are happening? These questions are difficult to be answered with a classical experimental approach. The necessity for a new approach brought us to consider a theoretical and computational Agent-Based Model (ABM), in order to respond to these scientific issues.

2 Agent-Based Techniques for Redesigning Human Behavior Models

Agent-Based Models (ABMs) are a simulation modeling technique that uses virtual agents interacting with other virtual agents within a virtual environment and with certain virtual resources (Bonabeau, 2002). The aim of this instrument is to simulate and predict possible scenarios based on one or multiple established behaviors. In Social

Sciences, ABMs are used to realize different virtual simulations regarding people interactions. One of the most interesting field is probably related to the decision-making and social psychology areas and, in particular, to the possibility of creating agents that better simulate human behaviors and social interactions. Most social and psychological phenomena occur not as the result of isolated decisions made by individuals but rather as the result of repeated interactions between heterogeneous individuals over time.

2.1 Individual Differences in ABMs

One characteristic of ABMs is the possibility of creating agents with unique characteristics capable of producing different behaviors. Once human behavior has been studied through empirical research that considers individual differences, it is possible to insert different coefficients into the model. Currently, several sophisticated multi levels methodologies allow to develop agents that can simulate human behavior by considering individual differences into the model. These multilevel models can describe many details and relationships among the constructs but, as a detailed photograph, they are not able to design dynamics of the behavior (Ceschi, Hysenbelli, Sartori, & Tacconi, 2013).

2.2 Capturing Dynamics

The particularity of ABMs is related to the possibility of obtaining dynamics systems which can reproduce simplified behaviors and trends over time. This instrument allows to overcome a limit of the descriptive approach which often lacks time series models. Most of the experiments based on the descriptive approach are static and just few of them can capture temporal dynamics.

ABMs allow also to register variations if during the experiment the environment or other variables change. This allows to create systems with agents able to learn from experience, which is a factor extremely important especially with agents dealing with economic transitions (Ceschi, Hysenbelli, & Slovic, 2013).

3 From Empirical Results to a Theoretical and Computational ABM: A Reinterpretation of the Prospect Theory

Several studies on behavioral finance have used ABMs to design more realistic agents by considering fallacies, heuristics and biases which can influence the human activity during trading or economic transactions in organizations (Janssen & Ostrom, 2006; LeBaron, 2000, 2006). Apart from very few studies, most ABMs in this field modeled phenomena like Overconfidence, Loss aversion or Status quo bias. In particular, these models are characterized by a comparison between markets in which all agents into the system are overconfident or loss averse and markets with hypothetical rational agents. The goal being, of course, to assess potential deviations from rationality.

We believe that a limit of these models is that they often fail to consider the presence of individual differences in agents and delicate relations among effects and constructs. It is the case of the relation presented between cost perception and the optimistic disposition that, in general, have been modeled as distinct phenomena present indistinctly in all agents.

In the next section we are going to present a possibly theoretical explanation based on the value function, which we believe can explain this relation and can be implemented in an ABM by considering differences between agents.

3.1 A Possible Implementation of the Value Function into Agents Considering Individual Differences

Considering the studies previously presented, it is possible to detect the presence of a negative relation in individuals between the optimism and the cost perception. With regards to values (y) of the Value Function of Kahneman and Tversky (1979), results for the positive and negative consequences of the choice have the characteristic of diminishing utility:

$$V(y) = \begin{cases} x^{\alpha}, & x > 0 \\ -\lambda(-x^{\alpha}), & x < 0 \end{cases}$$

The coefficient α in the equations above captures the trend of the decreasing function. Empirical studies estimate that α is on average equal to 0.88 and always less than 1.00. As it is well known, empirical results have also shown that humans are in general more sensitive to losses compared to gains (Kahneman & Tversky, 1979). In fact, the value function is steeper for losses than for gains. The coefficient λ, which is estimated at 2.25, indicates the difference in the slopes of the positive and negative arms of the value function.

Considering the evidence above, it is possible to assume that sensitivity to losses is not equal for all humans although this does not mean that λ should assume different values. It is more likely that variations in sensitivity depend more on the α coefficient than λ. This could explain why some people feel less losses and costs compared to others and, at the same time, that they are less attracted to gains. A variation of λ would change just the loss aversion disposition but not the perception on gains.

The α coefficient can be considered as opposite to the risk disposition, and to the availability or optimism in investments, because of the increase in the magnitude of the perceived value. This characteristic, as seen in the studies previously cited, is differently present in individuals and negative related to cost perception and Loss aversion.

Considering this explanation, investors can be designed considering a normal distribution of α (0;1;μ=0.88) in order to obtain agents with different disposition in managing value costs and gains (Figure 1).

The possibility of implementing this distribution in agents allow to replicate a relation among constructs already explored and proved by different empirical studies, capable of creating agents with different behaviors and dispositions in managing costs.

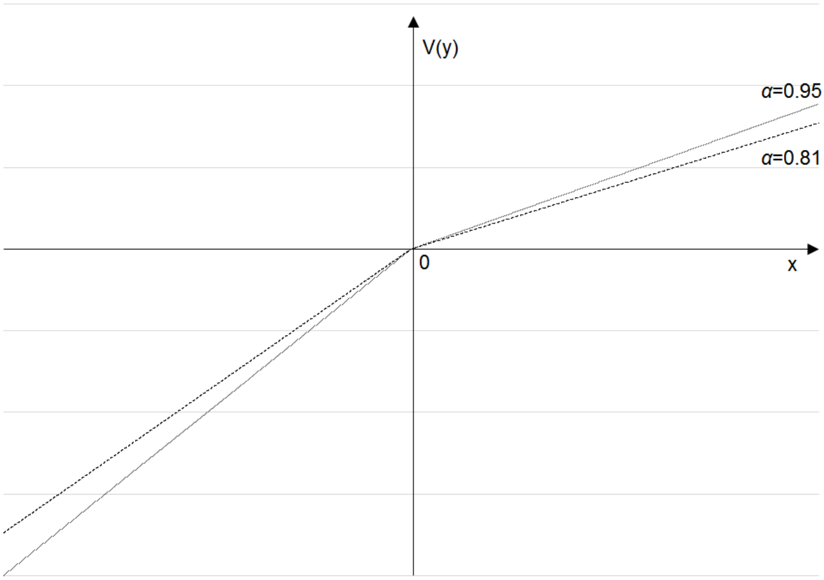


Fig. 1. Different α values in the Prospect Theory

The possibility of exploring different investment scenarios over time allows to study the development of different strategies based on learning from previous transactions. The decision to invest is conditioned by the Value function expressed over time which determines the emergence of learning strategies in agents.

4 Conclusion

The theoretical model here presented allows to rethink the role of the Value function in organizational economic models and to better understand human behavior based on the psychological, subjective perception of costs and gains. The Value function has been revisited on the basis of the additional empirical evidence presented earlier. The present study has shown a way to implement empirical theories through ABMs. We think that it could be a good example in order to show how "generative science" works and how to connect experiments that come from the descriptive approach to Agent Based Models.

Considering also that most of the studies which come from the descriptive approach miss of testing individual differences and effects over time, an ABM seemed to us the ideal solution to include them into the analysis of economic behavior over time. We encourage this way of making science which is more connected and integrated in order to develop relevant interdisciplinary studies.

References

1. Arkes, H.R., Blumer, C.: The psychology of sunk cost. Organizational Behavior and Human Decision Processes 35(1), 124–140 (1985)
2. Benartzi, S., Thaler, R.H.: Myopic Loss Aversion and the Equity Premium Puzzle. The Quarterly Journal of Economics 110(1), 73–92 (1995)
3. Bonabeau, E.: Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences of the United States of America 99(suppl. 3), 7280 (2002)
4. Bruine de Bruin, W., Parker, A.M., Fischhoff, B.: Individual differences in adult decision-making competence. Journal of Personality and Social Psychology 92(5), 938 (2007)
5. Ceschi, A., Hysenbelli, D., Sartori, R., Tacconi, G.: Cooperate or Defect? How an agent based model simulation on helping behavior can be an educational tool. Paper presented at the Paper Session Presented at the Meeting of the 3rd ebuTEL, International Workshop on Evidence Based and User Centred Technology Enhanced Learning, Trento, Italy (2013)
6. Ceschi, A., Hysenbelli, D., Slovic, P.: When awareness of those we cannot help demotivates us from helping those we can help: An agent-based simulation study of pseudoinefficacy. In: Dickert, S. (ed.) Paper presented at the Prosocial Responses to Donation Requests: Motivators and Demotivators. Symposium Conducted at the Meeting of the 55th TeaP Conference of Experimental Psychologists, Vienna, Austria (2013)
7. Ceschi, A., Sartori, R., Weller, J.: The factors of reasoning: Structures of belonging of heuristics and biases. Paper presented at the Poster Presented at the Meeting of the 24th SPUDM, Subjective Probability, Utility, and Decision Making. European Association of Decision Making, Barcelona (2013)
8. Chapman, D.A., Polkovnichenko, V.: First-Order Risk Aversion, Heterogeneity, and Asset Market Outcomes. The Journal of Finance 64(4), 1863–1887 (2009)
9. de Palma, A., Ben-Akiva, M., Brownstone, D., Holt, C., Magnac, T., McFadden, D., ... Wakker, P.: Risk, uncertainty and discrete choice models. Marketing Letters 19(3-4), 269–285 (2008)
10. Frisch, D.: Reasons for framing effects. Organizational Behavior and Human Decision Processes 54, 399–429 (1993)
11. Iyer, A., Lindner, A., Kagan, I., Andersen, R.A.: Motor preparatory activity in posterior parietal cortex is modulated by subjective absolute value. PLoS Biology 8(8), e1000444 (2010)
12. Janssen, M.A., Ostrom, E.: Empirically based, agent-based models. Ecology and Society 11(2), 37 (2006)
13. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. Econometrica 47, 263–291 (1979)
14. LeBaron, B.: Agent-based computational finance: Suggested readings and early research. Journal of Economic Dynamics and Control 24(5), 679–702 (2000)
15. LeBaron, B.: Agent-based computational finance. In: Handbook of Computational Economics, vol. 2, pp. 1187–1233 (2006)
16. Rubaltelli, E., Dickert, S., Slovic, P.: Response mode, compatibility, and dual-processes in the evaluation of simple gambles: An eye-tracking investigation. Judgment and Decision Making 7(4), 427–440 (2012)
17. Sartori, R., Ceschi, A.: Biases, reasoning and personality in finance. International Journal of Psychology 47(suppl. 1), 109–151 (2012)
18. Shafir, E., Diamond, P., Tversky, A.: Money Illusion. The Quarterly Journal of Economics 112(2), 341–374 (1997)

19. Simon, H.A.: A behavioral model of rational choice. The Quarterly Journal of Economics 59, 99–118 (1955)
20. Simon, H.A.: Bounded rationality and organizational learning. Organization Science 2(1), 125–134 (1991)
21. Soman, D., Cheema, A.: The effects of windfall gains on the sunk-cost effect. Marketing Letters 12, 51–62 (2001)
22. Strough, J.N., Karns, T.E., Schlosnagle, L.: Decision-making heuristics and biases across the life span. Annals of the New York Academy of Sciences 1235(1), 57–74 (2011)
23. Thaler, R.H.: From homo economicus to homo sapiens. The Journal of Economic Perspectives 14(1), 133–141 (2000)

Differences between Entrepreneurs and Managers in Large Organizations: An Implementation of a Theoretical Multi-Agent Model on Overconfidence Results

Riccardo Sartori, Andrea Ceschi, and Andrea Scalco

University of Verona, Verona, Italy

Abstract. The well-known research carried out by Busenitz and Barney (1997) exploring differences in the decision-making processes between entrepreneurs and managers in large organizations has been revisited and redesigned as a starting point to create a computational and theoretical Multi Agent Model (MAM) which shows differences in the decision-making processes. In the original study, researchers showed the presence of a different disposition in incurring in biases and in heuristics by entrepreneurs and managers. In particular, two interesting trend curves on the Overconfidence effect have been realized. Authors concluded by stating that the Overconfidence effect is significantly different in entrepreneurs and managers and helps distinguish between these two work categories. Starting from this conclusion and from their results, a computational and theoretical MAM has been designed, where, as suggested by the authors, different decision-maker agents can incur in the Overconfidence effect with different degrees.

Keywords: Multi Agent Models, Organizations, Entrepreneurs, Biases, Overconfidence.

1 An Introduction to the Individual Differences in Biases and Heuristics

Over the last five years, researchers in psychology and social sciences (Bruine de Bruin, Parker, & Fischhoff, 2007; Sartori & Ceschi, 2011, 2012) have strongly stressed what classical studies by famous Amos Tversky and Daniel Kahneman among the others found long ago: individual differences play an extremely important role in the availability to incur in biases and heuristics.

Individual differences are also reflected in in career choices, so it becomes of some interest to study people who take different professional roles, such as entrepreneurs and managers. These studies were conducted by means of several experiments in order to describe the decision-making processes taken by people and to recognize differences between the pure rationality of several economic theories.

In order to better understand the real processes of choice, researchers prefer to test when the rational choice is being violated rather than validating the human ability or

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_10, © Springer International Publishing Switzerland 2014

following several theoretical assumptions and axioms. In fact, the violation of rational choices can be measured experimentally, by carrying out the same experiments conducted in order to identify them.

In a sort of way, the degree of using the System 2 rather than the System 1 or the level of rationality (in strict meaning) is measured by analyzing the disposition of individuals in using several heuristics or incurring in cognitive biases. The ability to result in non-normative choices varies from individual to individual because heuristic responses are sometimes overridden by a non-autonomous analytic system of thought (Stanovich & West, 2000).

2 The Busenitz and Barney Research, a Summary

Research on differences between entrepreneurs and managers in large organizations has usually examined psychological and personal/demographic differences (Favretto & Sartori, 2007). After a great deal of research, it is now generally concluded that most of the psychological differences between entrepreneurs and managers in large organizations are small or nonexistent, although some exceptions exist. For example, such individual psychological attributes as locus of control and risk-taking have been shown not to vary significantly between entrepreneurs and managers in large organizations (Begley & Boyd, 1987; Sexton & Bowman, 1984), but some consistent psychological differences have been documented in need for achievement, tolerance for ambiguity and need for conformity (Begley & Boyd, 1987; Miner *et al.*, 1989). Despite the fact that very few studies have shown statistically significant differences between entrepreneurs and managers in large organizations in their risk-taking propensity (Brockhaus, 1980; Low & MacMillan, 1988), this individual psychological difference continues to be discussed as an important variable for understanding entrepreneurial behavior (Stevenson & Gumpert, 1985; Ray, 1994).

The study by Busenitz and Barney (1997) aimed at understanding why entrepreneurs and managers in large organizations may vary in the use of heuristics and biases, by measuring their disposition in incurring in Representativeness and in Overconfidence. Busenitz and Barney (1997) collected a sample from the two populations and, in order to measure the biases, they used two different tasks. To measure the Overconfidence, a task extracted from a study by Fischhoff, Slovic, and Lichtenstein (1977) was used. It was composed of a series of questions based on death rates from various diseases and accidents, such as: "Which cause of death is more frequent in the United States? A. Cancer of all types, B. A shark attack”. Respondents had to indicate the level of confidence that they had in their answers on a scale ranging from 50% to 100%. A response of 100% would indicate that they were totally confident that their choice was right.

Results confirm the hypothesis that entrepreneurs manifest the Overconfidence effect in decision-making processes more extensively than managers in large organizations. They showed how overconfident entrepreneurs are by drawing summary results and combining the correct percentage of responses given and their general level of overconfidence.

In accordance with the results found, we designed two exponential trend curves (Figure 1). Exponential curves of tendencies in Figure 1 indicate that entrepreneurs are in general more overconfident in their choices compared to the correctness of their responses. Instead, managers are less overconfident in their responses.

2.1 Feeling Overconfidence Under Uncertainty

Considering the results presented, it is possible to think about the presence of a different approach in feeling and managing overconfidence by these two categories. As previously showed, managers and entrepreneurs react differently in front of situations featured by different levels of uncertainty. If a low level of correctness is related to a low level of information necessary to individuate the right response, then the contextual level of uncertainty will be high. This brings individuals to respond to the situation with different levels of overconfidence depending on several factors, in part summarized to certain individual differences.

Considering the data here presented, it is possible to deduce that, in front of the same uncertain environment, managers will tend to respond with less overconfidence than entrepreneurs. Results allowed to think of a computational and theoretical Multi Agent Model (MAM) where decision-makers deviate from the strict econometric approach and where they can incur in overconfidence with different degrees (Figure 1).

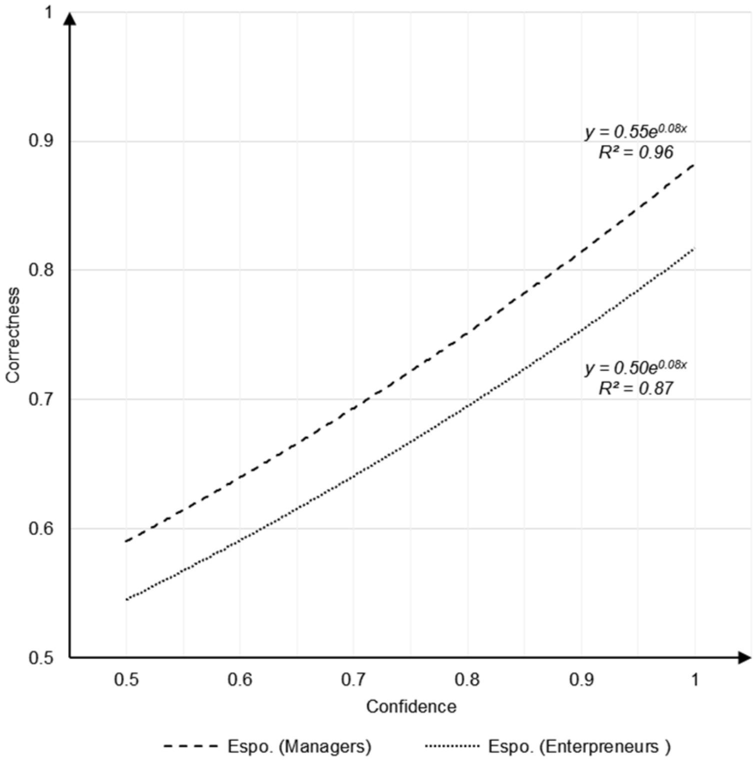


Fig. 1. Exponential trend curves of incurring in Overconfidence effect

3 Multi-Agent Models for Designing the Organizational Behavior

One characteristic of Multi-Agent Models (MAMs) is the possibility of creating different agents with unique characteristics capable of producing different behaviors once that a particular behavior has been studied through empirical research that considers individual differences.

In this way, it becomes possible to insert different coefficients into the model capable of simulating different agent behaviors (Ceschi, Hysenbelli, Sartori, & Tacconi, 2013; Ceschi, Hysenbelli, & Slovic, 2013). The next section will propose a computational and theoretical MAM starting from the empirical results presented above.

3.1 Creating Different Agents with MAMs

The results presented from the study by Busenitz and Barney (1997) allow to create agents with different capacity of managing events on the basis of their level of overconfidence. The possibility of simulating in organizations the behavior of entrepreneurs and managers should consider this empirical evidence, where the level of confidence grows logarithmically and differently on the basis of the uncertainty level present into the system, which can be resumed with the following equation:

$$c = ln(\frac{i}{\alpha})\,\gamma$$

where i is the level of information present into the environment (or at the opposite, the level of uncertainty) γ the empirical constant equal to 1.25, and α is the level of overconfidence. These parameters allowed to design two kinds of agents with different levels of confidence (c) based on a normal distributed α (Figure 1).

4 Conclusion

The present study has presented a methodological approach to implant empirical theories on MAMs. We think that it could be a good example in order to show how "generative science" works and how to connect experiments which come from the descriptive approach to agent based models. We encourage this way of making science which is more connected and integrated in order to develop relevant interdisciplinary studies.

References

1. Begley, T.M., Boyd, D.P.: Psychological characteristics associated with performance in entrepreneurial firms and smaller businesses. Journal of Business Venturing 2, 79–93 (1987)

2. Brockhaus, R.H.: Risk taking propensity of entrepreneurs. Academy of Management Journal 23, 509–520 (1980)
3. Bruine de Bruin, W., Parker, A.M., Fischhoff, B.: Individual differences in adult decision-making competence. Journal of Personality and Social Psychology 92(5), 938 (2007)
4. Busenitz, L.W., Barney, J.B.: Differences between entrepreneurs and managers in large organizations: Biases and heuristics in strategic decision-making. Journal of Business Venturing 12(1), 9–30 (1997)
5. Ceschi, A., Hysenbelli, D., Sartori, R., Tacconi, G.: Cooperate or Defect? How an agent based model simulation on helping behavior can be an educational tool. Paper Presented at the Paper Session Presented at the Meeting of the 3rd ebuTEL, International Workshop on Evidence Based and User Centred Technology Enhanced Learning, Trento, Italy (2013)
6. Ceschi, A., Hysenbelli, D., Slovic, P.: When awareness of those we cannot help demotivates us from helping those we can help: An agent-based simulation study of pseudoinefficacy. In: Dickert, S. (ed.) Paper presented at the Prosocial Responses to Donation Requests: Motivators and Demotivators. Symposium Conducted at the Meeting of the 55th TeaP Conference of Experimental Psychologists, Vienna, Austria (2013)
7. Favretto, G., Sartori, R.: Le età dell'impresa. Franco Angeli, Milano (2007)
8. Fischhoff, B., Slovic, P., Lichtenstein, S.: Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance 3, 552–564 (1977)
9. Low, M.B., MacMillan, I.C.: Entrepreneurship: Past research and future challenges. Journal of Management 14(2), 139–161 (1988)
10. Miner, J.B., Smith, N.R., Bracker, J.S.: Role of entrepreneurial task motivation in the growth of technologically innovative firms. Journal of Applied Psychology 74(4), 554–560 (1989)
11. Ray, D.M.: The role of risk-taking in Singapore. Journal of Business Venturing 9(2), 157–177 (1994)
12. Sartori, R., Ceschi, A.: Uncertainty and its perception: experimental study of the numeric expression of uncertainty in two decisional contexts. Quality & Quantity 45(1), 187–198 (2011)
13. Sartori, R., Ceschi, A.: Assessment and development centers: judgment biases and risks of using idiographic and nomothetic approaches to collecting information on people to be evaluated and trained in organizations. Quality & Quantity 47, 3277–3288 (2013)
14. Sexton, D.L., Bowman, N.: The effects of pre-existing psychological characteristics on new venture initiations. Paper Presented at the Annual Meeting of the Academy of Management, Boston, MA (1984)
15. Stevenson, H.H., Gumpert, D.E.: The heart of entrepreneurship. Harvard Business Review 63(2), 85–94 (1985)

The Empirical Microstructure of Agent-Based Models: Recent Trends in the Interplay between ACE and Experimental Economics

Paola D'Orazio[1,2] and Marcello Silvestri[2]

[1] Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, "G. D'Annunzio" University, Viale Pindaro 42, 65127 Pescara, Italy
[2] Research Group for Experimental Microfoundations of Macroeconomics (GEMM), Pescara, Italy
paola.dorazio@unich.it, marcello.silvestri@yahoo.it

Abstract. In this paper we discuss recent trends on the interplay between Experimental Economics and Agent-based Computational Economics (ACE). Experimental Economics proved useful in providing insights on human subjects' decision-making as well as microeconomic data to estimate artificial agents. Agent-based Computations Economics allows for observing the aggregate outcome of artificial agents' interactions and for replicating experiments at a larger scale.

1 On the Explanatory Power of Artificial Economies

Starting from the mid-1990s, researchers have begun acknowledging the explanatory power of agents' *heterogeneity* and *interaction* for the understanding of the economy as a *complex adaptive system* [18]. This in turn led to a methodological shift from mathematical-statistical models to computational models [22,19] which rely on the algorithmic approach in order to model economic phenomena [23].

Artificial economies are indeed a crucial tool in displaying different ways to understand what goes on in decentralized economies [24] and allow for modeling a large range of economic behaviors, which in turn entails the so-called "too many degree of freedom", i.e., how to calibrate artificial agents with the "proper" behavior [26].

In this paper we highlight some ongoing tendencies in the recent literature (2005-2013) on the interplay between Experimental Economics (EE) and Agent-based Computational Economics (ACE). Our survey shows a gradual shift of the interest from the aggregate framework to the *micro empirical structure*, thus accounting for a "complete" *heterogeneity* in modeling agents' behavior.

Additionally, the survey outlined in this paper suggests that EE and ACE are engaging in a stronger - than that of early works surveyed in [9] - *complementary* relationship. This complementarity is not new in the interplay between EE/ACE. Agent-based models have indeed been used as a "test bed" for behavioral specifications observed in the experimental lab. Besides the "traditional"

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_11,

application of the EE/ACE methodology (see [9] for a comprehensive survey), this review suggests that recently, Experimental Economists are gradually acknowledging the potential benefits that experiments run with human subjects can gain from simulations in an artificial environment. This in turn implies a kind of "reverse" relationship, i.e., ACE/EE.

The remainder of the paper is organized as follow. After a brief discussion on the importance of the algorithmic approach in Economics, section 2 recalls the early steps made in the design of artificial agents and section 3 reviews some methodological techniques used so far to fit Agent-based models. In section 4 we discuss the importance of the microspecification of ABM together with some future research directions.

2 Individual and Social Level in the Algorithmic Approach

Most works in Computational Economics have generally made use of *single-based algorithms* thus focusing on the learning at the social (aggregate) level [12].

Our review shows that recently many scholars are engaged in a different type of research, i.e., the *multi-population approach*. On one hand, it adds complexity to the model; on the other hand, it becomes eventually more useful and reliable in order to draw policy recommendations [25].

2.1 Parallel Artificial and Human Agents in Genetic Algorithms

Brian Arthur was among the firsts in exploring the idea of calibrating an algorithm to reproduce human behavior [3].

He called the attention on the need to go beyond the assumption of *rationality* by suggesting some ways to model economic choices by means of an algorithm that can be *"tuned to choose actions in an iterated choice situation the way humans would* " [3], p. 354]. To calibrate the algorithm in a way that could be defined as a "good indication" of human behavior, he used the results of an experiment performed by Robillard at Harvard University in 1952-53. Simulations' results - and tests of fitness - were very striking in that they showed that Arthur's automaton was able to replicate behaviors observed in the experimental lab also in different choice problems than those for which it was calibrated.

Starting from Arthur's contribution, *Evolutionary Algorithms* (EA) (the *Genetic Algorithm*, in particular) have been extensively used in Economics and a closer look to the works published in the 1990s and early 2000s suggests that, in order to calibrate artificial agents and artificial economies, researchers relied mainly on *aggregate data*. As discussed earlier, this in turn implies that Agent-based models have been focusing more on the *social* rather than on the *individual* level.

3 On the Interplay between Experimental Economics and ACE: Theory and Methods

3.1 From Experiments to Agent-Based Simulations

Autonomous agents play a crucial role in the ABM literature to explore the *emergent complexity* at the micro level.

However, artificial agents are usually considered as "equally smart" [8,7]. ABMs which parallel human and artificial agents behavior have indeed been used to *mirror* human subjects macroeconomic experiments results, as shown in the Arifovic's seminal paper [2]. The *aggregate* variables (prices, exchange rates, *etc.*) are firstly generated in the experimental laboratory by observing human subjects' behavior. Then artificial agents are calibrated in order to replicate experimental results

The agents' *calibration process* concerns the empirical "roots" of artificial agents. Arthur called the attention on some methodological rules-of-thumb on how to deal with the calibration. A first approach is the statistical parameterization of artificial agents by using available empirical datasets. A second approach consists in building human-like qualitative behavior and put it into the algorithm.

By adopting this perspective, Hommes explores the behavioral space of the heterogeneous expectations hypothesis [15,13]. Hommes runs lab experiments with human subjects and find support for *heterogeneity* among expectations. Additionally, he shows that a simple *heuristics switching model* fits different learning-to-forecast experiments (LtFEs). Hommes and Lux retook these results and propose a genetic algorithm (GA) as a way to model individual expectations and explain aggregate market phenomena [14]. In a more recent work, Anufriev et al. give better microfoundations to their model by using GA-based individual learning which eventually explains the *individual*, not just the aggregate result of the LtFEs [1].

In our opinion, this process is among the most important innovations in that it improves the building of the empirical microstructure in ABMs.

3.2 From Agent-Based Simulations to Experiments

The rapidly expanding research strand based on the interaction EE/ACE suggests that they are "natural allies" in that they help each other in coping with their *external validity* shorthcomings [9]. They indeed complement each other: EE helps ACE in dealing with its "degree of freedom" problem and ACE helps EE in controlling and providing benchmarks for experimental subjects' behavior. A closer look at the literature shows that several studies have also been conducted in the reverse direction, i.e., human subject experiments conducted in light of ACE results. Casari's article is the first exploring this issue. According to Casari, simulations with GA allow to: (a) make comparisons with experimental data and (b) make predictions about the effects of different experimental designs. The novelty of Casari's results lies mainly in the latter. He found that agents with identical goals and identical, although limited, levels of rationality behave in *different* manners, i.e., GA generates individually different patterns. Changes in the experimental design, as e.g. the restriction of the agents' strategy space, are then explored; the resulting predictions are also supported by experimental results. This particular use of GA simulations has been extended by Casari himself [4] and further studied by [6] and [5].

4 Conclusions and Future Research Directions

As Farmer an Foley pointed out [10], Agent-based models are able to fulfill both Keynes' purposes of building a financial complex economy [16] and the Lucas' *critique* [20] according to which macroeconomic models should account for "real" human adaptation and learning processes.

Our review suggests that the interplay between Experimental Economics and Agent-based models[21,17], as well as the "reverse" relationship (ACE-EE). proved to be fruitful. In the last 5-6 years in the Economics research there has been indeed a growing interest in the development of "new" microfundations for Agent-based models. The interaction between Experimental Economics and the use of modern heuristics [11] as well as algorithms [27] can help researchers to improve the models and the aggregate outcome of their simulations. Furthermore, researchers can set up experiments to closely observe agents behaviors and use experimental data to fill the gap of the availability of individual microeconomic data (namely, the lack of micro information that captures agents' dynamic decision-making).

According to the research methodology based on the integration of experimental insights and ACE techniques, artificial agents' behavioral rules can be thus identified and implemented by artificial agents. Although the use of experimental data is sometimes hinted at as a possibility to bind the very large parameter space of ACE models, to our knowledge there has not yet been developed any experimentally microfounded agent-based macroeconomic model. In our view, this methodology could help in the building of more real-like artificial macroeconomic models in that it provides a better empirical microstructure to the Agent-based model by encoding human decisions in the artificial agents (i.e., *avatars*).

References

1. Anufriev, M., Hommes, C., Makarewicz, T.: Learning to forecast with genetic algorithm. Tech. rep. (February 2013)
2. Arifovic, J.: The behavior of the exchange rate in the genetic algorithm and experimental economies. Journal of Political Economy 104(3), 510–541 (1996)
3. Arthur, W.B.: Designing Economic Agents that Act like Human Agents: A Behavioral Approach to Bounded Rationality. American Economic Review Papers and Proceedings 81(2), 353–359 (1991)
4. Casari, M.: Markets in equilibrium with firms out of equilibrium: A simulation study. Journal of Economic Behavior & Organization 65(2), 261–276 (2008)

5. Chen, S.H., Yu, T.: Agents learned, but do we? knowledge discovery using the agent-based double auction markets. Frontiers of Electrical and Electronic Engineering in China, 159–170 (2011)
6. Chen, S.H., Zeng, R.J., Yu, T.: Co-evolving trading strategies to analyze bounded rationality in double auction markets. In: Genetic Programming, Theory and Practice VI, pp. 195–213. Springer (2008)
7. Chen, S.H.: Varieties of agents in agent-based computational economics: A historical and an interdisciplinary perspective. Journal of Economic Dynamics and Control 36(1), 1–25 (2012)
8. Chen, S.H., Wang, S.G.: Emergent complexity in agent-based computational economics. Journal of Economic Surveys 25(3), 527–546 (2011)
9. Duffy, J.: Agent-based models and human subject experiments. In: Tesfatsion, L., Judd, K.L. (eds.) Handbook of Computational Economics, vol. 2, ch. 19, pp. 949–1011. Elsevier (2006)
10. Farmer, J.D., Foley, D.: The economy needs agent-based modeling. Nature 460, 685–686 (2009)
11. Gigerenzer, G., Hertwig, R., Pachur, T. (eds.): Heuristics: The Foundations of Adaptive Behavior. Oxford University Press, New York (2011)
12. Holland, J.H., Miller, J.H.: Artificial adaptive agents in economic theory. The American Economic Review 81, 365–370 (1991)
13. Hommes, C.: The heterogeneous expectations hypothesis: Some evidence from the lab. Journal of Economic Dynamics and Control 35(1), 1–24 (2011)
14. Hommes, C., Lux, T.: Individual expectations and aggregate behavior in learning-to-forecast experiments. Macroeconomic Dynamics 17, 373–401 (2013)
15. Hommes, C.: Bounded rationality and learning in complex markets. CeNDEF Working Papers 07-01, Universiteit van Amsterdam, Center for Nonlinear Dynamics in Economics and Finance (2007)
16. Keynes, J.M.: The general theory of employment. The Quarterly Journal of Economics 51(2), 209–223 (1937)
17. Kimbrough, E.O.: Heuristic learning and the discovery of specialization and exchange. Journal of Economic Dynamics and Control 35(4), 491–511 (2011)
18. Kirman, A.P.: Whom or What Does The Representative Individual Represent. Journal of Economic Perspective 6, 117–136 (1992)
19. LeBaron, B., Tesfatsion, L.: Modeling Macroeconomies as Open-Ended Dynamic Systems of Interacting Agents. American Economic Review 98(2), 246–250 (2008)
20. Lucas, R.J.: Econometric policy evaluation: A critique. Carnegie-Rochester Conference Series on Public Policy 1(1), 19–46 (1976)
21. Rouchier, J.: Data Gathering to Build and Validate Small Scale Social Models for Simulation. In: Rennard, J.P. (ed.) Handbook of Research on Nature Inspired Computing for Economics and Management, vol. 2, pp. 198–210. Idea Group (2007)
22. Tesfatsion, L., Judd, K.: Handbook of Computational Economics. Elsevier, Amsterdam (2006)
23. Velupillai, K.V.: Towards an algorithmic revolution in economic theory. Journal of Economic Surveys 25(3), 401–430 (2011)
24. Vriend, N.: Artificial intelligence and economic thoery. In: Many-agent Simulation and Artificial Life (Frontiers in Artificial Intelligence & Applications), pp. 31–47 (1994)

25. Vriend, N.J.: An illustration of the essential difference between individual and social learning, and its consequences for computational analyses. Journal of Economic Dynamics and Control 24(1), 1–19 (2000)
26. Windrum, P., Fagiolo, G., Moneta, A.: Empirical validation of agent-based models: Alternatives and prospects. Journal of Artificial Societies and Social Simulation 10(2) (2007)
27. Zhao, X., Lin, W., Yu, C., Chen, J., Wang, S.: A new hybrid differential evolution with simulated annealing and self-adaptive immune operation. Computers & Mathematics with Applications 66(10), 1948–1960 (2013); iCNC-FSKD 2012

Households Debt Behavior and Financial Instability: Towards an Agent-Based Model with Experimentally Estimated Behavioral Rules

Paola D'Orazio[1,2]

[1] Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, "G. D'Annunzio" University, Viale Pindaro 42, 65127 Pescara, Italy
[2] Research Group for Experimental Microfoundations of Macroeconomics (GEMM), Pescara, Italy
paola.dorazio@unich.it

Abstract. The present paper suggests the development of an experimentally microfounded Agent-based model in order to cope with the *complexity* and *instability* of the macroeconomic environment. The focus of the paper is on the *microspecification* of the ABM. For the micro level, I suggest to design an experiment in order to gain insights into households' behaviors. For the macro level, I plan to build an ABM where agents are estimated, rather than calibrated, by using data collected in the experimental laboratory.

1 The "Micro-to-Macro Mapping": Microfounding Agent-Based Models

The interest in this research originates from the increasingly widespread opinion among scholars about the helplessness of traditional economic theories (and the models they give rise to, namely, the DSGE models) in case of serious crises which emerge in real world economies. Many researchers thus claim for the need of new *tools* to cope with the *complexity* of socio-economic systems. The new tools must be able to allow for the building of *microfoundations* of Macroeconomics by considering also the *feedback* effect the macro system has on individual agents. This in turn implies going beyond analytically solvable pure theoretical macro models which proved to be of little help for policy guidance [1] because of a) their far from reality assumptions and (b) any consideration of interaction among heterogeneous agents [2].

In order to consider the macro economy as a complex *endogenously* organized system, researchers should be able to study both the micro level, i.e., *agents*, and the macro level, i.e., the *environment*.

Concerning the former, the "behavioral revolution" [3,4] is of crucial help, because it has increased the explanatory power of Economics by providing it with more realistic psychological foundations and "practical" ways in which behaviors can be incorporated into the models [5].

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_12,

Regarding the latter, models have to deal with the complexity and instability of the macro *environment*, which are indeed a central feature of the choices that economic agents face. It is thus crucial to consider the economy as a *Complex Adaptive System* [6], i.e., composed of many heterogeneous interacting agents. Agent-based models (ABM) allow indeed to study the "micro-to-macro mapping" [7] as well as social complexity. Furthermore, in an ABM observed dynamics are open-ended (not closed form as in DSGE models) and they allow for an *ergodic state* of the system, i.e., an equilibrium, which is an emergent and optional outcome [8]. While DSGE models are based on the centralized information processing structure, ABM takes a bottom-up approach that starts modelling realistic microfoundations and ends up analyzing the resulting aggregate behaviour. The dynamics of aggregate variables are the result of complex, continuously and endogenously changing micro-structure. This yields substantial advantages in modelling macroeconomic policies [9].

However, there are some methodological problems related to ABM. *Caeteris paribus*, the empirical model validation[1] and robustness checks of results. Indeed, the large flexibility of the setup (starting values) of agent-based models and the number of selected parameters give many "degrees of freedom" to the researcher. This in turn poses serious challenges to the use of ACE models for the *evaluation* and design of *economic policy* measures. This problem could be (partially) overcome by using the experimental method to build (i.e., estimate) agents that populate the artificial environment.

1.1 Accounting for the Consistency between Stocks and Flows

The dynamics of stocks in any economic model have a relevant explanatory power and can help understanding macroeconomic phenomena; thus adding a stock-flow consistency (SFC) requirement to the model helps to overcome the *fallacy of composition* critique.[2]

The SFC method proves to be appropriate to deal with financialization issues and to model the interaction between financial crises and real crises.

However, SFC models are *aggregate* models so that the researcher is guided more by the identities of accounting than by the individual economic behavior obtained using the deductive approach. Recently, SFC models are drawing new attention by researchers in that they are developing new ways to use the SFC framework as a development of the existing general aggregative models [12,13,14].

According to the "integrated" methodology I will use in my research, it is possible to build *behaviorally* richer SFC macroeconomic models by rooting them in individual financial behaviors. The combination of SFC and Agent-based modeling tools will yield more complete macro models. On one hand, the ABM can

[1] For a broader discussion on this issue see [10,11].

[2] It concerns the presumption that what is true of each single part of a whole is necessarily true of the whole as well. Basically, it questions the *aggregation process* used by "standard" macroeconomic models.

display consistency between stocks and flows; the result is a framework that ensures the compatibility of real and financial variables. On the other hand, the ABM provide explicit micro-foundations to macroeconomic relations.

1.2 Fitting Agent-Based Models with Experimental Data

Agent-based models have been widely used in Economics to study findings from human subject experiments. Brian Arthur was among the firsts in exploring the idea of calibrating an algorithm to reproduce human behavior [15]. He called the attention on the need to go beyond the assumption of *rationality*, suggesting some ways to model economic choices. He did not want to design a learning algorithm or automaton that maximizes some criterion. Rather, he aimed at designing an algorithm that can be *"tuned to choose actions in an iterated choice situation the way humans would* " ([16] p.354). To calibrate the algorithm in a way that could be defined as a "good indication" of human behavior, he used the results of an experiment performed in 1952-53 by Robillard at Harvard University. His results - and tests of fitness - showed that the automaton was able to replicate those behaviors also in different choice problems, than those for which it was calibrated.

In his review, [17] calls the attention on how the interaction between ABM and Experimental Economics has been recently used in Economics, concluding that they are "natural allies".[3]

Many simulations' results showed that GAs successfully replicate experimental behaviors in different environments (see [19,20,21] for a comprehensive survey). However, in the ABM literature artificial agents have usually been considered as "equally smart" [22] and they are built by relying on available theories on individual decision-making (see [14,23] among others). The exploration and *induction* of agents' behaviors by means of the experimental method is thus a more recent strand of research.

However, in order to account for a "complete" agents' heterogeneity and give the model better microfoundations, researchers need tools that allow to consider the *whole set of behavioral rules* found in experimental data. In the last decades, many scholars have been engaged in this line of research so that heterogeneous, interacting agents in ABM are designed in many different ways (for a comprehensive survey see [24]).

At the present stage of my research I am considering retaking Brian Arthur's main arguments and results and go *beyond the calibration process* by estimating the behavioral rules for artificial agents. Indeed, the main point of the present research project is that by means of the experimental method it is possible to

[3] He points out three main areas in which this interplay has been used. The first one is the so-called "*zero-intelligent*" agent approach based on the [18] seminal paper. It consists of models with very low rationality constraints which made evident that self-interested and rational behavior is observed mainly in highly structured and constrained markets. A second line of research focuses on *reinforcerment* and belief-based models of agents' behaviors. Finally, there are *Evolutionary Algorithms* where individual learning is more complicated.

gain insights into human behavior and collect information on heterogeneity of the population. Experimental data can thus be used to *estimate*, rather than "merely" calibrate, the behavioral rules that guide artificial agents' actions.[4]

In order to fit the ABM using experimental data, I am considering the implementation of a Genetic Algorithm [20,19], Classifier Systems [26,27] or a combination of both. In the evaluation of which evolutionary method and technique will best fit for my project, I am also concerned about the economic soundness of this tools. I am indeed considering both the concerns and the warning of [28,29] about the economic interpretation of evolutionary algorithms.

2 Research Design and Methodology

The methodology is based on several steps. Households' financial behaviors (in particular, their "debt love"/debt aversion) will be put under the "magnifying glass" by means of the *experimental method* in order to detect flaws in traditional theories of intertemporal consumption/saving decisions. The experimental method will allow also to explore actual decision-making processes and heterogeneity across agents. Data collected from experiments will be used to analyze the aggregate implications of households' behaviors and to eventually *estimate* the behaviors of artificial agents that will interact in the artificial environment.[5] By means of Object-oriented programming (OOP), different classes for each agent will be built. Each agent will be "endowed" with a balance sheet in order to keep track of their stock-flow transactions. Once the ABM will be set up, policy experiments will be performed in order to evaluate different macroeconomic policies.

2.1 Experimental Insights on Intertemporal Decision-Making

Several studies have been highlighting that individual preferences - and macro feedbacks - are observable and controllable in the laboratory such that economic modeling becomes richer and more realistic [31,32].

The designed experiment will address two issues in intertemporal consumption models. On one side, the experimental approach will allow to investigate the ability of subjects to solve the task of *intertemporal optimization* and the extent to which subjects behave according to standard optimization rules [33]. In this way, the experiment will take the theoretical predictions seriously and *test* if consumers are able to carry out a dynamic intertemporal (utility) optimization problem. On the other side, the experimental method will be used to study the *pervasiveness* of debt in the liabilities side of households' portfolios. The model will thus account also for the role financial innovations have been gaining in capitalistic economies in the last decades.

[4] [25] offer an interesting perspective on model calibration and estimation methodologies. It also discusses the related empirical "hidden dangers".

[5] As showed in [30], the use of experimental data allows to go beyond the standard parameters' *calibration* procedure by performing the *estimation* of artificial agents behavioral rules.

The experimental design is grounded on methods and results of the early experiments performed on intertemporal consumption. The first attempt of testing how closely the predictions of the optimality theory fit the actual behavior of subjects in an experimental setting is the paper by [34].

The Experimental Design: The Model. The starting point is the building of a benchmark intertemporal consumption model which is used to compute the optimal (theoretical) solution that will be compared to our experimental data in order to assess if there is a deviation from the optimal behavior. The major *innovation* introduced concerns the relaxation of the standard budget constraint by allowing borrowing, hence debt.

The solution for the intertemporal consumption problem is found by backward induction, following the method for microeconomic dynamic stochastic optimization problems. Among others, I considered the methods developed by [35] and [36]. Given that an explicit solution to the problem does not exist (it is implicitly characterised by an Euler equation), *numerical methods* are needed in order to find an optimal policy function (for saving and borrowing tasks) to be compared to subjects' decisions in the lab. This methodological choice is of particular importance because in the experiment we should be able to control for all possible confounding factors and focus on few variables of interest.

For the standard intertemporal consumption model set up, I compute the policy function for a consumption/saving task which in turn will be used as a benchmark for experimental data.

In order to have a simple and tractable model I do not consider the *discount factor*; a separate experiment is necessary in order to elicit - thus estimate - the discount parameter of subjects in the lab. I indeed decided to leave this investigation for another experiment [37,38,39].

The consumption decision takes the following functional form:

$$\max E \sum_{t=1}^{T} u(c_t) \tag{1}$$

subject to

$$w_{i,t} - c_{i,t} - b_{i,t} = \Delta W_{i,t} \tag{2}$$

$$W_{i,t} = W_{i,t-1} + \Delta W_{i,t} \tag{3}$$

$$W_{i,t} \geq -B^s_{i,t} \tag{4}$$

where w is wage, W wealth, B^s is the bank credit supply (exogenously set by the researcher), $b_{i,t}$ = represents the payback to the bank (if the household holds debt). The interest rate is set equal to zero.

$$b_{i,t} = \begin{cases} 0 & \text{if } W_t \geq 0 \\ -\beta W_t & \text{if } W_t < 0 \end{cases} \tag{5}$$

where β is the cost of debt.

The *intertemporal settings* that will be analyzed will be basically two. The *perfect certainty*, i.e., *deterministic*, model which would be the "candidate" framework to test the predictions of the model built on the rational expectations assumptions, and a *stochastic* version in which there is *uncertainty* about the future income stream. The latter will be useful to assess which kind of expectations (adaptive, etc) arise among experimental subjects.

Th experimental subject pool will be composed by students and workers - recruited throught the ORSEE software [40] - in an attempt to address the usual criticism about the *external validity* of experimental data and results.

3 Expected Outcome

The research presented in this paper represents a doctoral research project which is still in progress and it will complement a broader project of a macroeconomic ABM, in which the firm sector and the banking sector will be also considered. It benefits from the integration of different methodologies and modeling techniques: Experimental Economics method and Agent-based Computational Economics, enriched by the "informed intuition" of stock-flow consistency, will allow to develop more real-like macro models and more practical tools to guide policymakers. The ABM will be then used to perform some policy experiments, i.e., analyses on the effects and effectiveness of different fiscal and/or monetary policies.

The issues raised above make it clear that the *microspecification* of the ABM is crucial. Individual behaviors "observed" in the experimental laboratory become the solid foundations of the model so that the value added of having *experimental microfoundations* is that they allow for a more complete and empirical *heterogeneity* of artificial agents already at the initialization of the model, while heterogeneity is usually treated as an *emergent property* of the ABM. Moreover, information collected from experimental micro-systems differs from that obtained from empirical data [31] in that the former offers richer insights into the individual and collective dynamics of a model. Experimental data could thus represent a solution to the problem of availability of micro data which threatens the research on the microfoundations of macroeconomic models [33].

References

1. Velupillai, K.V.: Towards an algorithmic revolution in economic theory. Journal of Economic Surveys 25(3), 401–430 (2011)
2. Kirman, A.P.: Whom or What Does The Representative Individual Represent. Journal of Economic Perspective 6, 117–136 (1992)
3. Akerlof, G.A.: Behavioral macroeconomics and macroeconomic behavior. American Economic Review 92(3), 411–433 (2002)
4. Kahneman, D.: Maps of bounded rationality: Psychology for behavioral economics. American Economic Review 93(5), 1449–1475 (2003)

5. Kao, S.Y.F., Velupillai, K.V.: Behavioural economics: Classical and modern. ASSRU Discussion Papers 1126, ASSRU - Algorithmic Social Science Research Unit (2011)
6. Tesfatsion, L.: Agent-based computational economics: Modeling economies as complex adaptive systems. Information Sciences 149, 263–269 (2003)
7. Epstein, J.M.: Agent-based computational models and generative social science. In: Generative Social Science Studies in Agent-Based Computational Modeling. Introductory Chapters. Princeton University Press (2007)
8. Delli Gatti, D., Gaffeo, E., Gallegati, M., Giulioni, G., Palestrini, A.: Emergent Macroeconomics. Springer, Berlin (2008)
9. LeBaron, B., Tesfatsion, L.: Modeling macroeconomies as open-ended dynamic systems of interacting agents. American Economic Review 98(2), 246–250 (2008)
10. Fagiolo, G., Birchenhall, C., Windrum, P.: Empirical Validation in Agent-based Models: Introduction to the Special Issue. Computational Economics 30(3), 189–194 (2007)
11. Fagiolo, G., Moneta, A., Windrum, P.: A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. Computational Economics 30(3), 195–226 (2007)
12. Kinsella, S., Greiff, M., Nell, E.J.: Income distribution in a stock-flow consistent model with education and technological change. Eastern Economic Journal 37(1), 134–149 (2011)
13. Seppecher, P.: Flexibility of wages and macroeconomic instability in an agent-based computational model with endogenous money. Macroeconomic Dynamics 16(S2), 284–297 (2012)
14. Raberto, M., Teglio, A., Cincotti, S.: Debt, deleveraging and business cycles: An agent-based perspective. Economics - The Open-Access, Open-Assessment E-Journal 6(27), 1–49 (2012)
15. Arthur, W.B.: On designing economic agents that behave like human agents. Journal of Ecolutionary Economics 3, 1–22 (1993)
16. Arthur, W.B.: Designing Economic Agents that Act like Human Agents: A Behavioral Approach to Bounded Rationality. American Economic Review Papers and Proceedings 81(2), 353–359 (1991)
17. Duffy, J.: Agent-based models and human subject experiments. In: Tesfatsion, L., Judd, K.L. (eds.) Handbook of Computational Economics. Handbook of Computational Economics, vol. 2, pp. 949–1011. Elsevier (2006)
18. Gode, D., Sunder, S.: Allocative efficiency of markets with zero intelligence (z1) traders: Market as a partial substitute for individual rationality. Gsia Working Papers, Carnegie Mellon University, Tepper School of Business (1991)
19. Arifovic, J.: The behavior of the exchange rate in the genetic algorithm and experimental economies. Journal of Political Economy 104(3), 510–541 (1996)
20. Arifovic, J.: Evolutionary algorithms in macroeconomic models. Macroeconomic Dynamics 4, 373–414 (2000)
21. Dawid, H.: Adaptive Learning by Genetic Algorithms: Analytical Results and Applications to Economic Models. Springer Verlag New York, Inc. (1996)
22. Chen, S.H., Yu, T.: Agents learned, but do we? knowledge discovery using the agent-based double auction markets. Frontiers of Electrical and Electronic Engineering in China, 159–170 (2011)
23. Dosi, G., Fagiolo, G., Roventini, A.: Schumpeter meeting keynes: A policy-friendly model of endogenous growth and business cycles. Journal of Economic Dynamics and Control 34(9), 1748–1767 (2010)

24. Chen, S.H.: Varieties of agents in agent-based computational economics: A historical and an interdisciplinary perspective. Journal of Economic Dynamics and Control 36(1), 1–25 (2012)
25. Hansen, L.P., Heckman, J.J.: The empirical foundations of calibration. Journal of Economic Perspectives 10(1), 87–104 (1996)
26. Booker, L.B., Goldberg, D.E., Holland, J.H.: Classifier systems and genetic algorithms. Artificial Intelligence 40(1-3), 235–282 (1989)
27. Birchenhall, C.R.: Genetic algorithms, classifier systems and genetic programming and their use in the models of adaptive behaviour and learning. The Economic Journal 105(430), 788–795 (1995)
28. Waltman, L., Eck, N., Dekker, R., Kaymak, U.: Economic modeling using evolutionary algorithms: the effect of a binary encoding of strategies. Journal of Evolutionary Economics 21(5), 737–756 (2011)
29. Dawid, H., Dermietzel, J.: How robust is the equal split norm? responsive strategies, selection mechanisms and the need for economic interpretation of simulation parameters. Computational Economics 28(4) (2006)
30. Giulioni, G., Bucciarelli, E., Silvestri, M., D'Orazio, P.: Agent-based computational economics and experimental economics: a bridge to progress macroeconomics (2013) (manuscript submitted)
31. Smith, V.L.: Microeconomic Systems as an Experimental Science. American Economic Review 72(5), 923–955 (1982)
32. Smith, V.L.: Method in Experiment: Rhetoric and Reality. Experimental Economics 5, 91–110 (2002)
33. Carroll, C.D.: Representing consumption and saving without a representative consumer. In: Measuring Economic Sustainability and Progress. NBER Chapters, National Bureau of Economic Research, Inc. (July 2012)
34. Hey, J.D., Dardanoni, V.: Optimal consumption under uncertainty: An experimental investigation. Economic Journal 98(390), 105–116 (1987)
35. Carroll, C.: Lecture notes on solution methods for microeconomicdynamic stochastic optimization problems. Lecture Notes - John Hopkins University (2012)
36. Stachurski, J.: Economic Dynamics. Theory and Computation. MIT Press (2009)
37. Cubitt, R., Read, D.: Can intertemporal choice experiments elicit time preferences for consumption? Experimental Economics 10, 369–389 (2007)
38. Coller, M., Williams, M.: Eliciting individual discount rates. Experimental Economics 2(2), 107–127 (1999)
39. Harrison, G.W., Harstad, R.M., Rutstrm, E.E.: Experimental methods and elicitation of values. Discussion paper serie b. University of Bonn, Germany (1995)
40. Greiner, B.: An online recruitment system for economic experiments. Forschung und Wissenschaftliches Rechnen 2003. GWDG Bericht (2004)

Firm Size Distribution in Oblivious Equilibrium Model with Quality Ladder

Tetsushi Murao*

Department of Economic Engineering,
Faculty of Economics, Kyushu University,
Fukuoka 812-8581, Japan
murao@econ.kyushu-u.ac.jp

Abstract. In this article, we investigate a simulated firm size distribution in the model of Weintraub, Benkard, and Van Roy (Operations Research, 2010) which is a oblivious equilibrium (OE) model with a canonical quality ladder setting of Pakes and McGuire (Rand Journal of Economics, 1994). In previous research, validity of applying an OE model in a specific context have been assessed in two aspects: (i) how precisely the OE could replicate the MPE outcomes (light-tail condition); and (ii) whether restricting agents' information (so does strategy) could be reasonable. In contrast, we propose a new criterion for the validity of OE models: whether equilibria could replicate power law of firm size distribution that is typically observed in real world data. We find that, as the quality depreciation probability or the investment cost becomes higher, the distribution comes closes to power law. On the other hand, the entry cost have virtually no impacts on the curvature of log-log plots.

Keywords: Firm size distribution, Oblivious equilibrium, Industry dynamics.

1 Introduction

Research on dynamic industry models with heterogeneous firms is now blossoming and these models are served as a benchmark in the field of Macroeconomics and Industrial Organization ([10]).

As of now, there are three strands of these models ([16]); that is, stationary equilibrium (SE) models à la [9], Markov perfect equilibrium (MPE) models à la [6], and oblivious equilibrium (OE) models à la [15].

These models are primary classified according to whether the number of firms in the industry is finite or infinite: number of firms are infinite in SE models while finite in MPE and OE models. Owing to this property, the industry state in SE models is constant overtime since the law of large numbers holds. Thus, firms in SE models do not need to keep track of the evolution of industry state, and that is why SE models are computationally tractable.

Contrary to SE models, the number of firms in MPE models is finite: firms have strictly positive market share as in the real world industry. This implies that the profit

* We appreciate helpful comments from three anonymous referees. This research is partially supported by Kyushu University P&P Grant (No. 25334).

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_13,

of any firm is affected by the behavior of all other firms in the industry as well as own individual state. Since firms care about states of firms in all over the industry, an increase in the number of firms in the MPE models augments exponentially the dimension of industry state that firms have to keep track of.[1] As of now, applied research using MPE models is still limited due to this curse of dimensionality.

In response to the above intractability of MPE models, [15] proposes an equilibrium notion called *oblivious equilibrium* (OE). As with MPE models, the number of firms in OE models is finite. However, in contrast to MPE models, firms in OE models are assumed to have only incomplete knowledge about current industry state: they perceive the industry state as the constant long-run average industry state. Despite the positive market share of firms, this assumption inhibits the huge expansion of state spaces when number of firms in the industry is increased, due to exactly the same reason as in the case of SE models.

One virtue of OE models is that OE can provide a mean field approximation of Markov perfect equilibrium (MPE) of dynamic stochastic game with many players, when firm states obeys a "light-tail" condition ([14], [15]). In this regard, validity of applying an OE model in a specific context have been assessed in two aspects: (i) how precisely OE could replicate MPE outcomes (light-tail condition)[2] and (ii) whether restricting agents' information (so does strategy) could be reasonable.[3][4] However, in situations of relatively large number of firms, restricting agents' information on industry state can be rather more natural than full information assumption, regardless of the satisfaction of light tail condition: in these situations, OE models can be an appealing behavioral model in its own right ([15]).[5]

Based on the above insight, we propose an alternative criterion: whether equilibria could replicate the power law distribution that is typically observed in real world firm size data. Number of studies find that firm size (sales, employment) obeys the power law distribution, a linear relationship between firm size and its rank in a log-log plot.[6] In this research, we investigate a simulated firm size distribution in the model of [17] which is a oblivious equilibrium (OE) model with a canonical quality ladder setting of [13]. We find that, as the quality depreciation probability or the investment cost becomes higher, the distribution comes closes to power law; on the other hand, the entry cost has no effect on the curvature of log-log plots.

[1] For example, [15] pointed out that: "[M]ost industries contain more than 20 firms, but it would require more than 20 *million* gigabytes of computer memory to store the policy function for an industry with just 20 firms and 40 firm states."

[2] [15] proposes a formula for the error bound to assess how well OE could approximate MPE, which can be evaluated even without computing MPE itself.

[3] In a model of OE with dominant firms, [2] found the equilibrium firm behavior which is quite different from that in the original MPE model. They warned the possibility that the simplifying assumption about firms' industry knowledge causes unexpected results.

[4] [2] argues that, in this regard, OE can be viewed as a special case of *restricted experience based equilibrium* proposed by [7].

[5] See [16] and [1] for the relationship among SE, MPE, and OE models.

[6] See an exellent review of [8] and articles cited herein. [11] also introduces some empirical evidences.

Contribution of this research is summarized as follows. As of now, previous theoretical research investigating power law of firm size distribution is only limited to SE models.[7] Our paper is the first attempt to investigate this issue with non-atomic dynamic environment.[8]

2 Model

In the following analysis, we use the model of [15] which is a variant of the Pakes-McGuire model proposed by [13].

2.1 Timing

Time is discrete and infinite. In any period of time, there are some active incumbent firms in the industry. Each firm produces a differentiated product and also conducts process R&D to increase its quality. Firms are described by this quality of product which is the only individual state variable: we denote by positive integer $x_{it} \in [1, \bar{x}]$ the quality of a differentiated good produced by firm i in period t. The industry state is defined as a vector $s_t = (s_{1t}, \ldots, s_{\bar{x}t})$, where $s_{kt} \in \mathbb{N}$ is the number of incumbent firms at quality level k in period t.

In each period of time, events occur in the following order. (1) In the beginning of the period, an incumbent firm privately observes the realization of random sell-off value ϕ_{it}. If this value of exit exceeds that of continuation, the firm conducts no R&D investment and exits at the end of the period; otherwise the firm conducts R&D investment paying unit cost d; (2) Entry decision is conducted and the number of entrants is determined;[9] (3) Incumbent firms compete in the good market; (4) Exiting firms exit; (5) Investment outcomes are realized and new entrants enter.

2.2 Profit and Investment

There are M consumers in the market. Consumers face static discrete choice problem to purchase a product. Suppose that there are n firms in the market at time t and there is one outside good which gives zero utility. Each consumer either buy one product from firm $i \in \{1, \cdots, n\}$ or buy an outside good. In period t, consumer j receives utility u_{ijt} from good produced by firm i:

$$u_{ijt} = \theta_1 \ln\left(\frac{x_{it}}{\psi}\right) + \theta_2 \ln(Y_{jt} - p_{it}) + \nu_{ijt}, \quad i = 1, \cdots n; j = 1, \cdots, M,$$

where $Y_{jt} \equiv Y, \forall j, t$ is the income of the consumer; p_{it} is the price charged by the producer i; ν_{ijt} is the taste shock that is identically and independently distributed across

[7] See review articles of [8], [11], and studies cited herein.

[8] A companion paper ([12]) studies the firm size distribution in the MPE model of [13], by utilizing a continuous time formulation proposed by [4].

[9] In [15], the number of entrants λ is modeled as a Poisson arrival rate which can be pinned down by the zero profit condition.

products and consumers with the Type I extreme-value distribution; ψ, θ_1, and θ_2 is a parameter with positive value. Under these settings, the expected market share of each firm is given by:

$$\sigma(x_{it}, \mathbf{s_{-it}}, \mathbf{p}_t) = \frac{N(x_{it}, p_{it})}{1 + \sum_{j \in S_t} N(x_{jt}, p_{jt})},$$

where $\mathbf{p}_t = (p_{1t}, \cdots, p_{nt})$, $N(x_{it}, p_{it}) = \theta_1 \ln(x_{it}/\psi) + \theta_2 \ln(Y - p_{it})$, $\mathbf{s}_{-it}$ is a vector of states for i's competitors, that is, $(s_{1t}, \cdots, s_{i-1t}, s_{i+1t}, \cdots, s_{nt})$, and S_t is a set of indices of incumbent firms in time t.

[3] shows that there is the unique Nash equilibrium prices in above setting which we denote p^*_{it}. Profits are then computed as follows:

$$\pi(x_{it}, s_{-it}) = M\sigma(x_{it}, \mathbf{s}_{-it}, \mathbf{p}^*_t)(p^*_{it} - c).$$

Quality of products increases if R&D investment by firms is successful and decreases by a hit of idiosyncratic quality depreciation shock. Probability of successful innovation conditional on investment level ι is given by $a\iota/(1 + a\iota)$ while the probability of depreciation is given by δ, where $a \in \mathbb{R}^+$ and $\delta \in \mathbb{R}^+$ is a exogenous parameter. By combining innovation and depreciation processes, transitional probabilities are given by:

$$P[x_{it+1} = y | x_{it} = x, \iota] = \begin{cases} \frac{(1-\delta)a\iota}{1+a\iota} & \text{, if } y = x + 1 \\ \frac{(1-\delta)+\delta a\iota}{1+a\iota} & \text{, if } y = x \\ \frac{\delta}{1+a\iota} & \text{, if } y = x - 1. \end{cases}$$

2.3 Equilibrium

[15] considers the firms that only take care of its own state and long-run average industry state. Here, we introduce an equilibrium notion for the economy with such firms, which is called oblivious equilibrium proposed by [15].

Firstly, we present the Bellman equation for incumbent firms' problem. Prior to observing a realization of sell-off value ϕ, the exit decision rule can be expressed as the exit probability $\xi \in [0, 1]$ (see [5]). Then, Bellman equation for an incumbent obeying an OE strategy is expressed as follows:

$$V(x, \tilde{s}_{\mu,\lambda}) = \pi(x, \tilde{s}_{\mu,\lambda}) + \sup_{\xi \in [0,1],\ \iota \geq 0} \left\{ \begin{array}{c} \xi E[\phi | \phi \geq F^{-1}(1 - \xi)] \\ +(1 - \xi)\left[-d\iota + \beta \sum_{x'} V(x', \tilde{s}_{\mu,\lambda}) P(x' | x, \iota)\right] \end{array} \right\},$$

where β is the discount factor and $F(\cdot)$ is a cumulative distribution function for ϕ.

Next, we turn to the formal definition of the equilibrium.

Definition 1 (Oblivious Equilibrium). *Oblivious equilibrium consists of an oblivious strategy $\mu = (\iota, \xi) \in \tilde{\mathcal{M}}$ and an oblivious entry rate $\lambda \in \tilde{\Lambda}$ that satisfies following conditions, where $\tilde{\mathcal{M}}$ and $\tilde{\Lambda}$ is a set of oblivious strategies and oblivious entry rates, respectively:*

1. Firm strategies are a fixed point of the best response function $\mu' = \zeta(\mu)$, that is,

$$\sup_{\mu' \in \tilde{\mathcal{M}}} \tilde{V}(x | \mu', \mu, \lambda) = \tilde{V}(x | \mu, \lambda), \quad \forall x \in \mathbb{N},$$

where $\tilde{V}(x|\mu', \mu, \lambda)$ is defined as the expected net present value for a firm at state x given that the firm's strategy is $\mu \in \tilde{\mathcal{M}}$, its competitors' strategy is $\mu' \in \tilde{\mathcal{M}}$, and the entry rate is $\lambda \in \tilde{\Lambda}$.

2. The value of entry must be non-positive:

$$\beta\tilde{V}(x^e|\mu, \lambda) - \kappa \begin{cases} \leq 0, & \text{if } \lambda = 0, \\ = 0, & \text{if } \lambda > 0. \end{cases}$$

where x^e is the entry state and $\kappa > 0$ is a sunk entry cost. □

3 Results

We solve the above model numerically using the algorithm proposed by [15]. Our interest is whether equilibria could replicate power law of firm size distribution that is typically observed in real world data. Firm size are measured by sales of firms, as is common in the literature.

3.1 Parameterization

Table 1 shows the benchmark parameterization. These parameter values are basically selected according to [13] and [17]. In what follows, three experiments are conducted with these parameter values: (i) changing cost of investment d varying from 0.005 to 0.1; (ii) changing quality depreciation probability δ varying from 0.05 to 0.8; (iii) changing entry cost κ varying from 24.5 to 25.2. Through these experiments, other parameters are kept fixed at the values of Table 1.

Table 1. Benchmark parameterization for simulation experiments

β	d	δ	κ	m	c	ψ	θ_1	θ_2	Y	K	a
0.95	0.01	0.3	25	150	0.5	1	0.1	0.5	1	10	3

4 Simulated Firm Size Distribution

Fig. 1-3 is the firm size (sales) distributions in log-log plots corresponding each parameter value: the firm size and the probability of "size > x axis value" are both plotted in log scale. Fig. 1 and Fig. 2 show that, as investment cost d or quality depreciation rate δ becomes higher, the log-log plots come closer to linear; in particular, distribution almost exhibits power law when $\delta = 0.8$. On the other hand, we observe that changing entry cost κ have virtually no impacts on the curvature of plots.

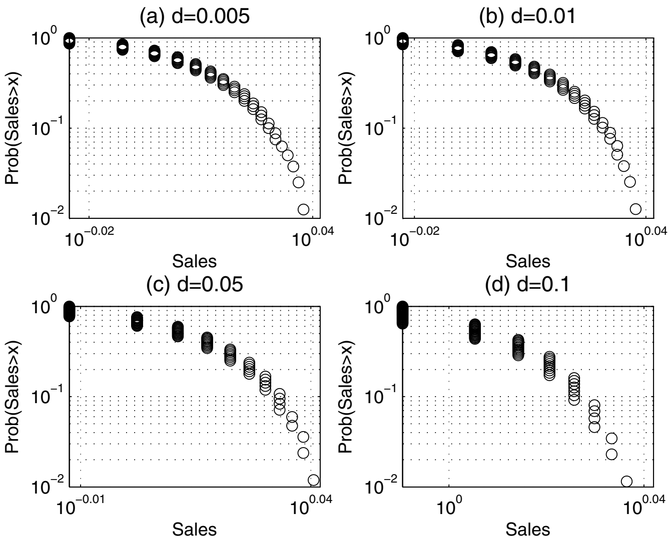


Fig. 1. Firm size distribution (varying investment cost)

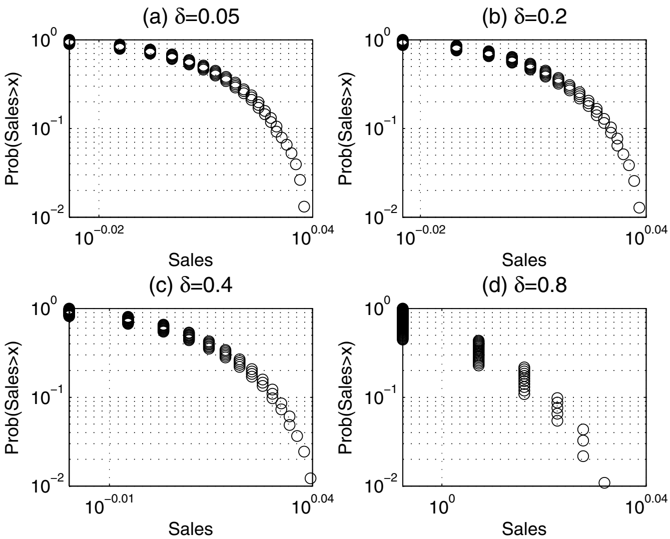


Fig. 2. Firm size distribution (varying depreciation probability)

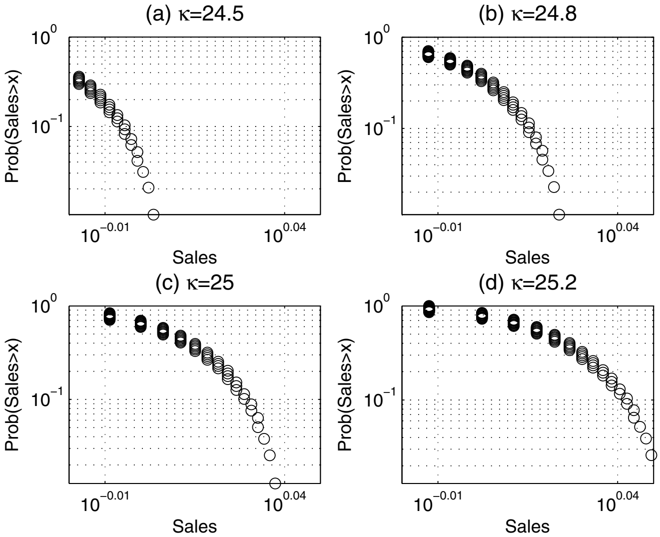


Fig. 3. Firm size distribution (varying entry cost)

5 Conclusion

In this article, we investigate a firm size distribution in the oblivious equilibrium (OE) model with Pakes-McGuire quality ladder setting. We find that the investment cost and the value of quality depreciation probability is the key to replicate the power law of firm size distribution.

A companion paper to this article ([12]) studies the firm size distribution in a MPE model of [13], by utilizing continuous time formulation of dynamic stochastic game proposed by [4].

References

1. Adalka, S., Johari, R., Weintraub, G.: Equilibria of Dynamic Games with Many Players: Existence, approximation, and market structure. Journal of Economic Theory (to appear)
2. Benkard, C.L., Jeziorski, P., Weintraub, G.Y.: Oblivious Equilibrium for Concentrated Industries. mimeo, Stanford University
3. Caplin, A., Nalebuff, B.: Aggregation and Imperfect Competition – On the existence of equilibrium. Econometrica 59(1), 25–59 (1991)
4. Doraszelski, U., Judd, K.: Avoiding the Curse of Dimensionality in Dynamic Stochastic Games. Quantitative Economics 3(1), 53–94 (2012)
5. Doraszelski, U., Satterthwaite, M.: Computable Markov-perfect Industry Dynamics. RAND Journal of Economics 41(2), 215–243 (2010)
6. Ericson, R., Pakes, A.: Markov-perfect Industry Dynamics: A framework for empirical work. Review of Economic Studies 62(1), 53–82 (1995)
7. Fershtman, C., Pakes, A.: Dynamic Games with Asymmetric Information: A framework for empirical work. Quarterly Journal of Economics 127(4), 1611–1661 (2012)

8. Gabaix, X.: Power Laws in Economics and Finance. Annual Review of Economics 1, 255–293 (2009)
9. Hopenhayn, H.: Entry, Exit, and firm Dynamics in Long Run Equilibrium. Econometrica 43, 1127–1150 (1992)
10. Hopenhayn, H.: Firm Microstructure and Aggregate Productivity. Journal of Money, Credit, and Banking 43, 111–145 (2011)
11. Luttmer, E.G.J.: Models of Growth and Firm Heterogeneity. Annual Reviews of Economics 2, 547–576 (2010)
12. Murao, T.: Firm Size Distribution in Pakes-McGuire Model: Mimeo, Kyushu University (2014)
13. Pakes, A., McGuire, P.: Computing Markov-perfect Nash Equilibria; Numerical implications of a dynamic differentiated product model. RAND Journal of Economics 25(4), 555–589 (1994)
14. Weintraub, G.Y., Benkard, C.L., Van Roy, B.: Oblivious Equilibrium: A Mean Field Approximation for Large-Scale Dynamic Games. In: Advances in Neural Information Processing Systems. MIT Press (2005)
15. Weintraub, G.Y., Benkard, L., Van Roy, B.: Markov Perfect Industry Dynamics with Many Firms. Econometrica 76(6), 1375–1411 (2008)
16. Weintraub, G.Y., Benkard, L., Van Roy, B.: Industry Dynamics: Foundations for models with an infinite number of firms. Journal of Economic Theory 146, 1965–1994 (2011)
17. Weintraub, G.Y., Benkard, L., Van Roy, B.: Computational Methods for Oblivious Equilibrium. Operations Research 58(4), 1247–1265 (2010)

Modeling Uncertainty in Banking Networks

Stojan Davidovic[1,*], Mirta Galesic[1], Konstantinos Katsikopoulos[1], and Nimalan Arinaminpathy[2]

[1] Max-Planck-Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94, 14195 Berlin, Germany
[2] Faculty of Medicine, School of Public Health, Imperial College London, London SW7 2AZ, United Kingdom
stojand@mpib-berlin.mpg.de

Abstract. Recent evidence from simulations of banking networks suggests that properties of the network design such as connectivity, bank size, or concentration affect networks' ability to withstand stress (Arinaminpathy et al. 2012; Gai at al. 2011). However, those studies typically assume that all banks have complete knowledge about the whole system. Here we introduce uncertainty into what banks know about other banks. We model uncertainty scenarios in which uncertainty is translated into asymmetric distribution of information among banks relative to the proximity of information source. Instead of knowing everything agents are faced with information delay and limited information availability. We show that when uncertainty is introduced, the system becomes more fragile to sudden shocks and a relatively small distress can push the system over the tipping point in which the whole banking network collapses.

Keywords: banking networks, uncertainty, simulation.

1 Introduction

The use of network approaches is gaining popularity among researchers studying financial systems. The need for a systemic study of financial markets stems from increasing concentration, connectivity and complexity of the rapidly growing financial system. In a real life demonstration of how disturbances in one part of the network can be diffused throughout the whole system, the ongoing financial crisis shows that there are no isolated financial markets but rather one global financial system. Furthermore, network approaches offer a general framework which enables access to insights from fields of research with more experience in network analysis such as ecology, epidemiology, engineering, and study of social networks.

A functioning banking network, in which banks borrow and lend money to each other efficiently, is essential for ensuring sufficient money supply and ultimately the stability of the economy. Todays' fractional-reserve banking entails long-term

[*] Corresponding author.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_14,

investing of large portion of depositors' funds and retaining only its small portion as liquid assets to service immediate needs. Such a scheme relies heavily on confidence of participants in the system: confidence that large portion of deposits will not be withdrawn within a short time, and confidence that enough liquid assets can be found at the interbank markets. In times of crisis, erosion of confidence can initiate a chain reaction within the interbank markets causing liquidity issues and upsetting the whole financial system.

Existing models of banking networks typically assume that all banks have complete and fully updated knowledge about parameters of other banks from the system, such as the state of their capital, deposits, loans, and liquid assets. However, in reality reporting is not real-time, the reports are not always fully reliable (e.g. as in the case of Lehman Brothers), all indicators are not included, and informal channels of communication facilitate further information asymmetries. All this factors contribute to the uncertainty of banks' knowledge about the system, and are amplified in the times of crisis when things are changing dramatically fast. Here we introduce uncertainty into what banks know about other banks. We start from the model of Arinaminpathy et al. 2012, and introduce uncertainty to investigate its consequences for the stability of the system. As opposed to risk, uncertainty entails a context in which not all relevant pieces of information are available in order to calculate probability distributions of potential outcomes. Real financial networks are much more complex than what we have considered here, but network approaches can be a helpful tool to start exploring those systems.

2 Model Overview

The present study is based on Arinaminpathy, Kapadia, and May (2012) model which investigates how banking network behaves under stress. The model is an agent based simulation which builds on dynamical network analysis developed in ecology to explore relationships between design properties of banking network and its resilience to sudden bank failures. Agents in the model are banks which are connected by borrowing and lending relationships established at the interbank market. Confidence in the system is included as a factor with a significant impact on banks' behavior. As stress progresses within the network, confidence drops, leading to more self-defensive decisions of banks, which become increasingly prone to cut their lending to other banks. This causes chain reaction known as liquidity hoarding, amplifying crisis effects. In addition to liquidity hoarding, the shocks are transmitted through two other channels of contagion: asset price contagion and the propagation of defaults through counterparty credit risk. The closing of failed banks entails liquidation of their external assets, and the increased supply of the assets pushes the prices at the asset market down. Banks sharing the same asset classes as failing banks suffer from the price drop (correlation between assets is neglected in the model). Finally, lenders run the risk that their borrowers if in default may not be able to pay back the loans, which is known as counterparty default.

2.1 Nodes and Edges

Nodes or banks in the network can be big (large) and small, and the size difference is fixed by the size ratio q ($q = \frac{large\ bank\ assets}{small\ bank\ assets}$). Banks are represented as simplified balance sheets (Figure 1). The liability side contains capital, retail deposits and interbank borrowing. The capital level measures how much stress on the assets side a bank can sustain before it becomes insolvent and suffers from a capital default. Interbank borrowing represents incoming loans from other banks and their number represents in-degree of an individual node. Retail deposits are taken to be external to the system. On the asset side there are n external asset classes. They are determined from a fixed number G of distinct asset classes distributed among banks. Liquid assets are a small fraction l of overall assets that banks hold in the most liquid form to meet immediate needs. They represent cash, central bank reserves or any similarly liquid financial instrument (e.g. high quality government bonds). Finally, interbank lending corresponds to outgoing loans to other banks in the system and their number is out-degree of a node. As with interbank borrowing, random half of interbank lending is assigned to be short term and the rest is long term ($p_s = 0.5$).

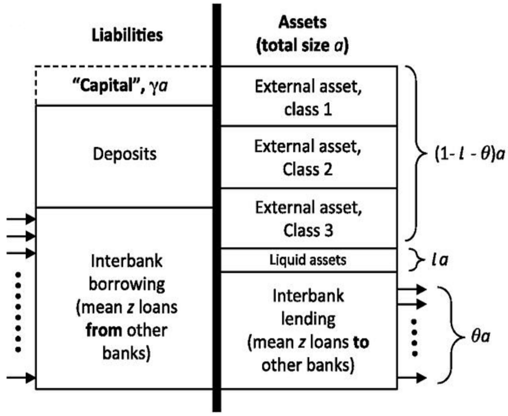


Fig. 1. A balance sheet representation of an agent (taken from Arinaminpathy et. al 2012)

2.2 Network

The network is a directed random graph with N banks (Arinaminpathy et. al reported no significant difference in the results for a more realistic preferential attachment graph). In-degree and out-degree of the nodes are determined by Poisson distribution with parameter z for small banks and $q * z$ for large banks. Each edge in the network is a single loan with direction from lender to borrower. The banks are also related to each other if they share one or more external asset classes. However, these relationships are not represented as edges but they are bases for asset price contagion.

2.3 Confidence and Individual Health

Confidence C is the first important determinant of banks' behavior. In Arinaminpathy et. al, it is calculated as a function of A and E, which are measures of solvency and liquidity of the system, respectively:

$$C = AE;$$

$$A = \sum_{i=1}^{N} A_i; \quad E = \sum_{i=1}^{N} E_i;$$

$$A_i = \frac{a_i}{\sum_{i=1}^{N} a_i^0}; \qquad E_i = \frac{e_i}{\sum_{i=1}^{N} e_i^0}.$$

At a given time point, A is the total value of all remaining assets in the system as proportion of its initial value; E is the fraction of interbank loans not withdrawn; A_i and E_i are the remaining assets and interbank loans of bank i as proportion of their total initial values in the system; a_i and e_i are the absolute values of remaining assets and interbank loans of bank i; and a_i^0 and e_i^0 are the initial absolute values of assets and interbank loans of bank i.

Unlike C which is a systemic parameter, h_i is individual health of bank i and is calculated as function of its indicators of solvency c_i and liquidity m_i:

$$h_i = c_i m_i; \qquad 0 < h_i < 1;$$

$$m_i = min[1, (A_i^{ST} + l_i)/L_i^{ST}].$$

c_i is the capital of bank i as a proportion of its initial value; m_i is the fraction of i's short-term liabilities that the bank can settle immediately, through its liquid and short-term assets; A_i^{ST} is the total value of bank's i short-term interbank assets; L_i^{ST} is the total value of bank's i short-term interbank liabilities; and l_i is the amount of liquid assets held by bank i.

2.4 Behavior of Banks

There are two possible actions that banks can take in the simulation: to shorten long-term interbank loans, and to withdraw short interbank loans. Only short-term loans can be withdrawn in a single time step, while shortening of long-term loans takes additional time step. These actions are fully determined by two decision rules:

$$h_i h_j < (1 - C); \tag{1}$$

$$h_i h_j < (1 - C)^2. \tag{2}$$

The shortening action is taken whenever the first condition is satisfied (1), and the withdrawing action is taken whenever the second condition is satisfied (2). There is simple logic behind the rules: if C is high (one or close to one) it is less likely that conditions will be satisfied. In contrast, a drop in C can cause hoarding as this affects the decision conditions of all banks. In addition, the shortening condition is easier to satisfy than withdrawing condition, which means that banks resort to withdrawal in only more urgent situations.

3 Modeling Uncertainty

To calculate C as defined in Arinaminpathy et. al, and take any action, banks have to know present and initial values of assets and interbank loans of all banks in the system. Such a condition is highly unrealistic, especially in times of crisis when system is changing rapidly. To explore how the network behaves in a more realistic setting we consider several scenarios of uncertainty. They mainly entail change of calculation of confidence C. In what follows we present one of these scenarios and briefly discuss other possibilities for introducing uncertainty in the discussion section. For ease of following we first present the notation template of model parameters:

$$P_{observed(optional)}^{time\ step\ (optional);observer}$$

For example, a_j^{0i} denotes i's judgment of j's initial (zero time step) absolute value of assets. Absence of *time step* indicator implies current value of a parameter without referring to a particular time step. Indicator of *observed* bank is missing when calculation of a parameter is not based on information from an individual bank.

In the basic uncertainty scenario, we assume limited information availability relative to the distance from the information source.

This is, bank i calculates C from the information about itself and its neighbors within the fixed maximal value of distance $d(i,j)$. Here, $d(i,j)$ is the minimal distance (shortest path between nods) between banks' i and j in the network, and it can take values $d(i,j) = 1,2, \dots d_{max}$. For instance, if $d_{max} = 1$, which is the default value of d_{max} in our scenario, then only i and i's immediate neighbors are included in the C calculation. In this view, each bank j in the network, except i itself ($j \neq i$), is considered as i's neighbor, but it may be distributed on the different distance from i. In general, C is calculated:

$$C^i = A^i E^i;$$

$$A^i = \frac{a_i + \sum_{j \in J_i(d_{max})} a_j^i}{a^{0i}}; \qquad E^i = \frac{e_i + \sum_{j \in J_i(d_{max})} e_j^i}{e^{0i}};$$

$$a^{0i} = a_i^0 + \sum_{j \in J_i(d_{max})} a_j^{0i}; \qquad e^{0i} = e_i^0 + \sum_{j \in J_i(d_{max})} e_j^{0i}.$$

Essentially, which banks are included in the calculation is determined by the set of i's neighbors $J_i(d_{max})$. To define $J_i(d_{max})$ we first define the set $J = 1,2, \dots, N$, which contains all banks in the network. Then, its subset $J_i(d_{max})$ is defined:

$$J_i(d_{max}) = \{ j \in J \mid d(i,j) \leq d_{max} \ \& \ j \neq i \}.$$

In contrast to Arinaminpathy et. al model, in which all banks share the same opinion about confidence in the system ($C^i = C$), here each bank calculate its own value of C, C^i. In addition, calculation of individual health stays the same.

4 Simulation

Once the network is established, we impose a shock by randomly selecting a single small or big bank and setting its capital to zero, causing its capital default. In the network all shocks eventually affect bank's capital. When a bank suffers from a shock that is equal or exceeds the level of its capital then the bank is in a capital default. Failing bank can settle its short interbank liabilities (L_i^{ST}) up to the level of the liquid assets and short loans ($A_i^{ST} + l_i$) that are at its disposal. If the liabilities exceed the assets ($L_i^{ST} > A_i^{ST} + l_i$) the rest is deducted from the lenders' capital. In addition, when shock exceeds the i's capital then exceeded value s is going to be shared among i's lenders (s/z), but only up to the level of the lending. On the other hand, borrowers gain from long term loans of a failing bank (their borrowing) that cannot be withdrawn. Furthermore, the external assets are sold at the market that is taken to be external to the model. However, it is assumed that the price of asset v is decreasing to a fraction exp $(-\alpha x_v)$ of its original value (modeled as in previous work (Cifuentes et. al, 2005; Gai & Kapadia, 2010)), where x_v is the proportion of that asset (v) being sold by the failing bank, and α is indicator of market liquidity which is directly related to C, $\alpha = 1 - C$.

Our network has $N = 100$ banks. The default value of q is 10 which, given other parameters of the model, leads to a system in which nine banks are big ($N_b = 9$) and the rest are small ($N_s = 91$). The value of parameter z is five ($z = 5$). This is, small banks on average have five incoming and five outgoing loans (edges), while for big banks the average number is 50 loans ($q * z$). Small banks have 10, and big 20 external asset classes ($n_s = 10, n_b = 20$). Given that we take that 10 banks on average share the same asset class ($g = 10$), this implies 109 distinctive external asset classes ($G = (N_b n_b + N_s n_s)/g$).

Regarding the balance sheets, the parameters are: $\theta = 0.2$, the proportion of total assets initially determined to be held in interbank loans; $l = 0.01$, the proportion of the total assets initially liquid; $\gamma = 0.04$, capital to asset ratio.

5 Results

The results are based on 100 repetitions of the simulation that contains 100 time steps. The figures (Figure 2-5 in the appendix) depict probability distribution of failing banks and dynamics of failing over time for complete knowledge and uncertainty scenarios, both including the initial shock imposed on a small and a big bank. We replicate the corresponding results from Arinaminpathy et. al. The percentages are not exactly the same (some default parameters in their model are different, such as: $N = 200$, $q = 24$, $n_s = n_b = 10$), but the pattern is identical. In Figures 3 and 5, we can see that the uncertainty scenario makes considerable difference for the results if compared with the complete knowledge scenario. In Figure 3a, which depicts the small bank failure case, the whole network collapses with a nonzero probability. To put it differently, a relatively small shock can lead to the crash of the whole system. In addition, Figure 3b shows sudden shifts in propagation of shocks in the system over time. Finally, Figures 4 and 5, with the initial shock applied on a large bank, display quantitative rather than qualitative difference between the complete knowledge and uncertainty scenarios. It is notable, however, that the uncertainty scenario is significantly more harmful for the system.

6 Discussion

The significance of the results is that uncertainty is a factor with important implications for the system stability. In the next step we aim to run further simulations considering other sources of uncertainty. Here, we briefly present two other possibilities.

Scenario 1. Introduction of a random error in the estimation of information as a function of the distance. In this scenario information about more distant banks is noisier. For the estimation of C assets of neighbors j are calculated:

$$a_j^i = a_j^j + (d-1)\varepsilon, \qquad d(i,j) = 1,2, \dots d_{max};$$

where parameter ε is determined by the normal distribution ($\varepsilon \sim N(0, \sigma^2)$).

Scenario 2. Introduction of information delay with the distance from the information source. This scenario entails an assumption that information about more distant banks is more outdated. Here, for the estimation of C assets of neighbors j are calculated:

$$a_j^i = a_j^{kj}, \qquad k = t - d + 1, \qquad d(i,j) = 1,2, \dots d_{max};$$

where k denotes the time-step of the data, and t denotes the current time-step.

Finally, we plan to thoroughly explore relationships between design properties of the network and the uncertainty scenarios. The ultimate goal is to find out policy implications for the design of a more resilient financial network.

References

1. Arinaminpathy, N., Kapadia, S., May, R.M.: Size and complexity in model financial systems. Proceedings of the National Academy of Sciences USA 109(45), 18338–18343 (2012)
2. Cifuentes, R., Shin, H.S., Ferrucci, G.: Liquidity risk and contagion. Journal of European Economic Association 3, 556–566 (2005)
3. Gai, P., Haldane, A., Kapadia, S.: Complexity, concentration and contagion. Journal of Monetary Economics 58(5), 453–470 (2011)
4. Gai, P., Kapadia, S.: Contagion in financial networks. Proceedings of the Royal Society A 466(2120), 2401–2423 (2010)

Appendix – The Results

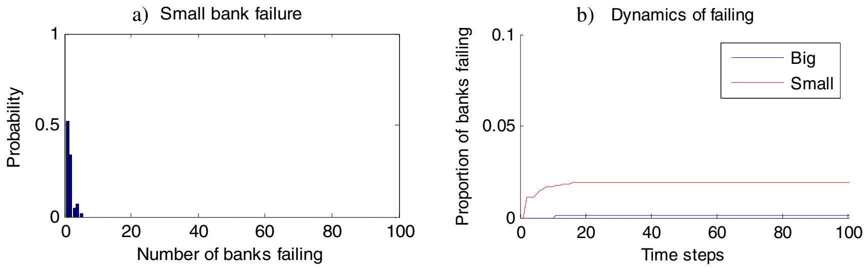


Fig. 2. Complete knowledge scenario of small bank failure: a) probability distribution for number of banks failing; b) dynamics of failing over time

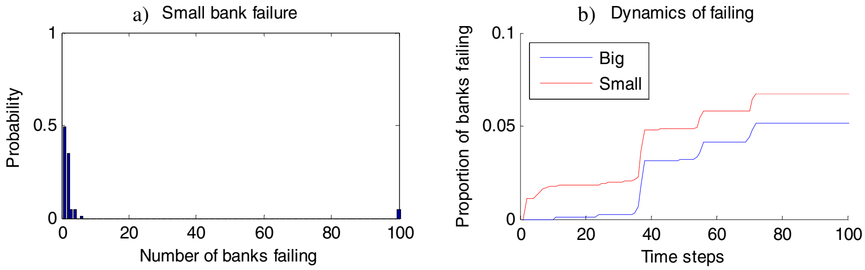


Fig. 3. Uncertainty scenario of small bank failure: a) probability distribution for number of banks failing; b) dynamics of failing over time

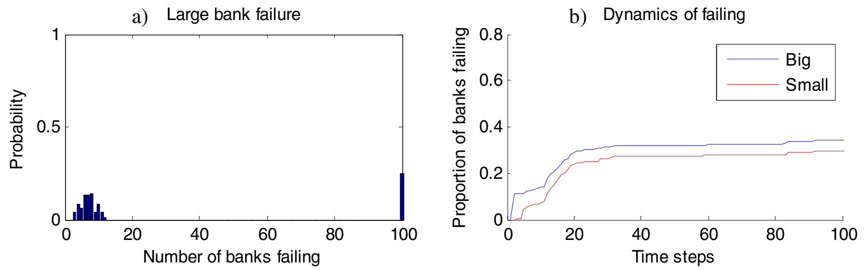


Fig. 4. Complete knowledge scenario of big bank failure: a) probability distribution for number of banks failing; b) dynamics of failing over time

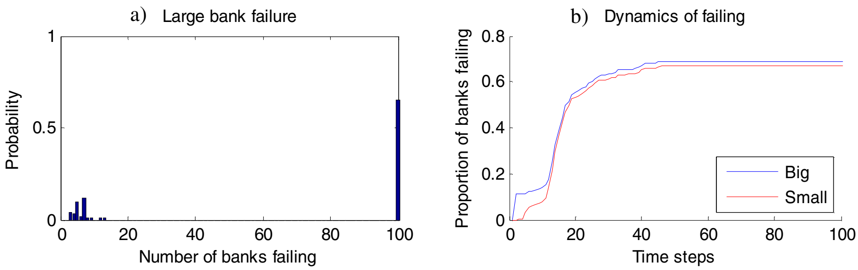


Fig. 5. Uncertainty scenario of big bank failure: a) probability distribution for number of banks failing; b) dynamics of failing over time

Changing the Hidden Rules - An Excel Template for Discussing Soccer's Competitive Balance

Joaquim Teixeira[1], Nuno Santos[2], and Paulo Mourao[3]

[1] BSc Economics/ Universidade de Tras-os-Montes e Alto Douro; 5000 Vila Real, Portugal
[2] BSc Mechanical Engineering / Universidade de Tras-os-Montes e Alto Douro; 5000 Vila Real, Portugal
[3] Department of Economics / University of Minho, 4700 Braga, Portugal
paulom@eeg.uminho.pt

Abstract. We constructed a (Microsoft Excel) Template for simulating the final tables of soccer leagues. In our Template, the user can insert different values for parameters that influence the soccer leagues (such as the importance of the sports history of each competing team, the influence of home advantages, or the relevance of the referees' decisions). The user is also able to choose among different systems of points awarded for victory. Then, the Template produces an index for the competitive balance of the entire league. We tested several combinations of parameters and concluded that the most balanced leagues are characterized by rewarding victory with 2 points and draws with 1 point and when the influence of the teams' history or budget is minimized.

Keywords: Simulator, Template, Soccer.

1 Introduction

Excel spreadsheets templates are interesting tools in order to let most of users simulate different kinds of scenarios, as argued by Gunga et al [1] or Cano et al [22]. This paper will introduce an Excel Template as a powerful tool for simulating soccer leagues and for discussing the competitive balance of these leagues. Therefore, this Excel Template can be used by professional organizers of soccer leagues (for instances, FIFA, UEFA, or the national leagues) in order to study which factors are influencing soccer competitive balance or which factors should be changed for having more interesting competitions.

We love sports for many reasons, from a cultural tradition that educates each of us to respect and admire sports events to professional reasons. One of these reasons relates to the ability of the best players and the best teams to catch our attention for lengthy periods. Deeply correlated with the quality of players and teams, it is important to have balanced sports competitions. Balanced competitions tend to disperse the probability of final victory across a large number of teams and thus maintain interest; in contrast, competitions whose winner is easily anticipated tend to

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_15, © Springer International Publishing Switzerland 2014

lose interest, fans and financial sustainability. This discussion is serious with regard to soccer, especially in the professional European soccer competitions. Season after season, we observe a concentration of the probability of winning the European professional leagues on a few teams, usually the ones exhibiting the largest budgets at the beginning of the competition. However, various authors claim that certain adjustments to the soccer competitions would produce more balanced professional leagues.

These authors claim that changing the system of rewards for victory (2 points replacing the current 3 points for most of the competitions), the number of rounds or the relevance of random factors (such as the referees' influence or the effect of home advantage) could improve the competitive balance of European professional leagues. However, some of the organizing institutions of soccer competitions (such as the national federations and the UEFA) face a lack of software programs that can simulate the expected values of the measures of Competitive Balance for a given number of editions of the competition. The remainder of the paper has four sections. Section 2 reviews the literature on the competitive balance in soccer. Section 3 shows the architecture behind our Template. Section 4 exhibits some figures produced by our Template. Section 5 presents our conclusions.

2 Discussing the Factors Behind the Competitive Balance of Soccer Leagues

The main factors that our review of literature identified as influencing the competitive balance of soccer leagues are as follows: the number of competing teams and the number of journeys/rounds, the points awarded for a victory, the home advantage, the size of the assistance, the influence of the referees, and the influence of the sport history.

Number of Teams

There are two main models of sports leagues: closed and open. The closed model is based on the North American closed leagues, which have a fixed number of members (and the addition of a new team or removal of an existing team is rare). The open system is common in European sports leagues and is based on a system of promotion and relegation. Several authors have studied how restructuring a league's format (modifying the number of teams and matches) can affect the annual assistance of stadiums and the competitive balance of the competition (Cairns [2]; Burkitt and Cameron [3]; Dobson [4]).

Points for a Win

In 1981, the English Football League introduced the system of three points for a victory, one for a draw and zero for a defeat, which came to replace the previous system present of two points for a victory, one for a draw and zero for a defeat. Starting in the 1990s, the new system became standard in most leagues and competitions. According to Newson (1984), the initial idea “was presumably that a greater reward for winning games would encourage more positive attitudes from teams and that the consequently more attractive and aggressive football would bring in bigger crowds”. If the teams are tied near the end of the match, they will launch themselves to the attack in search of the winning goal that would guarantee them the two additional points. However, the critics of this system consider that if a team is winning near the end of the game, that team will adopt a negative and defensive strategy to preserve not only the victory but also the two additional points compared to a draw. Other alternatives have already been studied in soccer leagues, such as the penalty shootout. Some leagues tried the penalty shootout after a draw match. In 1987, the Norwegian First Division used three points for a win at the end of 90 minutes, two points for a shootout win, 1 point for a shootout loss and zero points for a loss. Between 1996 and 2000, the Major Soccer League (USA) used three points for victory, one point for a shootout victory and zero points for a shootout loss or a loss at the end of 90 minutes.

Home Advantage

The existence of the home advantage in sports is a well-known and documented fact (Pollard [6]; Courneya and Carron [7]; Schnytzer and Weinberg [8]; Jamieson [9]). For instance, Jamieson [9] argued that the home team tends to win approximately 60% of all athletic contests. He studied several variables and concluded that the time era, season length, and game type all have a significant impact on the home-field advantage. Carron [10] also noted that the home advantage appears to be universal across all types of sports.

Assistance

There are many factors that might account for the home advantage, but the crowd factor and the social pressure exerted by the crowd is one of major importance (Neave and Wolfson [11]; Benrheim [12]; Becker and Murphy [13]). Wallsten and Barton [14], Wickens and Holland [15] and Nevil et al. [16] argued that crowd assistance produces a greater and more negative noise when visitor teams or players tackle, which may potentially (and most likely) serve as a biased cue for the referee’s decision.

Referees

Referees are important agents in a football match, employed to interpret the rules of football in an impartial way. However, they can exercise a considerable discretionary power that can have a very important influence in the final result of a football match, in particular when adding extra time, awarding penalties, allocating yellow and red cards or deciding on free-kicks or offside penalties (Garicano et al. [17]; Nevil et al. [16]).

Teams' History
However, in addition the previously identified dimensions, we have to account for what some authors argue is an even more influential factor: the team's history, which is highly correlated with the budget size and with the probability of winning matches and competitions (Mourao [18]).

Tactical Quality
Various authors (Lafuente [19] and Papahristodoulou [20]) also identify tactical quality as an important factor in winning. Lafuente (2008) developed an algorithm to ease the replacement of a player. Papahristodoulou [20] studied the optimal formations of two soccer teams (AC Milan and Barcelona) and how this optimal formation maximizes the probability of winning matches.

Luck
In our Template, we also considered "Luck" as a factor. Although the previously identified factors can be considered as endogenous factors for competing teams, there is always space for 'luck', even at residual levels, to complement the explanation of the determinants of victory.

3 Architecture of the Program

Using data obtained by numerical simulation, we also intend to study the competitive balance of soccer leagues. The main objective is to develop a simulator. We opted to construct this simulator using the format of an Excel Template for three main reasons: the low complexity of the language, its accessibility from the user perspective, and its flexibility. In this simulator, the user will have the opportunity to include official data from past competitions and then to select various parameters (see our flowchart at Figure 1). These parameters are initially defined, giving the user the option of assigning different "weights" to the parameters (that must total 1 at the end).

In the following step, the template will attribute random shocks (mean 0, standard deviation 0.5) to each team. These random shocks are weighted by the weights chosen by the user. Then, the template generates a matrix of results for all the matches of a season. Finally, the user receives numerical results (including an Index of Competitive Balance[1]) and graphics related to the desired number of simulated seasons. Full details are available upon request.

[1] We are going to follow Kupfer [21], and in this first version we provide the calculation of an Index of Competitive Balance that is essentially a Hirschman-Herfindahl Index. To calculate this index, Kupfer [21] suggests the following equation $HH = \sum_{i=1}^{n} s_i{}^2$ where s_i^2 represents the square of the percentage of points scored by each competing team at the end of a season.

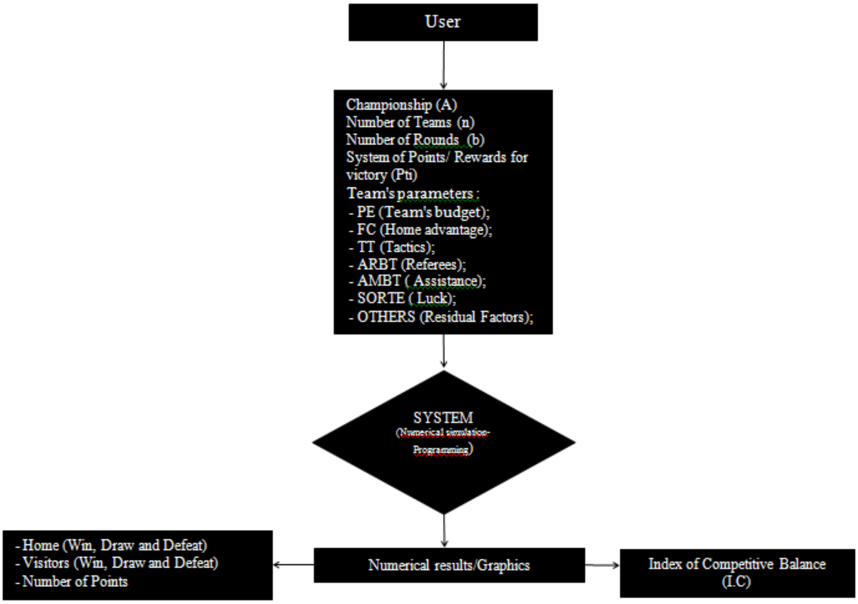


Fig. 1. Flowchart of the Template

4 Experiments – Some Observations

We have run some experiments using our Template to observe how different combinations of parameters lead to different values for the competitive balance of soccer leagues.

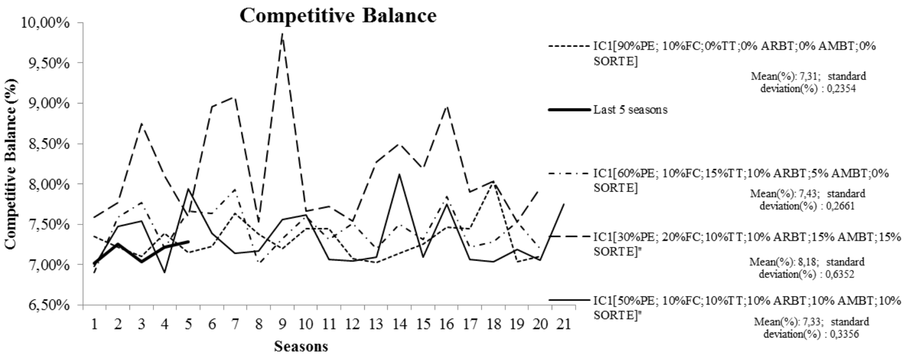


Fig. 2. Competitive Balance simulated for 20 seasons (16 teams, round-robin, 30 journeys. Victory 3 pts; Draw 1 pt)

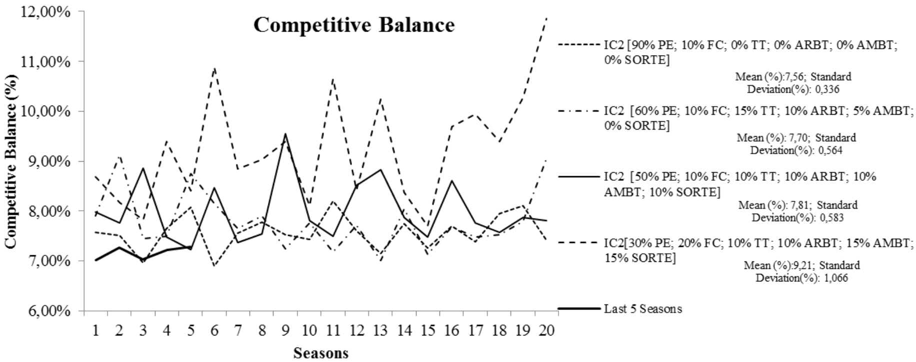


Fig. 3. Competitive Balance simulated for 20 seasons (16 teams, round-robin, 30 journeys. Victory 2 pts; Draw 1 pt)

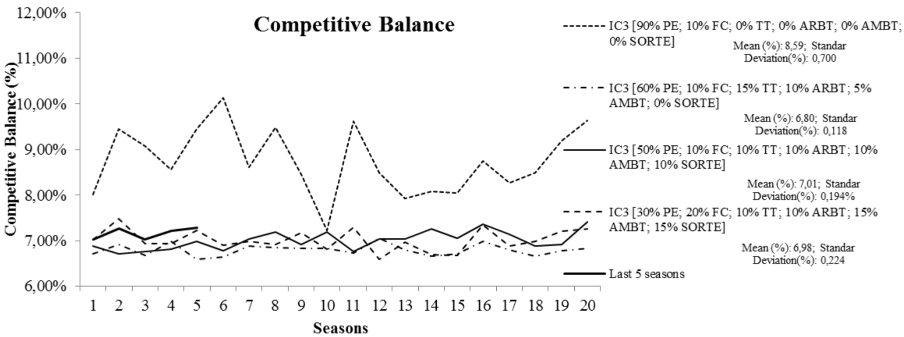


Fig. 4. Competitive Balance simulated for 20 seasons (16 teams, round-robin, 30 journeys. Victory 1 pts; Draw 0 pts)

We observe that the highest mean value (9.21, interpreted as the most balanced simulation) obtained for our measure of the competitive balance of 20 simulated seasons is observed in Figure 3 (in which case, each victory is awarded 3 points, each draw is awarded 1 point, and each defeat is awarded 0 points). This highest mean value is related to the combination of weights 30% (Teams' budgets), 20% (Home advantage), 10% (Tactical quality), 10% (Referees), 15% (Assistance), and 15% (Luck). The lowest mean value (6.80, the most unbalanced case) is observed in Figure 4 (in which case, each victory is awarded 1 point, and each draw and each defeat are awarded 0 points). This lowest value relates to the following combination of weights: 60% (Teams' budgets), 10% (Home advantage), 15% (Tactical quality), 10% (Referees), 5% (Assistance), and 0% (Luck).

5 Conclusion

This research paper reports the authors' work to develop a software solution that can examine how the competitive balance of soccer leagues can change due to different combinations of factors. We allow the user to choose parameters for a large set of dimensions that the literature has identified as influencing the path of a competition: the points awarded for victory, the number of teams and matches, the influence level of each team's history/budget, the influence of assistance, or the level of influence of the referees, among other factors.

We constructed an Excel Template that also produced an index of competitive balance for the simulated soccer leagues. Considering different combinations, we observed that awarding 2 points for each victory and 1 point for each draw and reducing the influence of the teams' budgets produced more balanced seasons.

As a further step, we intend to develop our template to consider the data for each match. We also intend to record as many (professional or youth) soccer leagues as possible to inform the user which soccer league is closest (or most correlated) to the league generated by the simulations, accounting for the parameters chosen by the user. This last step will provide to users many deeper details of their organized soccer leagues in such a way that they will have the possibility of being informed of the most interfering parameters in the competitive balance of their leagues.

References

1. Gunga, V., Kishnah, S., Pudaruth, S.: Design of a multi-agent system architecture for the scrum methodology. International Journal of Software Engineering & Applications (IJSEA) 4(4) (July 2013)
2. Cairns, J.: Evaluating changes in league structure: The reorganization of the Scottish Football League. Applied Economics 19, 259–275 (1987)
3. Burkitt, B., Cameron, S.: Impact of League Restructuring on Team Sport Attendeces: the case of rugby league. Applied Economics 24, 265–271 (1992)
4. Dobson, S.J., Wilson, J.: League Structure and Match Attendances in English Rugby League. International Review of Applied Economics 15(3), 335–351 (2001)
5. Newson, G.: Three points for a win: Has it Made any Difference? The Mathematical Gazzete 68, 87–91 (1984)
6. Pollard, R.: Home Advantage in football: A current of an unsolved puzzle. The Open Sports Science 1(1), 12–14 (2008)
7. Courneya, K.S., Carron, A.V.: The Home Advantage in Sport Competition: A Literature Review. Journal of Sport and Exercise Psychology 14, 13–27
8. Schnytzer, A., Weinberg, G.: Testing for home team and favorite biases in the Australian rules football fixed odds and point spread betting markets. Working Papers 2011-13, Department of Economics, Bar-Ilan University (2007)
9. Jamieson, J.P.: The Home Field advantage in Athletics: A Meta-Analysis. Journal of Applied Social Psychology 40(7), 1818–1848 (2010)
10. Carron, A., Loughhead, T.M., Bray, S.-R.: The Home Advantage in Sport Competitions: Courneya and Carron's (1992) conceptual framework a decade later. Journal of Sports Sciences 23(4), 395–407 (2005)
11. Neave, N., Wolfson, S.: Testosterone, territoriality, and the home advantage in soccer (2002)
12. Bernheim, B.D.: A Theory of Conformity. Journal of Political Economy 102(5), 841–877 (1994)

13. Becker, G.S., Murphy, K.M.: Social Economics: Market Behavior in a Social Environment. Belknap Press/ Harvard University Press, Cambridge, MA (2000)
14. Wallsten, T.S., Barton, C.: Processing probabilistic multidimensional information for decision. Journal of Experimental Psychology: Learning, Memory and Cognition 8, 361–384 (1982)
15. Wickens, C.D., Hollands, J.G.: Engineering Psychology and Human Performance, 3rd edn. Prentice Hall (2002)
16. Nevill, A.M., Balmer, N.J., Williams, A.M.: The influence of crowd noise and experience refereeing decisions on football. Psycohology of Sport and Exercise 3, 261–272 (2002)
17. Garicano, L., Palacios, I., Prendegast, C.: Favoritism Under Social Pressure, NBER Working Paper 8376 (2001)
18. Mourao, P.: The Indebtedness of Portuguese Soccer Teams – looking for determinants. Journal of Sports Science 30(10), 1025–1035 (2012)
19. Lafuente, J.G.: Automatismos y racionalidad en la toma de decisions para sustituir a un deportista en momentos decisivos. Cuaderno de Gestión 8(1), 39–58 (2008)
20. Papahristodoulou, C.: Optimal football strategies: Ac Milan versus Fc Barcelona. MRPA Paper No 35940 (January 2012)
21. Kupfer, D.: Economia Industrial: fundamentos teóricos e práticos no Brasil, Brasil. Elsevier ed. Campus, Rio de Janeiro (2002)
22. Cano, R., Sánchez, J., Pinzón, C.: Management ubiquitous of messages and documents organizational through intelligent agentes. In: Corchado, J.M., Rodríguez, S., Llinas, J., Molina, J.M. (eds.) International Symposium on DCAI. ASC, vol. 50, pp. 171–179. Springer, Heidelberg (2008)

Insider Trading, Earnings and Stock Based Compensation: A View to Speculation

Esther B. Del Brio[1], Ilidio Lopes-e-Silva[2], and Javier Perote[1]

[1] Faculty of Economics and Business, University of Salamanca, Spain
{edelbrio,perote}@usal.es
[2] Polytechnic Institute of Viseu, Business School, Portugal
isilva@estv.ipv.pt

Abstract. We design a laboratory experiment to study the relation between earnings management and insider trading and their effects on the stock markets. The experiment simulates a market where one insider and three outsiders trade on the stocks of a company. We show that if the insider affects the stock price with her earnings announcements, then she will use this power and increase her capital gains. These earnings management practices not only induce price inefficiencies but also negatively affect the profits of the other traders. We also show that the managerial stock-based compensation worsens these effects, seriously increasing stock prices and manager's profitability. We conclude that the lack of transparency in earnings management and the perverse incentives of the managerial compensation policies have an important role in generating price inefficiencies, bubbles and financial crises.

Keywords: Insider trading, earnings management, stock-based compensation, laboratory experiment.

1 Introduction

This article we shed some light on the opportunistic behavior of the insiders who not only possess private information and use it in their trading, but also deliver this information to the market with clear incentives to profitably manipulate the information. For this purpose we conduct a laboratory experiment designed to show how an insider might behave if her actions are not properly supervised and penalized.

We consider markets with one insider and three outsiders that trade on the stocks of a firm without knowing the final price at which their orders will be closed. The only information the traders have is the determinants of the price (price function) and the earnings figure reported by the insider (who knows the real earnings figure as well). The final payoff (after 40 rounds) depends on the final wealth the participants have on their dealings and investments (i.e. their portfolios are finally valued at the final stock price). We consider two alternative treatments: The baseline (T0), where all traders receive a dividend payout (a fixed amount per stock in their portfolios) every five periods; and a second treatment (T1), where insiders also receive a

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_16,

stock-based compensation depending on the increase of the stock price every five periods (see Bergstressera and Philipponb, 2006, for alternative compensation schemes).

Our results show that under these basic assumptions insiders misreport earnings figures so as to drive the stock price for their own benefit. In treatment T0 to maximize capital gains and in T1 to maximize their bonuses. As a consequence, insiders' returns are higher than outsiders' returns but also insiders' returns significantly increase when the stock-based remuneration policy is implemented. Therefore, earnings management practices and stock-based compensation induce two serious inefficiencies to the market: they affect positively/negatively the insider/outsider return and they distort the stock price, which sharply increases under stock-based compensation schemes.

2 Experiment

2.1 Setting

We create 10 different markets where the stocks of a firm are sold. At the beginning of the experiment each subject i is assigned an initial endowment of stocks (x_{it}) and money (m_{it}). Paticularly, they are initially endowed with 5 stocks and 100 experimental currency units (ECUs). The initial price of the firm's stocks (p_t) is 20 ECUs and thus subjects possess an initial wealth (ω_{it}) of 200 ECUs. In addition, subjects are also assigned to a group (market) of four people and a type (insider or outsider). Particularly, we refer to the insider as subject $i = 1$ and to the outsiders as subjects $i = 2,3,4$. These assignments remain during the entire experiment.

In each market there are one insider and three outsiders. The insider knows privately the information about the growth rate of firm's earnings (r_t), which is randomly chosen (with equal probability) from the following set $\{-20, -10, 0, 10, 20\}$ in each period. For example, if value 20 is drawn we assume that the earnings have risen 20 per cent in the current period. In every period, the insider has to report on the firm's earnings to the other three subjects in the market. This message ($\tilde{r}_t$) must also be a value of the set $\{-20, -10, 0, 10, 20\}$, but this reported value does not necessarily have to be the true one. Then, all four subjects in the market simultaneously make a decision about their trading orders. For the sake of simplicity, we restrict the operations to three possible cases: "buying a single stock of the firm", "selling a single stock of the firm" or "neither buying nor selling stocks". When making this decision, subjects do not know the closing price that buyers will have to pay and sellers will receive. This price is computed through the following equation:

$$p_t = p_{t-1} + 2(d_t - s_t) + \frac{1}{10}(\tilde{r}_t - r_t), \tag{1}$$

where d_t is the demand (number of subjects in the same market that are willing to buy a stock) and s_t is the supply (number of subjects in the same market that are willing to sell a share). Therefore, the price in period t is a function of its previous value, the excess of demand/supply and the surprises in the announced earnings with respect to the true ones. Therefore, three main assumptions are underlying the price formation.

First, prices follow a dynamic structure, which in case the number of agents in the market was very large and if earnings surprises were unpredictable would approximate a random walk (market efficiency hypothesis). Second, competitive forces in the market lead the prices up/down in case of excess of demand/supply. In spite of this assumption, in our game the market is not competitive since each subject has a certain market power and may affect the price movement to some extent. Third, subjects' expected earnings under truth-reporting are unbiased and thus this expectation does not significantly deviate from r_t. Then, if we interpret r_t as the expected earnings and assume rational expectations, a positive/negative surprise in the reported earnings triggers the share prices upwards/downwards.

Note that for the price function in equation (1) price movements in every period $(p_t - p_{t-1})$ are integers in the range [12,–12]. The upper/lower bound corresponds to the situation where all subjects are willing to buy/sell, the real earnings experiment a 20% decrease/increase and the insider reports a 20% increase/decrease of the earnings (e.g., $p_t - p_{t-1} = 12$ where $d_t = 4$, $s_t = 0$, $\tilde{r}_t = 20$ and $r_t = -20$).

After having received the subject's trading orders, the price is computed and the orders are closed at the corresponding price. Borrowing money to buy a share is not allowed and thus if a subject cannot pay the price after having chosen "buy a new stock" the computer reminds her that she does not have enough money and she is forced not to trade in this period. In this case the new price is recomputed assuming the new demand. In the same line, a subject can only decide to "sell a stock" if she has at least one stock in her account. These restrictions to the transactions help to stabilize the price throughout the experiment, because subjects cannot continue buying (selling) during more than 5 consecutive periods (approximately) since they do not have enough money (stocks) with which to trade.

The experiment lasts 40 periods and in every period subjects accumulate their wealth (the ECUs and the market value of their current stocks). Furthermore every 5 periods (i.e. in periods 5, 10, 15, 20, 25, 30, 35 and 40) agents received a dividend (D_{it}) of 2 ECUs for every share possessed at the end of the corresponding period (once the transactions of the period have been closed).

We considered two alternative treatments and a between-subjects design. In the baseline treatment (T0) subjects only receive the dividend remuneration and the capital gains from their market operations that are incorporated to their wealth. In the other treatment (T1) we additionally implement a stock-based remuneration policy for the insiders, which consists of an extra bonus every 5 periods (i.e. at the same time as the dividend payout). This bonus is gained only if the share price has increased during the last 4 previous rounds and in this case the insider bonus (b_{it}) is 5 times the price increase within this period. Therefore, a bonus is computed as

$$b_{it} = \begin{cases} 5(p_t - p_{t-4}) & if \ \ p_t - p_{t-4} > 0, \\ 0 & otherwise. \end{cases} \tag{2}$$

In period t=40 the final payoff of subject i (π_i) is computed as

$$\pi_i = m_{i40} + p_t x_{i40} + D_{i40} + b_{t40} \tag{3}$$

where $b_{i40} = 0$ in treatment T0 and for every subject $i = 2,3,4$ in treatment T1.

2.2 Procedures

We conducted the experiment, which was programmed within the z-Tree toolbox (Fischbacher, 2007), in the Laboratory for Research in Social and Economic Behavior (LINEEX), which is housed at the University of Valencia. For each treatment we organized a session with 40 subjects and thus a total of 80 undergraduates from various disciplines participated in the experiment. Participants were assigned into groups of four and received a role in the group: insider (one in every group) or outsiders (the other three of the group). These groups and roles remained unchanged during the entire experiment and their identities were never revealed. Then subjects were informed about their initial endowments (5 stocks and 100 ECUs) and the initial price of the stocks (20 ECUs), all the same in each market and for each subject.

Within each group the game was played for 40 periods and each period had two steps. In the first step the insider of each group received a message about the firm earnings that was randomly drawn. In both treatments we used the same sequence of earnings to ensure the comparability of the results. The real value of the earnings was never revealed to the outsiders of the group although they might infer ex post guesses based on the final stock price at the end of each period (since real earnings are part of the final price). With this private information the insider decided on the reported earnings figure that was publicly announced. In the second step each participant in the market decided privately and submitted their decision about either selling, buying or not participating in the market. With all the submitted orders and the reported and real earnings, the price was formed and the orders were satisfied accordingly. The results of the decisions on subjects' wealth and its components (stocks and liquidity) were recorded in a table, as well as the stock prices, reported earnings and their own trading decisions. This information was continuously updated and available on the screen during the experiment.

At the end of the experiment the total wealth of the subjects (the stocks valued at the final price of the game and the money possessed at the end of the experiment) was converted into Euros at a known exchange rate (50 ECUs = 1 Euro). Payment took place privately and the participants had to leave the laboratory immediately once they were paid. The maximum, minimum and average payoff was 22.38 (31.92), 12.54 (8.46) and 16.23 (17.07) Euros in treatment T0 (T1), respectively. A session lasted on average two hours and a half.

2.3 Hypotheses

We gather the hypotheses into two different categories:

(a) ***Hypotheses related to the earnings management and insider trading:***
In our framework insiders have access to private information about the firm performance and also have the privilege of disseminating the information to the market. Thus the first hypothesis under test is whether in this context insiders report the information truthfully. In this study we do not analyze either the ethics or the illegal considerations of misreporting (see Abdolmohammadi and Sultan, 2002). We neither consider possible controls or penalties on misreporting behavior or even collateral effects on reputation that might be introduced in alternative treatments.

Insiders, however, know that their misreporting may affect not only their own profitability but also that of their partners. In this case we wonder whether they behave honestly in this sense or they do not worry about such considerations and directly maximize their own profitability. We hypothesized this conjecture in H1.

H1: "Insiders systematically misreport information about earnings."

Assuming that H1 cannot be rejected, we go a step further and study whether manipulation is intended to affect market prices so insider's profits on their trading operations increase. For this purpose we revise the incentives' underlying equation (1). Given p_{t-1} and $d_t - s_t$, the best strategy for an insider that maximizes the profits in her trading operations is obtained by maximizing $|\tilde{r}_t - r_t|$ and thus reporting a 20% increase/decrease in earnings whenever earnings decrease/increase and, consequently, sell/buy an action when she expects an increase/decrease in the stock price. Of course the final effect on the stock price depends on the strategies of all players and their beliefs about the expected behavior of their opponents. Thus, it seems reasonable to test the hypothesis H2 about the information manipulation strategy of the insiders.

H2: "Insiders report high increases/decreases in earnings when earnings have decreased/increased and they intend to sell/buy stocks at a high/low price".

The insider strategic behavior hypothesized in H1 and H2 has negative externalities on others' profits, since outsiders decide about their trading based on false information and the insiders exploit their mistake-induced decisions profitably, then we are interested in testing hypothesis H3.

H3: "Earnings management positively/negatively affects insiders'/outsiders' returns. Thus, in the case of earnings management, insiders' returns are higher than average outsiders' returns".

(b) ***Hypotheses related to stock-based compensation:***

Our experiment also explores the insiders' behavior under the implementation of a stock-based remuneration policy and its effects on insiders' profitability and the firm value. The resulting hypotheses are stated in H4 and H5.

H4: "The stock-based remuneration policies positively affect insiders' returns (even without considering the bonus). Thus, Insiders' returns are higher in the presence of this manager compensation scheme".

H5: "The stock-based remuneration policies positively affect firm value. Thus, share prices rise more in the presence of this manager compensation scheme".

Hypothesis H2 is based on the idea that the main source of insiders' profit comes from their trading. Nevertheless, their optimal strategy might not be the same if we include another alternative manager remuneration policy. For example, under our stock-based compensation scheme the strategy that maximizes bonuses consists of trying to increase the stock price (e.g., by reporting high earnings and buying stocks) until they get the bonus and doing the opposite strategy just when they have received it in order to maximize the next bonus. Therefore, for this particular stock-based compensation scheme the following hypothesis seems plausible:

H6: "Under stock-based compensation earnings management is driven by bonus maximization, then insiders report increases/decreases on earnings before/immediately after getting the bonus".

3 Results

Table 1 displays some statistics at the group (market) level in T0 (baseline treatment). These data give a clear idea about the behavior of both insiders and outsiders during the experiment and, particularly, at the end of it. The data are disaggregated at the group level and the data for the three outsiders of every group are averaged. Consistently with hypothesis H3, the insider's wealth is higher (599 ECUs) than the average wealth of the outsiders (549 ECUs) despite the fact that the average number of stocks for both subject types are the same (8). This fact gives an intuition about the speculation source of the extra gains of the insider. On the other hand the stock price has sharply increased in all markets from 20 ECUs to 60 ECUs on average, as a consequence of the higher number of purchases than sales orders (530 and 413, respectively) and the messages about reported earnings, which accounts for misleading information 82% of the time (see hypothesis H1). It is also revealing that the majority of the messages (63%) correspond to the optimal manipulation strategies under hypothesis H2 (i.e. "20% earnings increase" and "20% earnings decrease").

Table 1. Descriptive statistics at the group level in T0 (without bonus)

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Total*
Insider's wealth at the end of the experiement	572	495	441	387	694	784	466	831	620	700	599
Outsider's average wealth at the end of the experiment	486,67	511,33	419	439	827	558,33	432,33	748,67	519,67	548	549
Average insider's weath during the experiment	322,45	355,73	341,28	366,73	441,23	455,8	367,85	540,48	426,65	457,65	407,585
Average outsider's wealth during the experiment	317,23	370,72	342,03	399,78	511,5	357,46	344,48	499,13	375,84	397,47	391,564
Insider's stocks at the end of the experiment	9	9	10	0	5	11	8	9	9	10	8
Outsider's average stocks at the end of the experiment	8	9	8	10	8	7	9	7	7	6	8
Average insider's stocks during the experiment	7	7	8	4	6	9	8	9	7	8	7
Average outsider's stocks during the experiment	6	8	7	8	7	6	7	7	6	6	7
Price at the end of the experiment	56	50	40	40	94	63	40	82	66	68	60
Average price during the experiment	35,25	41,1	36,85	43,35	59,55	39,55	36,08	57,4	48,28	49,6	44,7
Total number of purchases	64	57	52	49	49	61	51	44	44	59	530
Total number of sales	52	41	39	39	40	48	35	33	35	51	413
Total number of non-trade situations	44	62	69	72	71	51	74	83	81	50	657
Number of restrictions to purchases due to liquidity constraints	6	17	6	14	40	6	16	17	3	6	131
Times where the insider reports earnings truthfully	4	6	5	8	6	5	18	6	8	9	75
Times where the insider reports a 20% earnings increase	12	11	13	15	32	19	5	24	13	14	158
Times where the insider reports a 20% earnings decrease	8	9	13	13	4	8	17	13	6	4	95

* The 10 first rows are the average across groups. The last 7 rows are the sum for the 10 groups.

Table 2 displays descriptive statistics at the group (market) level in T1 (i.e. implementing a stock-based compensation scheme). This data highlights the much higher (931.7 ECUs) average wealth of insider's than that of the outsiders (614.97 ECUs). This comparison does not include the gains corresponding to the extra bonus (hypothesis H4). It is also noteworthy that these figures have increased considerably from treatment T0 as a consequence of the higher increase of the stock price (69 ECUs on average) (hypothesis H5). It is also revealing that in this case the insiders have on average more stocks (13) than the outsiders (8), while in the other treatment both had 8 stocks on average at the end of the experiment (this also explains the differences in profitability). It is also noteworthy that although the number of purchases has not increased in T1 (with respect to T0), the difference between purchases and sales is higher in this treatment. Furthermore, the misreporting behavior has been reduced from 82% of the cases to 73%, and also the leading message in T1 is a “20% earnings increase” (45% of all the cases). All of this is consistent with the optimal manipulation strategy in T1 (see hypothesis H6) directed to increase the stock price in order to maximize the bonus.

Table 2. Descriptive statistics at the group level in T1 (with bonus)

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Total*
Insider's wealth at the end of the experiment (with bonus)	992	987	919	869	667	816	1011	555	1435	1446	969,7
Insider's wealth at the end of the experiment (without bonus)	967	927	874	854	622	751	1011	555	1380	1376	931,7
Outsider's average wealth at the end of the experiment	607,33	750,67	587	695,33	410,33	473,67	779	435	727,33	684	614,97
Average insider's wealth during the experiment (with bonus)	552,73	589,08	401,03	560,8	434,9	386,1	608,33	396,05	703,38	636,5	526,89
Average insider's wealth during the experiment (without bonus)	548,85	583,58	396,9	557,05	431,65	381,23	600,33	394,05	697	629,63	522,03
Average outsider's wealth during the experiment	396,78	480,61	332,42	465,52	321,17	285,38	524,26	338,44	427,28	394,98	396,68
Insider's stocks at the end of the experiment	13	10	13	10	13	16	11	13	13	14	13
Outsider's average stocks at the end of the experiment	8	8	9	9	5	9	8	9	6	7	8
Average insider's stocks during the experiment	10	7	9	8	9	10	7	9	10	11	9
Average outsider's stocks during the experiment	7	7	7	8	6	7	7	8	5	6	7
Price at the end of the experiment	69	84	62	72	45	41	87	40	103	91	69
Average price during the experiment	46	56,1	35,68	52,25	36,58	25,73	60,8	34,95	59,33	49,05	45,65
Total number of purchases	42	52	52	38	53	51	47	61	49	52	497
Total number of sales	26	37	33	22	44	28	33	42	38	38	341
Total number of non-trade situations	92	71	75	100	63	81	80	57	73	70	762
Number of restrictions to purchases due to liquidity constraints	5	23	10	22	9	3	14	12	3	43	144
Times where the insider reports earnings truthfully	18	7	16	17	9	14	3	7	9	8	108
Times where the insider reports a 20% earnings increase	16	21	21	12	16	14	17	7	35	20	179
Times where the insider reports a 20% earnings decrease	3	1	13	2	7	18	0	9	4	4	61

* The 10 first rows are the average across groups. The last 7 rows are the sum for the 10 groups.

4 Conclusions

In this paper, we have designed a laboratory experiment to study the behavior of insiders in relation to their earnings management and trading practices. The experiment simulates a market were one insider and three outsiders trade on the stocks of a company. Their benefits come from the capital gains of their trading but also from the dividends and the increase in the price of their portfolio. We also assume that the stock price depends on the excess of supply/demand, but also on the difference between reported and real earnings of the firm (surprises). Under these basic assumptions if the insider reports the firm's earnings and the real earnings are not known by the outsiders (although ex post they may guess them from the realized prices) we show two main results. Firstly, insiders misreport the information for the purpose of affecting the price in the direction that allows them to increase their capital gains. Therefore, earnings management induces inefficiencies on the stock prices, but also the misleading information reduces the outsiders' profits, which are significantly lower than those of the insiders. Secondly, managerial stock-based compensation policies reinforce these results although the insiders' earnings management strategies are driven by bonuses rather than by capital gains maximization. This fact increases insiders' profits even discounting the effects of the bonuses but also raises stock prices.

The results are obtained in a very simple scenario, but highlight how managers behave if earnings management, insider trading and stock-based compensation are not accurately regulated and supervised. The lack of transparency in earnings management practices and the wrongly incentivized managerial compensation policies have an important role on generating price inefficiencies, bubbles and financial crises. In fact, the current economic-financial crisis is also (and perhaps above all) a crisis of confidence in corporations and securities markets. Recovering that confidence will inevitably require corporate information to be more credible and all the agents involved in the financial reporting process to be subject to scrutiny.

Acknowledgements. Javier Perote is grateful to the Junta de Castilla y León for financial support (project SA218A11_1). Esther B. Del Brio is also grateful to the Research Agency of the Spanish Government (Grant ECO2010-20741) for funding.

References

1. Abdolmohammadi, M., Sultan, J.: Ethical Reasoning and the Use of Insider Information in Stock Trading. J. Bus. Eth. 37, 165–173 (2002)
2. Bergstressera, D., Philipponb, T.: CEO Incentives and Earnings Management. J. Fin. Econ. 80, 511–529 (2006)
3. Fischbacher, U.: Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. Exp. Econ. 10, 171–178 (2007)

Service-Oriented Architectures: From Design to Production Exploiting Workflow Patterns

Maurizio Gabbrielli[1], Saverio Giallorenzo[1], and Fabrizio Montesi[2]

[1] Dipartimento di Informatica - Univ. di Bologna / INRIA, Bologna, Italy
[2] IT University of Copenhagen, Denmark

Abstract. In Service-Oriented Architectures (SOA), services are composed by coordinating their communications into a flow of interactions. Coloured Petri nets (CPN) offer a formal yet easy tool for modelling interactions in SOAs, however mapping abstract SOAs into executable ones requires a non-trivial and time-costly analysis. Here, we propose a methodology that maps CPN-modelled SOAs into Jolie SOAs (our target language), exploiting a collection of recurring control-flow patterns, called Workflow Patterns, as composable blocks of the translation. We validate our approach with a realistic use case. In addition, we pragmatically asses the expressiveness of Jolie wrt the considered WPs.

1 Introduction

Service-Oriented Computing (SOC) is a design methodology focused on the realisation of systems by composing autonomous entities called *services*. In a Service-Oriented Architecture [1] (SOA), services are composed by coordinating their communications into a flow of interactions. Several tools have been presented [2–4] to assist the process of SOA design, each focusing on a particular aspect of the system, e.g., the architectural composition, the interaction among components, etc. Coloured Petri nets [5] (CPNs) are a formal yet intuitive graphical tool, largely employed in business process modelling [6] and suitable for SOA specification. Although it is easy to understand the interactions of a CPN model, it is unclear which components form the system, which implement the described logic or whether it be spread among the components or centralised.

Therefore the aim of this work is to provide a methodology that allows the translation of CPN-modelled SOAs into executable ones.

The *Workflow Patterns Initiative* (WPI) studied and collected a comprehensive set of recurring patterns of process-aware information systems, dubbed *Workflow Patterns* [6] (WP). In particular we remark the exhaustive set of patterns of interaction, dubbed *Control-Flow Workflow Patterns* [7][1], modelled as CPNs. Since CPNs are composable, our idea, depicted by the scheme in Fig. 1, is that an SOA, modelled as a CPN, can be described in terms of the Workflow Patterns it is made of. Once the SOA is defined by a composition of WPs, the developer only has to refer to the implementation of each WP to build the whole system.

[1] Here referred as Workflow Patterns for simplicity.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_17,

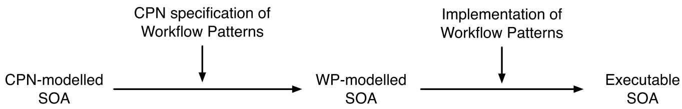


Fig. 1. The scheme of translation from abstract to executable SOAs

To realise our proposal, we provide the implementation of a substantial set of WPs. Such implementation is not immediate since WPs are abstract specifications and it is unclear how they map into executable code for service coordination. Moreover, although the same WP applies to different subnets of interactions, its implementation may differ sensibly depending on whether its logic is *centralised* in a single component or *distributed* among several ones. Centralised and distributed approaches suit different contexts. E.g., if a vendor wants to monitor its application he might prefer a single point of control to track the whole system. On the other hand, some scenarios strictly require a distributed approach, e.g., an interaction that comprehends different parties. In § 3 we consider a realistic use case that combines the two approaches.

We translate both the centralised and distributed versions of WPs as composable and executable SOAs. In order to provide a consistent translation we also define a *procedure* in § 2. Notably, such procedure might directly map a CPN-modelled SOA to an executable one, thus skipping the said in-between translation to a WP-modelled SOA. However, the behaviour of some WPs needs ad-hoc solutions (see Table 1) not directly mapped by the presented procedure. Thus, although providing an automatic procedure is an interesting challenge, in this work we focus on the practical implications of enabling developers translate CPN-modelled SOAs into executable ones by referring to our collection of Workflow Patterns. Our procedure applies to any service-oriented language, e.g., BPEL [8] but we choose to implement WPs in Jolie [9,10] for two main reasons. First Jolie supports several communication and serialisation protocols, thus the same implementation applies to different application domains. Second Jolie is based on a formal process calculus [11] which we plan to use to prove relevant correctness properties of translated SOAs. For reasons of presentation we delegate to the full paper [12] the analysis and the implementation of all considered WPs (*basic* and *advanced branching and synchronisation* patterns).

2 From Coloured Petri Nets to Jolie SOAs

Background. We assume the reader is familiar with the basic concepts of Petri nets on which Coloured Petri nets add expressions on arcs and types to tokens[2]. In order to understand our translation procedure we briefly describe the message-passing operations in Jolie. The language supports two kinds of message-passing operations which send or receive the content of a variable v. Sending operations

[2] [12] reports a comprehensive introduction on Coloured Petri nets.

specify the location L the message is sent to. *One-Way* operations (OWs) send or receive a message and immediately pass the thread of control to the subsequent activity in the process. The syntax of a receiving OW is simply $op\_name(v)$, whilst a sending OW adds the location of the request $op\_name@L(v)$. *Request-Response* operations (RRs) either send a request and keep the thread of control until they receive a response or they receive a message, do some computation, and send a response. The syntax of a sending RR is $op\_name@L(v)(v)$, whilst a receiving RR adds a pair of brackets that enclose the code run before sending the response to the invoker $op\_name(v)(v)\{code\}$. Jolie provides also an input choice with syntax $[\eta_1]\{B_1\}\ldots[\eta_n]\{B_n\}$. When an input operation η_i is triggered it disables all $\{\eta_1,\ldots,\eta_n\}\backslash n_i$ and executes. Then B_i executes.

Workflow Patterns in Jolie. In this section we present a procedure for translating a CPN-modelled Workflow Pattern into an executable Jolie SOA. The translation follows five principles: (*i*) transitions are services; (*ii*) places are message-passing operations (i.e., communications); (*iii*) communications carry typed messages, as coloured tokens do; (*iv*) arcs are properties on communications: they express the type of carried messages and the conditions that fire the communication; (*v*) a CPN models a WP composed by several services running in parallel. Following these principles, we translate CPN models of Workflow Patterns into Jolie SOAs as follows. We map input, internal, and output places into One-Way (OW) operations (principle *ii*). When it is compatible with the behaviour of the pattern, we coalesce two round-trip OW operations between two services into one Request-Response (RR) for brevity. Since in Jolie output operations define the service they communicate to, we map output places into OWs on default locations `DefaultOutput1,...`. This enables composition of the implemented patterns by binding their locations. As exposed in § 1 we translate both the centralised and the distributed versions of WPs. In the centralised approach a master service, called *orchestrator* encodes the whole behaviour of a WP coordinating the interactions among the services participating to the SOA. BPEL [8] is the most known technology for this approach, called *orchestration*. By convention the orchestrator of a WP is the only service that receives and sends messages outside the SOA. The distributed approach recalls that of *choreography* languages like WS-CDL [13]. Recent works [14,15] introduced automatic techniques to project executable services of an SOA from a choreographic specification. Following a similar approach in the distributed version of a WP we maintain a direct relation between transitions and services, imposing no restrictions on the scope of external input and output operations. Fig. 2 reports an example of a CPN in box **A** and informally depicts the architectural view of the translated centralised (**B**) and distributed (**C**) implementations. Listing 1.1 reports the corresponding code of, respectively, the orchestrator of the centralised version and of the services in the distributed one. In **B** the orchestrator (`O`) receives the input message as a OW operation `i1` (Line 1). `O` sends via RR the content `c` to service `A` which returns the condition `cond1` (Line 2). Based on `cond1`, `O` either redirects a request to `B` or `C` (Lines 4-8). `O` sends `c` to `D`, which returns its response. (Line 9) Finally `O` sends `c` as output to the `DefaultOutput1`.

(Line 10). Notably, `O` uses RRs to invoke operations on the services its composes, waiting for their responses. **C** maintains a direct relation between transitions and services which pass the thread of control using OW operations. Service `A` receives the input message `i1` (Line 1), evaluates condition `cond1` and redirects `c` to either service `B` (Line 3) or `C` (Line 4) which, in turn, forwards `c` to service `D`. Finally `D` receives `c` on operation `p3` and sends it to the output location `DefaultOutput1` (Line 10).

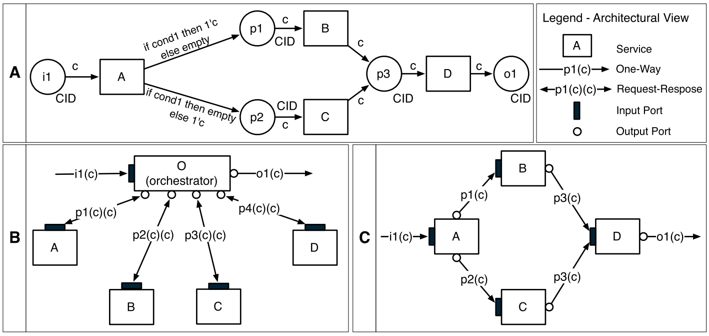


Fig. 2. (**A**) a CPN model, its centralised (**B**), and distributed (**C**) realisations

Listing 1.1. Centralised (right) and distributed (left) implementations of CPN **A**

```
1   //orchestrator
2   i1( c );
3   p1@A( c )( cond1 );
4   if( cond1 ){
5       p2@B( c )( c )
6   } else {
7       p3@C( c )( c )
8   };
9   p4@D( c )( c );
10  o1@DefaultOutput1( c )
```

```
1   //service A
2   i1( c );
3   if( cond1 ){
    p1@B( c ) }
4   else {  p2@C( c ) }
5   // service B
6   p1( c ); p3@D( c )
7   // service C
8   p2( c ); p3@D( c )
9   // service D
10  p3( c ); o1@DefaultOutput1( c )
```

3 The Upload Service Use Case

This section shows how a realistic SOA, modelled as a Coloured Petri net, translates into an executable SOA. The goal of this section is twofold. (*i*) We show the benefits of CPNs as SOA design tool. CPNs let the designer model the system focusing on interactions, whilst the control on message flow is left to the developer. (*ii*) We exhibit how a developer can easily map a CPN into a composition of WPs. With just the flow of interactions as model, the developer can mix distributed and centralised implementations of patterns and build the system.

The use case describes a typical interaction between a *User*, a file upload *Service Provider*, and an *Identity Provider*. In particular the Identity Provider offers an OpenID-like [16] authentication with a multi-factor mechanism [17]

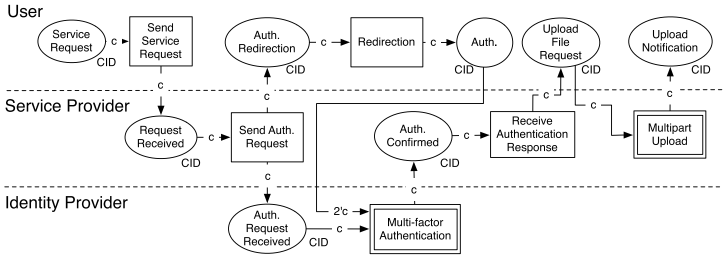


Fig. 3. The CPN model of the interactions in the Upload Service SOA

whilst the Service Provider offers a multipart upload procedure. Fig. 3 depicts the CPN model of the use case. The two double-line bordered boxes in the figure act as placeholders for the two subnets related to the multi-factor authentication and the Multipart Upload procedure, the latter modelled in Fig. 4.

Let us identify the WPs[3] that compose the CPN-modelled SOA. Interaction starts from the User that requests the service. This basic interaction is a distributed *Sequence* that passes the thread of control to the Service Provider. The Service Provider employs a distributed *Parallel Split* to start the multi-factor authentication and redirects the authentication request to the User. The Identity Provider allows users to identify themselves with a multi-factor authentication. Let us suppose that the Identity Provider demands a minimum number of different authentication procedures. Such *M-out-of-N* mechanism maps to the centralised version of the *Cancelling Partial Join* pattern being completely controlled by the Identity Provider. For reasons of presentation we choose not to discuss about the implementation of this pattern as it maps directly and does not show interesting interactions with other patterns. For the model and the code relative to the multi-factor authentication refer to [12]. After the successful authentication, the thread of control passes back to the Service Provider with another distributed *Sequence* which notifies the User (s)he can proceed to upload the file. The User and the Service Provider enter the Multipart Upload interaction whose behaviour results from the composition of several patterns. Fig. 4 depicts such interactions and highlights the most relevant WPs. Fig. 5 depicts the architectural view of the translation following the same informal representation used in Fig. 2. The User-controlled part of the interaction mixes centralised and distributed WPs. Listing 1.2 reports the code relative to the services `orchestrator` and `SendChunks` at User's side. When the `uploadRequest` arrives (Line 1), the orchestrator requires the User to select a file, passing the thread of control as a centralised *Sequence* to service `SelectFile` (Line 2). At file selection, the thread of control returns to the orchestrator which passes it to service `CreateChunks` (Line 3). The service employs a centralised *Thread Split* (A) to split the file into `n` chunks. Then the orchestrator implements a centralised *Thread Merge* (B) to collect triplets of chunks and send them to service `SendChunks` (Lines

[3] [12] reports the description of the considered patterns and their implementations.

5-7). Notably, since the orchestrator passes the thread of control to the invoked service and waits for its response, we can coalesce the OW operations between them into one RequestResponse. `SendChunks` implements a distributed *Parallel Split* to forward each chunk in parallel to the Service Provider (Lines 11-13). At Service Provider's side the service `StoreChunks` employs a centralised *Generalised AND-Join* (C) to receive the chunks (Listing 1.3 Lines 1-13). When the `n`th chunk reaches the service, it passes the thread of control with a distributed *Sequence* to service `ComposeFile` which employs a centralised *Thread Merge* (D) to restore the chunks into a single file. Finally a distributed *Sequence* returns the thread of control to the User, notifying the success of the upload procedure.

Listing 1.2. User's side

```
1  // orchestrator
2  uploadRequest(c);
3  selectFile@SelectFile(c)(c);
4  createChunks@CreateChunks(c)(c);
5  for( i=0, i<#c, i++ ){
6        r.c1=c[i++];r.c2=c[i++];r.c3=c[i];
7        sendTriplet@SendChunks(r)()
8  }
9  // SendChunks
10 sendTriplet(c)(){
11       sendFileChunk1@StoreChunk(c.c1)
12       | sendFileChunk2@StoreChunk(c.c2)
13       | sendFileChunk3@StoreChunk(c.c3)
14 }
```

Listing 1.3. Service Provider's side

```
1  // StoreChunks
2  [ sendFileChunk1(c) ]{
3      // store file chunk on queue 1
4      // check for upload completion
5  }
6  [ sendFileChunk2(c) ]{
7      // store file chunk on queue 2
8      // check for upload completion
9  }
10 [ sendFileChunk3(c) ]{
11     // store file chunk on queue 3
12     // check for upload completion
13 }
```

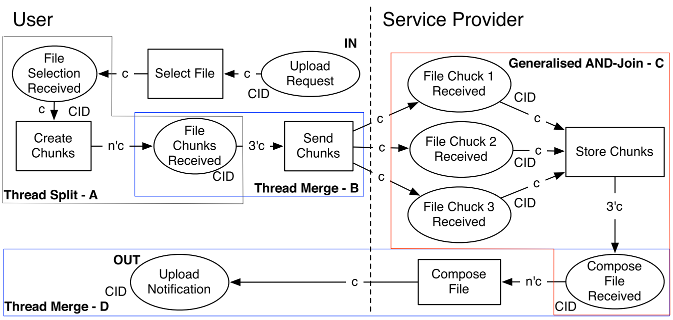


Fig. 4. Multipart Upload modelled as composition of Workflow Patterns

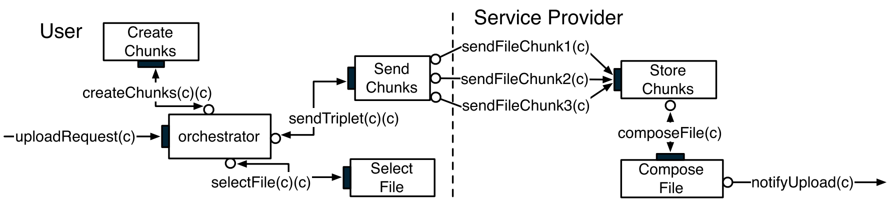


Fig. 5. The architectural view of Multipart Upload in Fig. 4

4 Conclusions

Contributions of this work are: (*i*) the definition of a methodology for translating CPN-modelled SOAs into composable and executable ones and (*ii*) the creation of a collection of implemented Workflow Patterns (reported in [12]). Such implementations follow both a centralised and a distributed approach to allow developers the flexibility to choose and mix them. A realistic use case proves our claim that the patterns obtained in this way can be used for building real SOAs starting from abstract specifications. In addition, (*iii*) our work also allows us to provide a pragmatic assessment on the expressiveness of the Jolie language. Table 1 summarizes the results of such an assessment. For each pattern, we indicate in the second column the kind of support offered by Jolie: "+" means direct support, i.e., the implementation of the pattern either uses some specific primitives provided by the language or is a composition of directly supported patterns. "+/–" indicates a "non direct" support, i.e., the translation of the CPN of the pattern does not completely follow the rules described in § 2 although it complies with the general structure of the pattern. In the third column of Table 1 we indicate the specific Jolie primitive and/or the other patterns used to implement a given pattern. Note that we report both the centralised and distributed implementations which, as expected, in some cases vary. As shown in Table 1 we can conclude that Jolie allows to implement most WPs.

Table 1. Support of *basic* and *advanced branching and synchronisation* WPs in Jolie

Workflow Pattern	Jolie Support	Supported by or main components	
		centralised	distributed
Sequence	+	`sequence` operator	
Parallel Split	+	`parallel` operator	
Synchronization	+	Parallel Split, scopes	
Exclusive Choice	+	`if … else`, input choice	
Simple Merge	+	Synchronization, `synchronized` scope	Sequence, `execution{sequential}`
Multi-Choice	+	Parallel Split, Exclusive Choice	
Thread Split	+	iteration, recursion, and Parallel Split, `spawn`	
Generalized AND-Join	+/-	Synchronization, input choice, and ad-hoc queues	
Multi-Merge	+	Synchronization	Simple Merge, `execution{concurrent}`
Thread Merge	+	iteration, multiple instances	
Structured/Local Synchronizing Merge	+	Multi-Choice, Synchronization	
Generalized Synchronizing Merge	+/-	Structured Synchronizing Merge	
Structured Patrial Join	+	Synchronization, Thread Merge	Thread Merge, Sequence
Blocking Partial Join	+/-	Generalized AND-Join, Structured Partial Join	
Cancelling Partial Join	+	Structured Partial Join	

Related Work. A close concept to Workflow Pattern is that of *service interaction pattern*, introduced in [18]. Service interaction patterns define recurring interaction patterns among services but, differently from Workflow Patterns, they are informally specified and therefore not employable in this work. Variants of Petri nets have been used for system modelling [19] and static analysis [20]. Finally WPI used WPs as a tool to evaluate the expressive power of business process languages, in particular for the cases of BPEL [21] and BPML [22].

Future Work. We plan to provide a formal definition of our technique for translating CPNs into Jolie code. Such a formalisation would enable to mechanically translate CPN-modelled SOAs into executable ones, also applying known methodologies of static analysis to assess properties of SOAs implemented in Jolie. We also plan to investigate the remaining patterns described by the WPI and to use the implemented WP to offer pattern composition as APIs [23].

References

1. Erl, T.: Service-Oriented Architecture: Concepts, Technology, and Design. PH (2005)
2. OMG, Service oriented architecture Modeling Language (2009)
3. OASIS, Reference architecture foundation for SOA version 1.0 (December 2012)
4. Mayer, P., Koch, N., Schroeder, A.: The UML4SOA Profile (July 2009)
5. Jensen, K., Kristensen, L.M.: Coloured Petri Nets - Modelling and Validation of Concurrent Systems. Springer (2009)
6. Van Der Aalst, W.M.P., Ter Hofstede, A.H.M., Kiepuszewski, B., Barros, A.P.: Workflow patterns. Distrib. Parallel Databases 14, 5–51 (2003)
7. Russell, N., Hofstede, A.H.M.T., Mulyar, N.: Workflow control-flow patterns: A revised view. Tech. Rep. (2006)
8. OASIS, BPEL, http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.html
9. Jolie Website, http://www.jolie-lang.org/
10. Montesi, F., Guidi, C., Zavattaro, G.: Service Oriented Programming with Jolie. Web Services Foundations, vol. 1
11. Guidi, C., Lucchi, R., Gorrieri, R., Busi, N., Zavattaro, G.: Sock: A calculus for service oriented computing. In: Dan, A., Lamersdorf, W. (eds.) ICSOC 2006. LNCS, vol. 4294, pp. 327–338. Springer, Heidelberg (2006)
12. Gabbrielli, M., Giallorenzo, S., Montesi, F.: Executable SOAs exploiting workflow patterns. Tech. Rep., http://www.cs.unibo.it/~sgiallor/p/eSOAs.html
13. W3C WS-CDL Working Group, Web services choreography description language version 1.0 (2004), http://www.w3.org/TR/ws-cdl-10/
14. Carbone, M., Montesi, F.: Deadlock freedom by design: multiparty asynchronous global programming. SIGPLAN Not. 48, 263–274 (2013)
15. Dalla Preda, M., Lanese, I., Mauro, J., Gabbrielli, M., Giallorenzo, S.: Deadlock freedom by construction for distributed adaptative applications. Tech. Rep., http://www.cs.unibo.it/projects/jolie/aioc.pdf
16. OpenID, Specifications, http://openid.net/developers/specs/
17. Multi-factor authentication, http://aws.amazon.com/iam/details/mfa/
18. Barros, A., Dumas, M., ter Hofstede, A.H.M.: Service interaction patterns. In: van der Aalst, W.M.P., Benatallah, B., Casati, F., Curbera, F. (eds.) BPM 2005. LNCS, vol. 3649, pp. 302–318. Springer, Heidelberg (2005)
19. Mendes, J., Leitao, P., Restivo, F., Colombo, A.: Composition of petri nets models in service-oriented industrial automation. In: INDIN 2010, pp. 578–583 (2010)

20. Lohmann, N., Kopp, O., Leymann, F., Reisig, W.: Analyzing bpel4chor: verification and participant synthesis. In: Dumas, M., Heckel, R. (eds.) WS-FM 2007. LNCS, vol. 4937, pp. 46–60. Springer, Heidelberg (2008)
21. Wohed, P., van der Aalst, W.M.P., Dumas, M., ter Hofstede, A.H.M.: Analysis of web services composition languages: The case of bpel4ws. In: Song, I.-Y., Liddle, S.W., Ling, T.-W., Scheuermann, P. (eds.) ER 2003. LNCS, vol. 2813, pp. 200–215. Springer, Heidelberg (2003)
22. van der Aalst, W.M.P., Dumas, M., ter Hofstede, A.H.M., Wohed, P.: Pattern-based analysis of BPML (and WSCI). FIT-TR-2002-05 (2002)
23. Guidi, C., Giallorenzo, S., Gabbrielli, M.: Towards a composition-based APIaaS layer. CLOSER 2014. SciTePress (to appear, 2014)

Reinforcement Learning Based on the Bayesian Theorem for Electricity Markets Decision Support*

Tiago M. Sousa[1], Tiago Pinto[1], Isabel Praça[1], Zita Vale[1], and Hugo Morais[2]

[1] GECAD – Knowledge Engineering and Decision-Support Research Center, Institute of Engineering – Politechnic of Porto (ISEP/IPP), Porto, Portugal
{rifer,gajls,icp,tmcfp,ifdsp,zav}@isep.ipp.pt
[2] Automation and Control Group – Technical University of Denmark, Denmark
morais@elektro.dtu.dk

Abstract. This paper presents the applicability of a reinforcement learning algorithm based on the application of the Bayesian theorem of probability. The proposed reinforcement learning algorithm is an advantageous and indispensable tool for ALBidS (Adaptive Learning strategic Bidding System), a multi-agent system that has the purpose of providing decision support to electricity market negotiating players. ALBidS uses a set of different strategies for providing decision support to market players. These strategies are used accordingly to their probability of success for each different context. The approach proposed in this paper uses a Bayesian network for deciding the most probably successful action at each time, depending on past events. The performance of the proposed methodology is tested using electricity market simulations in MASCEM (Multi-Agent Simulator of Competitive Electricity Markets). MASCEM provides the means for simulating a real electricity market environment, based on real data from real electricity market operators.

1 Introduction

The Bayes' Theorem has been applied in several scopes throughout the times, taking advantage on this probability theory's capabilities of supporting applications directed to the most alternative contexts. One of the many uses of this theorem is its applicability to be used as a reinforcement learning algorithm [1]. This applicability is based on the use of a probability estimation to determine which of the different action alternatives presents a higher probability of success in each context, therefore being considered as the most suitable action to be taken.

Bayesian networks have been developed to facilitate the task of prediction and abduction in artificial intelligence systems. Simplistically, these networks, also known as causal networks, or probabilistic networks, are graphic models for uncertainty based reasoning [2]. A Bayesian network is represented by a directional acyclic graph in which the knots represent domain variables, and the arches represent the

* This work is supported by FEDER Funds through COMPETE program and by National Funds through FCT under the projects FCOMP-01-0124-FEDER: PEst-OE/EEI/UI0760/2011, PTDC/SEN-ENR/122174/2010 and SFRH/BD/80632/2011 (Tiago Pinto PhD).

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_18, © Springer International Publishing Switzerland 2014

conditional or informative dependency between the variables. In order to represent the dependency strength, probabilities are used, associated to each network's group of parent-son knots. The values represented by the domain variables must be mutually exclusive and exhaustive, *i.e.* a variable must always assume only one of those values. Usually these values are boolean, enumerated, or numeric values. All values must efficiently represent the domain, but with enough detail to be distinguished.

Two knots must be connected directly, in case of one affecting the other, with the arch's direction pointing the effect direction. Once the network's topology is defined, the relationships between the connected knots must be quantified, specifying the conditional probability distribution for each knot [3], as in (1).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

This conditional probability provides the basis for the propagation of each knot's probability given the facts observed by each variable that influences them. The expected utility of the action each knot represents, is given by (2), being E the available evidences, A an action with possible outcomes O_i, $U(O_i|A)$ the utility of each of the outcome states given that action A is taken, $P(O_i|E,A)$ the conditional probability distribution over the possible outcome states, given that evidence E is observed and action A taken.

$$EU\left(A|E\right) = \sum_i P(O_i \mid E, A) \times U(O_i \mid A) \tag{2}$$

This paper proposes the application of the Bayes' theorem of probability as a reinforcement learning algorithm. The proposed reinforcement learning algorithm is integrated in the ALBidS. ALBidS uses several different strategies for providing decision support to market players. These strategies are used accordingly to their probability of success for each context. The approach proposed in this paper uses the adaptation of a Bayesian network to decide which is the action that presents the higher probability of success at each time, for each context, depending on past events. The proposed methodology is tested by providing decision support to an electricity market agent acting in a simulated electricity market, using MASCEM.

2 Adaptive Multi-Agent Simulation of Electricity Markets

Electricity markets worldwide suffered profound transformations. The privatization of previously nationally owned systems; the deregulation of privately owned systems that were regulated; and the upcoming integration of national systems, are some examples of such transformations [4]. In general, competitive environments, as is the present case of electricity markets, require good decision-support tools to assist players in their decisions. Relevant research is being undertaken in this field, namely in what concerns player modeling, simulation and decision-support.

The functioning of liberalized markets over the last years provides valuable information most of the times available to the community. Lessons can be learnt from the last years to improve knowledge about markets, to define adequate players' profiles and behaviors, but also to test and validate existing simulation tools, such as MASCEM [5, 6], making them suitable to represent the reality and provide the means for a coherent and realistic analysis of its evolution (or possible alternative pathways for the future of the electricity markets sector).

MASCEM [5, 6], is a tool that combines agent based modeling and simulation, with the characterization of agent objectives, dynamic strategies, and game theory for scenario analysis. MASCEM's purpose is to help understanding the complex and dynamic system behavior that emerges from the interactions of heterogeneous individuals, and support market players in their decisions in such an environment.

MASCEM's goal is to be able to simulate as many market models and players types as possible so that it can reproduce in a realistic way the operation of real electricity markets. This enables it to be used as a simulation and decision-support tool for short/medium term purposes but also as a tool to support long-term decisions, such as the ones taken by regulators.

There are several entities involved in the negotiations in the scope of electricity markets, MASCEM multi-agent model represents all the involved entities and their relationships, including: market facilitator agent, seller agents, buyer agents, VPP [7] agents, VPP facilitator agents, a market operator agent and a system operator agent.

Recently, a new multi-agent system, ALBidS was integrated with MASCEM [8] (Figure 1). ALBidS provides agents with the capability of analyzing different contexts of negotiation, such as the week day, the period, the particular market in which the player is negotiating, the economic situation and weather conditions, and automatically adapt their strategic behavior according to the current situation. In order to decide the best strategic behavior for the supported player to take in each situation, ALBidS uses reinforcement learning algorithms [9], to choose the most adequate from several different strategies according to each context. ALBidS' methodologies include: neural networks [10], data mining techniques [11], machine learning algorithms [8], Game Theory [12], competitor players' profiling, and strategies used by other simulators for market analysis and costs forecasts [9].

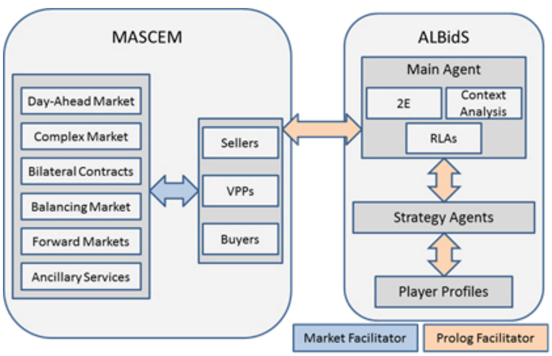


Fig. 1. Integration of ALBidS with MASCEM [7]

3 Bayes Theorem Reinforcement Learning Algorithm

The Bayes Theorem reinforcement learning algorithm applies the Bayes Theorem through the implementation of a Bayesian network. In this adaptation, each strategy (suggested action) is represented by a different knot. All strategy knots are connected to an output knot, which is responsible for the final decision on which is the most advantage action. This decision is based on the calculation of the success probability of each strategy, based on the observed events: if a strategy has accomplished to be the most successful amongst all in a certain negotiation period (event Yes), or not (event No). Figure 2 presents the topology of the Bayesian network, with an example considering three strategies.

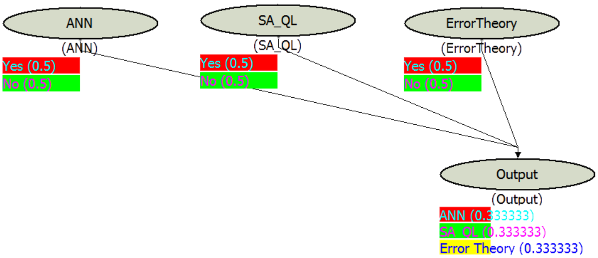


Fig. 2. Topology of the Bayesian network considering three strategies

At the initial point of a simulation all strategies must be initialized with the same probability of success, value that throughout the simulation shall be updated according to the performance of each. However, a Bayesian network does not represent a temporal probabilistic model. Therefore, it has been found necessary to use a dynamic Bayesian network. This dynamism has been achieved by contemplating the inclusion of the *Counting-Learning* algorithm.

The Counting-Learning algorithm considers, for each case about to be learned, an experience value, which defines in what degree the considered case will affect the conditional probability of each knot. This algorithm guarantees that all knots' probabilities are updated. For each new case, the updated experience value *exper'*, is calculated depending on the previous experience value *exper*, as presented in (3)

$$exper' = exper + degree \tag{3}$$

where *degree* represents the weight of the current case. The common approach (applied in this implementation) is defining the *degree* as assuming the value 1, meaning that each new case will be learnt as having the same importance as all the others. However, the *Counting-Learning* algorithm is prepared for other situations, in which the attribution of a higher value, *e.g.* 2, makes the algorithm learn this case as if it represented two similar cases at once (giving it the double of the importance). It can also be attributed a negative value, e.g. -1, so that it can "unlearn" a case.

The experience value and the *degree* are used in the updating process of a knot's probability. The probability update of a knot that has observed an event is done accordingly to (4). A new probability *probc'*, is calculated, taking into account the previous probability *probc*. The remaining knots *i* (for which the event was not observed) are updated according to (5), so that the probabilities vector (which contains all knots probabilities for each context) is kept normalized

$$probc' = \frac{(probc \times exper + degree)}{exper'} \tag{4}$$

$$probi' = \frac{(probi \times exper)}{exper'} \tag{5}$$

Once the network's dynamics are defined, the following step is to transform the network into the desired reinforcement learning algorithm, while guaranteeing the independence between probabilities concerning different contexts.

For the integration with the ALBidS system the proposed methodology requires the automatic creation of the knots (one input node for each strategy used by ALBidS, and the output node, which will determine the success probability of each strategy), and the connections between the nodes, so that its structure is built. Afterwards, the initial probabilities for each knot must be defined, in order to indicate the way each will affect the network. Once these initial definitions are completed, the network is able to be executed and interacted with, performing the probabilistic calculations and the probabilities propagation, including the *Counting-Learning* algorithm's application over the network.

The implementation of the presented models results in an enhanced reinforcement learning algorithm that learns from experience based on past events, taking into account the probability of success of each considered supporting strategy.

4 Case Study

In this case study three MASCEM simulations are presented, referring to 14 consecutive days, starting from Wednesday, 29th October. All three simulations use the same simulation scenario, a realistic scenario based on real data from MIBEL (Iberian market operator) [13], with the purpose of representing the MIBEL electricity market and its intervening players.

Agent Seller 2 is the test subject and will use a different strategy in each of the three simulations. In the first simulation it will use a regression on the data of the last 5 business days; in the second it will use a Neural Network (NN). Finally, in the third simulation, it will use the ALBidS system for decision support using the proposed reinforcement learning algorithm based on the Bayesian theorem of probability.

After the simulations, the incomes obtained by Seller 2 can be compared using each of the three tested strategies. This allows to take conclusions on the use of ALBidS equipped with the proposed reinforcement learning algorithm, over individual strategies, Seller 2 available offer capacity is kept constant at 550MW for each period throughout the simulations.

The evolution of the performance is analyzed for each period individually, for each period is considered by ALBidS as a different context. Therefore, the reinforcement learning algorithm's probabilities of success for each strategy are independent from one period to another. Figure 3 presents the incomes of Seller 2 in the first period of the day, along the 14 considered days. Figure 3 a) presents the results for the first simulation, Figure 3 b) presents the incomes of Seller 2 for the second simulation, and Figure 3 c) presents the incomes for the third simulation (Seller 2 using ALBidS using the proposed reinforcement learning algorithm).

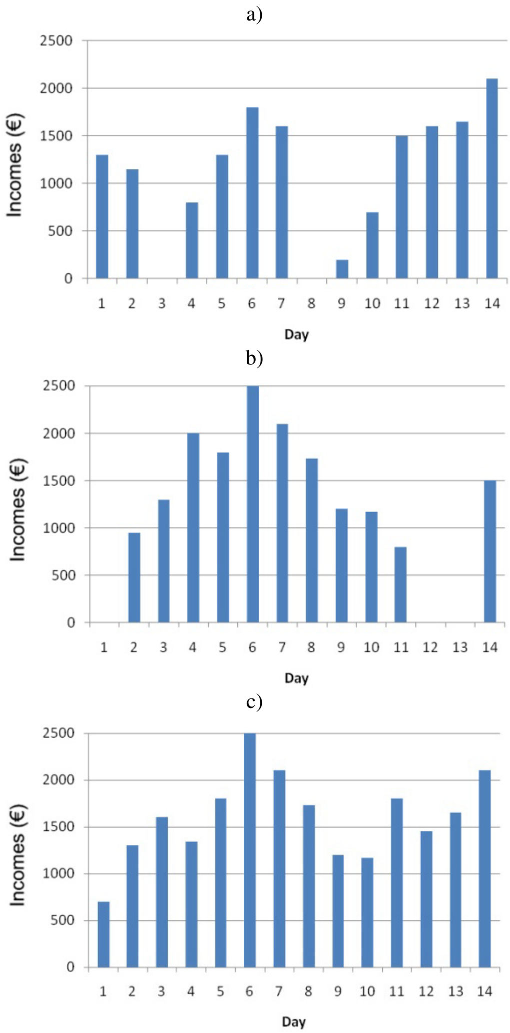


Fig. 3. Incomes obtained by Seller 2 in the first period of the considered 14 days, using: a) the regression on the data of the last 5 business days, b) the Neural Network, c) the ALBidS system with the proposed method

Comparing the charts presented in Figure 3, it can be seen that the third simulation was clearly the most profitable for Seller 2.

In the first day the income using the ALBidS system is located below the value of the income using the regression approach. That is because the probability values for all strategies have been initialized with equal values, and so the selection of the answer is made randomly. The selected algorithm answer originated a low income.

In the second day, after the reinforcement learning algorithm is updated, it chooses the strategy that had the greater success in the first day. Obviously the one that got the best reward, and it is visible that the income is located above both the other comparison strategies, as well as it happens on the third day.

In the fourth day its value is below the results of the NN; the algorithm is still selecting the other algorithm that got the best results in the first 3 days.

In the fifth day the value is equal to the NN, and it follows that trend until day 10, in spite of its lower values in days 9 and 10. In day 11 the chosen algorithm is no longer the NN, as the third simulation's value is higher than the ones of the other two considered strategies. Other algorithm passed the NN as the best result of success probability, as a result of the NN worst results in days 9 and 10.

In the last two days of the simulation, the selected algorithm is the regression, catching its high value tendency still on time to get two good final results.

Figure 4 presents the total amount of incomes obtained by Seller 2 in this period, using each of the three approaches.

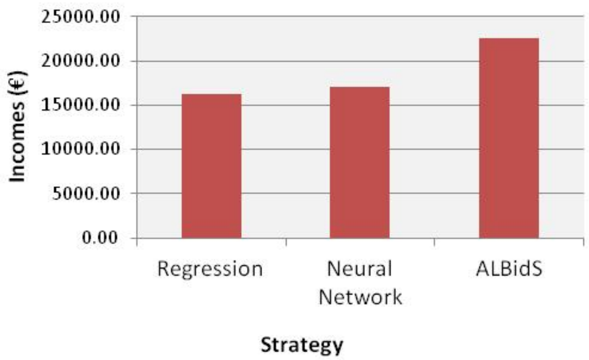


Fig. 4. Total incomes obtained by Seller 2 in the first period using each of the three approaches

From Figure 4 it is evident that from the considered approaches, ALBidS was the one that, by far, was able to provide the best results for Seller 2 in the considered period, using the learning capabilities provided by the success probability updating.

5 Conclusions

This paper proposed a new reinforcement learning algorithm based on the application of the Bayesian theorem of probability. The proposed approach was integrated in ALBidS, a multi-agent system that provides decision support to electricity market participating players; and was tested and validated using the MASCEM electricity markets simulator.

The proposed reinforcement learning algorithm implements a Bayesian network, in which each strategy (suggested action) is represented by a different knot. All strategy knots are connected to an output knot, which is responsible for the final decision on which is the most advantage action. This decision is based on the calculation of the success probability of each strategy, based on the observed events. At the initial point of a simulation all strategies are initialized with the same probability of success, value that throughout the simulation shall be updated according to the performance of each. In order to represent a temporal probabilistic model it has been found necessary to use a dynamic Bayesian network, which was achieved with the inclusion of the Counting-Learning algorithm.

The performed tests show the applicability and adequacy of the proposed method. As was shown, the methodology is able to adapt the confidence values of each different strategic proposal, and, as time progresses, suggest the alternative with the higher probability of success. This learning process is capable of leading a market agent into the achievement of higher profits, by choosing, at each time, the strategy, from those available, that presents the best guarantee of success.

References

1. Pinto, T., et al.: Multiagent System for Adaptive Strategy Formulation in Electricity Markets. In: IEEE Symposium on Intelligent Agents, SSCI-IA. IEEE Symposium Series on Computational Intelligence (2011)
2. Dore, A., Regazzoni, C.: Interaction Analysis with a Bayesian Trajectory Model. IEEE Intelligent Systems 25(3), 32–40 (2010)
3. Korb, K., Nicholson, A.: Bayesian Artificial Intelligence. Chapman & Hall/CRC (2003)
4. Shahidehpour, M., et al.: Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management, pp. 233–274. Wiley-IEEE Press (2002)
5. Praça, I., et al.: MASCEM: A Multi-Agent System that Simulates Competitive Electricity Markets. IEEE Intelligent Systems 18(6), 54–60 (2003)
6. Vale, Z., et al.: MASCEM: Electricity Markets Simulation with Strategic Agents. IEEE Intelligent Systems 26(2), 9–17 (2011)
7. Pinto, T., et al.: A new approach for multi-agent coalition formation and management in the scope of electricity markets. Energy 36(8), 5004–5015 (2011)
8. Pinto, T., Vale, Z., Rodrigues, F., Morais, H., Praça, I.: Strategic Bidding Methodology for Electricity Markets using Adaptive Learning. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K., Ali, M. (eds.) IEA/AIE 2011, Part II. LNCS (LNAI), vol. 6704, pp. 490–500. Springer, Heidelberg (2011)
9. Sun, J., Tesfatsion, L.: Dynamic Testing of Wholesale Power Market Designs: An Open-Source Agent-Based Framework. Computational Economics 30(3) (2007)
10. Amjady, N., et al.: Day-ahead electricity price forecasting by modified relief algorithm and hybrid neural network. IET Generation, Transmission & Distribution 4(3), 432–444 (2010)
11. Cao, L., Gorodetsky, V., Mitkas, P.: Agent Mining: The Synergy of Agents and Data Mining. IEEE Intelligent Systems 24(3), 64–72 (2009)
12. Bompard, E., et al.: A game theory simulator for assessing the performances of competitive electricity markets. IEEE St. Petersburg Power Tech (2005)
13. MIBEL - Operador del Mercado Ibérico de Energia – Iberian Electricity Market Operator (accessed on December 2013), http://www.omie.es

Distributed and Guided Genetic Algorithm for Humanitarian Relief Planning in Disaster Case

Fethi Mguis[1], Kamel Zidi[2], Khaled Ghedira[3], and Pierre Borne[4]

[1] Faculty of Sciences of Gabes, Tunisia
[2] Faculty of Sciences of Gafsa, Tunisia
[3] Higher Institute of Management of Tunis, Tunisia
[4] Ecole Centrale de Lille, France

Abstract. In this paper we propose a distributed and guided genetic algorithm for humanitarian relief planning in natural disaster case. It is a dynamic vehicle routing problem with time windows (DVRPTW), where customers should be served during a given time interval. This problem is an extension of classic vehicle routing problem. In the case of a disaster, emergency planning must be fast, consistent and scalable. For these reasons we opted for an improved genetic algorithm by adding some sort of guide to accelerate the convergence of the algorithm. Thus, the genetic algorithm can provide a population of solutions that can address the dynamic aspect of the problem. The objective of our approach is to provide a plan to meet all the demands with minimizing the total distance travelled. The proposed approach has been tested with theoretical data and showed high efficiency, which infers the possibility of applying for the management of emergency calls in the event of major disaster.

Keywords: Disaster planning, Disaster logistics, Vehicle routing problem with time windows, Dynamic VRP, Disaster relief, Discrete optimization, Multi-agents solving problem.

1 Introduction

The effectiveness of planning for the distribution of relief goods to disaster during the emergency phase should be done in minimal time. In addition it requires a prior determination of the optimal routes and schedules for the fleet to ensure that the total cost and time distribution are minimal, while the requirements and constraints such as capacity and availability vehicles and time windows on deliveries must be respected. So this is a complex and dynamic problem that is relevant to the operation process of the distribution of relief goods.

Although there have been some contributions in the field of vehicle dynamics turned on the problem of collection and distribution [CS], most of the available research has focused on the context commercial transportation and logistics planning in supply chain repetitive and routine treatments and environments orders [BBS]. Studies concerning vehicle routing problems in the context of dynamic planning disaster relief disaster are few and far between [MZ], [OEK], [MZGB1] and [MZGB2].

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_19,

The objective of this work is to propose an approximate resolution for planning disaster relief taking into account the new demands that may occur during the operation of the emergency. The objective is to avoid delays in the distribution of relief in the critical time and increase the useful life of the equipment which saves costs. The rest of the paper is structured as follows: Section 2 presents an overview of the literature about the vehicle routing problem in disaster case, section 3 highlights and describes the optimization approach in this study, the fourth section discusses testing of calculation and managerial implications arising from the study, while Section 5 summarizes the work, presents findings and suggests that future work.

2 Literature Review

Static routing problems have been extensively studied in the past and interest is now increasingly given to the problems of dynamic routing in profit, the business environment of the distribution [NS]. In this synthesis, some approaches solving vehicle routing problem are selected to be discussed. These are relevant to our dynamic approach to try to anticipate changes in travel time and possible changes arrivals emergency calls.

The travel time change over time are common in disaster areas due to, for example, the poor state of roads, safety and weather conditions. Therefore, the assumption of a constant travel time is unrealistic. Only time events or changes in travel time can be included using time dependent functions. One of the first approaches using this interpretation was approaching Malandraki and Daskin [MD], whose objective is to minimize travel time. However, the time-dependent problems may produce sub-optimal solutions if there are large uncertainties in travel times supported.

However, the solutions will increasingly sub-optimal if the events are not consistent with what was predicted. Therefore, approaches using more and more dynamic travel time. Previous results of numerical simulations suggest the superiority of a proactive policy in relation to route design that reacts to new information [FGS].

3 Proposed Approach

Our problem is laterally distributed, uncertain and dynamic. The dynamic aspect of the problem increases the complexity with respect to a vehicle routing problem classic that encourages us to use a set of reactive autonomous agents that adapt to their environment. In order to solve the above described problem, we propose a multi-agent model as developed by Zidi[Zi] to which we were add some features that we suspect necessary for the proper functioning of the system to adequately address the problem. Our model as described in Figure 1 was composed by a "System Agent" and a set of "Zone Subsystems". The "System Agent" has to create and supervise all other agents. Each "Zone Subsystems" was composed of a "Planning Subsystem", an "Area Manager Agent", "Information

Manager Agent", "Forecast Agent", "Disturbance Agent", and a local database. The "Zone Subsystem" was managed by the "Area Manager Agent" which provides communication with the "System Agent" and other "Zone Subsystems". The "Information Manager Agent" was responsible for the management of information necessary for the proper conduct of the rescue operation. In fact it provides information collection, their filtration, and updates the local database. The "Forecast Agent" allowed forecasting of emergency calls based on the information received and based on historical relief interventions. The "Disturbance Agent" will be used to detect disturbances that can occur, transmit "Planning Subsystem" to be processed. The "Planning Subsystem" was responsible for emergency planning, integration of new applications and correction planning in case of a disturbance. The "Planning Subsystem" was composed of three types of agents named "Supervisor Agent" who supervises the other components in addition to its roles in the "Planning Subsystem". The role of "Vehicle Agent" is finding a path for a vehicle. The "Interface Agent" handles communication with the external environment. The local database of "Zone Subsystem" will be used to store all data in the backup area. Each "Zone Subsystem" can communicate with the external environment mainly composed by: applicants emergency disaster, relief agencies, the geographic information system GIS, news and weather agencies.

This model requires minimal communication in order to optimize time resolution. This is essential for the problem of emergency management disaster. In addition, this model is extensible. Thus, this architecture can be adapted to different types of disasters (natural, accidental, etc.)

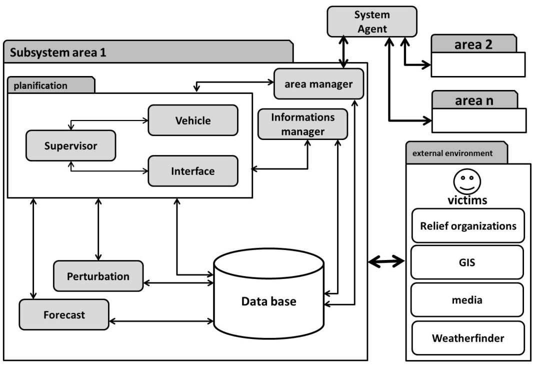


Fig. 1. Overview of the system

The subsystem of planning presented in Fig 2 was composed of three types of agents: Supervisor, Interface and Vehicle.

3.1 Supervisor Agent

This agent ensures the consolidation of emergency calls, the creation of tenders for the insertion of a new application, choose the best solution and update

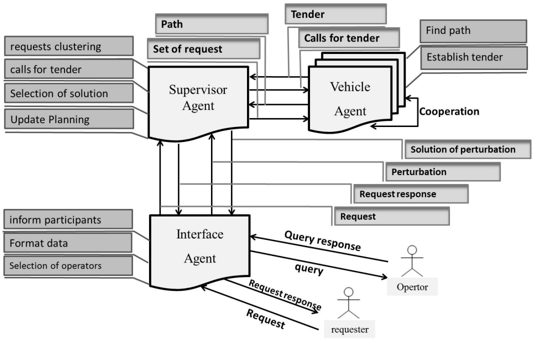


Fig. 2. Subsystem of planning

planning after any changes generated by the insertion of a new application or by the occurrence of a disturbance.

3.2 Interface Agent

This agent was responsible for all information exchanged between the "Planning Subsystem" and its external environment. It was responsible for receiving and formatting requests for help, negotiating with transport operators and informing stakeholders.

3.3 Vehicle Agent

Each vehicle agent was responsible for a vehicle. It has to ensure that the search path will be followed by the vehicle. In addition to receiving a bid for the insertion of a new request, the agent "Vehicle" checks can add this request and sets its offer, which sends it to the "Supervisor Agent".

3.4 Genetic Algorithm

In this paper, we use a guided genetic algorithms [ZMGB]. Indeed, we simulate the process of evolution of a population. One starts with an initial population composed of N solutions (individuals) of the problem. The degree of adaptation of an individual to the environment is expressed by the value of the cost function f (x), where x is the solution that is the individual. It is said that a person is much better adapted to its environment, the cost of the solution is lower or larger depending on test selected optimization. Within this population, occurs when the random selection of one or both parents, producing a new solution, through genetic operators such as crossover and mutation. The new population is obtained by the choice of N individuals among populations (parents and children) is called next generation. By iterating the process, we produce population richer individuals best suited. The process of the genetic algorithm is presented in Fig 3.

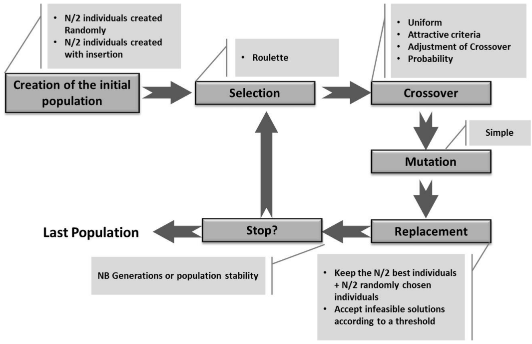


Fig. 3. Functioning of the genetic algorithm

Creation of the Initial Population. To build the initial population, we have developed a heuristic which builds the first half of the population randomly and the second half in a guided manner based on the method of insertion of Solomon[So].

Crossover. In our case we used the Uniform crossover. This method consists in determining the cost, the number of customers and the ratio of these last two for each round of two parents chosen at random. Then, each parent towers are sorted in ascending order according to the quotient: Cost/Number of customers. The third step is to establish a conflict matrix to describe the conflicts between the towers of both parents. The construction of the individual descending begins by inserting the first turn of one of the parents. The second step in the construction of the son is to exclude all rounds of the other parent are in conflict with the tour recently added. The same process is then repeated with the first round of the other parent and so on until there are more towers to insert. At this stage it is possible to customers not served. These customers are inserted into the shot so that the overall cost is minimal.

Mutation. In the process of evolution, mutation performs a broader exploration of the search space, to avoid premature convergence and diversity loss by bringing innovation in the population. In our approach we have adopted the method of simple random mutation. The principle of this method is to choose two nodes randomly. Result in generating a random value between 0 and 1. If this value is less than the mutation probability, then swap the two nodes, otherwise do nothing.

4 Experiments and Results

In this section, we report our approach results in comparison with other algorithms. The tests data were described in the section 4.1; while the experimental results of the different problems were summarized in the section 4.2.

4.1 Experimental Test Data

We do not have scientific benchmarks for the problem of emergency management disaster. For this reason, we chose Solomon's 100-customer benchmark problems [So] for the static vehicle routing problem with time windows (VRPTW). In these benchmark problems, 100 nodes are distributed in a Euclidean plane of 100*100 squares, and the travel times between nodes are equal to the corresponding Euclidean distances. There are six types of problems, named C1, C2, R1, R2, RC1 and RC2, each with 8-12 problems.

4.2 Experimental Results

The test results were shown in Table 1 for all the six type problems described in the section 4.1. NOV were the average of the used vehicles number. TD were the average of travel distance. Columns (1) and (2) were the computational results for the improved LNS algorithm (ILNS for short) which was proposed by Hong [Ho]. Columns (3) and (4) are the computational results for the Two-stage hybrid local search approach (HLS for short) which was proposed by Bent and Van Hentenryck [BH]. Columns (3) and (4) are the computational results for our approach Distributed and Guided Genetic Algorithm (DGGA for short).

In Table 2, we present a comparison of Distributed Guided Genetic Algorithm(DGGA) and Guided Genetic Algorithm(GGA) on the basis of their CPU time (in second). This comparison shows the advantage of using a distributed version of GGA relative to the centralized version. Based on these results, we can make the following conclusions on the effectiveness of the proposed approach to solve the problem studied:

– The solution quality generated by our approach can approximate to the best known solutions, the ILNS approach and the HLS approach. If the travel

Table 1. Comparison of our approach with ILNS and HLS based upon NOV and TD

	ILNS		HLS		DGGA	
Problem type	NOV (1)	TD (2)	NOV (3)	TD (4)	NOV (5)	TD (6)
C1	10	833.1	10	828.82	10	**828.65**
C2	3	590.31	3	**589.86**	3	643.94
R1	12.25	1218.28	11.92	1244.52	**11.9**	**1202.74**
R2	3.27	964.11	4	954.27	**2.9**	**913.6**
RC1	12.13	**1369.57**	**11.5**	1384.17	11.9	1375.89
RC2	3.75	1131.18	**3.25**	**1124.46**	3.9	1218.2

Table 2. Comparison of GGA and DGGA based on CPU time(s)

Problem Type	C1	C2	R1	R2	RC1	RC2
GGA	137.33	70.95	219.35	286.71	163.67	210.16
DGGA	102.37	55.32	187.17	239.32	134.39	149.90
Gain(%)	25.46	22.03	14.67	16.53	17.89	28.91

distances are only observed and the number of used vehicles were ignored, the average relative error of each type was less than 10%.

– Our approach provided an effective result within a reasonable time, which is very important in the disaster case. In fact, in the emergencies case, the solution should be provided quickly and efficiently. In addition our approach allowed at any time to provide an acceptable solution to stopping the execution of the algorithm after a given generation if necessary.
– Using multi-agent systems has allowed us to reduce the execution time which was very important for emergency planning in disaster case.

5 Conclusions and Future Works

In this paper, we have developed a solution to the problem of emergency planning for disaster. We have illustrated some of the complexities practices for emergency in disaster case. Despite the importance of emergency planning in several disaster cases, they have received few attention in the literature of the disaster and operations research. Thus, it appears that the extension of our approach to disaster relief, as described looks promising and with considerable value. Even if we know the boundaries explicitly construct models of reality and the unique nature of disasters, we suggest that operations research systems decision support can be beneficial in the disaster. Although several emergency relief organizations often hide their distribution planning for security reasons. Finally we conclude that the adaptation and the use of the approach described above can contribute to the improvement of the use of the vehicle fleet management as well as the routing and delivery of relief goods in disaster cases.

References

[BBS] Balcik, B., Beamon, B., Smilowitz, K.: Last mile distribution in humanitarian relief. Journal of Intelligent Transportation Systems 12(2), 51–63 (2008)

[BH] Bent, R., Hentenryck, P.: Scenario-based planning for partially dynamic vehicle routing with stochastic customers. Operations Research 52(6), 977–987 (2004a)

[CS] Campbell, A., Savelsbergh, M.: A decomposition approach for the inventory-routing problem. Transportation Science 38, 488–502 (2004)

[FGS] Fleischmann, B., Gnutzmann, S., Sandvob, E.: Dynamic vehicle routing based on online traffic information. Transportation Science 38(4), 420–433 (2004)

[Go] Goldberg, D.: Genetic algorithms in search, optimization, and machine learning. Advison-Wesley (1989)

[Ho] Hong, L.: An improved lns algorithm for real-time vehicle routing problem with time windows. Computers and Operations Research (2011)

[MD] Malandraki, C., Daskin, M.: Time dependent vehicle routing problems: formulations, properties and heuristic algorithms. Transportation Science 26(3), 185–200 (1992)

[MZ] Mete, H., Zabinsky, Z.: Stochastic optimization of medical supply location and distribution in disaster management. International Journal of Production Economics 126(1), 76–84 (2010)

[MZGB2] Mguis, F., Zidi, K., Ghedira, K., Borne, P.: Distributed approach for vehicle routing problem in disaster case. In: 13th IFAC Symposium on Control in Transportation Systems, Sofia-Bulgaria (2012a)

[MZGB1] Mguis, F., Zidi, K., Ghedira, K., Borne, P.: Modélisation d'un système multi-agent pour la résolution d'un problme de tournées de véhicules dans une situation d'urgence. In: 9ème Conférence Internationale de Modélisation, Optimisation et SIMulation, MOSIM 2012, Bordeaux, France (2012b)

[NS] Nagy, G., Salhi, S.: Heuristic algorithms for single and multiple depot vehicle routing problems with pickups and deliveries. European Journal of Operational Research 162, 126–141 (2005)

[OEK] Ozdamar, L., Ekinci, E., Kucukyazici, B.: Emergency logistics planning in natural disasters. Annals of Operations Research 129, 217–245 (2004)

[So] Solomon, M.: Algorithms for the vehicle routing and scheduling problems with time window constraints. Operations Research 35(2), 254–265 (1987)

[ZZ] Zeddini, B., Zargayouna, M.: Auto-organisation spatio-temporelle pour le vrptw dynamique. RJCIA (2009)

[Zi] Zidi, K.: Systme Interactif d'Aide au Dplacement Multimodal. PhD thesis, Ecole centrale de Lille France (2006)

[ZMGB] Zidi, K., Mguis, F., Ghedira, K., Borne, P.: Distributed genetic algorithm for disaster relief planning. Int. J. Comput. Commun. 8(5), 769–783 (2013)

FleSe: A Tool for Posing Flexible and Expressive (Fuzzy) Queries to a Regular Database

Víctor Pablos-Ceruelo and Susana Munoz-Hernandez

The Babel Research Group,
Facultad de Informática,
Universidad Politécnica de Madrid, Spain
{vpablos,susana}@babel.ls.fi.upm.es
http://babel.ls.fi.upm.es/

Abstract. We present FleSe, a tool for performing fuzzy and non-fuzzy queries to regular databases. The existing tools for querying databases have a syntax too complicate for normal users. What we present is a tool with a user-friendly interface that allows to perform any query that the underlying framework can solve. The framework is fully adaptable and configurable, so that introducing new knowledge and linking it to the information stored in databases can be done easily. We expect this work contributes to the development of more human-oriented fuzzy search engines.

Keywords: fuzzy logic, search engine, databases.

1 Introduction

The information stored in databases is most of the times crisp, i.e. non-fuzzy. We can study the technical reasons for deciding to use a non-fuzzy database, but there is at least a non-technical one which justifies it. This one is the inherent subjective character of fuzzy attributes. Take, for example, Victor's height: 1'81 cm. There is no problem in storing this crisp value (it is just a float number), but it is no so easy if we try to store if Victor is "tall", "very tall", "no tall at all" or any other fuzzy value, because it might not be true for all the people retrieving the value from the database. Elsa, whose height is 1'41 cm, might consider him very tall, while Susana, whose height is 1'72, might consider him just tall.

The drawback of storing information in a crisp way is that the retrieval is most of the times done in a crisp way. Suppose a database with flats and their prices. If we enter a query for retrieving the flats with a price lower than 200.000 € and there is a flat whose price is 200.300 € then it will not be retrieved, while it might be the one we are looking for (Is 300 € enough money for considering the flat too expensive?).

There are some tools for accessing regular databases (databases with non-fuzzy information) in a fuzzy way, as the one presented by P. Bosc and O. Pivert in [1], and they have a syntax similar to the SQL syntax (they are perfectly

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_20,

adequate for advanced users, developers or researchers), but they are not usable by normal users (the syntax is a little bit complicate for them).

The ideal tool would allow to use the information stored in non-fuzzy databases and, at the same time, provide a user friendly and flexible enough query syntax for representing any user question that it can answer.

We present here FleSe, a tool that suits all our needs. FleSe is a combination of a web interface for posing fuzzy and flexible queries and a framework for defining the links between the crisp information stored in a database and the fuzzy concepts that we can use in a query.

This combination makes FleSe 1) a very intuitive tool for normal users, since they can pose any query without learning the underlying syntax, 2) a general tool for developers, since just by defining the links between the crisp information stored in the database some fuzzy concepts they get a user-friendly interface for asking queries about those concepts and 3) a search engine that does not forget about the results that do not fully satisfy the conditions entered in the query, so for the flat's example before the flat with a price over 200.000 € would be in the valid results set.

2 Semantics

The structure used to give semantic to our programs is the multi-adjoint algebra, presented by J.Medina, M. Ojeda-Aciego, P. Vojtáš and J.M. Moreno in [2,3,4,5,6,7]. The interest in using this structure is that we can obtain the credibility for the rules that we write from real-world data, although this time we do not focus in that advantage nor in the existence of other mechanisms for this purpose, as the one proposed by A.M. Palacios, M.J. Gacto and J. Alcalá-Fdez in [8]. We simply highlight this fact so the reader knows why this structure and not some other one. Since the semantics details can be found in the papers cited, we only focus here in what matters to understand our contribution.

Suppose a program in the syntax the papers cited define and a query. In order to give an answer to the query the program is instantiated or grounded, the atoms are given an interpretation and this interpretation is extended to the formulas in the language. In bi-valued logic the interpretations are just $\{ \, true, \; false \, \}$ while in fuzzy logic it is a number $\mathsf{v} \in \; [0, 1]$, but in both cases the expected result is the maximum of the values obtained by the different rules whose head unify with the query.

The multi-adjoint algebra is an impeccable theoretical work but, as all theoretical works, needs adaption for modelling the real world. In our case, the necessity that is not satisfied is the definition of priorities for the rules. By defining priorities we can get as result a truth value which is not the maximum truth value, but the truth value of the rule with the maximum priority value. In this way the system could return the answers of the rules that the programmer considers to be more adequate, and not just the result of the rule that computes the higher value. This is why we take the works of V.Pablos-Ceruelo and S.Muñoz-Hernandez [9,10] as basis for ours: they extend the multi-adjoint semantics to deal with priorities. By using their semantics we have a framework with an easy syntax and a sound and complete semantics.

3 FleSe's Input and Output

In this section we describe the input that the tool receives, the syntax of the query, and how it returns the results (its output).

3.1 The Tool's Input

The queries syntax has been defined after studying multiple user queries and has the form shown in Eq. 1. Although at first sight it seems to restrict the user expressiveness, it allows to introduce all the queries we have studied.

$$\begin{array}{l} I'm\ \ looking\ \ for\ \ a/an\ \boxed{\text{individual}} \\ \left\{\begin{array}{l} \boxed{\text{not}}\ \boxed{\text{modifier}}\quad \boxed{\text{fuzz-pred}} \\ \qquad\qquad whose\ \boxed{\text{non-fuzz-pred}}\ \boxed{\text{comp-op}}\ \boxed{\text{val}} \end{array}\right\} \boxed{\text{AND}} \end{array} \tag{1}$$

In Eq. 1. *individual* is the element we are looking for (car, skirt, restaurant, ...), *not* is a negation mechanism, $modifier$ is a modifier (quite, rather, very, ...), $fuzz{-}pred$ is a fuzzy predicate (cheap, large, close to the center, ...), $non{-}fuzz{-}pred$ is a non-fuzzy predicate (price, size, distance to the center, ...), $comp - op$ is a comparison operand ("is equal to", "is different from", "is bigger than", "is lower than", "is bigger than or equal to", "is lower than or equal to" and "is similar to") and *val* is a crisp value (a number if $co - op$ serves to compare numbers or a string if it takes the values "is equal to", "is different from" or "is similar to"). The elements in boxes can be modified (and in some cases left blank) and the brackets symbolize choosing between the first line (a fuzzy predicate query) or the second one (a comparison between values). The "AND" serves to add more lines to the query, to combine multiple conditions. Some examples of use are "I'm looking for a restaurant not very close to the city center" (Eq. 2), "I'm looking for a restaurant whose food type is mediterranean and whose price average is lower than 30 €" (Eq. 3) and "I'm looking for a restaurant whose food type is similar to mediterranean and close to the city center" (Eq. 4). In the examples the empty boxes mean that we do not choose any of the available elements.

$$\begin{array}{r} I'm\ \ looking\ \ for\ \ a/an\ \boxed{\text{restaurant}} \\ \boxed{\text{not}}\ \boxed{\text{very}}\ \boxed{\text{close to the city center}}\ \boxed{} \end{array} \tag{2}$$

$$\begin{array}{l} \qquad I'm\ \ looking\ \ for\ \ a/an\ \boxed{\text{restaurant}} \\ \left.\begin{array}{l} whose\ \boxed{\text{food type}}\ \boxed{\text{is}}\ \boxed{\text{mediterranean}} \\ whose\ \boxed{\text{price average}}\ \boxed{\text{is lower than}}\ \boxed{30\text{ €}} \end{array}\right. \boxed{\text{AND}} \end{array} \tag{3}$$

$$\begin{array}{l} \quad\quad I'm\ \ looking\ \ for\ \ a/an\ \boxed{\text{restaurant}} \\ whose\ \boxed{\text{food type}}\ \boxed{\text{is similar to}}\ \boxed{\text{mediterranean}}\ \boxed{\text{AND}} \\ \square\ \square\ \boxed{\text{close to the city center}} \end{array} \tag{4}$$

In the queries in Eqs. 2 to 4 we can see that FleSe allows to enter fuzzy queries (Eqs. 2 and 4) and non-fuzzy queries (Eq. 3), but this is not all. We can ask for values similar to the one we are looking for (first line of Eq. 4) and we can use modifiers when the query is fuzzy, even the negation modifier[1](Eq. 2).

3.2 The Tool's Output

FleSe provides different sets of results for the query, so the user can chose the ones that suit best his/her expectations. By different sets of results we do not mean that they are disjoint, but that some users might prefer just to obtain the ten best results, others just the results that satisfy their query with a truth value over 0.7, others just the results that satisfy their query with a truth value over 0.5, others all the results whose truth value is over 0 and others might want to see everything in the database and how much they satisfy their query. This allows us to say that Flese covers the whole range of expectations of the user about the amount and quality of the results for his/her query.

4 Implementation Details

4.1 Used Technology

FleSe allows to combine the information stored in databases with the knowledge about fuzzy concepts. We decided to use Prolog [11] for representing this meta-information because it is one of the most successful programming languages for representing knowledge in computer science. Its main advantage with respect to the other ones is being a more declarative programming language.[2]

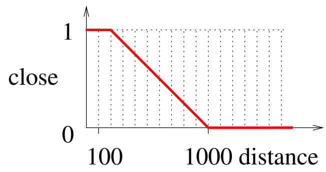
Prolog is based on logic. Fuzzy Logic (FL) is a subset of logic that allow us to represent not only if an individual belongs or not to a set, but the grade in which it belongs. Supposing the database contents and the definition for "close" in Fig. 1 and the question "Is restaurant X close to the center?" with FL we can deduce that Il tempietto is "definitely" close to the center, Tapasbar is "almost" close, Ni Hao is "hardly" close and Kenzo is "not" close to the center. We highlight the words "definitely", "almost", "hardly" and "not" because the

[1] The negation modifier predefined in FleSe gets as result for the negation one minus the value computed by the fuzzy predicate, but we can change it to have any other behaviour.

[2] We say that it is a more declarative programming language because it removes the necessity to specify the flow control in most cases, but the programmer still needs to know if the interpreter or compiler implements depth or breadth-first search strategy and left-to-right or any other literal selection rule.

name	distance	price avg.	food type
Il_tempietto	100	30	italian
Tapasbar	300	20	spanish
Ni Hao	900	10	chinese
Kenzo	1200	40	japanese
Zalacain	2000		
Don_Jamon			spanish

(a) Restaurants' database contents



(b) Close fuzzification function

Fig. 1. Restaurants' database contents and close fuzzification function

usual answers for the query are "1", "0.9", "0.1" and "0" for the individuals Il tempietto, Tapasbar, Ni Hao and Kenzo and the humanization of the crisp values is done in a subsequent step by defuzzification.

It is possible to develop FL programs in pure Prolog syntax, but there are libraries and extensions to reduce the development time. Although we know that there are others, as the Prolog-Elf system [12], the FRIL Prolog system [13], the F-Prolog language [14], the FuzzyDL reasoner [15], the Fuzzy Logic Programming Environment for Research (FLOPER) [16] and the Fuzzy Prolog system [17,18], we chose to take Rfuzzy [19] as basis. Rfuzzy [19] is a Prolog library developed as a Ciao Prolog [20] package. Since it is open source we modified it to suit our needs. The result is FleSe.

FleSe is a combination of a framework and a web interface. The framework allows us to define the links between the non-fuzzy information stored in the database and the web interface allows us to have a friendly interface. Although they can be seen as two different pieces, there is a strong link between them. The framework provides the web interface with the information it needs to fill each field, allowing us to present the user a web interface "intelligent", in the sense that it presents an input query form with which we can enter any query that the framework can solve with all the knowledge it has.

4.2 The Queries Interface

The queries interface is a web interface, mainly written in Java, JavaScript and HTML. To improve the usability we use Ajax to communicate asynchronously the user web navigator and the server. This, joint with the possibility to use Open Authentication for using the application allows any user to enter and start using the application without having to fill in any personal information.

The first selection that the user has to do is the database he/she will use for his/her query (Fig. 2). We provide this facility so any user can upload his/her own framework configuration file and choose it. The second selection (Fig. 3) is what we are looking for (restaurants in the example) and, when we select it, the framework presents us with the different predicates defined for it (Fig. 4). Once we choose a predicate the framework examines if it is a fuzzy predicate or not. If it is fuzzy two selection boxes appear to the left, for applying a modifier to the fuzzy predicate. If it is non-fuzzy one selection box appears to the right and, depending on the comparison we choose, another selection box or a free text field (Fig. 4 too).

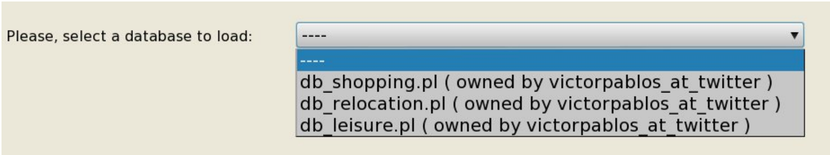


Fig. 2. Selecting the database

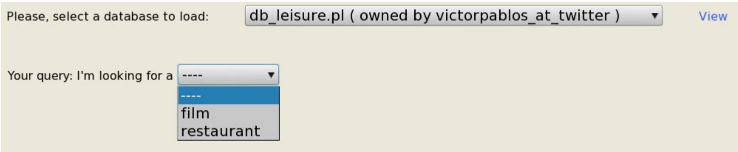


Fig. 3. Choosing what we are looking for

Fig. 4. Entering the conditions that we want the results to satisfy

4.3 The Answers Interface

As explained before, FleSe provides the ten best results (truth value always bigger than zero), the results whose truth value for the query is over 0.7, over 0.5, over 0 and all the results (Fig. 5). As we use Ajax for getting the results of each tab there is no waste of time computing the results shown in other tabs.

10 best results | Results over 70% | Results over 50% | Results over 0% | All results

restaurant	name	restaurant type	food type	years since opening	distance to the city center	price average	menu price	Truth Value
restaurant(meson del jamon, fast food, spanish, 8, 100, 20, 15)	meson del jamon	fast food	spanish	8	100	20	15	1.00
restaurant(museo del jamon, fast food, spanish, 8, 150, 20, 15)	museo del jamon	fast food	spanish	8	150	20	15	0.95

Fig. 5. Database selection

5 Conclusions and Present Work

We have presented FleSe, a tool for allowing a regular user to pose fuzzy and non-fuzzy queries to a regular (non-fuzzy) database. It allows to define the links between fuzzy concepts and non-fuzzy information stored in the database and uses this knowledge to present the user a user friendly interface for querying the system.

In our humble opinion, the presentation of the prototype of FleSe is a step forward in the direction of getting a search engine able to understand any user query an provide the results that he/she expects to obtain (and not just a list of web pages containing one or more of the words typed in the query).

Our current research focus in allowing each user to personalize the fuzzy concepts to suit their needs. As told in the introduction, the satisfaction of fuzzy concepts depends on each one's point of view and we think that a tool as FleSe should implement personalization capabilities.

References

1. Bosc, P., Pivert, O.: Sqlf: a relational database language for fuzzy querying. IEEE Transactions on Fuzzy Systems 3(1), 1–17 (1995)
2. Medina, J., Ojeda-Aciego, M., Vojtáš, P.: A multi-adjoint approach to similarity-based unification. Electronic Notes in Theoretical Computer Science 66(5), 70–85 (2002), uNCL 2002, Unification in Non-Classical Logics (ICALP 2002 Satellite Workshop), `http://www.sciencedirect.com/science/article/B75H1-4DDWJ13-37/2/bdc92744d6ddc8e888ea314efb711107`
3. Medina, J., Ojeda-Aciego, M., Vojtáš, P.: A completeness theorem for multi-adjoint logic programming. In: FUZZ-IEEE, pp. 1031–1034 (2001)
4. Medina, J., Ojeda-Aciego, M., Vojtáš, P.: Multi-adjoint logic programming with continuous semantics. In: Eiter, T., Faber, W., Truszczyński, M. (eds.) LPNMR 2001. LNCS (LNAI), vol. 2173, pp. 351–364. Springer, Heidelberg (2001)
5. Medina, J., Ojeda-Aciego, M., Vojtáš, P.: A procedural semantics for multi-adjoint logic programming. In: Brazdil, P.B., Jorge, A.M. (eds.) EPIA 2001. LNCS (LNAI), vol. 2258, pp. 290–297. Springer, Heidelberg (2001)
6. Medina, J., Ojeda-Aciego, M., Vojtáš, P.: Similarity-based unification: a multi-adjoint approach. Fuzzy Sets and Systems 146(1), 43–62 (2004)

7. Moreno, J.M., Ojeda-Aciego, M.: On first-order multi-adjoint logic programming. In: 11th Spanish Congress on Fuzzy Logic and Technology (2002), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.3800
8. Palacios, A.M., Gacto, M.J., Alcalá-Fdez, J.: Mining fuzzy association rules from low-quality data. Soft Comput. 16(5), 883–901 (2012)
9. Pablos-Ceruelo, V., Muñoz-Hernández, S.: Introducing priorities in rfuzzy: Syntax and semantics. In: CMMSE 2011: Proceedings of the 11th International Conference on Mathematical Methods in Science and Engineering, Benidorm, Alicante, Spain, vol. 3, pp. 918–929 (June 2011), http://gsii.usal.es/ CMMSE/index.php?option=com_content&task=view&id=15&Itemid=16
10. Pablos-Ceruelo, V., Muñoz-Hernández, S.: Getting answers to fuzzy and flexible searches by easy modelling of real-world knowledge. In: FCTA 2013: Proceedings of the 5th International Conference on Fuzzy Computation Theory and Applications (2013)
11. Sterling, L., Shapiro, E.: The Art of Prolog. MIT Press (1987)
12. Ishizuka, M., Kanai, N.: Prolog-elf incorporating fuzzy logic. In: IJCAI 1985: Proceedings of the 9th International Joint Conference on Artificial Intelligence, pp. 701–703. Morgan Kaufmann Publishers Inc., San Francisco (1985)
13. Baldwin, J.F., Martin, T.P., Pilsworth, B.W.: Fril- Fuzzy and Evidential Reasoning in Artificial Intelligence. John Wiley & Sons, Inc., New York (1995)
14. Li, D., Liu, D.: A fuzzy Prolog database system. John Wiley & Sons, Inc., New York (1990)
15. Bobillo, F., Straccia, U.: fuzzydl: An expressive fuzzy description logic reasoner. In: 2008 International Conference on Fuzzy Systems (FUZZ 2008), pp. 923–930. IEEE Computer Society (2008)
16. Morcillo, P.J., Moreno, G.: Floper, a fuzzy logic programming environment for research. In: de Oviedo, F.U. (ed.) Proceedings of VIII Jornadas sobre Programación y Lenguajes (PROLE 2008), Gijón, Spain, pp. 259–263 (October 2008)
17. Vaucheret, C., Guadarrama, S., Muñoz-Hernández, S.: Fuzzy prolog: A simple general implementation using CLP(R). In: Baaz, M., Voronkov, A. (eds.) LPAR 2002. LNCS (LNAI), vol. 2514, pp. 450–464. Springer, Heidelberg (2002)
18. Guadarrama, S., Muñoz-Hernández, S., Vaucheret, C.: Fuzzy prolog: a new approach using soft constraints propagation. Fuzzy Sets and Systems (FSS) 144(1), 127–150 (2004); possibilistic Logic and Related Issues
19. Muñoz-Hernández, S., Pablos-Ceruelo, V., Strass, H.: Rfuzzy: Syntax, semantics and implementation details of a simple and expressive fuzzy tool over prolog. Information Sciences 181(10), 1951–1970 (2011), special Issue on Information Engineering Applications Based on Lattices, http://www.sciencedirect.com/science/article/B6V0C-50PJWYR-2/2/26d8ff890f0effc98aa1c12225a5fb87
20. The CLIP Lab, The Ciao Prolog Development System WWW Site, http://www.clip.dia.fi.upm.es/Software/Ciao/

Software Fault Prediction Based on Improved Fuzzy Clustering

Golnoosh Abaei and Ali Selamat

UTM-IRDA Digital Media Centre, K-Economy Research Alliance &
Faculty of Computing, Universiti Teknologi Malaysia,
Johor Baharu, 81310, Johor, Malaysia
golnoosh.abaee@gmail.com, aselamat@utm.my

Abstract. Predicting parts of the software programs that are more defects prone could ease up the software testing process and helps effectively to reduce the cost and time of developments. Although many machine-learning and statistical techniques have been proposed widely for defining fault prone modules in software fault prediction, but this area have yet to be explored with high accuracy and less error. Unfortunately, several earlier methods including artificial neural networks and its variants that have been used, marred by limitations such as inability to adequately handle uncertainties in software measurement data which leads to low accuracy, instability and inconsistency in prediction. In this paper, first the effect of irrelevant and inconsistent modules on fault prediction is decreased by designing a new framework, in which the entire project's modules are clustered. The generated output is then passed to the next model in the hybrid setting, which is a probabilistic neural network (PNN) for training and prediction. We used four NASA data sets to evaluate our results. Performance evaluation in terms of false positive rate, false negative rate, and overall error are calculated and showed 30% to 60% improvement in false negative rate compared to other well-performed training methods such as naïve Bayes and random forest.

Keywords: Fuzzy Clustering, Fault Prediction, Probabilistic Neural Network.

1 Introduction

Since software projects play major role in nowadays industry, the accurate estimating of the software development cost is very important. According to the Standish group report [12], just 32% of software projects were on time and on cost in 2009, 44% of the projects are in challenged mode and 24% of projects have been cancelled. Designing, developing, testing, and all aspects of the software projects are affected by the relevant estimations and predictions. Software testing is known as a major factor in increasing the development cost. Early detection of fault-prone software components could enable verification experts and testers to concentrate their time and resources on the problematic areas of the system under development. During the recent decades, many methods for the software fault prediction have been proposed though, this area still

DOI: 10.1007/978-3-319-07593-8_21, © Springer International Publishing Switzerland 2014

poses many challenges and unfortunately, none of the techniques proposed within the last decade has achieved widespread applicability in the software industry. This is due to several reasons including the lack of software tools to automate this prediction process, the unwillingness to collect the fault data, and other practical problems [6], [9].

Usually a variety of statistical and machine learning techniques are used in software quality modeling. However, the relations between software metrics and quality factors are often complex and nonlinear, and thus they limit the accuracy of conventional approaches such as statistical methods. In addition, outliers and irrelevant data in training set can lead to the imprecise estimations. In fact, in many engineering problems, we encounter vagueness in information and uncertainty in training sets, so as these phenomena cause, we could not reach to certain results for our proposed solution. Soft computing on the other hand is a term that is recently become popular in the general area of prediction. It is a field within computer science that is characterized by the use of inexact solutions. Soft computing differs from conventional (hard) computing in that, unlike hard computing, soft computing deals with imprecision, uncertainty, partial truth, and approximation to achieve practicability, robustness, and low solution cost [11]. Components of soft computing include neural networks, support vector machines, fuzzy logics, evolutionary computation and so on. Our system, models the input information's vagueness through fuzzy clusters and then fault prediction is done based on probabilistic neural network (PNN) as it showed good and robust performance according to literature reviews [38].

The remainder of this paper continues with section 2, where a brief discussion on related works is presented. Prerequisite tasks including probabilistic neural network and fuzzy clustering are reviewed in section 3. Section 4 contains our proposed method. Experimental description and the results are described in section 5 and 6 respectively and finally, we summarize this paper in section 7.

2 Related Works

As the number of related works in this area is too much, we just presented number of recent and famous research papers in this section. Menzies et al [7] conducted several experiments based on different data mining algorithms with method level metrics on public NASA datasets. They evaluated their work with probability of false alarm and probability of detection, and balance. They reported the best performer as naïve Bayes and they used log-transformation with Info-Gain filters before applying the algorithms. They claimed that the best algorithm changes according to the dataset characteristics and numerous experiments should be performed for a robust prediction model. Catal and Diri [3] focused on the high-performance fault predictors based on machine learning such as random forests and algorithms based on artificial immune systems on public NASA datasets. They reported that random forests provides the best prediction performance for large datasets and naive Bayes is the best prediction algorithm for small datasets in terms of the area under receiver operating characteristics curve evaluation parameter. Rodriguez et al. [2] described two well-known subgroup discovery algorithms, the SD algorithm, and the CN2-SD algorithm to obtain rules that identify defect prone modules. The experiments performed on object-oriented metrics datasets from Eclipse repository showed that the EDER-SD algorithm performs well in most cases when compared to three other well-known SD algorithms.

Apart from the above related works, only few of them are using artificial neural network to build the prediction models. Kanmani et al. [1], introduced five-layer back propagation and three-layer probabilistic neural network models based on class level metrics to show that their approach is better than statistical methods. Their dataset were based on the undergraduate project developed in their institution. Khoshgofttar et al. [10] came up with a three-layer back propagation neural network after examining different combination of architecture. Both models used type-I, type-II, and the misclassification rate as their evaluation metrics.

3 Prerequisite Explanations

3.1 Probabilistic Neural Network (PNN)

The probabilistic neural network (PNN) was developed by Specht [16]. This network provides a general solution to pattern classification problems by following an approach developed in statistics, called Bayesian classifiers. The probabilistic neural network uses a supervised training set to develop distribution functions within a pattern layer. These functions are used to estimate the likelihood of an input feature vector being part of a learned category, or class. PNN is based on one pass learning with highly parallel structure. It is a powerful memory based network and able to deal with sparse data effectively. In PNN, the number of neurons in the hidden layers is usually the number of patterns in the training set because each pattern in the training set is represented by one neuron. The main advantage of PNN is the speed at which the network can be trained. Training a PNN is performed in one pass. The smoothing factor allows PNN to interpolate between the patterns in the training set.

3.2 Fuzzy Clustering

Clustering algorithms group the modules according to similarity of their software attributes. In fact, program modules with similar attributes are clustered together as they have similar quality characteristics; Furthermore, dissimilarity of data located in separate clusters should be as high as possible. Proper data clustering technique will enhance not only the efficiency of the training process, but also the performance of the model predictability precision. Accurate predictions obtained from such a good reliability model will be favorable toward higher software process efficiency and product quality [8]. Several parameters such as connectivity, intensively and distance among data characteristics determine the level of similarity. Usually in clustering methods data element belongs to exactly one cluster, which is famous as hard clustering, however, among them, a soft clustering method that is called fuzzy clustering calculates the relativity of each module (X = $x_1, x_2, \ldots, x_n$) to the specified clusters (C = $c_1, c_2, \ldots, c_c$) with membership values (M = $m_1, m_2, \ldots, m_n$) varies from zero to one. In this method, data elements belong to one or more clusters at the same time. The C-means clustering is one of the most important fuzzy clustering techniques developed in 1973 [14] and improved in 1981 [13]. Variety of different application has used this method to solve their problems. In this method, the final aim is to minimize a target function as shown in Eq. 1.

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \left\|x_i - c_j\right\|^2 \ , \ 1 \le m \le \infty \tag{1}$$

u_{ij} is the membership degree of x_i from the center of cluster j (c_j), and $\|x_i - c_j\|$ is the difference expressing the similarity between data (x_i) and the center of cluster j (c_j).

3.2.1 C-means Clustering Algorithm

In C-means clustering, first a set of random initial membership values (U_{ij}) are generated from each data module x_i for each cluster c_j. Then center value of each cluster is calculated based on Eq. 2 for k number of times. After that $u^{(k)}$ and $u^{(k+1)}$ is updated according to Eq. 3. Finally if difference between $u^{(k)}$ and $u^{(k+1)}$ is less than the threshold, the iteration stops, otherwise, new cluster's centers are employed based on Eq. 2

$$c^{(k)} = c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \ , \ c^{(k)} = \left[c_j\right] \quad with \quad U^{(k)} \tag{2}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \tag{3}$$

4 Proposed Method

One of the most important factors affecting the performance of neural network is the quality of data used in the training set. Existing of outliers and irrelevant data in the training set can lead to a less precise estimation. In such situation, predicting is not reliable because the process of training is performed using inconsistent and contradictory projects. Indeed, negative effect of outliers and irrelevant projects will divert the learning process from the correct direction and lead to the construction of an inaccurate estimation model. Clustering of the software projects is the key part of predicting method proposed in this section. To overcome the diversity and inconsistency of the projects collected in a dataset, it is required to separate the outliers and irrelevant projects from other ones. The modules clustering can increase the consistency of modules by putting similar modules in the same clusters. Instead of having a dataset, which includes numerous irrelevant and inconsistent modules, there will be several subsets comprising of consistent and similar modules. Clustering process is performed by analyzing the modules features to discriminate the most similar modules and putting them in the separate clusters. The proposed method is organized in two main phases called training and testing stages. In the training stage, the structure of the proposed hybrid method is configured and in the testing stage, the label of software modules is predicted. Training and testing stages are described as follows.

4.1 Training Stage

As it is shown in Fig.1.a., first 66% of any project is selected for the training purpose. Then fuzzy clustering is applied and cluster's centers are recorded. In addition, all produced clusters are also saved along with their center's weights. We should mention that this experiment is repeated 5 times to make sure the different combination of testing and training datasets are examined. When the clustering on training projects is finished, each cluster is trained with PNN if the number of modules in that cluster is more than the number of features for the reason that low population in a cluster can lead to incomplete training and inaccurate prediction. With this method, quality of training with PNN is improved. Therefore, if the cluster is not a low-population cluster, the cluster is trained with PNN and the weights are stored. On the other hand, low population clusters are stored without any training for the testing purpose.

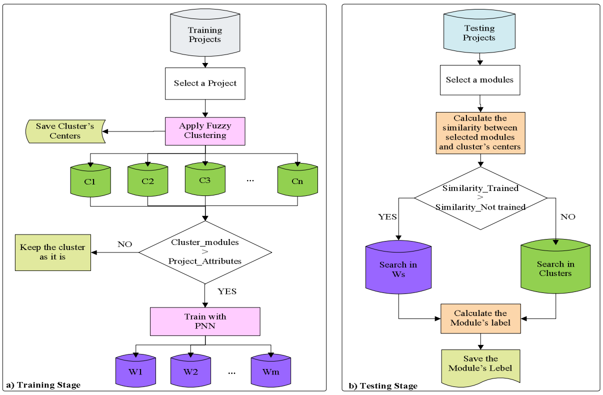


Fig. 1. a) Training stage of the proposed model, b) Testing stage of the proposed model

4.2 Testing Stage

According to Fig.1.b., testing stage consists of three main steps. At first, a module from test dataset is selected. Then similarity between that selected module and cluster's centers are calculated based on Euclidean distance to specify the most similar clusters to the selected test module. Second, after the cluster is identified, if number this cluster is not identified as low-population cluster, prediction based on training set obtained from PNN is conducted to predict the label of modules as faulty or non-faulty. On the other hand, if the selected cluster is from low-population clusters, the similarity between tested module, and each of the modules in a selected cluster is computed and label of the test module (faulty/non-faulty) is determined.

5 Experimental Description

5.1 Dataset Description

Four NASA data sets [4] are selected which are different in number of rows and defects' rate. The CM1 data set contains 498 modules and 10% of them are defected; this data belongs to NASA spacecraft instrument project. PC1 belongs to flight software for earth orbiting satellite; it has 1109 numbers of rows and 7% of the data are defected. Third data set is KC1, belongs to storage management project for receiving and processing ground data with 2107 models and 15% of the data is defected with one or more faults. The KC2 project belongs to the science data processing unit of storage management system used for receiving and delivering ground data. It contains 522 modules, of which 20% of them are defected.

5.2 Performance Evaluation

After specifying each module's label, each evaluation metrics is calculated based on the confusion matrix. If a module's label predicted as non-faulty but the actual label is faulty, we get the condition of false negative (FN). On the other hand, if the non-faulty modules was labeled as faulty, we call it false positive (FP). If the faulty module predicted as faulty and non-faulty module predicted as non-faulty, they are called true positive (TP) and true negative (TN) respectively. FNR (false negative rate) is the percentage of modules that were actually faulty but there were predicted as non-faulty. In contrast, FPR (false positive rate) is the percentage of modules that were actually non-faulty but there were predicted as faulty. FNR, FPR and errors need to be as small as possible in all the experiments. The following equations are used to calculate FNR, FPR, and overall error rate.

$$Overall\ Error\ Rate = \frac{FN+FP}{TP+FN+FP+TN} \tag{4}$$

$$FPR = \frac{FP}{FP+TN} \tag{5}$$

$$FNR = \frac{FN}{TP+FN} \tag{6}$$

6 Experimental Result and Analysis

In order to evaluate our work, we used three well-performed learning methods based on the literature reviews namely random forest, decision tree, and naïve Bayes [3], [7], [15]. Table 1 presents the prediction error results for the proposed fuzzy approach as compared to other approaches. As it can be seen in Table 1 and Figure 3, our proposed model (FCM-PNN) outperformed all methods with a vast distance in terms of false negative rate. Although our method does not improve false positive rate but a good method should be able to predict faulty modules as well as non-faulty modules with a stable rate. For example, in KC1, decision tree performed best among other classifiers in term of false positive rate (FPR), but if we consider the high value of false negative rate (FNR), we will find out that the model is not stable and trustable. According to the results, there is a big difference between our proposed method (FCM-PNN) and three other well-performed models in terms of false negative rate (FNR). Wrong predicting of modules that are actually faulty but there are predicted as non-faulty (FNR) is more important than wrong predicting the modules that are actually non- faulty but there are predicted as faulty. Higher value in FNR means that a large amount of fault-prone modules cannot be detected prior to the testing [5].

Table 1. Comparison results based on different methods

Dataset	Method	FPR	FNR	Error
KC1	FCM_PNN	0.19	0.21	0.17
	Random Forest	0.08	0.65	0.17
	Decision Tree, J48	0.04	0.81	0.16
	Naïve Bayes	0.08	0.64	0.18
KC2	FCM_PNN	0.25	0.12	0.23
	Random Forest	0.13	0.45	0.19
	Decision Tree, J48	0.07	0.58	0.18
	Naïve Bayes	0.51	0.55	0.16
CM1	FCM_PNN	0.04	0.45	0.28
	Random Forest	0.04	0.90	0.12
	Decision Tree, J48	0.01	0.98	0.11
	Naïve Bayes	0.08	0.71	0.14
PC1	FCM_PNN	0.17	0.38	0.19
	Random Forest	0.02	0.63	0.07
	Decision Tree, J48	0.01	0.86	0.07
	Naïve Bayes	0.05	0.73	0.11

7 Conclusion

In this paper, we proposed an improved fuzzy clustering model to enhance the prediction through effective uncertainty handling in order to achieve high accuracy while ensuring stability and consistency of results. In the proposed method, the quality of training in PNN is improved compared to other methods. In all cases, FNR has improved considerably. The results show that our proposed model could be a trustable model compared to other well-performed software prediction methods. Predicting lower value in FNR means that most fault-prone modules can be detected prior to the system testing which is useful for testers and project managers to allocate their time and resources on those parts while there is a limited time and budget for testing any software project.

Acknowledgement. The Ministry of Education Malaysia and Universiti Teknologi Malaysia (UTM) under research grant 03H02 have supported the research. Also, the Ministry of Science, Technology & Innovations Malaysia (MOSTI), under research grant 4S062 is hereby acknowledged for supporting this research work.

References

1. Kanmani, S., Uthariaraj, V.R., Sankaranarayanan, V., Thambidurai, P.: Object-oriented software fault prediction using neural networks. Information and Software Technology 49(5), 483–492 (2007)
2. Rodriguez, D., Ruiz, R., Riquelme, J.C., Harrison, R.: A study of subgroup discovery approaches for defect prediction. Information and Software Technology 55(10), 1810–1822 (2013)
3. Catal, C., Diri, B.: Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. Information Sciences 179(8), 1040–1058 (2009)
4. Promise Software Engineering Repository, `http://promisedata.org` (August 20, 2012)
5. Catal, C., Sevim, U., Diri, B.: Software fault prediction of unlabeled program modules. In: Proceedings of the World Congress on Engineering, vol. 1, pp. 1–3 (2009)
6. Hall, T., Beecham, S., Bowes, D., Gray, D., Counsell, S.: A systematic literature review on fault prediction performance in software engineering. IEEE Transactions on Software Engineering 38(6), 1276–1304 (2012)
7. Menzies, T., Greenwald, J., Frank, A.: Data mining static code attributes to learn defect predictors. IEEE Transactions on Software Engineering 33(1), 2–13 (2007)
8. Mahaweerawat, A., Sophatsathit, P., Lursinsap, C.: Adaptive self-organizing map clustering for software fault prediction. In: Fourth International Joint Conference on Computer Science and Software Engineering, KhonKaen, Thailand, pp. 35–41 (2007)
9. Catal, C., Sevim, U., Diri, B.: Clustering and metrics thresholds based software fault prediction of unlabeled program modules. In: Sixth International Conference on Information Technology: New Generations, ITNG 2009, pp. 199–204 (April 2009)
10. Khoshgoftaar, T.M., Allen, E.B., Hudepohl, J.P., Aud, S.J.: Application of neural networks to software quality modeling of a very large telecommunications system. IEEE Transactions on Neural Networks 8(4), 902–909 (1997)
11. Zadeh, L.A.: Fuzzy sets. Information and Control 8(3), 338–353 (1965)
12. `http://www.pmhut.com/the-chaos-report-2009-on-it-project-failure`
13. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers (1981)
14. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters (1973)
15. Abaei, G., Selamat, A.: A survey on software fault detection based on different prediction approaches. Vietnam Journal of Computer Science, 1–17 (2013)
16. Specht, D.F.: Probabilistic neural networks. Neural Networks 3(1), 109–118 (1990)

Facial Authentication before and after Applying the Smowl Tool in Moodle

Francisco D. Guillén-Gámez and Iván García-Magariño

Department of Computer Engineering and Industrial Organization,
Faculty of Technical Sciences. Open University of Madrid (UDIMA), Carretera de La Coruña,
KM.38,500, Vía de Servicio, nº 15, 28400 Collado Villalba Madrid, Spain
{franciscodavid.guillen,ivan.garcia-magarino}@udima.es

Abstract. In this work, a facial authentication system has been included examined, deployed and evaluated for students who study in distance higher education. Until today, there was hardly any technology that verified whether the students are themselves the ones in the realization of their activities within a LMS (Learning Management System) as Moodle. Currently, there are different technologies that identify people. One of them is biometrical facial authentication, which permits the authentication and verification of users based on facial features. Thanks to the new technologies' contribution within the educational community, there is a possibility to verify that there are not frauds and to avoid the identity theft while the students do their activities in the platform; in order to demonstrate that e-learning is as accepted and recognized as any other kind of education. The main objective is to check the functionality degree of facial authentication of the Smowl tool in Moodle platform, to identify the attention degree and opinions in the students' understanding, as well as to determine which kinds of activities have more acceptance before and after using the tool. In this work, 100 Master students were audited for the survey and the conclusions reached indicate a high acceptance of facial authentication as a technique to improve distance education, and that this perception of students improves even more after experiencing facial authentication when learning through Moodle rooms.

1 Introduction

Like Bliuc, Ellis, Goodyear et al. (2011) mention universities are on the way to transform a classic teaching model into a new mixed teaching model, combining distance learning and face-to-face learning. Distance learning is commonly also called distance education and e-learning. Race (1994) indicates that the process of continuous learning is characterized by being the student who chooses the rate, the moment and the place, and this is corroborated and experienced by other works (Ong & Lai, 2006; Cotton & Gresty, 2006).

However, the e-learning educational community confronts new aspects that should be solved in order to validate and recognize e-learning as well as face-to-face education. The question about the similarity of the effectiveness of e-learning along with face-to-face education continues being object of debates and researches throughout education history. According to Merisotis & Phipps (1999), the majority

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_22, © Springer International Publishing Switzerland 2014

of studies show that the learning results obtained using the Information and Communication Technologies (ICT) are similar to those obtained by traditional education.

One of the important aspects to keep in mind is the possibility for the students to be able to complete their tests and tasks avoiding identity theft. With the use of facial authentication within education, it appears the possibility to verify that there are not any frauds while students are doing their activities in the Moodle platform (Dougiamas, Taylor, 2003).

In this sense, the objective of the project presented below is to answer the question about how to minimize the mistake when doing a facial authentication and consequently, to be able to improve the results of the process of teaching-learning in distance education by this technology and using Moodle as platform. This work has successfully deployed a facial recognition tool, i.e. Smowl, which is introduced in its website (Smowl, 2014), into some virtual rooms of the Open University of Madrid. In particular, this work conducts and study with students before and after making use of the tool. The poll was taken of 100 students and the conclusions obtained show a high acceptance of facial authentication as a technique to improve e-learning.

The remainder of this article is organized as follows: the next section introduces the background research related to the current work, mentioning the improvements over the existing literature; section 3 presents the motivation of the research; section 4 presents the analysis of the perception of students about this proposal, describing the participants of the survey, and determining the questions of the surveys; section 5 presents the analysis of the results of the surveys, extracting the most relevant facts that need to be taken into account before actually deploying this proposal in an online university; finally section 6 mentions the conclusions and future work.

2 Related Works

Over the last few years facial biometrics have been integrated in education, therefore in E-Learning as well. We will focus on the best known articles about acceptance and perspectives that facial recognition can have about the Moodle platform.

The researchers Denhavi & Fard (2011) show a system that determines whether students are actually attending virtual lessons. In this model, two behavioral biometric characteristics (mouse and free-text keystroke movements) and one physical (facial features) are used. In the same line of research, Ullah, Xiao & Lilley (2012) suggest a facial authentication mechanism to ensure that students are not being impersonated in order to improve their grades in virtual tests. Our research differs from these projects on verifying and controlling that the students are who they say when they do their activities in different kinds of Moodle activities (also referred as Moodle tools from this point forward).

On the other hand, there are several researchers that use students' webcams in their methods to extract images of them in order to use facial authentication afterwards in these pictures. In the same vein, Pattanasethanon & Savithi (2012) show an effective technique concerning facial authentication from webcam images; or the work of the researchers Agulla, Rifón, Castro et al. (2008) that studies how to guarantee online students are themselves and to know exactly the amount of time that each student spends in front of their computer reading or doing their online tasks. In this line of research, the current work presents a study of the perception of students of being identified through facial recognition based on face biometrical features through several photographs taken through their webcams. This study allows universities and instructors to take well-informed decisions about whether to include the current biometrical face recognition approach in their education process.

In another context, the researchers Grafsgaard, Wiggins, Boyer et al. (2013) have developed an investigation about the software of facial expressions authentication and how to predict the effectiveness of online tutorial sessions through webcam. In this research, the students completed a preliminary test based on the content. After each session, the students answered a poll. The questions of the post-session poll were designed in order to measure several aspects of the participation and the cognitive load. This work shows similarities with ours, due to the existence of a poll before and after the process, in our case, before and after using Smowl in the tools and Moodle plugin. Thus, the analyzed software and the questions of the survey are different.

3 Motivation

This work has been done in the context of a research project whose objective is to have control over the students who study in distance higher education, in our case from the Open University of Madrid, in order to verify if the students are who they say when doing their activities within the LMS Moodle platform. This research will be accomplished in two different steps; the first one, before the implementation of the application in the LMS Moodle platform, in order to check the valuation degree that the use of this software in e-learning could achieve; and the second one, once the application is implemented in the platform, just to know the acceptance degree that it could have from the students once the technology has been used.

In order to develop the project, 10 questions were designed, related to the students' perception that they can have using facial authentication in their Moodle activities. The Smowl tool was applied as the facial authentication software for the research development. This software captures photos of students from webcams for their subsequent verification. Smowl was executed when students inserted contents in glossaries and performed tests of the different lessons in two subjects.

4 Analysis of the Perception of Students about Facial Recognition

The poll was taken from 100 students, 50% of them for the first part of the research (before using Smowl), and the other 50% was used for the second part (after using Smowl). For the first part of the research, the survey respondents were from the University of Málaga (UMA), while for the second part of the research the students were from the Open University of Madrid (UDIMA). The first part students' average age fluctuated around 21.9 years old, whereas the second part students' average age fluctuated around 33.9 years old.

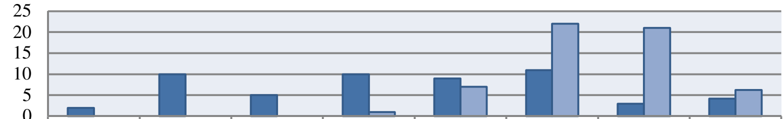
In most investigations, when assessing attitudes and opinions, the seven-point Likert scale is usually used: Totally disagree (1), Disagree (2), Slightly Disagree (3), Neither agree nor disagree (4), Slightly Agree (5), Agree (6) and Strongly Agree (7). The questions of the current research are answered with this scale. Some of these questions needed to be formulated slightly different, referring as (a) for the questions to students before using the Smowl tool, and denoting as (b) for the ones to the students after using the tool.

Table 1. Questions of the questionnaire

	Question
1	Do you think it is appropriate to use facial authentication in e-learning?
2	Do you think it is fair to monitor distance education in order to avoid cheating?
3	Would you be willing that your university include facial recognition if with this you are guaranteed that the prestige of your university will be higher, and will be easier to find a job?
4	(a) Do you think you would lose your privacy if Smowl was watching you while doing your course activities? (b) Do you think you have lost your privacy when Smowl was watching you while doing your course activities?
5	(a) Would you feel spied if Smowl was watching you while doing course activities? (b) Have you ever felt spied when Smowl was watching you while doing course activities?
6	(a) Would you feel embarrassed if Smowl was watching you while doing course activities? (b) Have you felt embarrassed when Smowl was watching you while doing course activities?
7	(a) Would you feel comfortable if Smowl was watching you while doing course activities? (b) Have you felt comfortable when Smowl was watching you while doing course activities?
8	Do you think it is appropriate to apply facial authentication to the tests?
9	Do you think it is appropriate to apply facial authentication to the activities of continuous assessment?
10	Do you think it is appropriate to apply facial authentication to the learning activities?

5 Results of the Analysis

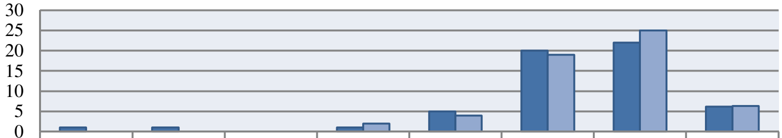
In relation to the knowledge the participants of the study have about the implementation of face authentication in e-learning in different moments, it is noticeable that students have changed their minds after testing Smowl. Also, it is necessary to have in mind that 'before' refers that students did not have experienced Smowl in Moodle when they were surveyed, and 'after' denotes that they were surveyed after experiencing Smowl in Moodle.



	Totally disagree	Disagree	Slightly Disagree	Neither agree nor disagree	Slightly Agree	Agree	Strongly Agree	Average in the seven-point Likert scale
Before	2	10	5	10	9	11	3	4.18
After	0	0	0	1	7	22	21	6.26

Fig. 1. Do you think it is appropriate to use facial authentication in e-learning?

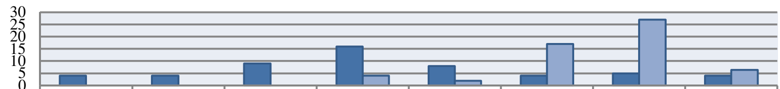
It is clear that the figure 1 is one of the most important of the research, because it provides with the view whether they are in favor or against the use of facial authentication in e-learning. If the two groups are analyzed separately, it is shown that before using Smowl, the students were not convinced and did not accept a hundred percent the tool usability in Moodle. The students changed their opinion after using the tool in their activities. If the average of each group is compared, then one observes that there is a difference greater than two degrees in the Likert Scale.



	Totally disagree	Disagree	Slightly Disagree	Neither agree nor disagree	Slightly Agree	Agree	Strongly Agree	Average in the seven-point Likert scale
Before	1	1	0	1	5	20	22	6.14
After	0	0	0	2	4	19	25	6.34

Fig. 2. Do you think it is fair to monitor distance education in order to avoid cheating?

During the data collection, question 2 came of being quite relevant because it is about the students' opinion about the fairness of the e-learning to be controlled in order to avoid cheating and to allow all the students be in the same conditions. As it can be seen in figure 2, the results are very similar, being the second group a bit higher. It is clear, for both groups, that students' control and verification in distance education is a need that has to be carried through.



	Totally disagree	Disagree	Slightly Disagree	Neither agree nor disagree	Slightly Agree	Agree	Strongly Agree	Average in the seven-point Likert scale
Before	4	4	9	16	8	4	5	4.04
After	0	0	0	4	2	17	27	6.34

Fig. 3. Would you be willing that your university include facial recognition if with this you are guaranteed that the prestige of your university will be higher, and will be easier to find a job?

The figure 3 shows the acceptance of the tool by the second group of students. As it is described the students from the first group are not convinced about the tool being a good mechanism and they think that it will not benefit the university to get prestige. However, the data shown from the second group show that they maintain an average of 6.34 out of 7. In other words, this means that, once students have used Smowl, they think the tool is very useful, as well as they assume it will help the university to get prestige and to show that distance universities have the same validity as regular universities.

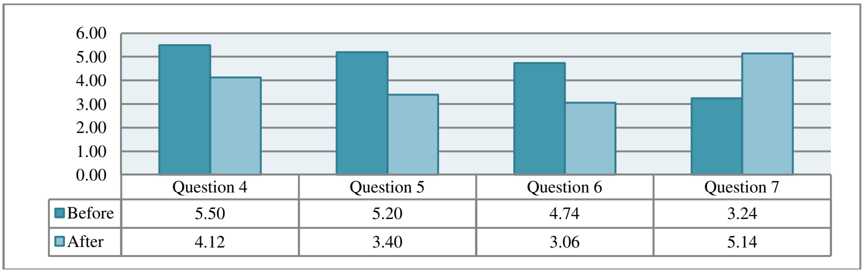


Fig. 4. Would you feel: Lose Privacy (Question 4), spied (Question 5), embarrassed (Question 6) and comfortable (Question 7)

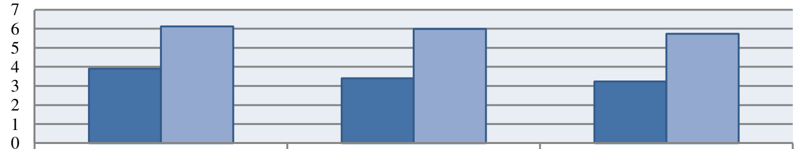
The figure 4 shows the students' perceptions when they are asked about different emotions. In the question 4, before using the tool, the students felt they have lost their intimacy with an average of 5.50 out of 7. This group thought that the fact of being photographed could contribute to the loss of their own privacy, due to the possibility of taking private aspects of their lives. On the other hand, after using the software, their opinions alter significantly considering that there is a drop of more than one point in the Likert scale achieving 4.12.

In the average of each group in the question 5 from the gathered data, it is shown a difference of almost two degrees in the Likert scale in relation to the two groups. That is, before using Smowl the students felt slightly agreed to be spied (5.20 out of 7), while once the students have tried the software in their online activities, they changed their minds positioning themselves as slightly disagree with 3.40 out of 7. Those data can be understood as that the students have improved their idea about the program and do not think it is not harmful to use it in e-learning, besides they do not really feel spied on.

The question 6, with the passage of time and through the use of Smowl, shows how the students have changed their minds about feeling ashamed when the tool was capturing images through their webcams. As the averages display, the students are now "Disagree" in the Likert scale, a positive fact related to the other group that remains standing almost "Slightly Agree". Overall, the students have changed their opinions when they have tried the tool several times in their activities, because they do not feel as ashamed as they believed before.

In order to conclude with the group of questions related to the perceptions the students have had before and after using Smowl, the question 7 describes how the average has improved with respect to the first group in terms of conformability. That is to say, the opinions of the students have improved after trying the software.

In the end, the figure 5 shows the opinions of the respondents that alter from before using the tool to after using the tool, for the different kinds of academic activities (tests, continuous evaluation activities, and learning activities). The improvement of the perception of appropriateness is quite notorious (about three points in the Likert scale) for all the kinds of activities, after using the tool.



	Question 8	Question 9	Question 10
Before	3.91	3.41	3.24
After	6.12	5.98	5.74

Fig. 5. Do you think it is appropriate to apply facial recognition to: tests, continuous evaluation activities, and learning activities? (Averages in the seven-point Likert scale)

It is also worth pointing out that the students accept facial authentication in tests more than in other kind of activities. Probably, this is because the tests have a short duration compared with the second type of activities more accepted by the students. The learning activities are the kind of activities less accepted by the students to be controlled, probably because in its learning process they do not want to be worried about being controlled with repetitive captures of images. In a nutshell, the perception of students has improved on each type of activity, but the order of preferences for each type of activity remains the same.

Therefore, one of the most relevant conclusions from this analysis is that the inclusion of facial authentication in e-learning within the platform LMS Moodle, should be done depending gradually on the activity and its type. In that way, if it is introduced gradually placing just the tests, in the second course or semester the perception will improve and it will be more tolerant to technology accepting its use in other kinds of activity as continuous evaluation activities and finally learning activities.

6 Conclusions and Future Work

Along this research, a series of objectives were set out about the usability and implementation of facial authentication within distance education, in order to avoid the students' fraud while they do their activities in the Moodle platform and at the same time avoiding the reject of students to the new technology. According to the data gathered, the research has reached the main objectives pursued.

On the other hand, it is important to wonder which the limit of privacy is for the student when using Smowl in their activities. As future work, it should be studied a legislation related to this relevant topic. The research is planned to be extended by introducing face-based authentication in the different courses of the next academic year, analyzing a higher number of students. In addition, our future research can determine whether this technology can finally replace face-to-face final exams.

On the other hand, in order to obtain the data, we have used two groups of students from different universities because the data collection was carried at different times. As future work, we shall use the same group of students to answer the survey before and after testing Smowl, in order to know what these same students think.

Finally, as future work, it a statistical test will be applied in order to collect and analyze the information. In particular, the analysis of variance (ANOVA) will determine whether the changes are statistically significant or not.

Acknowledgments. This work has been done in the context of the research project Desarrollo de un módulo de autenticación y monitorización biométrica de usuarios en entornos virtuales de aprendizaje, supported by the Open University of Madrid, with grant UD-019**.** We acknowledge the support of Smowl company for installing their software in the platform of our university. This research work has also been funded by the Spanish Ministry for Economy and Competitiveness through the project Social Ambient Assisting Living - Methods (SociAAL) (TIN2011-28335-C02-01).

References

1. Bliuc, A.M., Ellis, R.A., Goodyear, P., Piggott, L.: A blended learning Approach to teaching foreign policy: Student experiences of learning through face-to-face and online discussion and their relationship to academic performance. Computers & Education 56(3), 856–864 (2011)
2. Race, P.: Open learning handbook: promoting quality in designing and delivering flexible learning. Kogan Page, London (1994)
3. Ong, C., Yu Lai, J., Wang, Y.: Factors affecting engineers' acceptance of asynchronous elearning system in high tech companies. Information & Management 41, 795–804 (2004)
4. Cotton, D., Gresty, K.: Reflecting on the think aloud method for evaluating elearning. British Journal of Educational Technology 37(1), 45–54 (2006)
5. Merisotis, J.P., Phipps, R.A.: What's the Difference?: Outcomes of Distance vs. Traditional Classroom-Based Learning. Change: The Magazine of Higher Learning 31(3), 12–17 (1999)
6. Dougiamas, M., Taylor, P.: Moodle: Using learning communities to create an open source course management system. In: World Conference on Educational Multimedia, Hypermedia and Telecommunications, vol. 2003(1), pp. 171–178 (2003)
7. Smowl (2014) Smowl website, `http://smowltech.com/es` (last accessed April 03, 2014)
8. Dehnavi, M.K., Fard, N.P.: Presenting a multimodal biometric model for tracking the students in virtual classes. Procedia-Social and Behavioral Sciences 15, 3456–3462 (2011)
9. Ullah, A., Xiao, H., Lilley, M.: Profile based student authentication in online examination. In: 2012 International Conference on Information Society (i-Society), pp. 109–113. IEEE (2012)
10. Pattanasethanon, P., Savithi, C.: Human Face Detection and Recognition using Web-Cam. Journal of Computer Science 8(9), 1585 (2012)
11. Agulla, E.G., Rifón, L.A., Castro, J.L.A., Mateo, C.G.: Is My Student at the Other Side? Applying Biometric Web Authentication to E-Learning Environments. In: Eighth IEEE International Conference on Advanced Learning Technologies, ICALT 2008, pp. 551–553. IEEE (July 2008)
12. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. In: Proceedings of the 6th International Conference on Educational Data Mining (2013)

SOA Modeling Based on MDA

Haeng-Kon Kim[1] and Tai-Hoonn Kim[2]

[1] School of Information Technology, Catholic University of Daegu, Kyungbuk, 712-701, Korea
hangkon@cu.ac.kr
[2] Department of Convergence Security, Sung Shin Women's University, Korea
taihoonn@empal.com

Abstract. Along with the boom of Web services and the thriving Model Driven Architecture (MDA), we must consider the growing significance and utility of modeling in the development of software and solutions. The main advantages of MDA are the ability to transform one PIM into several PSMs, one for each platform or technology in which the final system will be deployed, and the automatic code generation that implements the system for those platforms from the corresponding PSMs. Service-oriented architectures (SOA) are also touted as the key to business agility, especially when combined with a model-driven approach. Model-Driven Architecture (MDA) is a well-developed concept that fits well with SOA, but until now it has been a specialized technique that is beyond practical application scope of most enterprises.

In this paper, we describe the initial investigation in the fields of MDA and generative approaches to SOA. Our view is that MDA aims at providing a precise framework for generative software production. Unfortunately many notions are still loosely defined (PIM, PSM, etc.). We propose here an initial exploration of some basic artifacts of the MDA space to SOA. Because all these artifacts may be considered as assets for the organization where the MDA is being deployed with SOA, we are going to talk about MDA and SOA abstract components to apply an e-business application. We also discuss the key characteristics of the two modeling architectures, focusing on the classification of models that is embodied by each. The flow of modeling activity is discussed in the two architectures together with a discussion of the support for the modeling flows provided by MDA. Our model of framework – a unified modeling architecture – is introduced which illustrates how the two architectures can be brought together into a synergistic whole, each reinforcing the benefits of the other with case study.

Keywords: Model-Driven Architecture (MDA), Domain Model, Service-oriented architectures (SOA), Software Process Improvement, Component Based Development, Repository.

1 Introduction

In order to find an appropriate solution to development and design of those systems an appropriate paradigm seems necessary. The object-oriented and component-based

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_23,

technology has not significantly met the needs of these systems, and may be considered as adding additional complexity to a domain that needs simplification. A new paradigm like service-oriented architecture is necessary. SOA is a paradigm that utilizes services as fundamental elements for developing applications. In order to gain the full benefits of such technology, an effective approach to modeling and designing these complex distributed systems is required. In fact there is not a suitable approach to SOA-based development and little works have been done on this area and most of them are for special applications and specific domains. To exploit the benefits of SOA effectively and duly, we propose an approach that involves MDA into the context [1,2,3,4].

Service-oriented architecture (SOA) is an approach to loosely coupled, protocol independent, standards-based distributed computing where software resources available on the network are considered as Services [3]. SOA is believed to become the future enterprise technology solution that promises the agility and flexibility the business users have been looking for by leveraging the integration process through composition of the services spanning multiple enterprises. The software components in a SOA are services based on standard protocols and services in SOA have minimum amount of interdependencies. Communication infrastructure used within an SOA should be designed to be independent of the underlying protocol layer. Offers coarse-grained business services, as opposed to fine-grained software-oriented function calls and uses service granularity to provide effective composition, encapsulation and management of services.

The problems of modeling solutions based on SOA have largely been resolved through the recognition of the importance of loose coupling and the consequent separation of concerns. Service Interfaces are shared amongst models showing the implementation and re-use of the services. Whilst the use of modeling within SOA is well established, it has suffered from the same issues as modeling in other architectures. The abstraction gap between the level of detail expressed in the model and the level of detail expressed in the code is a key issue. Yet it is the abstraction gap which is one of the key targets for the Model Driven Architecture. It seems likely, then, that if SOA and MDA can work together they will add value synergistically, leading to greater benefits than either architecture provides in isolation. Yet the two architectures are distant in terms of the way they address the issues surrounding modeling. SOA focuses on the stereotypical roles of models based on separation of concerns. MDA focuses on levels of abstraction, defining the role of models within a process. The question of the compatibility of these two model architectures remains open.

The service-oriented architecture (SOA) approach and the corresponding web service standards such as the Web Service Description Language (WSDL) [5] and the Simple Object Access Protocol (SOAP) [6] are currently adopted in various fields of distributed application development (e.g. enterprise application integration, web application development, inter organizational workflow collaboration). The service-oriented paradigm offers the potential to provide a fine grained virtualization of the available resources to significantly increase the versatility.

Model driven architecture (MDA) [7] has been proposed as an approach to deal with complex software systems by splitting the development process into three separate model layers and automatically transforming models from one layer into the other:

1. The Platform Independent Model (PIM) layer holds a high level representation of the entire system without committing to any specific operating system, middleware or programming language. The PIM provides a formal definition of an application's functionality without burdening the user with too much detail.
2. The Platform Specific Model (PSM) layer holds a representation of the software specific to a certain target platform such as J2EE, Corba or in our case the service oriented Grid middleware.
3. The Code Layer consists of the actual source code and supporting files which can be compiled into a working piece of software. In this layer, every part of the system is completely specified. MDA theory states that a PIM is specified and automatically transformed into a PSM and then into actual code, thus making system design much easier. The trick, of course, lies in the development of generic transformers capable of generating PSM and code layers from the PIM [8]. e-business application on SOA is a relatively young field of distributed computing and is currently lacking any form of tool support for a model driven approach to software development. This is unfortunate since we believe that due to its high complexity and the high rate of churn in the software technology market, a MDA approach is vital to the adoption of this new technology as in figure 1. Only if "business logic" (i.e. application functionality) developers can more or less effortlessly integrate a new middleware into their system, will a widespread adoption be possible. Furthermore, the developers responsible for the integration of the middleware into the overall system should be able to concentrate on middleware concerns and not have to cope with the business logic as well. This separation of concerns can be greatly facilitated by an appropriate MDA approach.

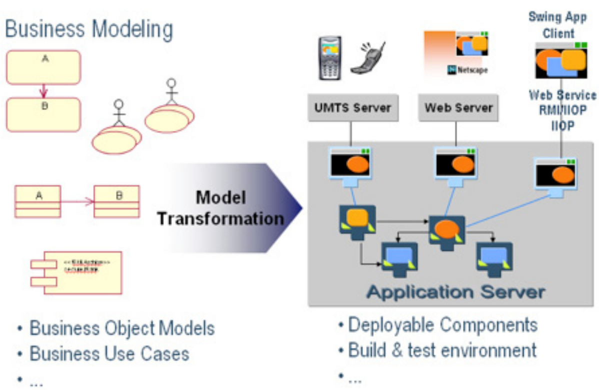


Fig. 1. MDA approaches

In this paper, we present a model-driven approach to SOA modeling and designing complex distributed systems based on MDA. MDA separates the Platform Independent Model (PIM) from the Platform Specified Model (PSM) of the system and transforming between these models is achieved via appropriate tools. The paper proposes a new approach to modeling and designing service-oriented architecture. In this approach the PIM of the system is created and then the PSM based on SOA is generated (this PSM is a PIM for next level). Then the final PSM based on a target platform (such as Web Services, Jini and so on) is generated. These models are generated with transformation tools in MDA and an approach to the model driven development for e-business applications on SOA is presented. The goal of the approach is to minimize the necessary human interaction required to transform a PIM into a PSM and a PSM into code for a SOA. To further separate the architectures specific components of the PSM from the business specific components of the PSM, a UML e-business Profile is introduced and a separation of the PSM layer into two parts is proposed which make the automated transformations from PIM to PSM to code easier to implement and more transparent for system designers, developers, and users. The separation of concerns introduced on the PSM layer is mirrored on the code layer by the use of Java annotations, allowing the same business code to run in different domains simply by exchanging the annotations and thus decoupling application code and SOA middleware.

2 Related Works

2.1 Modeling Web Services Metadata Based on MDA

Web services are emerging as the perfect framework for application-to-application integration or collaboration, to make these applications available as Web services. To standardize the use of Web services, the World Wide Web Consortium (W3C) proposed the Web Service Description Language (WSDL) standard, an XML-based language that describes Web service functionality. Essentially, a WSDL file is a language-independent XML-based version of an IDL (Interface Definition Language) file that describes the operations offered by a Web service, as well as the parameters that these operations accept and return. Thus, WSDL has become the standard that supports the description of Web services: What they do, how they should be used, and where they are localized [8,9].

The WS-Policy framework consists of two specifications: WS-Policy and WS-Policy Attachment.

- The **WS-Policy** specification describes the syntax for expressing policy alternatives and for composing them as combinations of domain assertions. The WS-Policy specification also describes the basic mechanisms for merging multiple policies that apply to a common subject and the intersection of policies to determine compatibility.

- The **WS-Policy Attachment** specification describes how to associate policies with a particular subject. It gives normative descriptions of how this applies in the context of WSDL and UDDI, (Universal Description, Discovery, and Integration), and it provides an extensible mechanism for associating policies with arbitrary subjects through the expression of scopes.

Along with the boom of Web services and the thriving Model Driven Architecture (MDA), we must consider the growing significance and utility of modeling in the development of software and solutions. MDA, which was proposed by the Object Management Group (OMG), is a model-driven framework for software development that proposes to model the business logic with Platform-Independent Models (PIMs) to later transform them on Platform-Specific Models (PSMs) by using transformation guides between the different models. The main advantages of MDA are the ability to transform one PIM into several PSMs, one for each platform or technology in which the final system will be deployed, and the automatic code generation that implements the system for those platforms from the corresponding PSMs.

Because Web services are software components, the development of Web services must exploit the advantages of MDA. To apply the MDA principles in the development of Web services, a modeling process must be considered. According to MDA principles, this modeling activity should result in automatic code generation. If we want to abstract from the platform in which the Web service will be deployed, the code that should be generated is the WSDL document that contains the Web service description in a standard format.

2.2 MDA Main Concepts

The main concepts of the MDA are beginning to be identified [6,7] A model represents a particular aspect of a system under construction, under operation or under maintenance. A model is written in the language of one specific meta-model. A meta-model is an explicit specification of abstraction, based on shared agreement. A meta-model act as a filter to extract some relevant aspects from a system and to ignore all other details. A meta-meta-model defines a language to write meta-models. There are several possibilities to define a meta-meta-model. Usually the definition is reflexive, i.e. the meta-meta-model is self defined. A meta-meta-model is based at least on three concepts (entity, association, and package) and a set of primitive types. The OMG MOF contains all universal features, i.e. all those that are not specific to a particular domain language. Among those features we find all that is necessary to build meta-models and to operate on them. Maintaining a specific tool for the MOF would be costly, so the MOF is aligned on the CORE part of one of its specific metamodels: UML. UML thus plays a privileged role in the MDA architecture. As a consequence, any tool intended to create UML models can easily be adapted to create MOF meta-models. MDA utilizes models and a generalized idea of architecture standards to address integration of enterprise systems in the face of heterogeneous and evolving technology and business domains. MDA combines computer-aided verification and machine intelligence during modeling to discover and remove design bugs before code reviews and testing. MDA Meta model acts as a filter to extract some relevant aspects from a system and to ignore for all other details. A meta-meta-model defines a language to write meta-models.

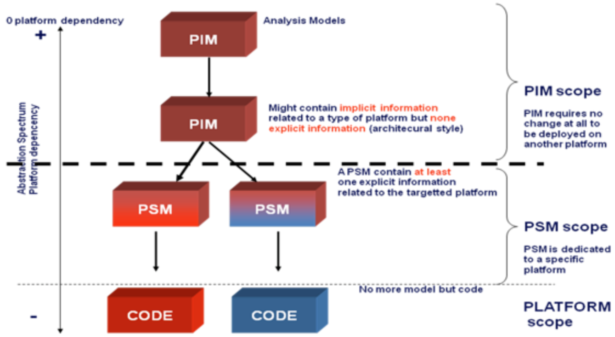


Fig. 2. MDA Development Process

The application of MDA to a use case begins by focusing on the development of the models. Figure 2 show the MDA process that includes: Computation Independent Model (CIM): describes concepts of a given domain but does not describe the software system. Platform Independent Model (PIM): describes software behavior that is independent of some platform. Platform Specific Model (PSM): describes software behavior that is specific for some platform. The first step in using MDA is to develop a CIM which describes the concepts for a specific domain. The CIM focuses on the environment and requirements of the system; the details of the structure and processing of the system are hidden or as yet undetermined. The next step involves developing the PIM. The term "platform" can have various meanings and can include one or more system aspects such as operating system, network configurations, and programming language. The meanings of PIM and PSM models are therefore relative to the definition of platform used in the use case. More important than the definition of platform is the recognition that PIMs and PSMs are supposed to separate aspects of program behavior from aspects of implementation. The third step is developing one or more PSMs which characterize a particular deployment of a software application. This could, for example, focus on the properties of a web application, whether the application should be generated in Java or Visual Basic, or whether the installation was for a standalone or networked machine. MDA requires development of explicit transformations that can be used by software tools to convert a more abstract model into a more concrete one. A PIM should be created, and then transformed into one or more PSMs, which then are transformed into code." The mappings between models are meant to be expressed by a series of transformation rules expressed in a formal modeling language. "A CIM is a software independent model used to describe a business system. Certain parts of a CIM may be supported by software systems, but the CIM itself remains software independent. Automatic derivation of PIMs from a CIM is not possible, because the choices of what pieces of a CIM are to be supported by a software system are always human. For each system supporting part of a CIM, a PIM needs to be developed first."

It is possible for concepts defined in a CIM to be automatically associated with properties defined in a PIM. For example, the concept "protein" defined in a CIM about proteomics experiments could be associated with PIM concepts such as a help feature that defined protein for users or a drop down list of protein names.

A meta-model in MDA defines a specific domain language. It may be compared to the formal grammar of a programming language. In the case of UML the need to define variants of the base language was expressed. The UML meta-model was then equipped with extension mechanisms (stereotypes, tagged values, constraints) and this allows defining specialization of the basic meta-models as so called profiles.

The MOF contains features to serialize models and meta-models in order to provide a standard external representation. The XMI standard defines the way serialization is performed. This is a way to exchange models between geographical locations, humans, computers or tools. When a tool reads a XMI serialized model (a UML model for example), it needs to check the version of the meta-model used and also the version of the XMI applied scheme.

2.3 SOA

SOA exposes real dependencies against artificial ones [11]. A real dependency is a state of affairs in which one system depends on the functionality provided by another. Beside real dependencies there are always artificial dependencies in which the system becomes dependent to configurations and various musts other systems expose. The target of SOA is to minimize artificial dependencies (although it can never be completely removed), and maximize real ones. This is done via loosely coupling, and the concept of service. A service is a coarse grain functionality objects, with interfaces expressed via a well defined platform independent language. When using services as computational objects, systems can register, find and invoke each other based on a well defined, every one accepted, language hence no one, highly becomes dependent to another system and a high degree of loosely coupling is achieved.

3 Applying MDA to E-Business Applications

3.1 Basic Ideas

The MDA organization may be viewed as a set of artifacts, some being standard building blocks, some being user developed. We may envision, in the not too far future, an organization starting with a hierarchical library of meta-models and extending it as an adaptation to its own local context (models as assets). Model reusability will subsume code reusability, with much more efficiency. This may be seen as orthogonal to code class libraries (e.g. Java, Swing, EJB, etc.). Inside a company, the various business and service models will be developed and maintained to reflect the current situation. Combining a service-oriented modeling architecture with MDA for e-business can bring many unique benefits. Firstly the clear organization of models and information based on the stereotypes derived from the service-oriented architecture and select perspective as development process. Secondly the productivity, quality and impact analysis benefits of the use of MDA with its emphasis on automation, transformation and synchronization. MS2Web solution for MDA in our approach is uniquely positioned to take advantage of the unified modeling architecture which results from bringing these two key architectures together. MDA combines a uniquely powerful implementation of the web services vision, together with the industry leading solutions for modeling service-based solutions.

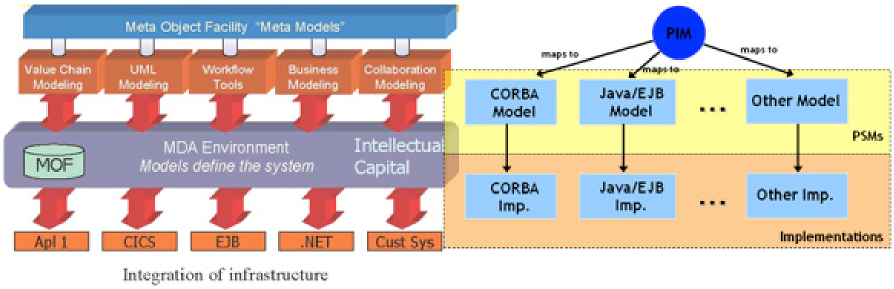


Fig. 3. Architecture for Applying MDA to e-business Development Process

Figure 3 shows our architecture for applying MDA to e-business and web service application in this paper. First it defines the language used for describing object-oriented software artifacts. Second, its kernel is synchronized with the MOF for practical reasons as previously mentioned. There is much less meta-modelers (people building meta-models) than modelers (people building models). As a consequence it is not realistic to build specific workbenches for the first category of people. By making the MOF correspond to a subset of UML, it is possible with some care to use the same tool for both usages. As a consequence the MDA is not only populated by first class MOF meta-models, but also with UML dialects defined by UML profiles for specific purposes languages. This is mainly done for practicality (widening the market of UML tools vendors) and there is some redundancy between UML profiles and MOF meta models (It is even possible to find conversion tools). There are many examples of profiles. Some are standardized by OMG working groups and other are independently defined by user groups or even by individuals. Examples of profiles are "UML for APL 1", " UML for CICS", "UML for Scheduling Performance and Time" (real-time applications), "UML for EJB", "UML testing", "UML for EAI", "UML for QoS and fault tolerance", "UML for Cust Sys".

One important kind of model that is being considered now is the correspondence model. A correspondence model explicitly defines various correspondences that may hold between several models. In the usual case, there are only two models: the source and the target. There may be several correspondences between a couple of elements from source and target. The correspondences are not always between couples of elements and they are strongly typed. There is not yet a global consistent view on correspondence models since this problem is appearing from different perspectives. When the notion of PDM and virtual machine is clarified we may then tackle the definition of a PIM, a model containing no elements associated to a given platform. In other times this was simply called a business model, but as for platform models we need to progress now towards a less naive and a more explicit view. The first idea is that the PIM is not equivalent to a model of the problem. We propose the architectural model for many elements of the solution that may be incorporated in a PIM as long as they don't refer to a specific deployment platform as in figure 4.

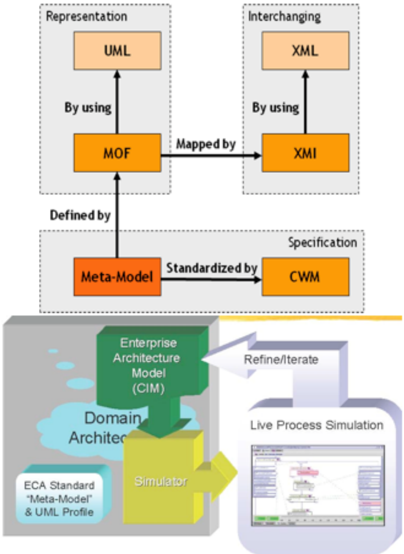


Fig. 4. Architecture for e-business Development in this paper

In our architecture model as in figure 3, the PIM of the system is created using UML diagrams by the analyst of the system. The PIM of the system will be designed simply without thinking about services that is pretty simple and is accomplished as CBD (Component-Based Development). The SOA-based PSM (which is a PIM for the next level) would be derived from the present PIM. The way which is used to identify this PSM must be quite different from the one used to identify PSM in component-based systems; because in component based systems the patterns which are used to determine the PSM of the system have a specific form. For each service in e-business applications, there is a single instance which manages a set of resources and consequently, unlike components, services are for the most part stateless that means need to view a service as a manager object that can create and manage instances of a type, or set of types. According to above discussion, in our approach after creating the PIM, this PIM is transformed -with a transformation tool- to another PIM based on SOA. In this transformation, for each class diagram in PIM for e-business, a Service Manager is created that manages the Instant Services. This management involves creation, deletion, updating a service and state management of services. To complete this transformation, we need some other special patterns for dealing with associations between classes. When this PIM based on SOA is created, the PSM of the system can be created based on a target platform such as Web Services, e-business and/or other platforms with transforming tools.

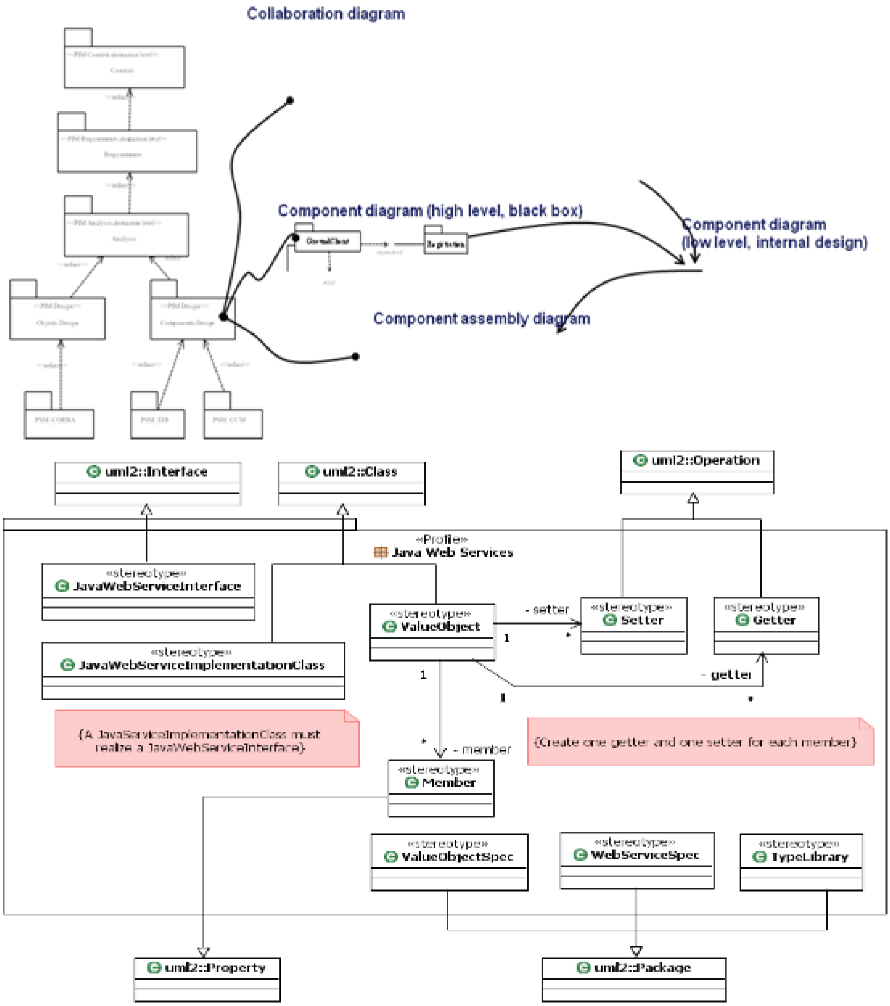


Fig. 5. PIM of e-business applications

Some operations apply on a single model and are called monadic by opposition to dyadic operations applying to two models. Operations applying on more than two models are rarer. Obviously the most apparent components in an MDA workbench are the precise tools composing this workbench. Fortunately in this context we should be able to propose a rather precise definition of a tool: it is an operational implementation of a set of operations applicable on specific models. The meta-models supported by a tool should be exhaustively and explicitly defined.

3.2 Generating the PIM for E-Business

PIM for e-business application is an abstract design of a computerized solution which does not include any platform specific elements. The core of the platform independent model (PIM) is a UML model – ranging from use cases through classes, interactions, states and other UML elements to the components as in figure 5.

3.3 Translation from PIM to PSM

While the PSM entities, i.e. Java classes or entity beans, bear the structure and deliver the behavior of inventory entities as described in the original inventory PIMs, end-users should not interact directly with these entities. Rather, entities should be accessed through a single interface that exposes a simple set of management methods and hides their complexity. This is a standard design guideline, which conforms to the related design pattern and influences the architectural design of components. In order to comply with the guideline, the case-study aims at implementing an application tool that allows users to manage the inventory content through a simple GUI. Example users of such a tool may be front-desk operators who respond to customer calls and access the inventory to setup a new or change the state of an existing product/service instance.

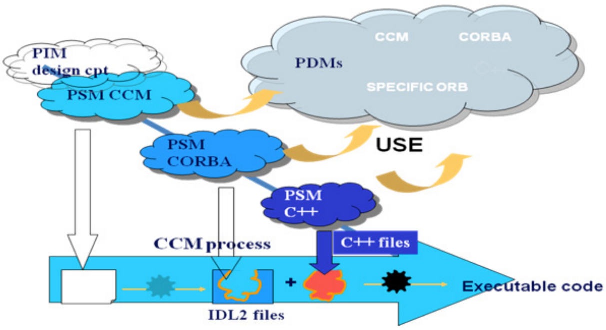


Fig. 6. Overall architecture for PIM to PSM translation

The case-study uses MDA to automatically generate the tool and associated GUI in Java and J2EE (session bean) in order to deliver the required embedded pattern and design guideline. Again, this paper only concentrates on the Java outputs.

Figure 6 shows the overall architecture for PIM to PSM translation. Transforming PSM based on SOA to the PSM based on e-business Services using WSDL is a straightforward task. In our approach, each value object and each interface in PIM will be transformed to WSDL Type and Port Type in the PSM respectively and the parameters of methods will be transformed to the Messages (Input/Output) in the PSM.

A transformation t transforms a model Ma into another model Mb t: Ma -> Mb. Model Ma is supposed based on meta-model MMa and model Mb is supposed based on meta-model MMb. We note this situation as: sem(Ma,MMa) sem(Mb,MMb). As a matter of fact, a transformation is like any other model. So we'll talk about the transformation model Mt. Mt: Ma->Mb. Obviously since Mt is a model, we postulate the existence of a generic transformation meta-model MMt, which would be similar to any other MOF based MDA meta-model:

In some cases the transformation takes some particular form if the source and target meta-models are in the relation of refinement like a CORBA and a CCM meta-model. Figure 7 shows the examples of translation interface.

```
mapping ParameterToInputPart (in UML2.Parameter) : WSDL.Part {
guard self.direction <> "return" {
        name := self.name;
        element := self.type.resolveByRule(
                "UMLTypeToWSDLElement", WSDL.Element);
        type := self.type.resolveoneByRule(
                "UMLTypeToWSDLType", WSDL.Type);
}

mapping ParameterToOutputPart (in UML2.Parameter) : WSDL.Part {
guard self.direction = "return" {
        name := self.name;
        element := self.type.resolveByRule(
                "UMLTypeToWSDLElement", WSDL.Element);
        type := self.type.resolveoneByRule(
                "UMLTypeToWSDLType", WSDL.Type);
}
```

Fig. 7. Example of translation Interface PIM to PSM

4 Conclusion

Service Oriented Architecture (SOA) is increasingly important in the business world as b2b transactions become ever more vital to business process out-sourcing and other co-operative activity. The problems of modeling solutions based on SOA have largely been resolved through the recognition of the importance of loose coupling and the consequent separation of concerns. Reinforced by the Supply-Manage-Consume concept, the separate modeling of solutions and services is a well established practice incorporated into advanced development processes that support SOA, including Select Perspective. Service Interfaces are shared amongst models showing the implementation and re-use of the services.

Whilst the use of modeling within SOA is well established, it has suffered from the same issues as modeling in other architectures. The abstraction gap between the level of detail expressed in the model and the level of detail expressed in the code is a key issue.

Yet it is the abstraction gap which is one of the key targets for the Model Driven Architecture.

Combining a service-oriented modeling architecture with MDA can bring many unique benefits. Firstly the clear organization of models and information based on the stereotypes derived from the service-oriented architecture and Select Perspective as development process. Secondly the productivity, quality and impact analysis benefits of the use of MDA with its emphasis on automation, transformation and synchronization. Select Solution for MDA is uniquely positioned to take advantage of the unified modeling architecture which results from bringing these two key architectures together.

In this paper we introduced an approach to modeling and design of complex distributed systems using SOA and MDA. In fact, to exploit the benefits of SOA effectively and duly, we propose an approach that involves MDA into the context. In this approach the PIM of the system is created and then the PSM based on SOA is generated. Then the final PSM based on a target platform is generated. These models are generated with transformation tools in MDA.

Acknowledgement. “This research was also supported by the International Research & Development Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning(Grant number: K 2012057499)”.

References

1. Bezivin, J., Hammoudi, S., Lopes, D., Jouault, F.: Applying MDA Approach for Web service Platform. In: Proceedings of the 8th IEEE Intl Enterprise Distributed Object Computing Conference (2004)
2. Huhns, M.N., Singh, M.P.: Service-Oriented Computing: Key Concepts and Principles. Journal of IEEE Internet Computing (2005)
3. Rahmani, A.T., Rafe, V., Sedighian, S., Abbaspour, A.: An MDA-Based Modeling and Design of Service Oriented Architecture. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006, Part III. LNCS, vol. 3993, pp. 578–585. Springer, Heidelberg (2006)
4. Gokhale, A., Natarajan, B.: Composing and Deploying Grid Middleware Web Services Using Model Driven Architecture. In: Meersman, R., Tari, Z. (eds.) CoopIS/DOA/ODBASE 2002. LNCS, vol. 2519, pp. 633–649. Springer, Heidelberg (2002), http://www.cydex21.com
5. Radhakrishnan, R., Wookey, M.: Model Driven Architecture Enabling Service Oriented Architectures. Whitepaper SUN Microsystems, pp. 1–13 (2004)
6. Skogan, D., Gronmo, R., Solheim, I.: Web Service Compositionin UML. In: Proceedings of the 8th IEEE Intl Enterprise Distributed Object Computing Conference, p. 111 (2004)
7. Mahmoud, Q.H.: Service-Oriented Architecture (SOA) and Web Services: The Road to Enterprise Application integration (EAI). Sun Developers Network (April 2005)
8. Corchado, J.M., De Paz, J.F., Rodríguez, S., Bajo, J.: Model of experts for decision support in the diagnosis of leukemia patients. Artificial Intelligence in Medicine 46(3), 179–200 (2009)
9. De Paz, J.F., Bajo, J., López, V.F., Corchado, J.M.: Biomedic Organizations: An intelligent dynamic architecture for KDD. Information Sciences 224, 49–61 (2013)
10. Rodríguez, S., de Paz, Y., Bajo, J., Corchado, J.M.: Social-based planning model for multiagent systems. Expert Systems with Applications 38(10), 13005–13023 (2011)
11. Bajo, J., De Paz, J.F., Rodríguez, S., González, A.: Multi-agent system to monitor oceanic environments. Integrated Computer-Aided Engineering 17(2), 131–144 (2010)

12. De Paz, J.F., Rodríguez, S., Bajo, J., Corchado, J.M.: Mathematical model for dynamic case-based planning. International Journal of Computer Mathematics 86(10-11), 1719–1730 (2009)
13. Corchado, J.M., Bajo, J., De Paz, J.F., Rodríguez, S.: An execution time neural-CBR guidance assistant. Neurocomputing 72(13), 2743–2753 (2009)
14. Závodská, A., ŠRamová, V., Aho, A.M.: Knowledge in Value Creation Process for Increasing Competitive Advantage. Advances in Distributed Computing and Artificial Intelligence Journal 1(3), 35–47 (2012)
15. Satoh, I.: Bio-inspired Self-Adaptive Agents in Distributed Systems. Advances in Distributed. Computing and Artificial Intelligence Journal 1(2), 49–56 (2012)
16. Agüo, J., Rebollo, M., Carrascosa, C., Julián, V.: MDD-Approach for developing Pervasive Systems based on Service-Oriented Multi-Agent Systems. Advances in Distributed Computing and Artificial Intelligence Journal 1(6), 55–64 (2013)

Intelligent Lighting Control System

Elena García[1], Sara Rodríguez[1], Juan F. De Paz[1], and Javier Bajo[2]

[1] Computer and Automation Department, University of Salamanca, Spain
[2] Artificial Intelligence Department, Polytechnic University of Madrid, Spain
{elegar,srg,fcofds}@usal.es, jbajo@fi.upm.es

Abstract. This paper presents an adaptive architecture that allows centralized control of public lighting and intelligent management, in order to economise on lighting and maintain maximum comfort status of the illuminated areas. To carry out this management, architecture merges various techniques of artificial intelligence (AI) and statistics such as artificial neural networks (ANN), multi-agent systems (MAS), EM algorithm, methods based on ANOVA and a Service Oriented Aproach (SOA). It performs optimization both energy consumption and economically from a modular architecture and fully adaptable to the current lighting systems possible. The architecture has been tested and validated successfully and continues its development today.

Keywords: Light sensors, intelligent systems, distributed systems, Autonomous control, Street lighting.

1 Introduction

Nowadays, the concept of Smart Cities is increasingly a common trend in technological projects. The balance with the environment and natural resources is a practical and responsible key for these paradigms, which aim to achieve a state of comfort for citizens and institutions based on sustainable development. In this respect, talk about energy efficiency is paramount, not only to reduce energy costs, but also to promote environmental and economic sustainability.

One of the main costs faced by councils in towns and cities is the lighting bill. According IDEA [1], throughout the 2010 in Spain there were about 4,800,000 points of light with an average power of 180 W and 4,200 hours of annual use. Representing an electricity consumption of 3,630 GWh / year, this is excessive consumption. The technological advances that are experiencing external lighting installations along with its intelligent use will allow reducing that as high consumption.

This research appears from a greater project by the research group of BISITE (Bioinformatics, Intelligent Systems and Technology Education) of the University of Salamanca, which is to build a system that allows centralized control street lighting as well as an intelligent management, in order to economize on illumination and maintain maximum comfort status of the illuminated areas. This is to avoid excessive illumination of areas, as there are many times that it is not necessary to maintain maximum light intensity for an optimal service.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_24, © Springer International Publishing Switzerland 2014

To validate the system in an experimental research level, a small test case is available, consisting of a portable installation of 5 luminaires with specific hardware. The functionality to be achieved is divided into two modules.

One module shall be responsible for the direct management of the various installations and control panels and will serve as a communication layer with each site, so as to allow as much control and monitoring in almost real-time of each facility and even luminaire. In this sense, the system must provide a service interface that can be accessed for each installation using a standardized interface independent of the underlying technology and hardware of each installation.

The other module, explained in this article, will have as its main objective the management of the lighting schedule for each installation, consumption management, and prediction. In this regard, a light planning is defined as light output level for each installation offered hourly. This light planning must be possible by programming user preferences, or by observing different decisive environment factors in determining the appropriate level of brightness for each site at each time. Thus, different factors come into play: astronomical clock, weather and traffic and pedestrian flow. Moreover, it will be interesting to make a prediction of consumption by the light patterns assigned to each zone, depending on its economic rate.

To carry out this management, the built system combines different statistics and artificial intelligence (AI) techniques such as artificial neural networks (ANN), multi-agent systems (MAS), EM algorithm, methods based on ANOVA and a Service Oriented Approach (SOA) [5].

The article is structured as follows: Section 2 shows a state of the art concerning projects and research conducted in the field of Smart cities and light control, showing the most commonly used techniques in this field and carry out a comparison between them and the system presented. Section 3 shows the presented system, its operation and details of the techniques used. Section 4 describes the case study developed for system validation and finally, Section 5 some results and conclusions of this work.

2 Background

The concept of smart cities, smart environments, or smart homes [2] itself is still emerging in our society. Make a "smart" city is one of the objectives currently most often heard at the research as a strategy to mitigate the problems caused by the rapid growth of the urban population. Problems such as lack of resources, pollution, traffic congestion and deteriorating infrastructure are some of the many problems that increasingly large urban populations face [3].

One of the many definitions of Smart Cities is: "*The use of smart computing technologies to make city services more intelligent, interconnected and efficient - which includes administration, education, health care, public safety, real estate, transportation and utilities.*"[4]. It seems clear that the purpose of these is sustainable economic development, based on new technologies (ICT) to provide better quality of life and prudent management of natural resources through the engagement of all citizens.

Today, more and more cities around the world are committed to develop pilot projects related to this movement, some even in Spain, such as SmartSantander [1]: for now the city has a great display of parking sensors to indicate to drivers the free sites. They also have a municipal Wi-Fi network that aims to cover the entire village, and even augmented reality applications to boost tourism. Málaga Smart City[2]: the project aims at saving energy by micro power management: energy storage in batteries for use in buildings, street lighting and electrical transport, promoting the use of electric cars, etc. Smart City Valladolid-Palencia[3]: considers two cities, adding transport between them as a problem and has smart meter network, integration of electric cars, energy efficiency in buildings, traffic organization, etc.

The current research works include the implementation and control of distributed lighting systems to facilitate the implementation of new infrastructure in a city or the optimization of existing infrastructure; further integration with other control systems and optimization of heating, cooling or controlling air quality. For instance, in [6] it presents a systematic approach to the modeling, optimization, control, and adaptation in a color-tunable LED lighting control system. Through light sensor feedback, the control system is able to achieve significant energy savings without substantially sacrificing lighting quality. The key techniques used here are an appropriate choice of cost function based on color metrics and the trade-off between quality of light and energy consumption for LED lighting systems. The authors in [7] employ formal methods for design a graph model, accompanied by means of control, including AI methods (rule-based systems, pattern matching) to design and control an outdoor lighting system. In this case, the work is focused only on the design phase and the control phase designing features such as dynamic, sensor-based control, multiple luminaire states and complex geometries. Other research on lighting control systems base their operation in image processing [8], fuzzy systems [9], cooperative methods and wireless sensor network (WSN) [10] or simulation algorithms [11] and predictive control [12] for energy optimization.

There are also some tools already developed as Lites[4], that has temperature sensors, ambient light, power, motion detection; CityLight[5], that allows remote management of lighting, fault detection and planning lighting patterns manually or Tvilight[6] that regulates the lighting based on presence sensors and maintains minimum brightness in inactive hours.

[1] http://www.smartsantander.eu/

[2] http://www.lacatedralonline.es/innova/system/Document/attachments/12351/original/IDCCiudadesinteligentes.pdf

[3] http://www.valladolidadelante.es/lang/modulo/?refbol=adelante-futuro&refsec=smart-city-vyp&idarticulo=79302

[4] http://www.lites-project.eu/lites-led-based-intelligent-street-lighting-energy-saving

[5] Teliko http://www.teliko.com/en/

[6] http://www.tvilight.com/

This paper presents an adaptive architecture that allows centralized control of numerous public lighting installations. Specifically, it allows distributed and real-time intelligent control based on prediction and analysis techniques of lighting, one of the main shortcomings of the systems listed above. From a modular architecture fully adaptable to the current lighting systems possible, an energetic and economic optimization is possible. The architecture has been tested and validated successfully and continues its development today. The following sections describe the operation and technologies used in it and the results currently obtained.

3 Proposed Architecture

The system presented aims to frame the intelligent management of all public lighting, including monitoring and real-time control of the lights, and the establishment of lighting patterns that fit with the use of the public highway installations.

The following figure shows the context of the system, mainly composed of the control software (Intelligent Street Lighting Software) and the set of public-lighting installations, accessible via the internet. Facilities include special hardware for global and individual control of each luminary, while communication between devices is done by PLC. The control software is composed of three modules:

- The hardware abstraction layer allows communication with facilities regardless of the underlying hardware.
- The management server contains both device management and intelligent algorithms for efficient energy management. The "*Data sources*" module captures information related to pedestrian and traffic flow, weather data, and data about the monitoring of the facility. The "*Data analysis*" module deals with the study of information collected for the detection of foot traffic patterns, management of neural networks to predict consumption from light intensities, and estimates of consumption online. Finally, the *"Luminosity patterns generator"* module allows the creation of adequate light planning suitable for the specific facility lighting depending on the standards of pedestrian flow and weather conditions of each case.
- The web application provides access to all functionality for configuring lighting schedules, monitoring and control of facilities.

The ideal goal of the street lighting architecture design is that it can work well and provide safe and stable street lighting control for our daily life without human intervention. But human users should know whether the system is working normally or not. So the interaction between the system and human users is necessary. The system should also be controlled by human users manually in some particular situations. The system includes the ability to automatically interact ("smartly") or manually according to the lighting used and the needs of the specific case study.

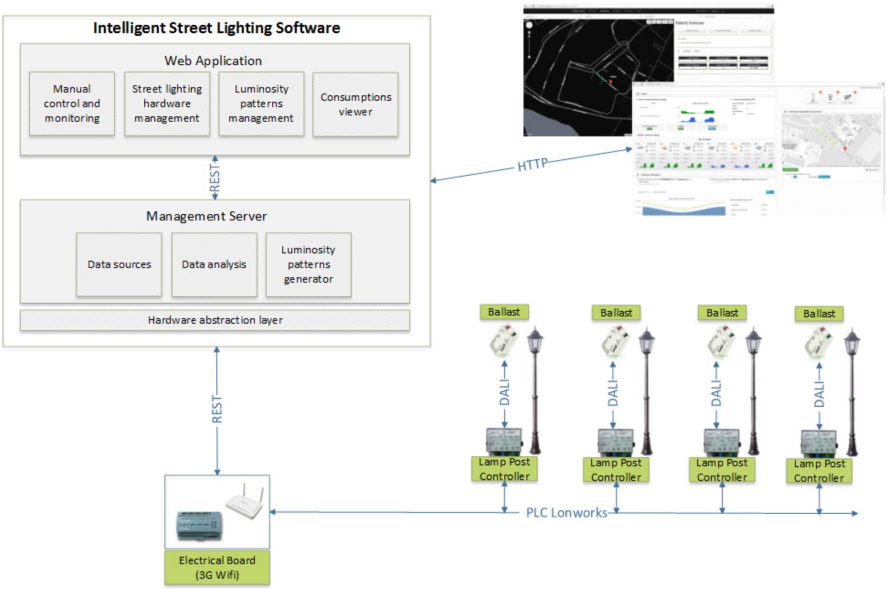


Fig. 1. Main system components

3.1 Work Flow

The establishment of adequate lighting configuration for each environment is one of the objectives of this project, meaning light configuration like a set of times at which the area is illuminated and the brightness levels associated with each time. This will save on lighting consumption, maintaining the state of maximum comfort in the lighted areas, as there are many occasions where it is not necessary to maintain a maximum level of light intensity to provide optimal service to the area, causing excessive consumption.

In the presented system the light designs can be set in a manual or smart way. In the first, the user is encouraged to plan the time slots (in hours) and the luminous flux of each time slot. In the second way, we proceed to the observation of different environmental factors that may be influential in determining adequate lighting for the particular area, such as flow or pedestrian traffic, or weather conditions each time, which influence the level of ambient light, especially near the hours of sunrise and sunset times.

The flow diagram of Figure 2 shows the procedures to complete the light patterns depending on ambient factors and the different user preferences. It is possible to observe, with common parts, two different workflows, which correspond to the process of generating light patterns for a given period of time. One of the flows can generate patterns without establishing a maximum estimated expenditure, and the other, setting it. Maximum expenditure means the maximum amount to spend on lighting bill for the period over which the light patterns are concluded.

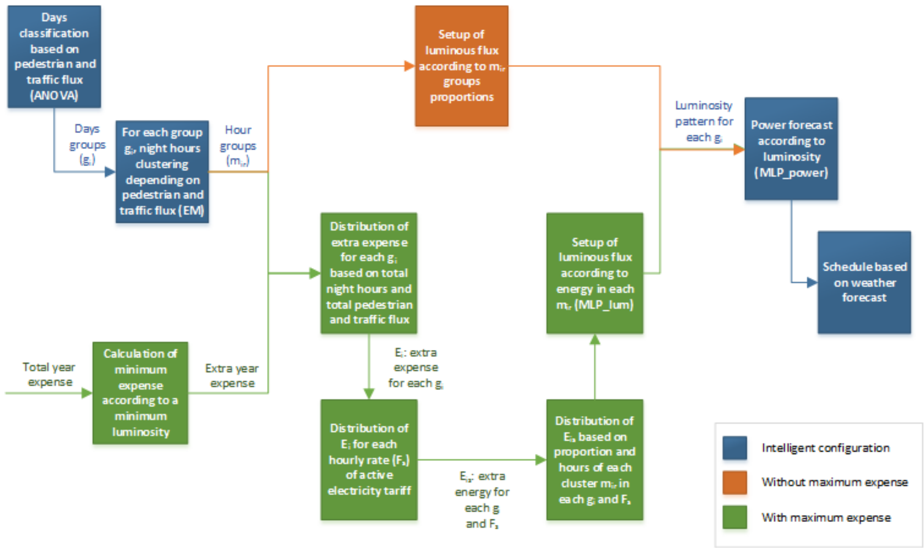


Fig. 2. Workflow procedures followed for lighting

Both processes share the initial logic. As a first step, historical data of pedestrian and traffic flow of several weeks are taken. Subsequently, a classification is performed by analysis of variance (ANOVA) [13] to determine what day of the week patterns share pedestrian and traffic flow according to the different hours of the period.

The periods (days) that share the same characteristics, according to this analysis, also will share luminous pattern. To determine the set of similar days is necessary to take into account different variables such as day of the week, time of day and the volume of people / traffic that is located in that time slot and day of the week. Not being quantitative variables is necessary to apply cluster techniques such as ANOVA to draw similarity between groups. Here is to be applied two-way ANOVA with repetition. The factors are the day of the week and the time slot, the time slot is considered factor group.

After obtaining the groups of days, for each group is applied a clustering algorithm (Expectation-maximization EM) [14][15] to determine at what time of night usually spend a similar number of pedestrians or traffic. These clusters of hours generated for each group of days will result in different light levels appropriate to pedestrian or traffic flow.

After these two steps, the flowchart shows a bifurcation. The left branch corresponds to the process to follow if you do not set a maximum expenditure, which mainly follows a simple process of adjusting lighting levels based on the proportion of generated clusters after EM technique for each group of days. Moreover, the right branch serves logic followed when a maximum flow is established. The steps followed in this last branch focus primarily on optimal distribution of the amount with which it has to provide the same light levels in scenarios with similar environmental characteristics. At all times a minimum configurable brightness is guaranteed, because in order to comply with the appropriate legislation, the area will not be left with insufficient lighting although the amount entered by the user is less than the cost of this. The distribution amount is performed based on features such as hourly rate, with or without time restrictions, the evening hours that can affect and the proportion of clusters based on traffic and pedestrian flows.

The two branches obtain different light patterns depending on the groups of days generated with the ANOVA technique. The penultimate step in the flowchart is shared and consists of the prediction of the spending of the lighting design completed, which in the case of the branch with maximum expenditure, will coincide with a small margin of error depending on the expenditure estimation technique used. The estimate of expenditure is performed by a neural network MLP (Multi-Layer Perceptron) [16][17] that predicts power level in function of lumens and is trained with historical data of the luminaire type used in each installation.

The other shared step in the workflow, optional for users, is a replanning that is performed periodically to adjust light patterns established previously to climatic conditions. This process consists in checking the prediction of the weather to advance or delay the time on and off lights in the hours of dawn and sunset. This is for that the lighting design conforms to the lighting conditions of the place in which the system is installed. In this way, for any day in which bad weather (rain, fog, etc.) is expected, the system will turn off the luminaries sooner or later that the usual hours, coinciding with the hours of dawn and sunset each day. This process is repeated weekly, so that the light patterns are sent weekly to the control node of the area.

3.2 Distribution of Expenditure

To calculate the distribution of the maximum E_T in Z time entered by the user is first necessary to calculate the minimum amount of expenditure E_{min} to a minimum brightness L_{min}. This time period has Nh overnight hours. One MLP network is used to predict the spending power of the luminaires used depending on the required level of brightness. The additional expenditure E will be distributed for generating light patterns.

$$\begin{gathered} L_{min} \rightarrow RNA_{pow} \rightarrow Pow_{min} \\ E_{min} = Pow_{min}\, Nh \\ E = E_T - E_{min} \end{gathered} \tag{1}$$

The first step taken is to distribute the amount E between groups of days gi generated in the ANOVA process. This distribution is based on the number of hours of night Nhi of each group i and pedestrian and traffic flow P_i that exists in that group.

The calculation of traffic and pedestrian flow P_i of each g_i is done by average people who go through every night $\overline{Pd_\iota}$ and the number of days D_i that belongs to g_i, taken the average of the historical data used before (Fig 1). The number of people is limited, and is equal to the upper bound in case of exceeding this bound.

$$\begin{gathered} P_i = \overline{Pd_\iota}\, D_i \\ \overline{Pd_\iota} = \frac{1}{J\,K} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk} \,/\, \mathrm{j} \in \mathrm{g}_i \end{gathered} \tag{2}$$

Both Nh and P variables can have different degrees of influence ρ at the time of allocation of the amount E. The extra expense of each group gi is given by equation (3).

$$E_i = \frac{Nh_i}{Nh} E\,\rho_{Nh} + \frac{P_i}{P} E\,\rho_P \tag{3}$$

Where:

$$Nh = \sum_{i=1}^{I} Nh_i \;\; ; \;\; P = \sum_{i=1}^{I} P_i$$
$$0 \le \rho_{Nh} \le 1 \;\; ; \;\; 0 \le \rho_P \le 1 \;\; ; \;\; \rho_{Nh} + \rho_P = 1 \tag{4}$$

Once the distribution of expenditure for each group is done, E_i, a process recurs distribution thereof, in this case between different times Fs of stating the electricity tariff associated with the area to be illuminated. To do this, we take into account the proportion (w_{ir}) and time of use in hours (n_{ir}) in each group m_{ir} (which represents the values that are classified in the EM_r cluster of each group g_i ANOVA), and the price of the energy in each time slot, L_s.

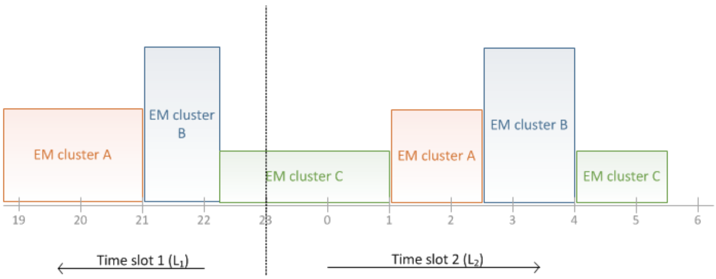


Fig. 3. Area of time slots

The figure above shows graphically a possible deployment scenario clusters m_{ir} result of the EM algorithm, where each cluster is represented by a rectangle. The x axis represents time in hours, while the y-axis represents the proportion of clusters (w_{ir}), which is determined by the average of pedestrian and traffic flow determined for each group m_{ir}. The distribution of expenditure E_i is done by calculating the total area of each time slot F_s weighted price of energy in these slots. Thus, a fair distribution of the expenditure is insured for, thus able to illuminate with the same light flow spaces schedules with similar environmental factors (3).

$$E_{is} = \left(\sum_{r=1}^{R_i} n_{ir}\, w_{ir} \right) L_s\, x$$
$$E_i = \sum_{s=1}^{S} E_{is} \tag{5}$$
$$w_{ir} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk}}{IJK} \;\; / \; \mathrm{j} \in \mathrm{g}_i \, , \mathrm{i} \in \mathrm{m}_{ir}$$

Once the distribution of expenditure for each time slot F_s of each group of days g_i is made, the price of energy does not vary for each slot of the group. In this step, each expendture E_{is} becomes equivalent to the energy consumed in the slot s of the days that belong to the group i during the period of time Z, that the user chose at the beginning of the process. The distribution of extra power (it will continue denoting as E_{is}) for each hour $h \in H_{irs}$ (set of hours that belong to F_s slot grouped in subgroup m_{ir} hours, of the initial group g_i) was performed similarly to the previous step manner, to continue to ensure a fair distribution. It will consider both the size of cluster w_{ir} as the number of hours used n_{ir}.

$$p_h = w_{ir}\, x \;\; / \, h \in H_{irs}$$
$$E_{is} = \left(\sum_{r=1}^{R_i} n_{ir}\, w_{ir} \right) x \tag{6}$$

The expression (6) denotes p_h as the extra power to supply each hour $h \in H_{irs}$. In this way, the power to supply in each hour to all luminaries (p_{th}) will be the minimum power (1) of each hour more extra power calculated p_h.

$$p_{th} = \frac{Pow_{min}}{Nh} + p_h \tag{7}$$

At this point, the power supply to each luminaire in every hour (simply dividing p_{th} by the number of luminaires) is known. For the resulting light output of the power supplied a MLP network that predicts light output from input power is utilized. The equivalence between luminous flux and power tends to be linear, so this approximation is quite accurate: in most cases, if you double the power is spent, it will get twice the luminosity. However, the use of MLP networks to predict luminosity based on power or vice versa, is performed for the sections in which it is not linear: each model of luminaire can light to a minimum power; moreover the expenditure of each luminaire can be influenced by the facility in which it is (because it is necessary more hardware, etc.). Thus, the system achieves more accurately approximate the luminosity that is going to have with a certain power. Or conversely, the cost of having the lights burning with a certain luminous flux. Knowledge of the luminous flux which has every hour for each group g_i presupposes already done the lighting pattern to be followed for each group of days. Whereupon, planning lighting design is approximated to the expenditure that the user wants to spend and to the pedestrian and traffic flow patterns in the area, depending of the desired degree of influence set.

4 Case Study

To develop the first prototype a hardware solution installed in a briefcase was purchased. Fig.4 shows a photo of the prototyping environment that emulates an installation of five street lamps with a control node. Four lamps interior and one exterior were used to test the control system while maintaining low cost hardware. Each lamp is controlled by an adjustable ballast that is behind it, hidden in the top panel. A luminaire controller regulates and monitors each ballast and liaises with the control node.

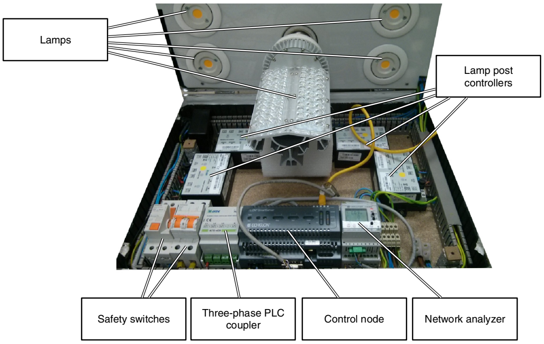


Fig. 4. Prototyping environment that simulates an installation of five streetlights

Luminaire controllers (ISDE brand ASL-510-TCH) are placed inside each lamp and communicate via PLC with the control node, but also can be placed at the beginning of a line of street lamps. These controllers interpret commands received through the line to regulate the output of the ballasts of the lamps using the DALI protocol. Also they monitor the status of the lamps, the instantaneous consumption and power supply of each lamppost.

The chosen control node is an Echelon Smart Server, a general purpose controller for automation of non-critical processes. In street lighting systems, is able to control and monitor up to 192 single or double lamps head through PLC. It offers a SOAP interface for configuration and remote management that has been used for integration with the developed system. The PLC signal injected by the control node replicates in three phases using a phase coupler [three-phase coupler PLC]. The network analyzer is the CVM-MINI brand model Circutor. It connects via a parallel port RS-485 to Echelon SmartServer, with which it communicates using the MODBUS protocol. For the prototype system for estimating pedestrian flow IP camera TP-LINK DSC-942L interior placed on a window was used.

5 Results, Conclusions and Future Work

Figure 5 shows some of the results obtained by using the system. The upper panels show data of pedestrian flow in two consecutive weeks (one week in purple and the other blue). After applying the analysis of variance traffic patterns are detected and rated day is done, resulting in two groups: green, weekdays, and blue, on weekends. The lower graphs show the generated light designs for each group of days. Using the EM algorithm, hours with similar traffic are detected, adjusting a luminosity level in each group.

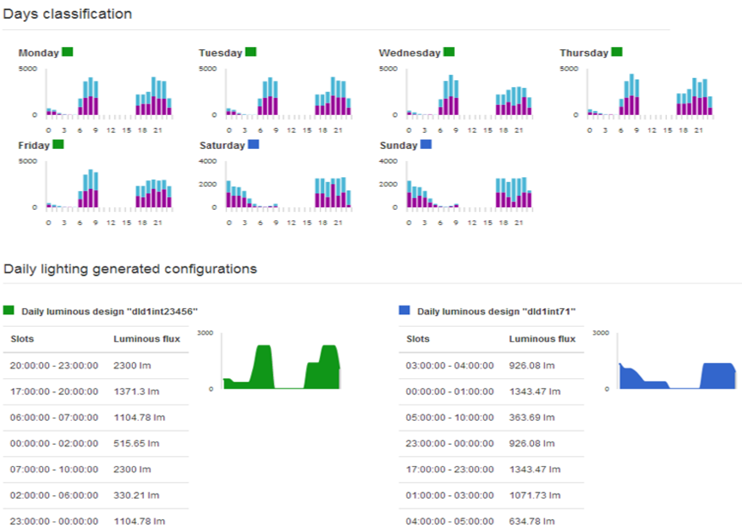


Fig. 5. Results

Figure 6 shows a prediction of both daily consumption and annual calendar designed. The green line represents the expense of having all lights full brightness during nighttime hours. The blue area represents the estimated model to the application of consumption. In this case, an approximate savings of 25% is achieved while maintaining the maximum light intensity at peak traffic and pedestrian flow.

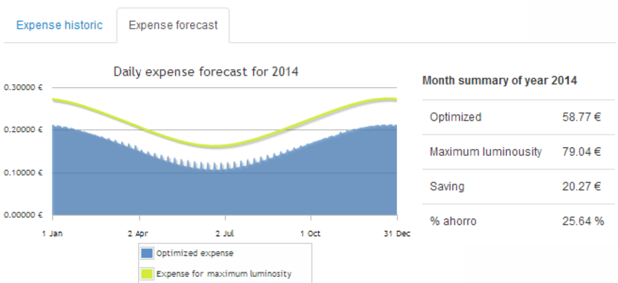


Fig. 6. Prediction of daily and annual consumption

The system is able to set lighting schedules all public lighting installations controlled and can say that this measure contributes to energy savings achieved by regulating the light intensity. The user can define their own light patterns, where lighting pattern is the hours in which the lights are on and at what level of brightness. Likewise, it is possible to assign different light patterns each day.

Furthermore, not only the lighting schedule can be established according to the user preferences, but the system has the necessary mechanisms to automatically adjust light levels based on traffic or pedestrian flow provided for each zone. To achieve this, the system is based on historical information of people flow and makes a clasification of days to find patterns of pedestrian/traffic flow. Based on these patterns, the system establishes an appropriate lighting design for each type of day. In this sense, we have applied intelligent techniques and algorithms (ANOVA, EM clustering technique, MLP) correctly and a process that fuses all together for the conclusion of the lighting schedule is made. In addition, a distribution algorithm that reduces spending and complies with the minimum luminosity and brightness levels at all times is presented. Finally, the application also allows the user to query historical data related to the luminance calendars that have been established on site, and the historical use of them.

In conclusion say that it is very difficult to find prototypes that are based on historical data of pedestrian and traffic flow to adjust the luminosity of the areas. The systems are often reactive, not predictive. The main reason for developing the system is based on the prediction of pedestrian / traffic flow is the savings in hardware. Place a camera in the area for pedestrians and vehicles spot for a while, is much cheaper than having every few luminaires a presence sensor that regulates the brightness depending on the passage of pedestrians and vehicles, in addition to the constant change light intensity emitted by the luminaires could punish excessively. Future work will focus on the following three aspects. (1) Add other sensors to the lamp member and investigate how to use sensor fusion to further improve intelligence level of the system. (2) Develop a system of alerts that happen in the real-time hardware: cast a light, overvoltage on the line, etc. (3) Develop new algorithms to make the lamp members cooperate with each other.

Acknowledgements. This work has been carried out by the project *PIRSES-GA-2012-318878*.

References

[1] IDAE, Instituto para la Diversificacion y Ahorro de la Energía, http://www.idae.es/ (last access January 12, 2014)

[2] Gómez-Romero, J., Serrano, M.A., Patricio, M.A., García, J., Molina, J.M.: Context-based Scene Recognition from Visual Data in Smart Homes: An Information Fusion approach. ACM/Springer Journal of Personal and Ubiquitous Computing. Special Issue on Sensor-driven Computing and Applications for Ambient Intelligence 16(7), 835–857 (2012)

[3] Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J.R., Mellouli, S., Nahon, K., Pardo, T.A., Scholl, H.J.: Understanding Smart Cities: An Integrative Framework. In: 45th Hawaii International Conference on System Sciences, pp. 2289–2297 (2012)

[4] Washburn, D., Sindhu, U., Balaouras, S., Dines, R.A., Hayes, N., Nelson, L.E.: Helping CIOs Understand "Smart City" Initiatives. Growth 17 (2009), http://c3328005.r5.cf0.rackcdn.com/73efa931-0fac-4e28-ae77-8e58ebf74aa6.pdf

[5] Daigneau, R.: Service Design Patterns: Fundamental Design Solutions for SOAP/WSDL and RESTful Web Services, 1st edn. Addison-Wesley Professional (2013)

[6] Afshari, S., Mishra, S., Julius, A., Lizarralde, F., Wason, J.D., Wen, J.T.: Modeling and control of color tunable lighting systems. Energy and Buildings 68(Pt. A), 242–253 (2014)
[7] Wojnicki, I., Ernst, S., Kotulski, L., Se¸dziwy, A.: Advanced street lighting control. Expert Systems with Applications 41(4 Pt. 1), 999–1005 (2014)
[8] Carrillo, C., Diaz-Dorado, E., Cidrás, J., Bouza-Pregal, A., Falcón, P., Fernández, A., Álvarez-Sánchez, A.: Lighting control system based on digital camera for energy saving in shop windows. Energy and Buildings 59, 143–151 (2013)
[9] Yang, C., Fan, S., Wang, Z., Li, W.: Application of fuzzy control method in a tunnel lighting system. Mathematical and Computer Modelling 54(3-4), 931–937 (2011)
[10] Zhang, J., Qiao, G., Song, G., Sun, H., Ge, J.: Group decision making based autonomous control system for street lighting. Measurement 46(1), 108–116 (2013)
[11] Bourgeois, D., Reinhart, C., Macdonald, I.: Adding advanced behavioural models in whole building energy simulation: A study on the total energy impact of manual and automated lighting control. Energy and Buildings 38(7), 814–823 (2006)
[12] Barra, K., Rahem, D.: Predictive direct power control for photovoltaic grid connected system: An approach based on multilevel converters. Energy Conversion and Management 78, 825–834 (2014)
[13] Miller Jr., R.G.: Beyond ANOVA: basics of applied statistics. CRC Press (1997)
[14] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical 39, 1–38 (1977)
[15] Little, R., Rubin, D.: Statistical Analysis with Missing Data. Wiley & Son, New York (2002)
[16] Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice Hall (1998) ISBN 0-13-273350-1
[17] Collobert, R., Bengio, S.: Links between Perceptrons, MLPs and SVMs. In: Proc. Int'l Conf. on Machine Learning, ICML (2004)

Norm's Benefit Awareness in Open Normative Multi-agent Communities: A Conceptual Framework

Al-Mutazbellah Khamees Itaiwi[1], Mohd Sharifuddin Ahmad[2], Moamin A. Mahmoud[2], and Alicia Y.C. Tang[1]

[1] College of Graduate Studies, Universiti Tenaga Nasional,
[2] College of Information Technology, Universiti Tenaga Nasional,
43000 Kajang, Selangor, Malaysia
mutazhiti@yahoo.com, {sharif,aliciat}@uniten.edu.my,
moamin84@gmail.com

Abstract. In open normative multi-agent communities, agents adopt new norms to increase their utilities. Several studies have developed mechanisms for agents to adopt new norms. These mechanisms are based on sanction, imitation, or social learning. The limitation of these mechanisms is that over time all agents follow the new norms, while in a real situation, usually there is a number of agents that persistently violate the norms for their benefits. We consider that intelligent agents should adopt new norms based on their awareness of the norms' expected benefits on their utilities and not only by sanctions or imitating other agents. Consequently, this paper presents a conceptual framework for agents' awareness of norms' benefits in open normative multi-agent communities. In the proposed framework, four components constitute agents' awareness of norms' benefits which are Norm's Adoption Ratio; Norm's Yields; Norm's Trust, and Norm's Morality. Using these components, however, agents would be able to evaluate the benefits of detected norms and subsequently determine whether the norms increase or decrease their utilities for eventual adoption or rejection.

Keywords: Social Norm, Normative Systems, Norms Benefit Awareness, Normative Multi-agent Systems, Software Agents.

1 Introduction

Norms and normative multi-agent systems have become the subjects of interest for many researchers [1, 2]. The term norm is used to characterize the behaviors of community members and is generally accepted as efficient means to normalize their behaviors [2]. Recently, numerous models on multi-agent systems have been investigated and this includes norms in agent architectures [3, 4, 5]. Boella et al. [6] claimed that researchers of moral and legal philosophy have studied traditional normative systems [7].

In open normative multi-agent communities, agents adopt new norms to increase their utilities [8]. Several studies have developed mechanisms for agents to adopt new

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_25, © Springer International Publishing Switzerland 2014

norms, which are based on sanction [9, 10], imitation [11, 12, 13], or social learning [14, 15]. The limitation of these mechanisms is that over time all agents comply with the norms. Such agents possess inferior autonomies as their decisions are influenced by external powers, which they subsequently comply. In real situations, there is always a number of agents that persistently violate the norms for their benefits (i.e., deviant agents). For more advanced ability, we stress that intelligent agents should adopt new norms based on their awareness about the expected benefits on their utilities and not only by sanctions [9, 10] or imitating other agents [11, 12, 13]. In the real world, we occasionally see some people violate the norms despite the influence of such external powers due to their awareness of the expected benefits of such violations. Consequently, this paper addresses this issue for investigation by introducing a conceptual framework of agents' awareness of norms' benefits in open normative multi-agent communities where norms are adopted based on their expected benefits.

The concept of norms' benefits awareness enables agents to recognize the benefits or gains arising from complying or violating the norms. Agents' that interact with internal and external environments need to be aware of the norms' benefits and use their abilities and motivations to exploit these benefits for achieving their goals. However, to achieve coordination and cooperation in open communities, Criado et al. [16] suggested that norms should represent an effective tool and agents must be able to adopt norms autonomously based on the yields gained.

This paper, which reports the work-in-progress of our research in norms, presents a conceptual framework for agents' awareness of norms' benefits. In this framework, we discuss, first, the components that constitute the norms' benefits awareness and subsequently present the proposed framework. Four components are observed from the literature that constitute norms' benefits awareness which are Norm's Adoption Ratio (they referred to it as norm strength) [13]; Norm's Yields [17]; Norm's Trust [18], and Norm's Morality [9]. Using these components, agents evaluate the benefits of the norms and subsequently determine whether the norms increase or decrease their utilities for eventual adoption or rejection.

The objectives of this study are, (i) to highlight the issues of norms' benefits awareness in open multi-agent communities of online-designed norms, and (ii) to develop a norms' benefit awareness framework. This paper is the first research attempt of its kind that analyzes and formulates the phenomenon of norms' benefits awareness in a multi-agent community. Our contribution in this paper is three-fold. Firstly, we analyze the components that constitute norms' benefits awareness. Secondly, we develop a framework for norms' benefits awareness. Thirdly, we propose an approach that can be used to calculate the benefits of a particular norm.

The next section dwells upon the related work on norms adoption. Section 3 discusses the components of norms' benefits awareness. Section 4 details out the framework of norms' benefits awareness and propose a norms' benefits calculation method. Section 5 concludes the paper.

2 Related Work

Norms and normative systems have received greater attention due to their ability to coordinate agents' interactions [1, 2]. In general, norms are integrated with multi-agent systems to guide agents to behave in a socially harmonious way. Consequently, research on norms identification and detection have recently appeared in the literature expounding issues and techniques on this research domain [9, 13, 19, 20, 21]. For example, Savarimuthu et al. [19, 20] proposed a norm identification technique in which a visitor agent infers the norms of an agents' community without the norms being explicitly given to the agent. The agent infers the norm based on a sanctioned agent and after inferring, the agent adopts the norm. In another work by Mahmoud et al., [9] they proposed an algorithm to detect obligation and prohibition norms which they called the Obligation and Prohibition Norms Mining algorithm (OPNM). The algorithm exploits the resources of the host system, implements data formatting, filtering, and extracting the exceptional events, i.e. those that entail rewards and penalties of the obligation and prohibition norms.

Mahmoud et al. [13] presented a self-enforcing agent which detects norms based on beliefs that are triggered by emotions of imitating the majority and the agent extracts the norms using a norms mining algorithm.

Sen and Airiau [14] proposed a social learning theory, in which every agent in a community learns simultaneously from repeated interactions with randomly selected neighbors. Bosse et al. [15] presented a dynamic agent-based approach to simulate and formally analyze the process of social learning of agents' behaviors. The general mechanism is based on behavior changes by influence of peers. The approach involves the influence of three types of agents groups which are peers, parents, and school.

From the above works, we notice some deficiencies in the approaches that are based on sanctions, imitations or interactions. In this paper, we extend a further refinement to these approaches by enabling agents to evaluate the benefits of norms' awareness that would improve agents' decisions in adopting or rejecting a particular norm.

3 Norm's Benefit Awareness Components

We propose the concept of norms' benefits awareness to enhance the ability of normative agents in norms compliance or violation. Such ability to recognize the norms' benefits would greatly improve the agents' performance in achieving their goals. We define, initially, the underlying elements that constitute this concept.

Definition 1: Norm's Awareness is the ability of an agent to recognize the norms in open agent communities, in which norms are not explicitly given to agents.

Definition 2: Norm's benefits awareness is the ability of an agent to determine the gains or losses from adopting or violating a particular norm in open agent communities.

Definition 3: An agent's utility is a measure of its usefulness in a particular domain.

From the literature, we identify the components that constitute the norms' benefit awareness, which are, Norm's Adoption Ratio, Norm's Yield, Norm's Trust, and Norm's Morality. We justify the significance of these parameters in developing the framework by assessing the influence of each parameter on the decision of agents to adopt or reject a norm.

- Norm's Adoption Ratio (N_{AR}): A Norm's Adoption Ratio is the ratio of agents practicing a particular norm to the population of agents in a community. If P is the agents' population, and N_a is the number of agents practicing a particular norm, then $N_{AR} = N_a{:}P$. This ratio is high when a norm is enacted by a majority of agents, which experience the norm's benefits. Such experience reinforces an agent's decision to enact the norm and gain the expected benefits or violate the norm to avoid expected losses. For example, in an elevator scenario, if the majority practices the norm of *excusing* oneself when exiting the elevator, an agent expects that the benefits from adopting such norm avoids it from a sanction and/or increases its reputation.
- Norm's Yield (N_Y): To adopt a norm, an agent should consider the expected yield of that norm. A norm's yield refers to the gain received from adopting a norm arising from the norm's return on an agent's utility. When an agent discovers the yield value of a particular norm, it infers the benefits of adopting the norm. If the norm possesses high yield, it attracts agents to adopt it. For example, reading news online becomes the norm of many communities because it is inexpensive and convenient.
- Norm's Trust (N_T): Another parameter that motivates an agent to adopt a norm is when the agent is able to evaluate a norm's trust value. A norm's trust refers to the degree of an agent's belief in a norm that influences other agents to adopt the norm. If the trust value of a particular norm is high, it increases the possibility of adopting the norm. Andrighetto et al. [18] provided an example of a bus stop scenario of a particular community, in which when people arrive at the bus stop, the norm is, they do not form a queue but instead they sit on a bench and memorize who came earlier than them. In such situation, because people highly trust the norm, they adopt the norm.
- Norm's Morality (N_M): A norm's morality refers to the state of a norm (good or bad) in comparison with a morality reference. The morality value of a norm allows an agent to check whether the norm conforms to its morality reference. If it conforms, the probability of adopting the norm is high and vice versa. For example, talking loudly or shouting is a low morality norm for many communities. But if it is detected as a strong norm in a particular community, in this case, the agent has the option to accept or reject the norm basing on its awareness about the norm's expected benefits.

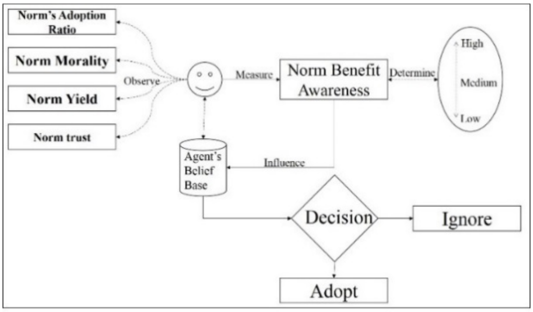


Fig. 1. Evaluating the Norm's Benefit Awareness

If an agent is able to determine the values of the above parameters, it can evaluate the Norm's Benefit Awareness. As shown in Figure 1, the agent becomes aware of the norm's benefit from observing and evaluating the parameters values (Norm's Adoption Ratio, Norm's Yield, Norm's Trust, and Norm's Morality). Having determined the parameters' values, e.g. high; medium; or low, the agent's belief is influenced by these values, which in turn influence its decision to adopt or reject the norm.

4 A Conceptual Framework for Norm's Benefit Awareness

In the first section we present the conceptual design of this framework. The second section discusses the framework formalization.

4.1 Conceptual Design

In this framework, we introduce the concept of Norm's Benefit Awareness in an open normative multi-agent community, in which agents recognize the benefits of adopting the norms and subsequently decide whether to adopt or reject the norm. In this work, we propose a Norm's Benefit Evaluator, N_BE that enables agents' awareness of the benefits of adopting or rejecting a particular norm. The Norm's Benefit Evaluator, evaluates the parameters, Norm's Adoption Ratio, Norm's Yield, Norm's Trust, and Norm's Morality. Having evaluated the parameters, the Norm's Benefit Evaluator determines whether the result increases or decreases an agent's utility. The agent's belief is influenced by the output of the Norm's Benefit Evaluator (increase; decrease). When the agent's belief is updated with the expected benefit, it decides whether to accept or reject the norm.

As shown in Figure 2, an agent first detects a norm (1). The agent then launches the Norm's Benefit Evaluator (2). The Evaluator evaluates the parameters' values, Norm's Adoption Ratio, Norm's Yield, Norm's Trust, and Norm's Morality (3). It then determines the effect of the detected norm on the agent's utility (increase; decrease) (4) which in turn influences the agent's belief (5). From its belief, the agent is aware of the benefits of the detected norm (6). The agent then decides (7) whether to accept and comply with the detected norm (8a) or reject and violate the norm (8b).

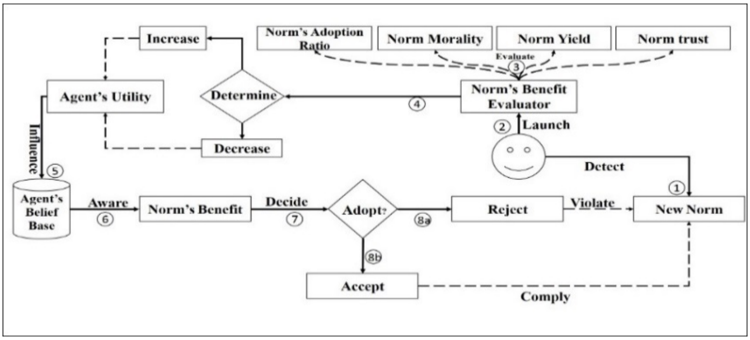


Fig. 2. A Proposed Framework for Norm's Benefit Awareness

4.2 Formalization

We introduce here the proposed approach for evaluating the Norm's Benefit Awareness (N_BA) via the Norm's Benefit Evaluator (N_BE). As mentioned earlier, for an agent to determine the N_BE, it needs to evaluate the parameters, Norm's Adoption Ratio, N_{AR}; Norm's Yield, N_Y; Norm's Trust, N_T; and Norm's Morality, N_M. The output of the N_BE either increases or decreases the agent's utility.

Let δ be the value of adoption ratio parameter, N_{AR}; λ be the value of yield parameter, N_Y; σ be the value of trust parameter, N_T; and μ be the value of morality parameter N_M. The parameters' values (δ, λ, σ, μ) are assumed to be positive, neutral, or negative. If the highest value of N_BE is 1, and the lowest value is 0, there is a range of threshold value, T, bounded between a and b; a maximum value, Max > a; and a minimum value, Min < b. If X is the result of N_BE, then,

$$N_BE = \begin{cases} \text{Max (Increase)} & 1 \geq X \geq a \\ \text{T (Threshold Value)} & a > X \geq b \\ \text{Min (Decrease)} & b > X \geq 0 \end{cases}$$

which means that the Norm's Benefit Evaluator (N_BE) equals Max (increase an agent's utility) when its result, X, is bounded between 1 and a; equals T when X is bounded between a and b (increase or decrease an agent's utility), or equals Min when X is bounded between b and 0 (decrease an agent's utility).

The parameters' (N_{AR}, N_Y, N_T, N_M) proposed evaluation procedure is as follows. Let E_P be a norm's positive effect, E_N the neutral effect, E_G the negative effect; n the norm; and α an agent with a utility, u_α. If E_P is bounded between 0 and 1 (positive effect), E_N equals 0 (neutral effect), E_G is bounded between 0 and -1 (negative effect), Table 1 shows the parameters' N_{AR}, N_Y, N_T, and N_M evaluation.

Table 1. The Parameters' (N_{AR}, N_Y, N_T, N_M) Evaluation

Parameters	Positive Effect (E_P)	Neutral Effect (E_N)	Negative Effect (E_G)
N_{AR}	$1 \geq \delta > 0$	$\delta = 0$	$0 > \delta \geq -1$
N_Y	$1 \geq \lambda > 0$	$\lambda = 0$	$0 > \lambda \geq -1$
N_T	$1 \geq \sigma > 0$	$\sigma = 0$	$0 > \sigma \geq -1$
N_M	$1 \geq \mu > 0$	$\mu = 0$	$0 > \mu \geq -1$

As mentioned earlier,

$N_BE = \delta(N_{AR}), \lambda(N_Y), \sigma(N_T), \mu(N_M)$
For the norm n, $N_BE = \delta(n), \lambda(n), \sigma(n), \mu(n)$

The parameters' $\delta(n), \lambda(n), \sigma(n), \mu(n)$ evaluations are as follow:

$$\delta(n) = \begin{cases} E_P(u_\alpha) & 1 \geq \delta > 0 \\ E_N(u_\alpha) & \delta = 0 \\ E_G(u_\alpha) & -1 \geq \delta > 0 \end{cases} \qquad \lambda(n) = \begin{cases} E_P(u_\alpha) & 1 \geq \lambda > 0 \\ E_N(u_\alpha) & \lambda = 0 \\ E_G(u_\alpha) & -1 \geq \lambda > 0 \end{cases}$$

$$\sigma(n) = \begin{cases} E_P(u_\alpha) & 1 \geq \sigma > 0 \\ E_N(u_\alpha) & \sigma = 0 \\ E_G(u_\alpha) & -1 \geq \sigma > 0 \end{cases} \qquad \mu(n) = \begin{cases} E_P(u_\alpha) & 1 \geq \mu > 0 \\ E_N(u_\alpha) & \mu = 0 \\ E_G(u_\alpha) & -1 \geq \mu > 0 \end{cases}$$

If $\delta(n), \lambda(n), \sigma(n), \mu(n) = \text{Max} \Rightarrow N_BE = \text{increase}(n, u_\alpha)$
If $\delta(n), \lambda(n), \sigma(n), \mu(n) = \text{T} \Rightarrow N_BE = (\text{increase}(n, u_\alpha) \vee \text{decrease}(n, u_\alpha))$
If $\delta(n), \lambda(n), \sigma(n), \mu(n) = \text{Min} \Rightarrow N_BE = \text{decrease}(n, u_\alpha)$

Table 2 shows the sample results of N_BE.

Table 2. Sample Results of NBE

δ	λ	σ	μ	N_BE	Agent's Utility (u_α)
E_P	E_P	E_P	E_P	Max	increase
E_P	E_P	E_G	E_G	T	critical (increase \ decrease)
E_P	E_G	E_G	E_G	T	critical (increase \ decrease)
E_G	E_P	E_P	E_G	T	critical (increase \ decrease)
E_G	E_G	E_G	E_G	Min	decrease

Hence, we define the Norm's Benefit Awareness (N_BA) of the agent, α, on the norm, n, as follow:

$$N_BA(\alpha, n) = \begin{cases} (\text{adopt } (\alpha, n)) \Leftrightarrow \text{increase } (n, u_\alpha) & N_BE = \text{Max} \\ (\text{ignore } (\alpha, n)) \Leftrightarrow \text{decrease } (n, u_\alpha) & N_BE = \text{Min} \end{cases}$$

This means that Norm's Benefit Awareness for the agent, α, on norm, n, is either α adopts the norm, n, if and only if it increases its utility (N_BE = Max), or α ignores the norm, n, if and only if it decreases its utility (N_BE = Min).

5 Conclusion and Future Work

In this paper, we present initial findings on norm's benefits awareness in an open multi-agent community where norms are adopted based on their expected benefits.

We observe that four components constitute the norm benefit awareness which are Norm's Adoption Ratio, Norm's Yields, Norm's Trust, and Norm's Morality.

This paper also proposes an approach to calculate the norm's benefits awareness via a norm's benefit evaluator. For an agent to determine the norm's benefit evaluator, it needs to evaluate the parameters' values, i.e., adoption ratio, yield, trust, and morality. The output of the $N_B E$ either increases or decreases the agent's utility.

In our future work, we shall develop a method to calculate the absolute values of the parameters. From these values, an agent is able to calculate the norm's benefit and decide whether to adopt or reject a detected norm. When the norm's benefit is high, it motivates the agent to adopt it and vice versa.

References

1. Hollander, C., Wu, A.: The Current State of Normative Agent-Based Systems. Journal of Artificial Societies and Social Simulation 14(2), 6 (2011a)
2. Alberti, M., Gomes, A.S., Gonçalves, R., Leite, J., Slota, M.: Normative Systems Represented as Hybrid Knowledge Bases. In: Leite, J., Torroni, P., Ågotnes, T., Boella, G., van der Torre, L. (eds.) CLIMA XII 2011. LNCS (LNAI), vol. 6814, pp. 330–346. Springer, Heidelberg (2011)
3. Broersen, J., Dastani, M., van der Torre, L.W.N.: Resolving Conflicts Between Beliefs, Obligations, Intentions, and Desires. In: Benferhat, S., Besnard, P. (eds.) ECSQARU 2001. LNCS (LNAI), vol. 2143, pp. 568–579. Springer, Heidelberg (2001)
4. Sadri, F., Stathis, K., Toni, F.: Normative KGP agents. Computational and Mathematical Organization Theory 12(2), 101–126 (2006)
5. Ahmad, A., Ahmed, M.M., Yusoff, M.Z., Ahmad, M.S., Mustapha, A.: Resolving Conflicts between Personal and Normative Goals in Normative Agent Systems. In: The Seventh International Conference on IT in Asia 2011 (CITA 2011), Kuching, Sarawak, July 12-14, pp. 153–158 (2011)
6. Boella, G., Hulstijn, J., van der Torre, L.: Interaction in normative multi-agent systems. Electronic Notes in Theoretical Computer Science 141(5), 135–162 (2005)
7. Alchourron, C.E., Bulygin, E.: Normative Systems. Springer (1971)
8. Mahmoud, M.: Norms Detection and Assimilation in Open Normative Multi-agent Communities. PhD Thesis, Universiti Tenaga Nasional (2013)
9. Mahmoud, M.A., Ahmad, M.S., Ahmad, A., Mustapha, A., Yusoff, M.Z., Hamid, N.H.: Obligation and Prohibition Norms Mining Algorithm for Normative Multi-agent Systems. In: Barbucha, D., Le, M.T., Howlett, R.J., Jain, L.C. (eds.) Advanced Methods and Technologies for Agent and Multi-Agent Systems. Frontiers in Artificial Intelligence and Applications, pp. 115–124. IOS Press (2013)
10. Savarimuthu, B.T.R.: Mechanisms for norm emergence and norm identification in multi-agent societies (Thesis, Doctor of Philosophy). University of Otago (2011)
11. Epstein, J.: Learning to be thoughtless: Social norms and individual computation. Computational Economics 18(1), 9–24 (2001)
12. López, y López: Social Powers and Norms: Impact on Agent Behaviour. PhD thesis, Department of Electronics and Computer Science, University of Southampton, United Kingdom (2003)

13. Mahmoud, M.A., Ahmad, M.S., Ahmad, A., Mustapha, A., Yusoff, M.Z., Hamid, N.H.: Building Norms-Adaptable Agents from Potential Norms Detection Technique (PNDT). International Journal of Intelligent Information Technologies (IJIIT) 9(3), 38–60 (2013)
14. Sen, S., Airiau, S.: Emergence of norms through social learning. In: Proceedings of IJCAI 2007, pp. 1507–1512 (2007)
15. Bosse, T., Gerritsen, C., Klein, M.C.A.: Agent-Based Simulation of Social Learning in Criminology. In: Proc. of the Int. Conf. on Agents and AI, ICAART 2009, pp. 5–13 (2009)
16. Criadoa, N., Argentea, E., Noriega, P.B., Botti, V.: MaNEA: A distributed architecture for enforcing norms in open MAS. Engineering Applications of Artificial Intelligence 26(1), 76–95 (2013)
17. Hexmoor, H., Venkata, S.G., Hayes, D.: Modelling social norms in multiagent systems. Journal of Experimental and Theoretical Artificial Intelligence 18(1), 49–71 (2006)
18. Andrighetto, G., Governatori, G., Noriega, P., van der Torre, L.W., Balke, T., da Costa Pereira, C., Krueger, I.: Normative Multi-Agent Systems (2013)
19. Savarimuthu, B.T.R., Cranefield, S., Purvis, M., Purvis, M.: Norm Identification in Multiagent Societies. Discussion Paper, department of Information Science. University of Otago (2010a)
20. Savarimuthu, B.T.R., Cranefield, S., Purvis, M., Purvis, M.: Obligation Norm Identification in Agent Societies. Journal of Artificial Societies and Social Simulation 13(4) (2010b)
21. Oren, N., Meneguzzi, F.: Norm Identification through Plan Recognition. In: Proceedings of COIN 2013@AAMAS (2013)

The *Geranium* System: Multimodal Conversational Agents for E-learning

David Griol, José Manuel Molina, and Araceli Sanchis de Miguel

Computer Science Department
Carlos III University of Madrid
Avda. de la Universidad, 30, 28911 - Leganés (Spain)
{david.griol,josemanuel.molina,araceli.sanchis}@uc3m.es

Abstract. Many e-learning applications use conversational agents as means to obtain enhanced pedagogical results such as fostering motivation and engagement, incrementing significant learning and helping in the acquisition of meta-cognitive skills. In this paper, we present *Geranium*, a multimodal conversational agent that helps children to appreciate and protect their environment. The system, which integrates an interactive chatbot, provides a modular and scalable framework that eases building pedagogic conversational agents that can interact with the students using speech and natural language.

Keywords: Conversational agents, multimodal interaction, chatbots, speech, e-learning.

1 Introduction

Conversational agents [1] have became a strong alternative to enhance multi-agent systems with intelligent communicative capabilities and provide a more natural access to multiagent systems [2]. To successfully manage the interaction with users, these agents are usually developed following a modular architecture, which generally includes the following tasks: automatic speech recognition (ASR), spoken language understanding (SLU), dialog management (DM), database management (DB), natural language generation (NLG), and text-to-speech synthesis (TTS).

According to Roda et al. [3], educative technologies should i) accelerate the learning process, ii) facilitate access, iii) personalize the learning process, and iv) supply a richer learning environment. These aspects can be addressed by means of multimodal conversational agents by establishing a more engaging and human-like relationship between the students and the system. For this reason, this kind of agents have been employed to develop a number of educational systems in very different domains, including tutoring [4], conversation practice for language learners [5], pedagogical agents and learning companions [6], dialogs to promote reflection and metacognitive skills [7], or role-playing actors in simulated experiential learning environments [8], etc.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_26,

Due to this variability and the huge amount of factors that must be taken into account, these systems are difficult to implement and typically are developed ad-hoc, which usually implies a lack from scalability. In this paper we describe the *Geranium* system, a web-based interactive software with a friendly chatbot that can be used as a learning resource for children to study about the urban environment. The proposals for the development of the different modules of the system eases the construction of educative conversational by isolating pedagogic from the technical detail, so that teachers and parents can add new contents without having a technical background at the same time as the software includes these new data for the interaction with the students.

The developed system, which is accessible using desktop and mobile devices, provides multimodal interaction instead of usually mediated simple text-based forms interaction, including spoken access and a visual representation through an animated bot with gestures and emotional facial displays. Also, the system infers a knowledge level for the students based on their answers, and encourages learners to engage in a dialog to reflect on their self-assessment and any differences between their belief and the expressed by the system.

2 The *Geranium* Pedagogical System

The *Geranium* system has been developed with the main aim of making children aware of the diversity of the urban ecosystem in which they live, the need to take care of it, and how they can have a positive impact on it. The system has a chatbot named *Gera*, a cartoon that resembles a geranium, a very common plant in the Spanish homes.

Figure 1 shows two snapshots of the system. As can be observed, it has a very simple interface in which the chatbot is placed in a neighborhood. There are several buttons to select the type of questions, the chatbot has a balloon that shows the text, image or videos corresponding to the questions and possible answers, and there is a "push-to-talk" button that enables the oral input. As the chatbot changes its expressions, the background also changes. For example, Figure 1 shows the response of the system to an incorrect response, as can be observed, *Gera* is "sad" and the background has a grey color.

The chatbot poses questions to the children that they must answer either orally or using the graphical interface. Once an answer is selected, the system checks if it is correct. In case it is, the user receives a positive feedback and *Gera* shows a "happy" (usual case) or a "surprised" (in case of many correct questions in a row) face. If the answer selected is not correct, Gera shows a "sad" expression and provides a hint to the user, who can make another guess before getting the correct response. *Gera* has 7 expressions: happy, ashamed, sad, surprised, talking, waiting and listening, shown in Figure 2, which can also be extended by adding new resources to the chatbot expressions database.

The activities are grouped in four topics: "in the market", "animals and plants", "waste", and "water and energy". In the first topic, the children are asked about fruits and vegetables, the plants where they come from and the

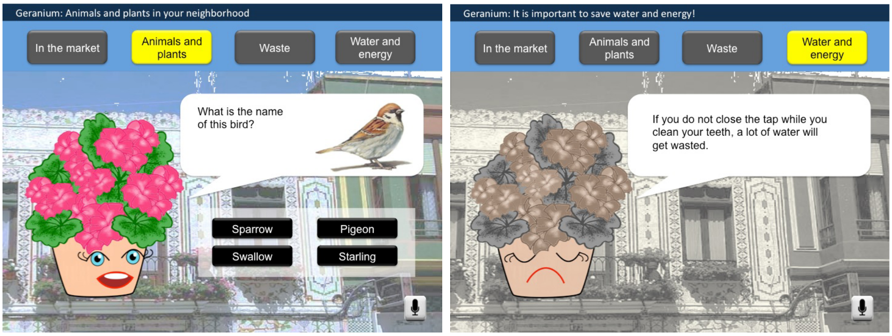


Fig. 1. Snapshots of the response of the *Geranium* system to a correct and an incorrect answer

seasons when they are collected. The second topic comprises questions about animals and plants that live in the city, showing photographs, drawings and videos of birds, flowers, trees and leaves and how they change or migrate during the year. In the third topic, the children are asked about recycling, differentiating between the different wastes and the suitable containers. Finally, the fourth topic deals with good practices to save water and energy at home. Currently, there are 20 questions per category (a total of 80 questions), although the system can be extended adding new questions with their respective answers to the database.

The speech input is activated with a push-to-talk button and the recognition hypotheses generated by the recognizer are passed to the SLU module for processing. Once the input has been processed, the dialog manager chooses the next system action for which a system answer is generated and synthesized. During oral communication, along with the speech and textual response, the chatbot provides feedback with its expressions. When it is listening to the user, it shows the "listening" face, and if the answer selected was not understood, it shows an "ashamed" expression (see Figure 2). There are two additional expressions: "talking" and "waiting", to resemble the chatbot talking and waiting for the user to provide a response to a question, respectively.

The natural language understanding and dialog management modules have been developed according to the Voice Extensible Markup Language (VoiceXML, `www.w3.org/TR/voicexml20`), defined by the W3C as the standard for implementing interactive voice dialogs for human-computer interfaces. VoiceXML applications are usually based on the definition of grammars for the SLU module. In our system, grammars are encoded following the Java Speech Grammar Format (JSGF, `www.w3.org/TR/jsgf/`), supported by any VoiceXML platform.

In the *Geranium* system, for each question type there is a grammar template with the usual structure of the responses, and a new grammar is dynamically generated that makes use of the template and contains the exact response options for the actual question. Each of the options has an assigned code which is used also in the GUI and makes it possible to easily control the synchronization

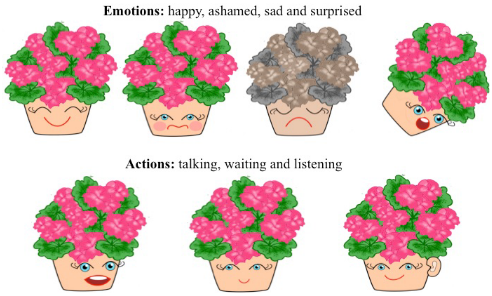


Fig. 2. Facial expressions of the *Gera* chatbot

between the different modalities employed. Also the template makes it possible to maintain the same structure for the responses to similar questions (e.g. to ask for the name of a bird, a plant or a fruit in a photograph) even when they belong to different categories (e.g. birds are in the category "animals and plants" and fruits are in the category "in the market"). This facilitates the system usage, as it is easier for the users to know what the system expects.

The inclusion of static and dynamic grammars makes it possible to implement flexible dialogs with a wide range of possibilities from system-directed initiative to mixed initiative. Static grammars deal with information that does not vary over time, including the templates for the different question types and the grammars to control the exercises flow (e.g. to repeat a question, ask for help or select an option in the menu). Dynamic grammars include information that varies with time and make it possible to easily update and increment the learning contents.

This way, the SLU and dialog manager modules are simplified by generating a VoiceXML file for each specific question in the system, including the corresponding system prompt and the grammar that defines the valid user's inputs for the prompt. Regarding dialog management, all the events in the application are controlled using JavaScript. The dialog manager selects the next system prompt (i.e. VoiceXML file) by following the JavaScript program that determines the order for the set of questions, which is based on the transitions summarized in Figure 3 and is based on VoiceXML finite states.

Figure 4 shows an example to generate a VoiceXML file and grammars corresponding to the snapshot shown in Figure 1, in which the student is asked to tell the name of a bird. As it can be observed, the VoiceXML file corresponding to each one of the questions can include more than one system prompt. To do this, a prompt counter is defined to track the number of times the prompt has been used since the form was entered. The values for the properties are computed dynamically taking into account the dialog history. The question template is *what_is_it.jsgf*, whereas the exact options for the response are in the *question_ex1b3.jsgf* grammar, which is generated at run time. Thus, the student utterances can be short, but also more elaborated, as for example: "Starling",

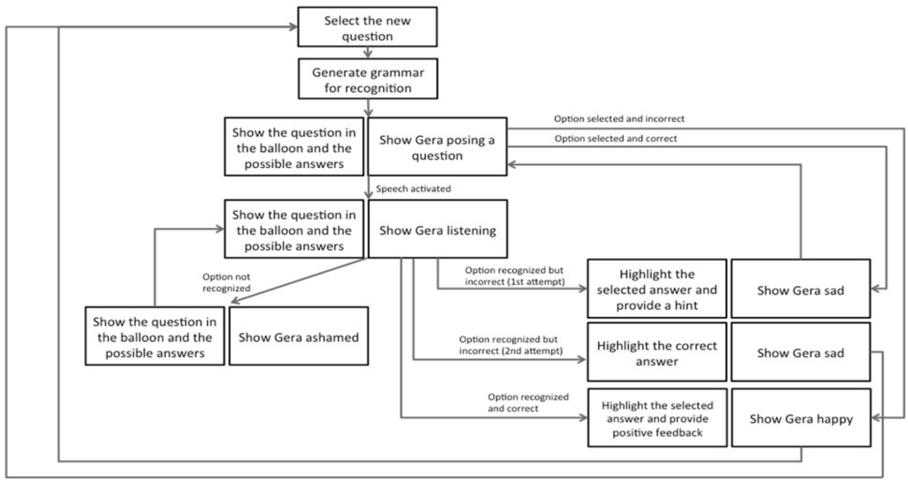


Fig. 3. Transitions to render the chatbot responses in *Geranium*

"A sparrow", "It looks like a pigeon", "I think it is a sparrow", or "I am not sure but it can be a swallow".

In addition, we have considered different functionalities that allow the adaptation of the system taking into account the current state of the dialog as well as the characteristics of each user. We have captured the main VoiceXML events: *noinput* (the user does not answer in a certain time interval or it was not sensed by the recognizer), *nomatch* (the input did not match the recognition grammar or was misrecognized), and *help* (the user explicitly asks for help).

Regarding the graphical user interface, the system answer generator produces the HTML output for the GUI and the template to be used by the natural language generator to obtain the lexical form of the next system prompt, which is then synthesized. With respect to the input, the visual and oral modalities are synchronized by means of the codes assigned to the answers for each question, both in the HTML form and in the VoiceXML grammars.

3 Evaluation

A preliminary evaluation of the *Geranium* system has been already completed with the participation of 6 primary school teachers of the levels for 8, 9 and 10 years old children, who rated the naturalness and pedagogical potential of the system. Teachers were told to bear in mind that the system was aimed at children of the same age as their students. The questionnaire shown in Table 1 was defined for the evaluation. The responses to the questionnaire were measured on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The experts were also asked to rate the system from 0 (minimum) to 10 (maximum) and there was an additional open question to write comments or remarks.

```
File question_ex1b3.jsgf:
#JSGF V1.0;
grammar question_ex1b3;
public <question_ex1b3> =
<.../grammar_templates/what_is_it.jsgf#
what_is_it> {out.opt = rules.options;};
<options> = sparrow {this.out=‘‘0’’;} | pigeon
{this.out=‘‘1’’;} | swallow {this.out=‘‘2’’;} |
starling {this.out=‘‘3’’;};
File what_is_it.jsgf:
#JSGF V1.0;
grammar what_is_it;
public <what_is_it> = [<pre_answer>] <options>;
<pre_answer> = [<certainty>] [<belief>] [
<phrase>];
<certainty> = ‘‘Of course’’ | ‘‘For sure’’ |
‘‘I am sure’’ | ‘‘I know’’ | (‘‘I am not sure’’
[but]) | (‘‘I do not know’’ [but])
<belief> = I (think | believe);
<phrase> = [it (is | ‘‘can be’’ | ‘‘might be’’ |
‘‘could be’’ | ‘‘may be’’ | ‘‘looks like’’ |
‘‘seems [to be]’’)] (a | an)
```

```
<?xml version=‘‘1.0’’ encoding=‘‘UTF-8’’?>
<vxml xmlns=‘‘www.w3.org/2001/vxml’’
xmlns:xsi=‘‘www.w3.org/2001/
XMLSchema-instance’’
xsi:schemaLocation=‘‘www.w3.org/2001/vxml
www.w3.org/TR/voicexml20/vxml.xsd’’
version=‘‘2.0’’ application=‘‘Gera.vxml’’>
<form id=‘‘ex1b3_form">
<grammar type=‘‘application/x-jsgf’’
src=‘‘/grammars/mainGera.jsgf’’/>
<field name=‘‘question_ex1b3’’>
<grammar type=‘‘application/x-jsgf’’
src=‘‘/grammars/question_ex1b3.jsgf’’/>
<prompt> What is the name of this bird?
</prompt>
<prompt count=‘‘1’’> Tell me the name of
the bird.
</prompt>
<prompt count=‘‘2’’>
If you think that the bird in the picture
is an eagle, just say eagle.</prompt>
<help> It is a little brown bird that lives
in your neighborhood and eats seeds and
insects. </help>
<noinput> Have a go! You are very familiar
to this little bird.</noinput>
<filled>
<nomatch> Ohh ohh! I did not get that.
Remember the options are: sparrow, pigeon,
swallow or starling. Please try again!
</nomatch>
<return namelist=‘‘follow_question.js’’/>
</filled>
</field>
</form>
</vxml>
```

Fig. 4. Example VoiceXML document and corresponding grammars

Also, from the interactions of the experts with the system we completed an objective evaluation of the application considering the following interaction parameters: i) Question success rate (SR). This is the percentage of successfully completed questions: system asks - user answers - system provides appropriate feedback about the answer; ii) Confirmation rate (CR). It was computed as the ratio between the number of explicit confirmations turns and the total of turns; iii) Error correction rate (ECR). The percentage of corrected errors.

The results of the questionnaire are summarized in Table 2. As can be observed from the responses to the questionnaire, the satisfaction with technical aspects was high, as well as the perceived didactic potential. The chatbot was considered attractive and adequate and the teachers felt that the system is appropriate and the activities relevant. The teachers also considered that the system succeeds in making children appreciate their environment. The global rate for the system was 8.5 (in the scale from 0 to 10).

Table 1. Percentage of different dialogs obtained

Technical quality
TQ01. The system offers enough interactivity
TQ02. The system is easy to use
TQ03. It is easy to know what to do at each moment.
TQ04. The amount of information that is displayed on the screen is adequate
TQ05. The arrangement of information on the screen is logical
TQ06. The chatbot is helpful
TQ07. The chatbot is attractive
TQ08. The chatbot reacts in a consistent way
TQ09. The chatbot complements the activities without distracting or interfering with them
TQ010. The chatbot provides adequate verbal feedback
TQ011. The chatbot provides adequate non-verbal feedback (gestures)
Didactic potential
DP01. The system fulfills the objective of making children appreciate their environment
DP02. The contents worked in the activities are relevant for this objective
DP03. The design of the activities was adequate for children of this age
DP04. The activities support significant learning
DP05. The feedback provided by the agent improves learning
DP06. The system encourages continuing learning after errors

Table 2. Results of the evaluation of the system by experts

	Min / max	Average	Std. deviation
TQ01	3/5	4.17	0.69
TQ02	3/4	3.67	0.47
TQ03	4/5	4.83	0.37
TQ04	5/5	5.00	0.00
TQ05	4/5	4.67	0.47
TQ06	4/5	4.83	0.37
TQ07	4/5	4.83	0.37
TQ08	4/5	4.50	0.50
TQ09	4/5	4.83	0.37
TQ10	4/5	4.67	0.47
TQ11	3/5	4.50	0.76
DP01	5/5	5.00	0.00
DP02	4/5	4.67	0.47
DP03	4/5	4.83	0.37
DP04	5/5	5.00	0.00
DP05	4/5	4.67	0.47
DP06	4/5	4.83	0.37

SR	CR	ECR
96.56%	13.00%	93.02%

Although the results were very positive, in the open question the teachers also pointed out desirable improvements. One of them was to make the system listen constantly instead of using the push-to-talk interface. However, we believe that this would cause many recognition problems, taking into account the unpredictability of children behavior. Also, although they considered the chatbot attractive and its feedback adequate, they suggested creating new gestures for the chatbot to make transitions smoother.

The results of the objective evaluation for the described interactions show that the developed system could interact correctly with the users in most cases, achieving a question success rate of 96.56%. The fact that the possible answers to the questions are restricted made it possible to have a very high success in speech recognition. Additionally, the approaches for error correction by means of confirming or re-asking for data were successful in 93.02% of the times when the speech recognizer did not provide the correct answer.

4 Conclusions and Future Work

In this paper we have described the *Geranium* conversational agent, a web-based interactive software with a friendly chatbot that can be used as a learning resource for children to study about the urban environment. The system has been developed using an architecture to cost-effectively develop pedagogic chatbots. This architecture is comprised of different modules that cooperate to interact with students using speech and visual modalities, and adapt their functionalities taking into account their evolution and specific preferences. We have carried out an evaluation of the *Geranium* system with primary school teachers to assess its ease of use and its pedagogical potential. The study showed a high degree of satisfaction in the system appearance and interface, and the results were very positive with respect to its pedagogical potential. For future work, we plan to replicate the experiments with children to validate these preliminary results, incorporate the suggestions provided by the teachers, and also compare the developed system with other pedagogical tools.

Acknowledgements. This work was supported in part by Projects MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485).

References

1. Pieraccini, R.: The Voice in the Machine: Building Computers that Understand Speech. The MIT Press (2012)
2. Pinzón, C., Bajo, J., de Paz, J., Corchado, J.: S-MAS: An adaptive hierarchical distributed multi-agent architecture for blocking malicious SOAP messages within Web Services environments. Expert Systems with Applications 38(5), 5486–5499 (2011)
3. Roda, C., Angehrn, A., Nabeth, T.: Dialog systems for Advanced Learning: Applications and Research. In: Proc. of BotShow 2001, pp. 1–7 (2001)
4. Pon-Barry, H., Schultz, K., Bratt, E.O., Clark, B., Peters, S.: Responding to student uncertainty in spoken tutorial dialog systems. IJAIED Journal 16, 171–194 (2006)
5. Fryer, L., Carpenter, R.: Bots as Language Learning Tools. Language Learning and Technology 10(3), 8–14 (2006)
6. Cavazza, M., de la Camara, R.S., Turunen, M.: How Was Your Day? a Companion ECA. In: Proc. AAMAS 2010, pp. 1629–1630 (2010)
7. Kerly, A., Ellis, R., Bull, S.: CALMsystem: A Dialog system for Learner Modelling. Knowledge Based Systems 21(3), 238–246 (2008)
8. Griol, D., Molina, J., Sanchis, A., Callejas, Z.: A Proposal to Create Learning Environments in Virtual Worlds Integrating Advanced Educative Resources. JUCS Journal 18(18), 2516–2541 (2012)

DiSEN-AlocaHR: A Multi-Agent Mechanism for Human Resources Allocation in a Distributed Software Development Environment

Lucas O. Teixeira and Elisa H.M. Huzita

Informatics Department, State University of Maringá, Maringá, Paraná, Brazil
lucasxteixeira@gmail.com, emhuzita@din.uem.br

Abstract. The success or failure of a project is directly related to individual talent of the participants and, most important, how they are assigned to the tasks in a project. This paper presents a context-aware multi-agent mechanism to support the human resource allocation in globally distributed software projects. This mechanism performs the human resource allocation to tasks of a project taking into account the participants contextual information, the requirements of the tasks and the interpersonal relationship among the human resources. The participants contextual information includes culture, idiom, temporal distance and previous experience. The mechanism is composed by three elements: (i) capture and inference of information, (ii), validation and consolidation of knowledge, and (iii) human resources allocation.

Keywords: context-awareness, ontology, global software development, human resources allocation.

1 Introduction

The resource that most affects the quality and cost of software development is the human [1]. Thus, the success of a software project can be related to the talent of the team and, more importantly, the way that the participants are allocated to the tasks. This paper presents a multi-agent mechanism that performs human resource allocation in a globally distributed environment, called DiSEN-AlocaHR (DiSEN-Allocation of Human Resources) taking into account participants cultures, idioms, localizations, experiences, tasks requirements and the interpersonal relationship among the human resources. DiSEN-AlocaHR is part of a bigger project called DiSEN (Distributed Software Engineering Environment). DiSEN is a tool that offers features to support communication, persistence and collaboration among teams geographically distributed [2].

Global Software Development (GSD) is characterized by the cooperation of globally distributed teams in decentralized units. GSD brings some challenging issues such as geographic, temporal and socio-cultural distance, complex communication, control and coordination [3]. Therefore, increasing the similarity of the project participants cultures, idioms and temporal distance will reduce the impact of GSD issues.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_27,

Characteristics such as participants cultures, idioms, localizations and interpersonal relationship can be considered contextual information. Contextual information is any information that can be used to characterize the relevant entities of a interaction. The dissemination of contextual information ease the communication, collaboration and coordination in GSD, once that the participants of the project are aware of other participants characteristics.

A model called DiSEN-CSE (DiSEN-Context Sensitive Environment) was designed in order to DiSEN use properly contextual information. It was designed considering context information acquisition, representation, persistence, processing and sharing [4]. Also, an ontology called OntoDiSEN was designed to represent GSD's contextual information [4]. Then, a knowledge-based multi-agent homonymous architecture called CAKMAS (Context Awareness and Knowledge Management Multi-Agent System Architecture) was developed to provide a consistent multi-agent architecture for DiSEN [5].

This paper is structured as follows. Section 2 introduces some related works. Section 3 presents the CAKMAS framework. Section 4 discuss the project and development of the mechanism. Section 5 show a example scenario of the mechanism Finally, in the conclusions section the contributions and future works are discussed.

2 Related Work

Some very interesting techniques and approaches to solve the problem of human resource allocation were found in the literature and are presented in sequence. The evolutionary algorithms are the main approaches used. It includes genetic algorithms [6], particle swarm optimization [7] and ant colony optimization [8]. There are other approaches, such as multidimensional model [9], sociometric analysis [10] and a framework for virtual organization creation [11].

The literature reviewed showed that the most common approaches consider the problem simply as an optimization one, focusing mainly in the process of optimization instead of focusing on which characteristics can be taken into account. In this study we focus our attention on designing and developing a effective and practical mechanism to perform human resources allocation in a globally distributed software development context while considering some subjective characteristics such as participants cultures, idioms, localizations, experiences, tasks requirements and interpersonal relationship. The main objective of this work is the increase of similarity of the cited characteristics in order to minimize the issues in GSD.

3 CAKMAS

The DiSEN environment is still in development and since its conception the idea was that it would be mostly an agent-based system. Thus, based on [4], the multi-agent applications ContextP-GSD and DiSEN User Rating were developed and then CAKMAS was conceived as a refactoring of the reusable components of both [5].

The main entity of CAKMAS is the Ontology Agent (OA) which is responsible for maintaining a consistent global knowledge base (KB) as well as persisting and publishing the KB in the environment. Thus, the OA acts as a broker between all agents in the environment and the KB [5].

DiSEN-AlocaHR can be seen as one, among many, CAKMAS application agents. Thus, the mechanism interacts with the OA directly to persist knowledge. Therefore, there is no need to do consistency check, once that this is OA responsibility. Also, the mechanism captures contextual information via its sensors, maintain a local knowledge base and reason about the domain using its knowledge and inference rules.

4 DiSEN-AlocaHR

DiSEN-AlocaHR is a context-aware multi-agent mechanism that performs the allocation of human resources into globally distributed software projects taking into account subjective characteristics. The mechanism conceptual architecture, shown in Figure 1, is composed by: (i) knowledge capture, (ii) knowledge consolidation, and (iii) allocation of human resources elements. Each element is an independent agent. Thereby it is hoped to distribute responsibilities and so, reduce the workload of each agent.

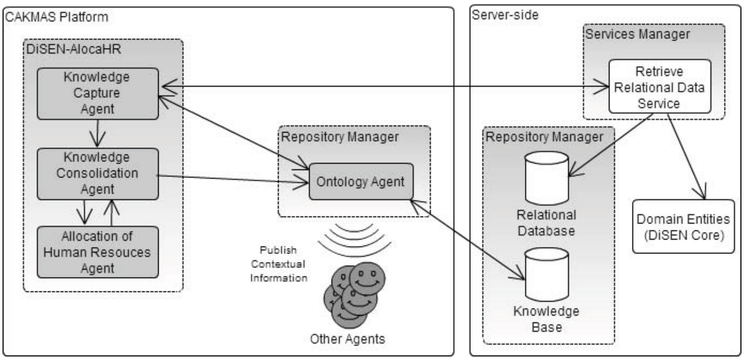


Fig. 1. DiSEN-AlocaHR conceptual architecture

4.1 Knowledge Collection

The first element is responsible for calculating scores for participants culture, idiom, localization and experience. The real challenge is to correctly quantify the subjective characteristics. By a design decision, all scores used in the mechanism are at least zero and at most one, zero indicates a perfect similarity, which is better, and one indicates no similarity.

It is considered that each project has a site and a standard idiom established. This decision was made to reduce the amount of computation needed.

The interpersonal relationship among the human resources will be considered later. We choose to postpone it because it is easier to compute it when all possible teams that can be created are known.

The participant culture, idiom, temporal and experience score are calculated as follows:

- **Culture Score:** The work develop by [12] presents a analysis of global cultures divided by countries[1]. Thus, the culture score is the difference between the user culture, from the user country, and the project reference culture, from the project reference site. The formula to calculate the culture score is

$$C_u = |CR_{P_R} - CR_u| / max_c r \tag{1}$$

  which C_u is the culture score for user u, CR_{P_R} is project R reference culture, CR_u is user u culture e $max_c r$ is the greatest difference between two cultures, this was done so that the scores would be at least zero and at most one.
- **Temporal Score:** The temporal score is simply the difference between the user timezone, from the user site, and the project timezone, from the project reference site. The formula to calculate the temporal score is

$$T_u = MIN(|TR_{P_R} - TR_u|/12, ((12 - |TR_{P_R}| + (12 - |TR_u|)) + 1)/12) \tag{2}$$

  which T_u is the temporal score for user u, TR_{P_R} is project R timezone and TR_u is user u timezone.
- **Idiom Score:** We considered two aspects in this score: if the idioms are different we established a similarity between them; if the idioms are the same we considered the user competence. The work develop by [13] presents a tree of languages by its origins, we used such tree to calculate the similarity of different idioms. The competence is classified in five levels, one through five. The formula to calculate score for a user u idiom s is

$$I_u^s = \begin{cases} (1 - (IR_{us}/5))/2 & \text{If } ID_{P_R} = s \\ 0.5 + (dist(ID_{P_R}, s)/(2 \times max_i d)) & \text{If } ID_{P_R} \neq s \end{cases} \tag{3}$$

  which I_u^s is the idiom score for user u idiom s, ID_{P_R} is the project R reference idiom, IR_{us} is the user u competence level of idiom s, $dist$ is a function that returns the distance of two idioms in the tree and $max_i d$ is the greatest difference between two idioms.

  The idiom score for user u is given by

$$I_u = MIN(\underset{s \in S_u}{\forall} I_u^s) \tag{4}$$

  which I_u is the user u idiom score, S_u is the set of all idioms known by user u.

[1] http://www.geerthofstede.nl/dimension-data-matrix

- **Experience Score:** To establish the experience score we determined which project that the user participated is the most similar to the current project, which the user is being allocated, and how successful such project was. To determine the most similar project the work develop by [14] was used. To determine the success of a project we determined some critical success factors relevant in GSD through a questionnaire applied to specific GSD study groups, taking into account some factors discovered by [15]. The formula to calculate the experience score is

$$E_u = \frac{TP_{P_m} + FS_{P_m} + CI_{P_m} + SS_{P_m} + UF_{P_m} + TF_{P_m}}{6} \tag{5}$$

  which E_u is the experience score for user U, TP_{P_m} represents the training of the project m development team, FS_{P_m} is the project m financial resources score, CI_{P_m} represents the client involvement, in other words represents the presence of the client in the development team, in project m, SS_{P_m} is the project m schedule score, UF_{P_m} is the project m user feedback e TF_{P_m} is the project m development team feedback. The feedback is classified in five levels, one through five. The project m represents the most similar project to the current project.

  The financial resources score is given by

$$FS_{P_m} = \frac{|FE_{P_m} - FR_{P_m}|}{MAX(FE_{P_m}, FR_{P_m})} \tag{6}$$

  which FS_{P_m} is the project m financial resources score, FE_{P_m} is the project m estimated expense, FR_{P_m} is the project m real expense.

  The schedule resources score is given by

$$SS_{P_m} = \frac{|SE_{P_m} - SR_{P_m}|}{MAX(SE_{P_m}, SR_{P_m})} \tag{7}$$

  which SS_{P_m} is the project m schedule score, SE_{P_m} is the project m estimated schedule, SR_{P_m} is the project m real schedule.

4.2 Knowledge Consolidation

This element is responsible for the persistence of the knowledge discovered and inferred, including the scores computed, in both components. This is accomplished by interacting with CAKMAS Ontology Agent.

4.3 Human Resource Allocation

This element is responsible for doing the allocation of human resources taking into account the aggregation of all scores.

To compute the interpersonal relationship score it is needed to discover all possible teams. In order to discover all possible teams we considered the knowledge and skills required by each task and the knowledge and skills of the users.

By doing that each task has a list of users that can be allocated to it. Finally, by doing all combinations of those users all possible teams that can be formed are discovered.

To quantify the interpersonal relationship among the participants in a team we used the sociometric test [16]. The sociometric test finds mutual acceptance among the participants. The formula to calculate the interpersonal relationship score is

$$R_k = 1 - \frac{MA_k}{n_k} \tag{8}$$

which R_k is the team k interpersonal relationship score, MA_k is the team k mutual acceptances and n_k is the team k number of participants.

Finally, with all scores computed and the tasks requirements satisfied it is possible to compute the final score for each possible team. Such score indicate which team is the most appropriate for the project. The formula to calculate the final score is

$$F_k = \frac{\dfrac{\left(\sum_{u \in U_k} C_u + T_u + I_u + E_u\right)}{n_k} + R_k}{5} \tag{9}$$

which F_k is the team k final score, U_k is the set of team k users, C_u is the culture score for user u, T_u is the temporal score for user u, I_u is the idiom score for user u, E_u is the experience score for user u, n_k is the team k number of participants and R_k is the team k interpersonal relationship score.

5 Example Scenario

This section presents a usage scenario of DiSEN-AlocaHR containing hypothetical users and projects. On Tables 1 and 2 are presented the hypothetical users and projects. Note that, P stands for proficient, H for high, M for medium, L for low and B for beginner.

Table 1. Hypothetical Users with Knowledge and Skills

User	Current Location	Birthplace	Idioms	Knowledge	Skills
Arnold	Pretoria	Arab Countries	English (P)	Ontology Design (H) UML (M) Unit Test (M) Context-aware Systems (H)	Design Patterns (P) Object Oriented Programming (M) SPARQL (M)
George	Maringá	Brazil	Portuguese (P)	Ontology Design (H) RUP (L) SCRUM (P) Context-aware Systems (H)	Design Patterns (H) Object Oriented Programming (M) SPARQL (M)
Mike	Miami	China	Chinese (H) English (P)	Software Agents (M) RUP (H) UML (M) Unit Test (L)	Design Patterns (P) Object Oriented Programming (M)

On Table 3 is illustrated the score calcutated for culture and timezone, using the formulas explained in Section 4.

Finally, Tables 4 and 5 presents the pre-allocation of human resources and the best team that can be allocated to the project considering that only Mike and Arnold had a mutual acceptance and also the scores explained in Section 4.

Table 2. Hypothetical Project and Tasks with Required Knowledge and Skills

Project	Reference Location	Reference Idiom	Activities	Tasks	Knowledge Required	Skills Required
Travel Agency	Miami	English	Analysis	Requirements Engineering	Software Agents (M) RUP (H)	Design Patterns (P) Object Oriented Programming (L)
			Project	Ontology Design	Context-aware Systems (M) Ontology Design (H)	Object Oriented Programming (L) SPARQL (M)
			Development	Coding	UML (M) Unit Test (L)	Object Oriented Programming (M) Design Patterns (H)

Table 3. Culture and Timezone Score

Project	Reference Location	Project Culture	User	Birthplace	Current Location	User Timezone	User Culture	Culture Score	Timezone Score
Travel Agency	Miami (USA) (UTC -4)	333	Arnold	Arab Countries	Pretoria	UTC +2	296	0.193	0.5
			George	Brazil	Maringá	UTC -3	335	0.01	0.084
			Mike	China	Miami	UTC -4	307	0.135	0

Table 4. Pre-allocation of Human Resources

Project	Task	User
Travel Agency	Requirements Engineering	Mike
	Ontology Design	Arnold George
	Coding	Mike Arnold

Table 5. Allocation of Human Resources

Project	Task	User	Interpersonal Relationship Score	Final Score
Travel Agency	Requirements Engineering	Mike	0	0.1055
	Ontology Design	Arnold		
	Coding	Mike		

6 Conclusions and Future Works

GSD brings communication, control and coordination difficulties. Thus, by increasing the similarity in the participants culture, idiom and localization those difficulties might be reduced. DiSEN-AlocaHR is a mechanism that performs human resource allocation taking into account culture, idiom, localization, experience, tasks requirements and the interpersonal relationship.

The main difference between DiSEN-AlocaHR and the other approaches found in the literature is that almost none of them consider subjective characteristics that may have a great impact in some contexts, GSD is a great example.

As future work, we identified that the experience score needs further study and that more characteristics would increase the efficiency of the mechanism.

We would like to thank the support from Fundação Araucária.

References

1. Elonen, S., Artto, K.A.: Problems in managing internal development projects in multi-project environments. International Journal of Project Management 21(6), 395–402 (2003)
2. Huzita, E.H., Tait, T.F., Colanzi, T.E., Quinaia, M.A.: Um ambiente de desenvolvimento distribuído de software-disen. In: I Workshop de Desenvolvimento Distribuído de Software, pp. 31–38 (2007)
3. Holmstrom, H., Conchuir, E.O., Agerfalk, P.J., Fitzgerald, B.: Global software development challenges: A case study on temporal, geographical and socio-cultural distance. In: Proceedings of the IEEE International Conference on Global Software Engineering, pp. 3–11. IEEE Computer Society, Washington, DC (2006)
4. Chaves, A.P., Wiese, I.S., da Silva, C.A., Huzita, E.H.M.: Um modelo baseado em context-awareness para disseminação de informações em um ambiente de desenvolvimento distribuído de software. In: XXXIV Conferencia Latinoamericana de Informáica (CLEI 2008), Santa Fe, Argentina, pp. 1365–1374 (2008)
5. Monte-Alto, H.H., Biasão, A.B., Teixeira, L.O., Huzita, E.H.: Multi-agent and context-aware solutions for a global software development environment. International Journal of Artificial Intelligence 11(A13), 115–129 (2013)
6. Osman, M.S., Abo-Sinna, M.A., Mousa, A.A.: An effective genetic algorithm approach to multiobjective resource allocation problems (moraps). Applied Mathematics and Computation 163(2), 755–768 (2005)
7. Wang, S.Q., Gong, L.H., Yan, S.L.: The allocation optimization of project human resource based on particle swarm optimization algorithm. In: Proceedings of the 2009 IITA International Conference on Services Science, Management and Engineering, pp. 169–172. IEEE Computer Society, Washington, DC (2009)
8. Wang, Q., Liu, M.: Optimization of task allocation and knowledge workers scheduling based on ant colony algorithm. In: International Symposium on Intelligence Information Processing and Trusted Computing (IPTC), pp. 386–389. IEEE (2010)
9. Junjie, C., Shating, W., Chen, C.: Method research for dynamic multi-project human resource allocation based on multidimensional model. International Conference on Information Management, Innovation Management and Industrial Engineering 3, 78–81 (2011)
10. Ballesteros-Pérez, P., González-Cruz, M.C., Fernández-Diego, M.: Human resource allocation management in multiple projects using sociometric techniques. International Journal of Project Management 30(8), 901–913 (2012)
11. Camarinha-Matos, L.M., Afsarmanesh, H.: A framework for virtual organization creation in a breeding environment. Annual Reviews in Control 31(1), 119–135 (2007)
12. Hofstede, G., Hofstede, G., Minkov, M.: Cultures and Organizations: Software of the Mind, 3rd edn. McGraw-Hill Companies (2010)
13. Katzner, K.: The Languages of the World. Taylor & Francis (2002)
14. Azzeh, M., Neagu, D.C., Cowling, P.I.: Software project similarity measurement based on fuzzy c-means. In: Wang, Q., Pfahl, D., Raffo, D.M. (eds.) ICSP 2008. LNCS, vol. 5007, pp. 123–134. Springer, Heidelberg (2008)
15. Nasir, M., Sahibuddin, S.: Critical success factors for software projects: A comparative study. Scientific Research and Essays 6(10), 2174–2186 (2011)
16. Moreno, J.: Sociometry, experimental method and the science of society: an approach to a new political orientation. Sociometry monographs. Beacon House (1951)

Multi-Agent Web Recommendations

Joaquim Neto[3,4] and A. Jorge Morais[1,2,4]

[1] Faculty of Engineering of the University of Porto, Porto, Portugal
[2] Laboratory of Artificial Intelligence and Decision Support (LIAAD – INESC TEC L. A.), Porto, Portugal
[3] National Laboratory of Civil Engineering (LNEC), Lisbon, Portugal
[4] Universidade Aberta (Portuguese Open University), Lisbon, Portugal

Abstract. Due to the large amount of pages in Websites it is important to collect knowledge about users' previous visits in order to provide patterns that allow the customization of the Website. In previous work we proposed a multi-agent approach using agents with two different algorithms (associative rules and collaborative filtering) and showed the results of the offline tests. Both algorithms are incremental and work with binary data. In this paper we present the results of experiments held online. Results show that this multi-agent approach combining different algorithms is capable of improving user's satisfaction.

1 Introduction

Recommender systems technology usually applies technologies and methods from other information system areas, such as Human Computer Interaction and Information Retrieval. On most of those kinds of systems was possible to find algorithms and techniques used in Data Mining and Knowledge Discovery areas. The motivation of these areas is related to the analysis and processing of large datasets [1]. Considering the large number of pages in the Web, it became natural to apply this concept to the Web scope, resulting in the new area of Web Mining [2][3]

Recommender systems [3] have had several improvements over the last decade and are increasingly present in the websites. The recommendation of a set of Web pages of potential interest for the visitor of a website, based on the users' access history, aims to improve the users' experience, providing them with effective search tools. One of the current solutions that are being proposed to address this problem is using autonomous agents. Multi-Agent Systems [4] is a research area that has been in great development over the last decade and, given their characteristics, allow the combination of multiple recommendation algorithms, increasing the chances of the suggested recommendations are effectively of users interest. The multi-agent approach is being proposed because of its flexibility and its capability of dynamic adaptation to the Web applications needs [5]. Moreover, Multi-Agent Systems are already used for automatic retrieval and update of information in Websites [6]. In previous work [7] we presented the implementation of a multi-agent recommender system, and the results of offline experiments.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_28, © Springer International Publishing Switzerland 2014

In this paper we present the results of online experiments of the multi-agent recommender system, using a Website with photographs that was built for this purpose. Different incremental algorithms based on binary data produce item-based recommendations and make bids to provide the next set of recommendations to the user. Agents are cooperative in the sense they base their bids on client's satisfaction instead of their own revenue and they share the same data. However, their results are not combined in order to provide recommendations. Our goal is to show that online experiments confirm the results of the ones taken offline.

The remaining of the paper starts by presenting previous approaches and applications in the area of recommender systems and multi-agent systems, followed by the description of our approach. The results of the experiments online, and some conclusions and future work complete the paper.

2 Previous Approaches and Applications

In [8] is described a global vision on adaptive Web sites, which is based on user interaction analysis. Some approaches were proposed, such as reorganization of the Website [9], use of recommendations in the pages [10], automatic categorization of user actions [11], or seek of relevant Web sequence paths using Markov models [12].

Recommendation systems include several algorithms and techniques, such as the combination of clustering with nearest neighbour algorithms [13], Markov chains and clustering [14], association rules [15], and collaborative and content-based filtering [16]. Web dynamics has been controlled, for instance, by efficient incremental discovery of sequential Web usage patterns [17], and on-line discovery of association rules [18]. Data-driven categorization of Website usability may be done by typical usage patterns visualization [11] or with objective metrics [19].

Some platforms already implemented, like WebWatcher, use previous users' knowledge to recommend links [20], while AVANTI implements an adaptive presentation based on a model constructed from user actions [21], and WUM infers a tree structure from log records enabling experts to find patterns with predefined characteristics [22]. In [23] it was proposed an integrated tool (HDM) to discover access patterns and association rules from log records in order to automatically modify hypertext organization.

Multi-agent approaches for developing complex systems, like Web adaptation, were defended in [24]. Intelligent agents may also be an important contribution for autonomic computing [25]. Such systems main characteristics are being complex systems with self-administration, self-validation, self-adjustment and self-correction. Web adaptation systems should also have these characteristics, because Website environment dynamics requires either a high degree of system automation or high allocation of human resources. The use of autonomous agents in a similar context can be found in [5], where a multi-agent platform for personalization of Web-based systems was proposed, due to its flexibility and its dynamic adaptation to Website needs. Another important usage of multi-agent systems in this context is the automatic collection and update of information in Websites [6].

Some related work including a web adaption platform [26] and an implementation of collaborative filtering using an incremental approach [27] were also important for the architecture of our system.

3 Multi-Agent Approach

The multi-agent system recommender [7] was implemented taking into account that agents should answer rapidly to any request from another agent and prepare in advance for the next request, and tasks that involve a large amount of time (like updating the model) should not interfere with the performance of the system.

Recapitulating what was described in that paper, two recommender agents were created. The first one generates single-condition association rules and the second one uses a collaborative filtering algorithm. Both algorithms share a matrix $A_{n \times n}$, where n is the number of items (Webpages) and each $a_{ij} \in A$ registers the total number of co-occurrences of items i and j in the same session. The matrix is updated each time a session ends.

The single-condition association rules agent checks all possible rules $i \rightarrow j$, where i and j are items, taking into account two values (k number of sessions):

$$Support_{i \rightarrow j} = \frac{a_{ij}}{k} \qquad Confidence_{i \rightarrow j} = \frac{a_{ij}}{a_{ii}}$$

Therefore, if a set of n recommendation is requested, the n best recommendations according to the confidence that satisfy minimum confidence and support requirements are proposed.

The collaborative filtering agent uses the same matrix to compute similarity, returning the top n most similar items:

$$sim(i, j) = \frac{a_{ij}}{\sqrt{a_{ii}}\sqrt{a_{jj}}}$$

Agent biddings are based on an accumulated score for each given item obtained from previous ratings – the best N are sorted and if the next selected item was in that set it receives a score N-p+1, where p is the ordered position of the item. To this score we add the percentage of the overall score to untie equal biddings:

$$Bid_{agent,item} = Score_{item} + \frac{1}{\#requests * N} * \sum_{i \in Items} Score_i$$

The multi-agent approach was implemented in Java, using the JADE platform [28]. The communication with the browser is implemented using AJAX [29], using XMLHTTPRequest interface, so that the user can consult the Web page without losing interest. The interaction between the user and the recommender system, as well as the recommender system architecture is presented in figure 1.

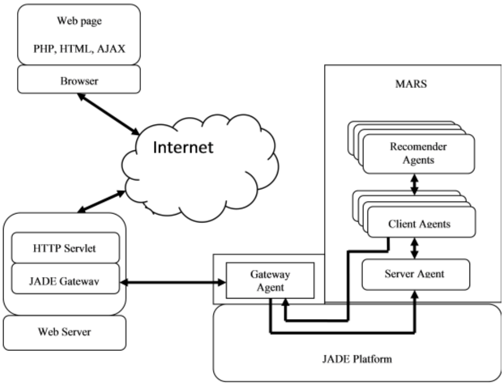


Fig. 1. Website and recommender system interaction

4 Experimental Results

In [7] we presented offline experiments focused on four datasets (obtained from real Web data records). In this paper we present and compare the results of online experiments with a photographs website [31] to the offline experiments.

Like in [7], we consider each time recommendations are made four possibilities:

a. No item was followed (discarded, no implicit knowledge – end of session).
b. The set of recommendations was empty.
c. An item not in the recommendation set was followed.
d. One of the recommendations was followed.

For evaluation of performance, since the algorithms are incremental, which means we do not have a fix split for train and test sets, the evaluation that fits better to our case is a per-user variant, where predictions are computed and the ranking metrics are calculated for each recommendation, and the average over all recommendations gives the final value [32].

There are two measures that we will use for evaluating recommendation: precision and recall [33]. Precision is the ratio of relevant items selected to number of items selected – it represents the probability that a selected item is relevant. Recall is the ratio of relevant items selected to total number of relevant items available. In our case, precision and recall are given by the following formulas (given N recommendations, and considering b, c and d of the list of possible situations above):

$$Recall = \frac{\#d}{\#b + \#c + \#d} \qquad Precision = \frac{1}{N} Recall = \frac{1}{N} \frac{\#d}{\#b + \#c + \#d}$$

This measure is also applied to the recommendation system, which combines agents' algorithms. When the recommendation set is incomplete or inexistent (because it is the first time the item appears, so there are no correlations yet), the system completes it with the most popular items.

In our online experiments we used a photographs website [31] that was built for this purpose. The website includes 344 photographs, distributed by 8 galleries, in the total of 353 Web pages. For recommendation we only consider the dataset of the 344 items with photographs. We decided to present 10 recommendations to the user, each time a Web page is being consulted.

The online experiments were undertaken in two different phases:

1. Recommendations were turned off, meaning that the system did everything as usual except presenting the recommendations to the user.
2. Recommendations were presented to the users and could be followed.

Table 1 shows the previous datasets used for offline experiments and new dataset used for online experiments – e1 and e2 refer to the respective phases.

Table 1. Datasets characteristics

Dataset	#items	# records	#sessions	#records/#session	#records/#items
e-com	335	1409	413	3.411622	4.20597
pe200	200	2042	200	10.21	10.21
e1	344	2514	364	6.9066	7.3081
e2	344	2377	205	11.5951	6.9098

In order to analyse the results of our experiments for the two phases, we consider the tables 2 and 3, where the results are expressed in terms of the evaluation metrics (EM) recall and precision, considering association rules (AR), collaborative filtering (CF), the winner of auction (W) and the best possible result if the best recommendation was always chosen (B).

Table 2. Results of experiments online – considering all records, even those where the user follows an item that does not belong to the item set. The end of session records were discarded.

EM	AR		CF		W		B	
	e1	e2	e1	e2	e1	e2	e1	e2
Recall	10,60%	30,62%	13.30%	27.49%	12.05%	34.02%	15.30%	35.13%
Precision	1.06%	3.06%	1.33%	2.75%	1.20%	3.40%	1.53%	3.51%

Table 3. Results of experiments online – considering records where an eligible item was followed. All other records are treated as end of session.

EM	AR		CF		W		B	
	e1	e2	e1	e2	e1	e2	e1	e2
Recall	27.91%	52.20%	35.01%	46.86%	31.70%	58.01%	40.27%	59.89%
Precision	2.79%	5.22%	3.50%	4.69%	3.17%	5.80%	4.03%	5.99%

Also of interest for analysis are the results presented in table 4, where we show the impact of the presence of the multi-agent recommender system (MARS) in terms of the percentage of views of items eligible for recommendation.

Table 4. Results of experiments online

	MARS	
	e1	**e2**
%viewed items eligible for recommendation	47,33%	62,66%
%viewed items that follows a recommendation	23.76%	50.44%

The dispersion graph of Figure 2 shows the distributions of item views obtained on online experiments.

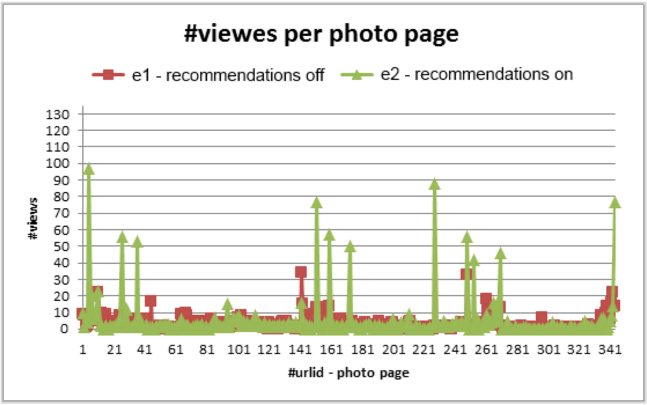


Fig. 2. Item views distribution

5 Discussion and Future Work

Comparing the two phases of the experiments, we observe that for phase one, with the recommendations turned off, the CF based agent gets a better result than MARS. However, for the phase two of experiments, with the recommendations turned on, the MARS is able to obtain better results than the individual algorithms. From table 3 we conclude that, with MARS, the times the users follow a recommended item grows from 31.70% to 58.01%. The score for MARS outperforms the AR based agent in 5.8% and the CF based agent in 11.15%.

Table 4 shows that with the recommendations turned on, the percentage of items (photographs) viewed, grows from 47.33% to 62.66%, the same happened with the percentage of those viewed items that followed a recommendation, 50.44% for e2 while for e1 is 23.76%.

Figure 2 shows that with recommendations turned on is possible to detect more peaks, as a consequence of the users followed the presented recommendations, however, besides this, we could see that the page view distribution is similar with that obtained with de recommendations turned off. So, the MARS does not significantly restrict the range of viewed items.

As we can observe, the MARS is able to clearly outperform the individual algorithms when the recommendations were presented to the user and could be followed. We may also conclude that the increase in the items viewed, as well as in the followed items, when the recommendations were presented to the user, represents a benefit for the users as well for the website.

As future work, we intend to make more online experiments with this and other Websites, and to make some improvements to the MARS.

Acknowledgements. This work was funded by: Fundação para a Ciência e a Tecnologia (FCT), Ref. SFRH/BD/27571/2006; ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), through FCT within project FCOMP -01-0124-FEDER-022701; PRODEI – FEUP.

References

1. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park (1996)
2. Cooley, R., Mobasher, B., Srivastava, J.: Web mining: Information and patterns discovery on the world wide Web. In: Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, California, pp. 558–567 (1997)
3. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender systems: an introduction. Cambridge University Press (2011)
4. Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley & Sons (2002)
5. Ardissono, L., Goy, A., Petrone, G., Segnan, M.: A multi-agent infrastructure for developing personalized web-based systems. ACM Trans. Inter. Tech. 5(1), 47–69 (2005)
6. Albayrak, S., Wollny, S., Varone, N., Lommatzsch, A., Milosevic, D.: Agent technology for personalized information filtering: the pia-system. In: SAC 2005: Proceedings of the ACM Symposium on Applied Computing, pp. 54–59. ACM Press, New York (2005)
7. Morais, A.J., Oliveira, E., Jorge, A.M.: A Multi-Agent Recommender System. In: Omatu, S., Paz Santana, J.F., González, S.R., Molina, J.M., Bernardos, A.M., Rodríguez, J.M.C. (eds.) Distributed Computing and Artificial Intelligence. AISC, vol. 151, pp. 281–288. Springer, Heidelberg (2012)
8. Perkowitz, M., Etzioni, O.: Towards adaptive web sites: Conceptual framework and case study. Artificial Intelligence 118(2000), 245–275 (2000)
9. Ishikawa, H., Ohta, M., Yokoyama, S., Nakayama, J., Katayama, K.: Web usage mining approaches to page recommendation and restructuring. International Journal of Intelligent Systems in Accounting, Finance & Management 11(3), 137–148 (2002)
10. El-Ramly, M., Stroulia, E.: Analysis of Web-usage behavior for focused Web sites: a case study. Journal of Software Maintenance and Evolution: Research and Practice 16(1-2), 129–150 (2004)
11. Berendt, B.: Using Site Semantics to Analyze, Visualize, and Support Navigation. Data Mining and Knowledge Discovery 6(1), 37–59 (2002)
12. Borges, J.L.: A Data Mining Model to Capture User Web Navigation Patterns. PhD thesis, University College London, University of London (2000)
13. Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery 6(1), 61–82 (2002)
14. Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S.: Model-Based Clustering and Visualization of Navigation Patterns on a Web Site. Data Mining and Knowledge Discovery 7(4), 399–424 (2003)

15. Jorge, A., Alves, M.A., Grobelnik, M., Mladenic, D., Petrak, J.: Web Site Access Analysis for A National Statistical Agency. In: Mladenic, D., Lavrac, N., Bohanec, M., Moyle, S. (eds.) Data Mining And Decision Support: Integration And Collaboration. Kluwer Academic Publishers (2003)
16. Basilico, J., Hofmann, T.: Unifying collaborative and content-based filtering. In: Proceedings of ICML 2004, Twenty-first International Conference on Machine Learning. ACM Press, New York (2004)
17. Masseglia, F., Teisseire, M., Poncelet, P.: HDM: A client/server/engine architecture for real time web usage mining. Knowledge and Information Systems (KAIS) 5(4), 439–465 (2003)
18. Lin, W., Alvarez, S.A., Ruiz, C.: Efficient Adaptive-Support Association Rule Mining for Recommender Systems. Data Mining and Knowledge Discovery 6, 83–105 (2002)
19. Spiliopoulou, M., Pohle, C.: Data mining for measuring and improving the success of web sites. Journal of Data Mining and Knowledge Discovery, Special Issue on E-commerce 5(1-2), 85–114 (2001)
20. Armstrong, R., Freitag, D., Joachims, T., Mitchell, T.: WebWatcher: A learning apprentice for the world wide web. In: Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, California, pp. 6–12 (1995)
21. Fink, J., Kobsa, A., Nill, A.: User-oriented adaptivity and adaptability in the AVANTI project. In: Designing for the Web: Empirical Studies. Microsoft Usability Group, Redmond, Redmond (1996)
22. Spiliopoulou, M., Faulstich, L.C.: WUM: a tool for web utilization analysis. In: Atzeni, P., Mendelzon, A.O., Mecca, G. (eds.) WebDB 1998. LNCS, vol. 1590, pp. 184–203. Springer, Heidelberg (1999)
23. Masseglia, F., Teisseire, M., Poncelet, P.: Real Time Web Usage Mining: a Heuristic Based Distributed Miner. In: Second International Conference on Web Information Systems Engineering (WISE 2001), vol. 1, p. 0288 (2001)
24. Jennings, N.R.: An agent-based approach for building complex software systems. Communications of the ACM 44(4), 35–41 (2001)
25. Kephart, J.O.: Research challenges of autonomic computing. In: ICSE 2005: Proceedings of the 27th International Conference on Software Engineering, pp. 15–22. ACM Press, New York (2005)
26. Domingues, M.A., Jorge, A.M., Soares, C., Leal, J.P., Machado, P.: A data warehouse for web intelligence. In: Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA 2007), pp. 487–499 (2007)
27. Miranda, C., Jorge, A.M.: Incremental Collaborative Filtering for Binary Ratings. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2008), vol. 01, pp. 389–392. IEEE Computer Society, Washington, DC (2008)
28. JADE (Java AgentDEvelopment Framework) Website, `http://jade.tilab.com` (access date: January 31, 2014)
29. Asynchronous JavascriptAnd XML (AJAX), Mozilla Developer Center, `https://developer.mozilla.org/en/ajax` (access date: January 31, 2014)
30. Apache Derby Website, `http://db.apache.org/derby/` (access date: January 31, 2014)
31. 1000 Palavras Fotografia, `http://www.1000palavras.pt/` (access date: January 31, 2014)
32. 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)
33. Cleverdon, C., Kean, M.: Factors Determining the Performance of Indexing Systems. Aslib Cranfield Research Project. Cranfield, England (1968)

Designing Strategies for Improving the Performance of Groups in Collective Environments

L.F. Castillo[1,4,⋆], M.G. Bedia[2], C. Lopez[3], F.J. Seron[2], and G. Isaza[4]

[1] Department of Engineering Systems, National University of Colombia, Colombia
[2] Department of Computer Science, University of Zaragoza, Spain
[3] Department of Plastic and Physical Expression, University of Zaragoza, Spain
[4] Department of Systems and Informatics, University of Caldas, Colombia
luis.castillo@ucaldas.edu.co

Abstract. *Capture the Flag* is a well-known game mode that appears in numerous gaming platforms. It consists in a turn-based strategy game where players compete to capture the other team's flag and return it to their base. In order to win, competitive teams must use a great deal of teamplay to generate a successful strategy. Both teams must defend the own base from incoming attackers and get into the other team's base, then take the flag and go back home. The strategy will be a particular case of the well-known "exploration vs. exploitation dilemma" – a recurrent paradox that emerge in all systems that try to get a balance between two types of incompatible behaviors. In this paper, we will show how to apply a "group strategy", based on the "exploration vs. exploitation dilemma" that improves the behavior of a teamplay in a videogame platform.

1 Introduction

The main objective of this paper is to show a series of experiments in the first-person shooter videogame domain (*Capture the Flag* game) and it will be demonstrated how multi-agent systems solve problems of organizing and strategy. We have used a combination of bio-inspired algorithms and organizing roles analyzing potential interactions between bots. Our experiments will be a particular case of the "exploration vs exploitation dilemma", a paradox that appears in numerous situations where systems needs being adaptable and learnable at the same time [2] and solutions of the dilemma are built evolving balances between parts. Our experiments will be programmed in the Unreal tournament 2004 (UT2004) environment. This is first-person shooter enviroment where the target consists in trying to catch the enemy flag. Unreal Tournament 2004 is the tool used in 2kBotPrize2 worldwide contest held every year, where researchers from around the world present their bots in order to overcome an evolution of the

⋆ Corresponding author.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_29,

Turing test. In general, a multi-agent system [3] presents the following characteristics: (i) each agent maintains its own abilities (data acquisition, communication, planning and action, etc.) but (ii) the multiagent system has a global goal. A multiagent system must be able to assign to each of its components one or more specific tasks taking into account the global goal. Several methods have been developed to solve problems from a multiagent-system approach [3] but, in general, they systematically show a pair of incompatible aspects: (1) in order to maximize strategies, agents must solve tasks they know how to deal, but on the other hand, (2) they should try to discover new possibilities. If we focus on the first aspect, we would only exploit the information available with difficulties to be adaptable to new situations. For that reason, the ideal behavior will be built by a balance between two opposite strategies. Traditionally, the fact of finding a balance between these two factors is called the "exploration - exploitation dilemma" (EE). Some authors have remarked that different versions of this dilemma appear in different engineering domains [1]. In general, we could define both strategies as: (1) Exploration, including strategies as searching, experimentation, play, flexibility, innovation, etc., (2) Exploitation, where we can find refinement, choice, production, efficiency, selection, implementation and execution. In this paper, we will show how to solve an instantiation of the EE dilemma in a videogame environment [5]. We will create a team of bots in UT2004 environment with capabilities to solve the problem of capturing the flag, using techniques of long-distance communication. We will use a genetic algorithm to find combinations of bots that create the most efficient system. In the last part, we will evaluate the consequences and advantages of the model implemented.

2 *Capturing the Flag* as a Multiagent Environment

Our model consists of a set of agents that try to "get the flag" and communicate to other partners (See Figure 1). Each bot can listen the transmissions from the others but only when he/she is not evolved in any other tasks. So, as a particular case of a EE dilemma, bots can do two different behaviors: moving around the map while they are looking for the flag, or waiting for other bots to capture the flag and being informed.

Therefore, in the framework of our environment, each bot has two states: (1) *Listening:* the bot is able to receive transmissions from colleagues. If other bot indicates the place where the flag is situated, the agent goes; (2) *Tracking:* In this stage, the bot is randomly searching where the flag is placed. Eventually, the agent can find the flag. However, if other bot is trying to get in contact to it, the transmission does not exist and the message is not received until the agent's state is modified.

2.1 Description of the Problem

Bots are able to get the flag because they can detect the NavigationPoint in which the flag is. They can navigate between NavigationPoints becuas Pogamut

Fig. 1. Snapshot of the game (spanish version): teamplay in the UnrealTournament platform

[1], the Java middleware that enables controlling virtual agents in the Unreal-Tournament environment, uses the A * algorithm for calculating routes. This module contains a PathNavigator which is programmed to avoid obstacles, open doors, wait for elevators, etc. Once the agent has started the movement, every situation that happens along its way, will be captured with a listener. To model the two possible states of the bot (listening and tracking) we have defined a parameter called index of hyperactivity (IH) that is defined by the percentage of time that a bot is moving (with values that go from 0 to 1, with reference to a fixed unit of time "Tc"), which gives us the minimum time interval in both states:

- Time tracking : $IH \cdot\ Tc$
- Listening time: $(1 - IH) \cdot\ Tc$

In every cycle (T_c), every bot walks freely around the map, totally randomly, when the status is 'moving '. Otherwise, in the othr state ('listening') the bot is waiting for information (from other bots) about where the flag is.

2.2 Analytical Approach

From the previous section, it is known that there will be two factors for determining when a bot is able to find a flag: i) the structure of the team (i.e., the IH values for every member) and (ii) the structure of the environment (i.e., the map). In this sense, the probability for finding the flag will be a mapping:

$$p(flag) = p(x_0, N, IH)$$

where x_0 is the point where the bot starts, N is the total number of points in the mesh navigation and IH is the index of hyperactivity that was described in

Fig. 2. Snapshot of the experimental map: 25 x 25 NavPoints

the previous section. Our experiments were made using the following values: 25 Navpoints in the map (See Figure 2), 3 possible states to start, and a team of three bots, so the independent parameter is only the IH value of every bot. It is easy to calculate the probability of finding the flag in terms of IH using notions of markov chains.

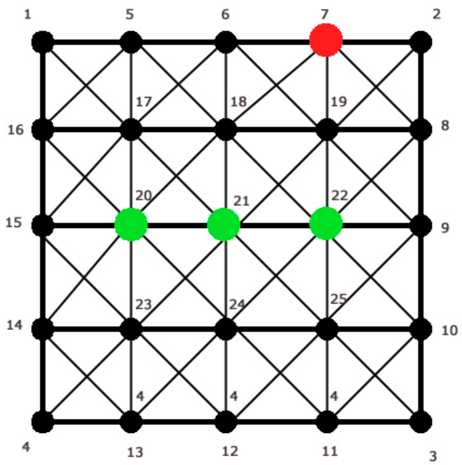


Fig. 3. Environmental and experimental conditions

Movements of our bots are programmed to connect adjacent points. Once they move between two points, without relevance of where they come from, the next step only depends of where they are now. It fulfills the conditions to be modeled by a a Markov chain. First we will see how it is built our navigation mesh points. In Figure 3, a 25 x 25 mesh corresponding to the map created for simulations of this experiment is shown. Exit points for bots are marked in green (points numbers 20, 21, 22) and goal in red (point number 7). Once we have the matrix of the initial probability, a successful strategy according to the number of steps that the bot has to be performed, and will be expressed by: $M^n = M \cdot M \cdots M$, where n is the number of steps. For example, in Figure 4 is shown the probability of reaching the destination point (7) per cycle, in different situations, and depending on the value of IH. In order to calculate the state after several cycles, we must multiply these probabilities a number of cycles we want to check. So, we are able to know the probability of capturing the flag (per bot). In the following, we will model how it changes when we include other bots to solve the problem once the first has already found. It can be modelled as follows:

$$Bot1 : T_t = (Tc \cdot IH3 + MAX(T_{nav2}; T_{nav3})) \cdot n_{ciclos}$$

$$Bot2 : Tt = (Tc \cdot IH3 + MAX(T_{nav1} - Tc \cdot IH3; T_{nav3})) \cdot n_{ciclos}$$

$$Bot3 : Tt = (MAX(IH2 \cdot Tc; T_{enc3}) + MAX(T_{nav1} - IH3 \cdot Tc; T_{nav2})) \cdot n_{ciclos}$$

where

$$IH1 \leqq \ IH2 \leqq \ IH3.$$

If we call Tt the time when the problem is solved, IH as the hyperactivity index bot, Tc is the cycle time, $Tnav_i$ is the time that the bot (i) takes to go from its current position to the flag objective, and Tencies the time it takes to find the bot i the flag in n cycles. We can deduce that the way to minimize the total time resolution of the problem is that the bot 3 finds the flag in the

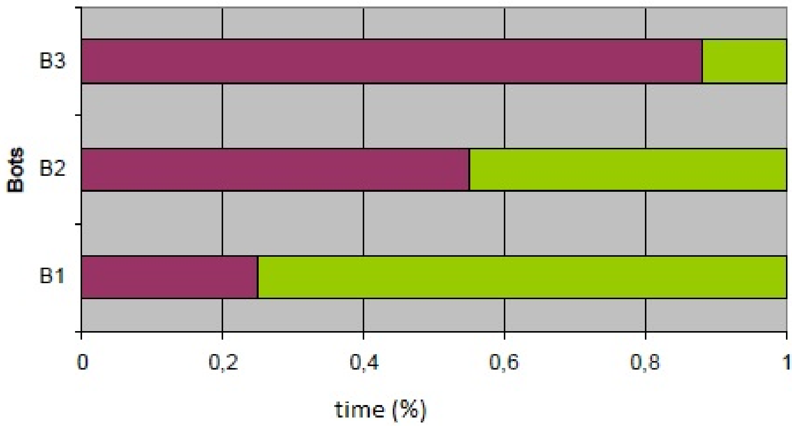


Fig. 4. Graphical picture expressing de relationship of a multiagent system with differente IH values along a cycle time (B1 means Bot1, B2 means Bot 2, and B3 means Bot 3)

first cycle and before the bot 2 to pass the state ' listening', which can be seen graphically in Figure 4.

Once the bots 'embodies' the form of the analytical solution, we implement a genetic algorithm able to evolve bots and try different solutions to see which best fits the optima. In the evaluation process will use a fitness function, which is defined as the time it takes all the bots to reach the flag, so its value is inversely proportional to the efficiency of the team (Fitness = min Ttotal). For each simulation, we limit the maximum time to 35 seconds, so any team that exceeds this time bots will be scored with a fitness of 35, which will be considered as failed goal.

3 Results and Evaluation

After 2900 generations, we have obtained an ordered list of the best fitness, as well the genotypes that have generated these solutions (See in Figure 5). In Probability theory, logistic regression is a regression model for dependent variables or binomially distributed response. The primary objective that solves this technique is to model and influences the probability of occurrence of an event, usually dichotomous, under the presence or absence of various factors.

In our case, the event will be the threshold that our bots must overcome; the parameters will the different structures and combination of genotypes. In the logistic expression, p is the probability of the event occurring, and Bx the parameters calculated by regression. The purpose of the logistic regression is to predict the probability that a problem is solved below a certain threshold (in seconds) for a given genotype. The objective is determining what (hyperactivity index) weight more to increase or decrease the likelihood of having a succesfull task.

The results obtained with a logistic regression for the 2900 data versus fitness thresholds {35, 25 , 15 and 10} are represented in Figure 5. In this table we can see both the weights to be applied to each input as well as their respective p- values. A positive coefficient indicates that a higher value of the input will make the system more ability to get the goal. The opposite occurs with negative coefficients. In a logistic regression , a p- value greater than 0.05 indicates that the coefficient on the applied is not significant (marked in red).

	B1		B2		B3	
	Coef.	P-val	Coef.	P-val	Coef.	P-val
35	1.00155	0.0	0.982621	0.0	0.184483	0.453
25	0.705250	0.003	1.01927	0.0	0.583642	0.023
15	0.742842	0.012	1.16226	0.001	0.996409	0.010
10	0.0120865	0.988	-0.046455	0.956	-2.18946	0.003

Fig. 5. Logistic regression with different teams /thresholds (35,25,15,10)

From these results, we check if the regression has been effective and has helped us to set up teams of bots that solve more quickly the problem of finding the flag versus the restriction of the thresholds . To do this we will create a team of bots for each threshold and we follow the following rules: (1) If the p- value indicates that the coefficient of a particular bot is not significant, we will take the average value of IH of the bot; (2) if the coefficient is positive, and the p-value indicates that it is relevant, we take the largest value within its range; and (3) if we find that the coefficient is negative and the p- value reveals that it is significant, we take the smallest value within the range obtained with the mean and standard deviation of the index of hyperactivity.

With the teams built from the previous instructions, we make 3000 simulations to see if the results provided by the logistic regression are valid and allow us to obtain configurations of bots more efficient in tackling the task at different thresholds. The results of the simulations are shown in Figure 6, where green mark are related to those that got the goal and in red are the ones that did not get the objective.

35	60.88 %	71.11 % (*)	67.20 %	64.52 %
25	42.46 %	53.45 % (*)	51.33 %	43.67 %
15	15.56 %	20.18 %	20.98 %	10.62 %
10	2.59 %	0.76 %	0.49 %	4.75 % (*)

Fig. 6. Results of the teamplays using data from the logistic regression. Every collum es realted to one bot (bot 1, bot 2, and bot 3).

4 Conclusions

We can summarize the results as follows:

- Threshold 35 : For the experimental multiagent-system, the fact of solving the problem in less than 35 seconds is not complicated, as we can infer from the results of the general simulation (nearly 61 %).
- Threshold 25-15: In this case, the system has a fewer number of cycles to solve the problem so we will need more active bots, even assuming the risk of having communication problems. For this reason, we take the highest value of values B3 (and as well with B1 and B2).
- Threshold 10: Overcoming this threshold is very complicated for any multiagent system, and it is shown because a very low percentage of teams get it. In this case everything depends on the most active bot (B3) (with a high value to find the flag and get in contact with the others). In this case, the bot may not hear the transmission while continues to explore. Therefore, we are interested in a low value of B3 to any of the bot that make the unit capable of receiving the transmission and reach the flag in less than 10 seconds.

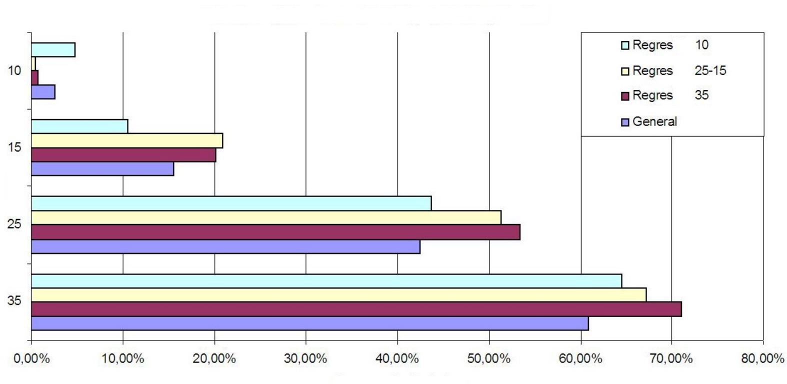


Fig. 7. Graphical results of simulations from logistic regression models

We are facing a very uncertain environment and the fitness function does not punish or reward any behavior: it is only dependent on the time at which the task is solved. This becomes latent in the results, since for the same team of bots rarely they obtain the same result (See Figure 7).

References

1. Gemrot, J., et al.: Pogamut 3 Can Assist Developers in Building AI (Not Only) for Their Videogame Agents. In: Dignum, F., Bradshaw, J., Silverman, B., van Doesburg, W. (eds.) Agents for Games and Simulations. LNCS (LNAI), vol. 5920, pp. 1–15. Springer, Heidelberg (2009)
2. Holland, J.: Adaptation in Natural and Artificial Systems. MIT Press (1992)
3. Thrun, S.B.: Efficient exploration in reinforcement learning. Technical Report CMU-CS-92-102, School of Computer Science, Carnegie Mellon University (1992)
4. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading (1989)
5. Van Lent, L.J.E., Buckman, J., Hartford, J., Houchard, S., Steinkraus, K., Tedrake, R.: Intelligent Agents in Computer Games. In: Proceedings of the National Conference on Artificial Intelligence, Orlando, FL, pp. 929–930 (July 1999)
6. De Oliveira, D.F., Canuto, A.M.P., De Souto, M.C.P.: The diversity/accuracy dilemma: an empirical analysis in the context of heterogeneous ensembles. In: Proceedings of the Eleventh Conference on Congress on Evolutionary Computation (CEC 2009), pp. 939–946. IEEE Press, Piscataway (2009)
7. Lemke, A., Zilmer-Perdersen, L.: Virtual Evacuation Training using Interactive Cognitive Agents. Master's Thesis. Technical University of Denmark (2008)
8. Busemeyer, T., Jessup, R.K., Johnson, J.G., Townsend, J.T.: Building bridges between neural models and complex decision making behaviour. Neural Netw. 19(8), 1047–1058 (2006)
9. Spivey, J.: The continuity of mind. Oxford University Press, New York (2007)

Multiagent Application in Mobile Environments to Data Collection in Park Zones

María Navarro, Fernando de la Prieta, Gabriel Villarrubia, and Mohd Saberi Mohamad

Department of Computer Science, University of Salamanca Plaza de la Merced s/n, 37008, Salamanca, Spain
{mar90ali94,fer,gvg}@usal.es, mohd.saberi@gmail.com

Abstract. This paper presents an application of automatic parking which assists users. For this, recognition techniques are used, such as the algorithm to recognize a license plate from a photograph or the use of NFC protocol to identify an object in the environment. As well, multiagent system integration will be necessary in order to communicate the different devices and data providers involved in this system. We obtain a system that assists the users in his payments and vehicle identification in an efficient way.

Keywords: Virtual organizations, NFC protocol, OCR techniques, neural nets, morphologic operations, agent system, clustering.

1 Introduction

Nowadays, the car is essential for our lives, not only for travel but also for everyday tasks like going to work or shopping. Due to this increased use of vehicles, there are time constraints and parking space. In cities, there are zones, called blue and green zones. Here, the payment system is based on a machine in which a ticked is obtained. This ticket depends on the time you have paid for, but it is a pre-payment form. There exists a guard, who supervises that every user is paying for the parking service. This task consists of looking at the ticked inside the car and writing fines in case there are some infringement or if the ticket is not visible. So, we can consider that the system of parking in these zones is inefficient and uncomfortable for users as well as guards. The proposed model will seek to overcome these difficulties and improve its efficiency.

The main goal of this research is to develop a system that facilitates the client and the guard tasks. Nowadays, the mobile device is an essential tool daily used. That's why a mobile model is implemented so that an user manages his payments, receive notification if any infringement has incurred or know the position of the car at any time. As well, the guard can access via his mobile device to the vehicle identification mechanisms.

This identification was developed by NFC technology and by image recognition. Although there are previous studies that try and solve this problem efficiently [1], there are still difficulties to recognize plates, such as image resolution,

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_30,

color, lighting or computing load. Some ways to partially overcome these problems are the use of infrared camera to know about the position of plates and easily read digits [2]. In our case, the camera position is always variable because it is a mobile device. This is a difficulty that provokes the system not to work properly, so we consider essential the research and improvement of the techniques proposed for license plate recognition using a mobile device, and the integration of this system into a multiagent system to facilitate the communication among the server, the device and the data provider among others.

The paper is structured as follows: Section 2 explains a brief review about agents, NFC technology and image recognition. Section 3 describes the system, integrated with agents and with OCR techniques. Finally, section 4 presents preliminary results of NFC and image identification, studying several algorithms and their results as well as conclusions and future improvements in the system.

2 Background

Although agents field was studied widely to develop this system [3] and the utility of MAS solving Ambient Intelligence problems [4–6], our efforts in the background was dedicated to identification techniques. In particular, we research about states of the art in NFC and plate recognition.

2.1 NFC

NFC is the acronym for the concept *Near Field Communication*. The NFC is a wireless communication interface with a limited working distance of about 10 cm. The interface can operate in several modes. The modes differ if a device creates its own field of radio frequency (RF) or a device gets the power from the RF field generated by another device.

In the field of research, there are several studies on the possibility of applying NFC technology for monitoring certain medical-related parameters [7], or work related to the welfare of the people in the field of ambient intelligence [8] or studies focused on the security vulnerabilities that the NFC protocol can present and possible solutions [9].

2.2 Plate Recognising

Although there are commercial software, ANPR is being extensively investigated [1, 2, 10]. Commercial systems can get higher success rates to 95 % in controlled environments. In environments where lighting conditions are not controlled (e.g. by infrared cameras) or varies the orientation or foreshortening of the plate size, the success rate drops to 75 %.

In the ANPR we can distinguish two phases: 1) Location of the plate within the image, critical step for subsequent identification and 2) the recognition and validation of license plate characters. The methods used to locate the license plate within the image are varied. In any case, the result would be each character separately. Subsequently, the recognition of the segmented characters is often implemented using neural networks. The entrance of the networks may be a direct representation of the characters or indirect, in which the characters suffer from some prior processing [10]. Others opt for the correlation, just as in [11], although in this case the procedure is modified with certain heuristic rules to improve their hit rate.

3 System Description

As discussed above, the system to be built is based on agents [12]. Note that the most important part of our research is to manage the identification, as it is the place where the studies have focused on image recognition, studying OCR with other clustering algorithms and neural networks. We study the NFC identification protocol as well.

The first step in this section include descriptions of agent-based system, to focus after on one target system: identification of an environmental object. We have employed GORMAS [13], and as anticipated in section 3, PANGEA [14] has been our ultimate development platform. In developing the design of the virtual organization and establish the model is necessary to analyze the needs and expectations of potential users of the system. The result of this analysis will be the roles of the entities involved in the proposed system. In particular, the following roles were found:

- User. Represents the user or client system. In this case, the user will be vigilant or client. The agent will separate because different functionality depending on the role you acquire.
- Identify Vehicle: This role is responsible for carrying out the process ID of the vehicle, it can run various methods according to the input parameters.
- Informer: It is responsible for presenting system communications. You can carry out several notifications, depending on what role or agent communicates to him. This role allows the user to interact with the system, and is responsible for collecting the ratings that the user can perform.
- Locator: This process is responsible for locating certain objects in the environment, as the vehicle or the user.
- Infringements: Its responsibility is to manage the whole system of infringements that occur when a client does not have appropriate behavior.
- Communication: It is responsible for communication between the virtual organization of the central server and the organization. This communication results in the response of certain agents according to the input parameters from the server OV.
- Supplier: This role is an agent that accesses and stores all or most of the information needed to manage the actions that govern this system. This information can be extracted through services such as data collection system masCloud [15], or by objects in the environment such as NFC tag, images or information provided by user roles.

– Supervisor: An agent that exercises this role will have overall control of the system. Analyze the structure and syntax of all messages in and out of the system.

A fully description of this techniques will be presents on the following sections.

3.1 Car Identification Techniques

Managing vehicle identification has been a central part of our research. In this process, various techniques have been explored and chosen the most adapted to the needs. We will explore each of them, commenting on the study process and the development.

The NFC recognition is one of the techniques that allowed us to recognize an object in the environment, in our case the vehicle. The system is design in order to recognize habitual vehicles, that is, vehicles registered in our system. Unregistered vehicles are identified by using image recognition.

In NFC, the goal is to communicate an active device (mobile device) with a passive device (NFC tag). The advantage of these *tags* is that require no maintenance or batteries, are activated by the electromagnetic field itself. Each vehicle has a unique number that identifies your car. Thus, it can be read the card with its own mobile and identify the vehicle.

Furthermore, to identify the vehicle by image research about different programs already in operation was performed, but also on the facilities that support the library OpenCV [16]. Currently, OpenCV implements a module to interact with Android more efficiently, but can use native library algorithms such as the algorithm of K-means [17] and Multilayer Perceptron neural networks, without significantly increasing the runtime. Therefore, we chose this as the main tool to analyze and process the image to extract the text.

The next step was to process the image to extract the basic features of it and be able to recognize the digits. Specifically, we were ahead 4 main stages:

1. Remove the plate
2. Extract each digit separately
3. Recognize each digit
4. Check the accuracy of the recognized plate

To remove the plate, the same composition was studied, considering only the license plates of cars (most of our clients). Its design is unique. On the left is wearing a blue strip with the symbol of the European Union and the letter E for Spain, and contains three letters ranging from BBB to ZZZ, and 4 digits ranging from 0000 to 9999. This character set is the necessary key to the identification of each vehicle. Being a white plate with black letters, it facilitates to some extent the image processing. The purpose of preprocessing is to obtain a binarized image, where there are only black and white colors.

There are many techniques to improve image quality and convert it in black and white. In fact, many times the recognition algorithm success depends heavily on this pre-processing and extraction of digits. In this study we have tried various forms of image processing, though the best results have been given, as described below.

To minimize calculus within the mobile device, three steps are followed:

1. Perform a transformation of the color image to grayscale image
2. Eliminate certain noise in the image, using an smoothing technique contours [18].
3. Eliminate certain imperfections in the image, using some morphological transformation.

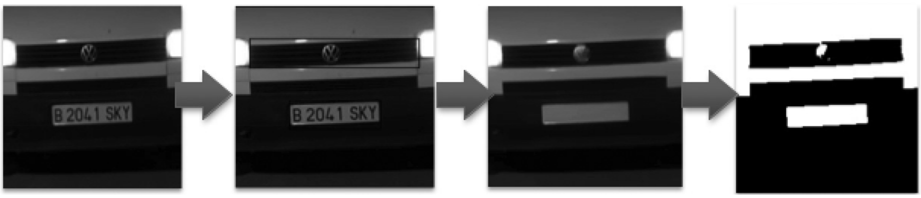


Fig. 1. Outlining the transformation of the image to identify the license plate

To extract the digits, a similar process was followed. Image 2 show the final result.

Fig. 2. Outlining the transformation of the image to identify the plate digits

To recognize the digit two techniques were studied. On one hand, various clustering algorithms are contemplated and finally opted for the study of the nearest neighbour algorithm. Furthermore, a multiperceptron neural network was trained, and the results were observed. In particular, we consider 50 neighbors for the clustering algorithm, and a multilayer perceptron net in the case of neural nets, with only one hidden layer of 150 neurons, 144 entries (one for each pixel of the image), and 28 output neurons, one for each letter. A tansig function is considered as the activation function in hidden layer, and a logsig function for output layer.

In both cases it has had to create a set of training data, and test data. For training, eight examples of each letter, considering that may have different inclinations (included in the training set) are created. With this training set, the neural network and the clustering algorithm was trained and then tested with images collected from real examples, to study what level of precision and total error committed.

4 Preliminary Results and Conclusions

The degree of accuracy in identifying the vehicle with NFC is always 100 %. Tests have been carried out with four types of NFC tags of different technologies, depending on the capacity (0'5K, 1K or 2K), always with the same results. Possible security issues have not been taken into account due to the brevity of the communication between the tag and the mobile device. Still, it is desirable to study new innovations to protect this type of communication. This study is already in future lines.

There were two difficulties when collecting data. The first was to develop an agent collecting data correctly by the necessary protocols. Fortunately, there are interfaces that facilitated the exchange of information between the active and passive device. The second difficulty was to study what was the maximum distance at which the mobile device detects the tag. This was an important functionally because the guard wants the best possible comfort to read labels. Therefore, it was investigated on different ways to increase the recognition distance. In the end, the conclusion was that we relied mainly on hardware and type of active device used.

Test were made to study the results with license plate recognition in an image, with special attention to the type of lighting and features, and the type of algorithm or structure implemented. The nearest neighbor algorithm was set to 50 neighbors to consider when classifying each character. Neural networks were a little more complex to define and train, so it was considered only a Multilayer Perceptron network with a hidden layer. The first problem was precisely determine the most appropriate network structure. From the outset, we consider a network with an input pixel. It was determined that each image must be 12 pixels high by 12 pixels wide, which left us with a neural network of 144 entries. The number of outputs had to be equal to the total points (28 output neurons, one for each letter).

After several tests, the number of neurons in the hidden layer 150, with a tangential activation function (the most commonly used to recognize characters [19]) was fixed.

We present an illustrative table below (Table 1) with the results obtained using the nearest neighbor algorithm, compared with neural networks. The success rate is also presented in recognizing the license plate and each of the digits. The data have been calculated at 25 test plate images.

Upon closer analysis, it was revealed that some characters such as M or W are recognized in most cases, while others were easily confused with each other. Still, the error level increased if tested were distorted or in low resolution. In particular, we note that the letter B was often recognized as the number 8, the G and the number 6. This confusion between numbers and letters could be

Table 1. Comparative table of results at each stage of recognition

Stage	Clustering	Neural Nets
Plate Defining	84% (21/25)	
Digits Defining	86'5% (173/200)	
Characters Recognition	60% (120/200)	80'5% (161/200)

overcome by introducing a condition. If the recognized digit was one of the first four, would be recognized as a letter. Otherwise, be recognized as a number. Thus, if for example, detect an 8 on the second digit, we would consider a B Of course, an improvement of this solution is possible if the recognizer is improved with some fuzzy logic patterns or more training sets.

Moreover, some characters with similar patterns are not correctly recognized. For example, X, K and Y, or B and G. For the latter problem the only plausible solution was to fatten the number of training patterns, and improve the recognizer.

The execution times of license plate recognition depended on the different characteristics of the image. Still, we can set the average execution time of about 4 seconds. It is an acceptable time, if a little slow for the mobile environment in which it is developed. Improving this time and the results shown is considered as future lines.

Another highlight was the successful development of virtual organizations to achieve the objectives of this work. The process of identifying and organizing roles helped improve the management and thus to improve efficiency. PANGEA as platform allowed fluid communication between agents, which is evident in the design of the application.

5 Conclusions

The research presented herein aims to bring new perspectives in the field of image recognition using the multiagent systems to communicate data server and camera from a mobile device. Also research in the field of virtual organizations was essential for the proper management of the implementation of a multiagent system. In this regard, it has been studied and used a development-oriented organization method that covers the analysis and design of the structure of de virtual organization, showing that the system is feasible to solve our problem of parking automation.

With this approach, the expected benefits are to improve the users' comfort. On one hand, the user is able to safe his money, because he only pays the time he was parked. On the other hand, the payment, the payment renovation after the limit of time imposed by the enterprise is made only touching a button. Also, the model assists the guards by identifying the vehicle in a very easy way, only aproximating the mobile device to the stick in the car glass. This might provoke that the enterprise can have less employeers. Also, the enterprise can take advantage of the users' vehicle localization to advertise restaurants or shops near the user.

So, could you discuss the expected profits estimation as well as possibility of direct application of the results provided. What about a real-life verification of the approach proposed?

We have shown that there is a viable way of recognizing digits using a camera of a mobile device. However, the results have not been as good as expected at first. As a future line, we propose to improve results not only on accuracy, but at runtime, essential because of the limited resources of the mobile device and the nature of it (easy to use and fast). Some proposals to reduce this time could be to use another clustering algorithm, another neural network structure or even change the network. This should never be detrimental to the accuracy of the results. Anyway, it is viable to use this identification system in a real system, allowing the guard to correct manually if there is a mistake in the plate recognition.

In addition to these improvements, this application can be analyzed for future research in other fields of research. For example, the application could tell the user what places are free, looking for a new algorithm fuzzy logic to decide in terms of traffic and a number of certain parameters. However, this work has not been explored yet.

Acknowledgements. This work has been partially supported by the Spanish Government through the project iHAS (grant TIN2012-36586-C01/C02/C03).

References

1. López Fernández, J.M.: Software para el reconocimiento automático de matrículas (2006)
2. Vázquez, N., Nakano, M., Pérez-Meana, H.: Automatic system for localization and recognition of vehicle plate numbers. Journal of Applied Research and Technology (2002)
3. Rodrguez, S., de Paz, Y., Bajo, J., Corchado, J.: Social-based planning model for multiagent systems. Expert Systems with Applications 10(38), 13005–13023 (2010)
4. Corchado, J., Fyfe, C.: Unsupervised neural method for temperature forecasting. Artificial Intelligence in Engineering 4(13), 351–357 (1999)
5. Tapia, D., Abraham, A., Corchado, J., Alonso, R.: Agents and ambient intelligence: case studies. Journal of Ambient Intelligence and Humanized Computing 2(1), 85–93 (2010)
6. Fdez-Riverola, F., Corchado, J.: Cbr based system for forecasting red tides. Knowledge-Based Systems 5(16), 321–328 (2003)
7. Patrick, K., Griswold, G., Raab, F., Intille, S.S.: Health and the mobile phone. Technical report (2008)
8. Morak, J., Kumpusch, H., Hayn, D., Modre-Osprian, R., Schreier, G.: Design and evaluation of a telemonitoring concept based on nfc. Enabled Mobile Phones and Sensor Devices. IEEE Transactions on Information Technology in Biomedicine

9. Haselsteiner, E., Breitfuss, K.: Security in near field communication. strengths and weakness. Technical report (2012)
10. Shyang-Lih, C., Sei-Wan, C.: Automatic license plate recognition. IEEE Trans. on Intelligent Transportation Systems
11. Martín, F., Borges, D.: Automatic car plate recognition using partial segmentation algorithm. In: SPPRA
12. Venturini, V., Carbo, J., Molina, J.M.: Methodological design and comparative evaluation of a mas providing ami. Expert Systems with Applications 39(12), 10656–10673 (2012)
13. Argente, E., J. Botti, I.V.: GORMAS: Guía para el desarrollo de sistemas multiagente abiertos basados en organizaciones. PhD thesis (2008)
14. Zato, C., et al.: PANGEA - Platform for Automatic coNstruction of orGanizations of intElligent Agents. In: Omatu, S., Paz Santana, J.F., González, S.R., Molina, J.M., Bernardos, A.M., Rodríguez, J.M.C. (eds.) Distributed Computing and Artificial Intelligence. AISC, vol. 151, pp. 229–239. Springer, Heidelberg (2012)
15. De la Prieta, F., Rodríguez, S., Bajo, J., Corchado, J.M.: A Multiagent System for Resource Distribution into a Cloud Computing Environment. In: Demazeau, Y., Ishida, T., Corchado, J.M., Bajo, J. (eds.) PAAMS 2013. LNCS (LNAI), vol. 7879, pp. 37–48. Springer, Heidelberg (2013)
16. Bradski, G., Kaehler, A.: Computer vision with the opencv library. Technical report (2008)
17. Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 551–556. ACM (2004)
18. Wahl, F.M.: Image preprocessing procedure for noise removal US Patent 4,747,156 (May 24, 1988)
19. Holley, R.: How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine 15(3/4) (2009)

Organizational Architectures for Large-Scale Multi-Agent Systems' Development: An Initial Ontology

Markus Schatten

Artificial Intelligence Laboratory,
Faculty of Organization and Informatics,
University of Zagreb, Pavlinska 2, 42000 Varaždin
markus.schatten@foi.hr

Abstract. An initial conceptualization of organizational design techniques for large-scale multi-agent systems (LSMAS) based on the paradigm of organizational architecture is introduced. Seven perspectives (organizational structure, processes, culture, strategy, individual agents, organizational agents and inter-organizational context) are analyzed and defined. By using a graph theoretic approach and recursive definitions the ontological framework allows for modeling all perspectives on an arbitrary level of detail and complexity from the global system to low level individuals. The provided conceptualization is the foundation of a to be established ontology on organizational design methods for designing, developing and maintaining LSMAS.

Keywords: ontology, large scale multi agent systems, organizational design, organizational architecture.

1 Introduction

Due to their complexity and high-level of concurrency large-scale multi-agent systems (LSMAS) have to be modeled using a higher level of abstraction to be comprehensible at a global level [1]. (Human) organization theory provides us with the necessary methodology to approach complex systems and through the process of organizational design allows us to develop, analyze and optimize different perspectives of organization: structure, processes, culture, strategy, individual behavior and dynamics [2].

Due to the less formal approach to this field of research[1] in order to be usable in LSMAS development, we find it necessary to establish an ontology of organizational design methods that will allow for modeling complex systems having a formal semantic. The ultimate goal is for the ontology to act as a meta-model

[1] Descriptions from organization theory often include metaphors and vague descriptions which cannot directly be formalized mathematically. Also there are various schools of organization theory which often differ significantly in their conceptualizations.

DOI: 10.1007/978-3-319-07593-8_31,

for a to be established LSMAS development framework that would allow for organizational design of such systems on a high level of abstraction.

The idea of applying organizational design methods to multi-agent systems (MAS) is not new, and there has been a substantial amount of research in this area.[2] We have chosen to limit our approach to a specific worldview in organization theory, namely the paradigm of organizational architecture, which is put forward by [6–10]. From this perspective an organization consists of its organizational structure, organizational culture, its business processes, strategy and individual agents (human or artificial) which together form a complete mutually interconnected system.

Due to the highly unstructured domain to be formalized we used a three step approach in formalizing the needed knowledge about organizational architecture. Firstly we have developed a wiki about organizational design methods[3] in order to provide a (as much as possible) state-of-the-art literature review on organizational architecture in a collaborative environment. Afterwards this wiki was annotated with semantic descriptors to yield a semantic wiki [11–13] in order to make a first step towards formalizing the domain. The last and final step is to use the experience gathered during the semantic wiki development in order to establish a formal ontology of organizational design methods which are applicable to LSMAS. The first part of this step is presented in this article.

The rest of the paper is organized as follows: in section 2 we shall introduce the paradigm of organizational architecture as the main theoretical foundation for the to be established ontology. In section 3 we establish a number of formal concept definitions and their interrelations that capture the most important aspects of our approach. In 4 we draw our conclusions and provide guidelines for future research.

2 Organizational Architecture as a LSMAS Framework

Organizational architecture represents a well established paradigm for describing (human) organizations [6–10, 14]. We will adopt the interpretation of [2] that organizational architecture is a complex organizational system which can be modeled from different perspectives: organizational structure, organizational culture, business processes, strategy and individual agents (human or artificial) which are mutually intertwined. Additionally we will add two important aspects of organizations which are not clearly captured by these perspectives: (1) organizational dynamics and (2) contextual and inter-organizational aspects. We are now able to provide an informal definition of these perspectives:

Organizational structure defines the decision and information flows of an organization.

[2] Due to space constraints, we are unable to provide a comprehensive literature review, but the interested reader is advised to consult [3–5] for an extensive review.

[3] Available at `http://ai.foi.hr/oovasis/wiki`, in Croatian.

Organizational culture defines important intangible aspects of an organization including knowledge, norms, reward systems, language and similar.
Strategy defines the long term objectives of an organization, action plans for their realization as well as tools on how to measure success.
Processes define the activities and procedures of an organization.
Individual agents define the most important asset of any organization - the individuals actually performing the work.
Organizational dynamics define organizational changes including reorganization of any of the above mentioned components.
Context and inter-organizational aspects define organizational behavior towards its environment including strategic alliances, joint ventures, mergers, splits, spinouts and similar.

In the following we will focus on the organizational perspectives leaving individual agent modeling to future research.

3 Conceptualization

In the following we will give a number of more formal definitions that shall outline the main concepts of the to be established ontology of organizational design methods. These definitions are in a way a continuation of the work in [15–17]. While establishing these definitions we were guided by a set of simple but subtle objectives: (1) we want to keep the number of (basic) concepts at a minimum in order to avoid redundancy, but not at the cost of limited expressivity; (2) we will tend to use conceptualizations that are compatible with well established agent development approaches in order to allow for their introduction in the to be developed modeling language; (3) we will tend to establish well founded relations between the perspectives of organizational architecture (as outlined in section 2) to allow compatibility between different forms of organizing regardless of perspective; (4) we will have established organizational design and engineering methods in mind when providing the definitions in order to allow the establishment of not only a best-practices knowledge base, but also a familiar framework for modeling and design, (5) we shall have the interorganizational perspective in mind when defining concepts so not to limit definitions to the design of only one organization, but a multitude of them, as well as the design of their mutual interplay; (6) concepts should be atomic in order to allow for (inter-)organizational dynamics including exchange of perspectives and reorganization.

Herein we will take a graph theoretic approach to defining the various concepts in the perspectives whereby as an inspiration we shall use the fractal organization principle [18][4] which allows us to define organizational concepts recursively. Let us first consider the organizational structure perspective.

Definition 1. *An organizational unit is defined as follows:*

[4] Please bare in mind that this is a (human) organizational principle, not necessarily a mathematical structure.

– *Any agent is an organizational unit.*
– *If $OU = (O, R, C)$ is a labeled graph in which O is the set of organizational units (nodes), R is a labeled set of roles (relations, arcs) and C is a criteria of organizing then OU is an organizational unit.*

This definition has an important implication: it allows us to deal with agents, groups and teams of agents, organizations of agents, networks of organizations of agents (or organizations of organizations) as well as virtual organizations of agents (as overlay structures in the sense of [19]) in the same way. This in particular means that organizational units may form a lattice structure in which each unit can belong to several super-units and/or be composed of several subunits. The criteria of organizing could for example be an objective, function, goal, mission, unit name, higher-order role etc. Note that the fixed point in this definition is the individual agent, the implications of which are twofold: (1) it allows us to connect the perspective of structure with the perspective of individuals (directly), and (2) it allows us to model structure as a network of organizational units, which is defined over the relation of a criteria of organizing. This further implies connections to all other perspectives which might be the particular criteria. Due to the recursive nature of the definition, in a complex organizational hierarchy we are actually dealing with layered hypergraphs (in which edges actually represent graphs). In this way a zooming facility can be implemented that would allow a developer to comprehend the organizational structure on an arbitrary level of detail. Thus, an organizational unit, in a way, represents an aggregation of the underlaying complexity similarly to holonic, swarm-like, aggregate, multi-level and other conceptualizations. Note here also, that the definition of an organizational unit is detached from the definition of an organization, and thus allows us to model interorganizational structure in which organizational units might represent different organizations which might have mutual relations (for example competitor, strategic partner, spinout etc.).

The processes perspective can be modeled in a very similar manner.

Definition 2. *An organizational process is defined as follows:*

– *Any atomic activity performed by some individual agent is an organizational process.*
– *If $OP = (P, R, C)$ is a labeled directed graph in which P is a set of organizational processes, R is a labeled set of ordered relations between processes and C is criteria of organizing, then OP is an organizational process.*

The given definition allows for modeling organizations as networks of processes which can be defined in a number of ways. For example, the criteria for organizing might be that one process uses inputs from another or that two processes are using the same resources, or even that two processes are performed by the same organizational unit or that they are crucial for the same organizational goal. Again this definition is recursive, and inherits the same properties as the definition of organizational units above - possible zooming facilities and detachment of actual organization (thus it is possible to model interorganizational processes). Another possibility here that emerged is to model the same process from various perspectives (e.g. criteria of organizing). Note that the fixed point in this definition is an atomic activity or service performed by some individual agent. This means that we have one possible connection to the individuals perspective as well as to all other perspectives if different criteria are selected.

We will closely bound the strategic perspective to the paradigm of the balanced scorecard [20], but adapt it by again using the fractal principle.

Definition 3. *An organizational strategy is defined as follows:*

- *Any measurable objective that can be achieved by an atomic activity is a strategy.*
- *If $OS = (S, R, C)$ is a labeled directed graph in which S is a set of strategies , R is a set of relations between the strategies and C is a criteria of connections, then OS is an organizational strategy.*

Such a definition provides us with the possibility to model agent organizations as networks of objectives which might be defined in a number of ways depending on the criteria of connections. Such criteria might be influence (the outcome of one strategy influences another, e.g. a mathematical function), responsibility (two strategies are under the responsibility of the same organizational unit), achieveability (two strategies can be achieved by the same organizational process), etc. Note that we deliberately excluded two very important aspects of strategy: action plans and decisions. Both plans and decisions are part of the dynamics perspective, since they represent active changes in this or other perspectives. Due to the recursive definition we are again able to model strategy on a number of different levels of detail, and, due to the detachment of a particular organization, we are able to model interorganizational strategy (for strategic partnerships for example). The fixed point in this case is an atomic objective that can be achieved by an atomic activity which allows the connection of this perspective to the process perspective.

Organizational culture in human organizations is a complex cybernetic system that deals with various intangible aspects of organizational behavior including but not limited to language, symbols, rituals, customs, methods of problem solving, knowledge, learning etc. Due to this complex nature, formalizing culture (for agent organization and even more for human organizations) is a quite hard and non-trivial open research question. In [17] we used cultural artifacts (for example knowledge in some agent's knowledge base, written norms of behavior, language protocols, learning processes etc.) as the fixed point of definition, allowing us to model organizational culture as networks of cultural artifacts. The problem with this definition is that it remained quite vague and didn't quite fit into the rest of the proposed framework since it overlapped with the other perspectives (for example procedures and protocols might be modeled as processes).

In order to provide a more applicable conceptualization of the organizational culture perspective we decided to introduce give a more detailed view by defining only one special concept which is of greatest importance - knowledge artifacts, while other (possibly valuable) concepts have been excluded for the time being.

These excluded concepts might be included in the final version of the ontology if they show to be valuable enough and applicable to a wide range of LSMAS applications. We will model knowledge in terms of organizational memory [21] again in a recursive manner.

Definition 4. *Organizational knowledge artifacts are defined as follows:*

- *Any knowledge artifact*[5] *that is accessible to agents is an organizational knowledge artifact.*
- *If $OK = (K, R, C)$ is a labeled graph in which K is a set of organizational knowledge artifacts, R is a set of relations between these artifacts and C is an organization criteria, then OK is an organizational knowledge artifact.*

Thus, agent organizations can be seen as a network of knowledge artifacts which are accessible by particular agents. Special cases of knowledge artifacts are norms which establish the rules of interaction between agents and values which influence decision making and selection of objectives.

The individual agent perspective does not need much additional explanation since the design and implementation of agents is a well researched field. Still we do need to point out the implications of the aforementioned perspectives on agents. Since agents have to be aware of their organizational context depending on the organizational model the various perspectives will introduce context knowledge into the particular knowledge-bases of agents. Also since agents have to behave in accordance to the model, it will introduce agent behaviors and protocols which agents will have to adhere to.

In the organizational dynamics perspective there are three important concepts: (1) time, (2) event and (3) change. In [16, 17] we have introduced active graph grammars (AGGs) to model changes in the organizational structure of agent organizations. AGGs are an active database theory inspired formalism that allow reactive behavior in graph structures. Since all of the above concepts (organizational units, processes, strategy and knowledge artifacts) are defined in terms of graphs, AGGs (or any similar formalism that allows for event detection and graph transformation) can be used to introduce changes in any of the perspectives. In this way reorganization can occur on any level. AGGs (or similar formalisms) will allow us to implement a best practices knowledge base of organizational design techniques which will be individuals in the to be established ontology.

Some drawbacks of AGGs is that they are defined to be local to agents and that they allow only for reactive behavior. The first problem could be approached by defining AGGs as shared knowledge artifacts to which all agents comply to a certain criteria (are part of an organizational unit for example) have access to. The drawback regarding reactive behavior might be solved using a BDI approach to agents in which the plans are consecutive executions of active graph rewriting rules.

[5] By knowledge artifact we understand a wide range of explicit knowledge in which we assume that it is queriable by the agent, including but not limited to data and knowledge bases, neural networks and machine learning architectures, various information services etc.

This last remark brings us to another important aspect of organizational dynamics and that is the continuous performance of the LSMAS that is being modeled. All previous perspectives only dealt with the static (structural) aspects of organizational architecture. These static aspects set the stage of performance: for example a defined strategy is a referent point in decision making and execution of action. This situation is fairly compatible with the usual BDI approach: beliefs are knowledge artifacts, desires are strategies and intentions are plans for executing processes. While this is a well known procedure for individual agents, the recursive definition of perspectives (especially organizational units and strategy) introduce some additional complexity: here a multitude of agents or even multitude of agent organizations have to reach consensus about a collective process to be performed. Beliefs, reasoning techniques and possible actions of agents might differ considerably, thus there is need for distributed consensus making techniques like abstract argumentation. Still extending the BDI approach to agent organizations is an open research question.

In the end the contextual and inter-organizational perspective has already been partially addressed in the other perspectives, mostly regarding inter-organizational aspects. To model the environment of agent organizations we will use the usual approach to introduce a special individual agent to which all other agents have access.

4 Conclusions and Future Work

In this work-in-progress paper we gave an initial conceptualization for a to be established organizational design methods ontology based on the paradigm of organizational architecture. We defined seven important perspectives that shall allow us different views of a complex system: organizational structure, processes, strategy, culture, individual agents, organizational dynamics and inter-organizational context. By using recursive definitions of most important concepts (organizational units, processes, strategy, knowledge) we established a framework that shall allow us to view parts of agent organizations on an arbitrary level of detail. Since each of these definitions has a criteria of organizing we allowed for modeling different views of the same underlying instances.

Our future work is oriented towards the implementation of this ontology in OWL with the goal of establishing a meta-model for a graphical modeling language of LSMAS. This language should then be implemented in a modeling tool together with a best-practices knowledge base of organizational design techniques.

References

1. Lamarche-Perrin, R., Demazeau, Y., Vincent, J.M.: How to build the best macroscopic description of your multi-agent system? In: Demazeau, Y., Ishida, T., Corchado, J.M., Bajo, J. (eds.) PAAMS 2013. LNCS (LNAI), vol. 7879, pp. 157–169. Springer, Heidelberg (2013)

2. Žugaj, M., Schatten, M.: Arhitektura suvremenih organizacija. In: Tonimir and Faculty of Organization and Informatics, Varaždinske Toplice, Croatia (2005)
3. Horling, B., Lesser, V.: A Survey of Multi-Agent Organizational Paradigms. The Knowledge Engineering Review 19(4), 281–316 (2005)
4. Argente, E., Julian, V., Botti, V.: Multi-agent system development based on organizations. Electron. Notes Theor. Comput. Sci. 150(3), 55–71 (2006)
5. Schatten, M., Ševa, J., Kudelić, R., Tomičić, I.: Organizational Aspects of Large Scale Multi-agent Systems - A Roadmap, pp. 1–23 (in press, 2014)
6. Nadler, D.A., Gerstein, M.S., Shaw, R.B.: Organizational Architecture, Designs for Changing Organizations. Jossey-Bass, San Francisco (1992)
7. Merron, K.: Riding the wave: Designing your organization's architecture for enduring success. Van Nostrand Reinhold (1995)
8. Galbraith, J.R.: Designing organizations: An executive briefing on strategy, structure, and process. Jossey-Bass (1995)
9. Henning, J.P.: The Future of Staff Groups: Daring to Distribute Power and Capacity. Berrett-Koehler Store (1997)
10. Churchill, C.: Managing growth: The organizational architecture of microfinance institutions. In: USAID Microenterprise Best Practices Project Paper, pp. 7–26, 81–87 (1997)
11. Schatten, M., Čubrilo, M., Ševa, J.: A semantic wiki system based on f-logic. In: Aurer, B., Bača, M. (eds.) 19th Central European Conference on Information and Intelligent Systems – CECIIS 2008 Conference Proceedings, Faculty of Organization and Informatics, pp. 57–61 (2008)
12. Schatten, M., Čubrilo, M., Ševa, J.: Dynamic queries in semantic wiki systems. In: Aurer, B., Bača, M. (eds.) Proceedings of the 20th Central European Conference on Information and Intelligent Systems, Faculty of Organization and Informatics, pp. 13–20 (September 2009)
13. Schatten, M.: Knowledge management in semantic social networks. Computational & Mathematical Organization Theory 19(4), 538–568 (2013)
14. Miciunas, G.: Cre/fm organizational architecture: Structuring staff success. The Environments Group (2002), `http://www.envgroup.com/browse/presentations/IFMA2002_staffsuccesspaper.pdf`
15. Žugaj, M., Schatten, M.: Otvorena ontologija organizacijske arhitekture u funkciji upravljanja znanjem. Ekonomski Vjesnik XX(1-2), 39–45 (2007)
16. Schatten, M.: Active graph rewriting rules for modeling multi-agent organizational dynamics. In: Ivković, M., Pejić Bach, M., Šimićević, V. (eds.) Proceedings of the IBC 2012, 1st International Internet & Business Conference, pp. 180–185. BIT Society, Rovinj (2012)
17. Schatten, M.: Reorganization in multi-agent architectures: An active graph grammar approach. Business Systems Research 34(1), 14–20 (2013)
18. Warnecke, H.J.: Die fraktale fabrik - produzieren im netzwerk (the fractal company - production in the network). In: GI Jahrestagung, pp. 20–33 (1992)
19. Barnatt, C.: Office space, cyberspace & virtual organization. Journal of General Management 20(4), 78–91 (1995)
20. Kaplan, R.S., Norton, D.P.: The Balanced Scorecard: Translating Strategy Into Action. Harvard Business School Press, Boston (1996)
21. Walsh, J.P., Ungson, G.R.: Organizational memory. Academy of Management Review 16(1), 57–91 (1991)

Exploring the Role of Macroeconomic Mechanisms in Voluntary Resource Provisioning in Community Network Clouds

Amin M. Khan and Felix Freitag

Department of Computer Architecture,
Universitat Politècnica de Catalunya,
Barcelona, Spain
{mkhan,felix}@ac.upc.edu

Abstract. Internet and communication technologies have lowered the costs of enabling individuals and communities to collaborate together. This collaboration has provided new services like user-generated content and social computing, as evident from success stories like Wikipedia. Through collaboration, collectively built infrastructures like community wireless mesh networks where users provide the communication network, have also emerged. Community networks have demonstrated successful bandwidth sharing, but have not been able to extend their collective effort to other computing resources like storage and processing. The success of cloud computing has been enabled by economies of scale and the need for elastic, flexible and on-demand provisioning of computing services. The consolidation of today's cloud technologies offers now the possibility of collectively built community clouds, building upon user-generated content and user-provided networks towards an ecosystem of cloud services. We explore in this paper how macroeconomic mechanisms can play a role in overcoming the barriers of voluntary resource provisioning in such community clouds, by analysing the costs involved in building these services and how they give value to the participants. We indicate macroeconomic policies and how they can be implemented in community networks, to ease the uptake and ensure the sustainability of community clouds.

Keywords: community cloud, cloud computing, community networks, economic mechanisms, collaborative resource sharing.

1 Introduction

Recent developments in communication technologies like Internet, email and social networking have significantly removed the barriers for communication and coordination for small to large groups bringing down the costs that obstructed collaborative production before the era of Internet [1]. The ICT revolution ushered in group communication and collaborative production with popular applications now widely adopted, like social networking, social bookmarking, user-generated

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_32,

content, photo sharing, and many more. Even infrastructures based on a cooperative model have been built, for example community wireless mesh networks gained momentum in early 2000s in response to limited options for network connectivity in rural and urban communities [2]. Using off-the-shelf network equipment and open unlicensed wireless spectrum, volunteers set up wireless networks in their local communities to provide network and communication infrastructure. These wireless networks have proved quite successful, for example Guifi.net[1] provides wireless and optical fibre based broadband access to more than 20,000 users. Community networks successfully operate as IP networks, since the nodes' bandwidth is shared among all the members in a reciprocal manner.

Despite achieving sharing of bandwidth, community networks have not been able to extend this sharing to other computing resources like storage. There are not many applications and services used by members of community networks that take advantage of resources available within community networks. Community networks are based on voluntary contributions of participants, and economic or social incentives to encourage this have been crucial to achieve the sustainability of the community networks [3]. Apparently the current incentives in community networks are not sufficient enough to overcome the barriers for realising the sharing of other computing resources besides just bandwidth.

Sharing of computing resources in the Internet is now commonplace because of the wide adoption of cloud computing model [4]. Cloud computing provides on-demand, elastic, flexible and cost-effective access to computing resources. Today's clouds are mainly provided upon a pay-per-use model, where the cloud services are offered to the consumers as a utility and by commercial providers. Cloud computing allows enterprises and individuals to reduce significantly the time and capital investment in setting up their own infrastructure. Instead, they can request resources on demand from the cloud services providers, which not only lowers the total cost of ownership for consuming resources because of economies of scale, but leaving low level details to the service providers focus can be shifted towards building and using high level applications. This also applies that an individual or organisation is no longer limited by the resources present locally and owned directly. When demand exceeds the current capacity, more resources can be requested on the fly from one or more cloud services providers. This has relevance for community networks as the members in aggregate boast much more resources than owned by a single individual or a small group. When members of community network can share and trade resources based on a cloud computing model, they can sell their excess capacity as the demand fluctuates and in return can take advantage of services and applications that were not possible earlier due to the limited resources locally.

The concept of community clouds has been introduced in its generic form before, e.g. [5, 6], as a cloud deployment model in which a cloud infrastructure is built and provisioned for an exclusive use by a specific community of consumers with shared concerns and interests. We refer here to a specific kind of a community cloud in which sharing of computing resources is from within community

[1] `http://guifi.net`

networks, using the application models of cloud computing in general. Members of community network can share and trade resources, they can sell their excess capacity as the demand fluctuates and in return can take advantage of services and applications that the community cloud enables, which were not possible earlier due to the limited resources on the users' local machines. Realising community cloud involves a lot of challenges both in technological and socio-economic context, but also promises interesting value proposition for communities in terms of local services and applications.

Our main objective in this paper is to explore the macroeconomic mechanisms that can help in adoption and growth of community cloud model. We contribute first a cost-value proposition describing the conditions under which community clouds should emerge. Secondly, we propose a set of macroeconomic policies that, if placed in community networks, should accelerate the uptake and help the sustainability of community clouds. We elaborate on this in the rest of the paper as follows. Section 2 introduces possible cloud scenarios in community networks. Section 3 discusses our cost-value proposition of community clouds, and section 4 proposes different macroeconomic mechanisms for community clouds. Section 5 concludes and indicates future work.

2 Cloud Scenarios in Community Networks

We consider clouds in community networks, a community cloud that provides services built from using resources available from within the community networks and owned and managed by the members of the community networks themselves. Such a community cloud infrastructure that is deployed in real community networks needs to be designed according to the conditions and characteristics of community networks, which also determine the most likely scenarios for these community clouds.

2.1 Background on Community Networks

A community network like Guifi.net is organised into zones where a zone can be a village, a small city, a region, or districts of a larger city. Mostly, the detailed technical support for the members is only available within the community of their zone [7], so we identify a zone to have the highest social strength within the community network. The computer machines or nodes in a community network vary widely in their capacity, function and capability, as illustrated in Figure 1. Some hardware is used as super nodes that have multiple wireless links and connect with other super nodes to form the backbone of the community network [7]. Others act just as clients and are only connected to the access point of a super node. As depicted in Figure 1, resources for the community cloud can be attached to the networking nodes.

2.2 Local Community Cloud

The cohesive nature of zones gives rise to the scenario of the *local community cloud*, interpreting the characteristics of the social networks existing within zones

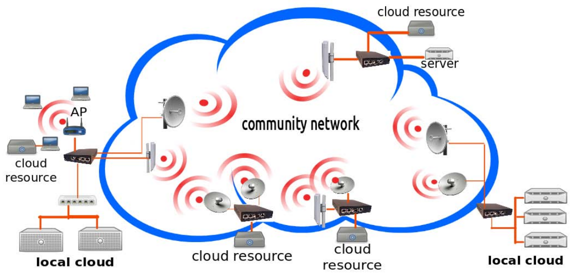


Fig. 1. Nodes in a community network with cloud resources

and the topology of the community network. In this scenario, some super nodes with their better connectivity and high availability are responsible for the management of a set of attached nodes that are contributing cloud resources.

2.3 Federated Community Cloud

Local community cloud can provide services for the users within its zone. Multiple cloud nodes from different zones in a community network, however, can participate together in a *federated community cloud* to support greater functionality and higher capacity. The nodes in a given zone are directly managed by a super node in that zone but they can also consume resources from other zones, given that there is a coordination mechanism among zones in place. Within an economic context, the local community cloud is an example of a virtual organisation, and the federated scenario represents the peering agreements between multiple virtual organisations.

3 Cost and Value Relationships in Community Cloud

The community clouds can be seen as private enterprises with private provisioning of public goods. This model can suffer from social dilemmas, like the tragedy of the commons, meaning that free riding and under-provisioning will destroy the system in the absence of any mechanisms to overcome these issues. The socio-economic context of community networks implies that mechanisms that foresee social exclusion can be effective to direct the users' behaviour [8].

Figure 2 shows the desired relationship between the cost and value proposition as the community cloud evolves and gets adopted by wider audience. In the nascent stage, the community cloud will not be able to provide much value until a critical mass of users are using the system. After that threshold, still the relative cost to achieve a little utility will be significant, which means that the early adopters of the system remain highly motivated and committed to the success of community cloud and continue to contribute resources even though they receive

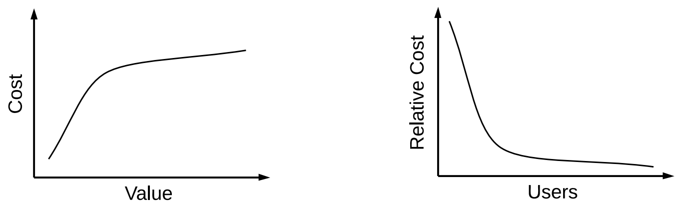


Fig. 2. Relationship between cost and value in evolution of community cloud

little value from the system in return. But once a significant proportion of the overall population has joined the community cloud, the relative cost to obtain value from the system tumbles and in the longer run the system is able to sustain itself with contributions that may be small in size but are made by a large number of users. The objective of the economic mechanisms and the social and psychological incentives is to let the system transition from inception through early adoption to finally ubiquitous usage.

3.1 Costs for Participation

The initial costs for setting up nodes in the community cloud involves hardware costs including the price of the computing and networking equipment, and installation costs including the manual labour needed. The continuous operation of the cloud node requires additional costs including network costs given by donating network bandwidth and any other subscription fees, energy costs to pay for electricity bills to run the computer equipment as well as cooling apparatus, maintenance cost to fund any technical support and replacements for parts, and hosting costs to provide storage space for the equipment. Besides these costs at the individual level, there are also the transaction costs [9] or management overheads to direct the group coordination and collaborative production efforts necessary for the operation of community cloud.

3.2 Value Proposition

The individuals in community cloud act as private enterprises where they offer services to generate revenue. The revenue for the community cloud users include tangible benefits like the services and applications that they will be able to consume, and intangible benefits like the sense of belonging to the community and personal satisfaction because of their contributions. The services can range from infrastructure to platform to software services meeting a spectrum of different needs of the users. Once community cloud gets adopted by a critical mass, community may also generate revenue by offering computing resources to commercial enterprises, similar to selling excess power capacity in the case of Smart Grid. For example, community can get into partnership agreements with the ICT providers where community can buy network bandwidth in return for providing access to the computing resources of the community cloud.

3.3 Comparison with Commercial Services

We discuss the community cloud cost and value in comparison with two popular commercial services that are also based in part on the idea of reciprocal sharing, Spotify[2] and Skype[3]. Spotify is a subscription-based music streaming service which reduces its infrastructure and bandwidth costs by serving cached content from users' devices as well as its own servers. Skype is a communication service which uses caches on users' devices for storing and processing information required for managing the underlying infrastructure. Both Spotify and Skype offer free as well as paid services. Why do users agree to contribute resources, and even when they are paying for the service?

An argument is that the costs for users are minimal. Both services mostly consume storage, computation time, power and bandwidth on the users' devices. Since these resources are not very expensive and the services' usage remains relatively low, the users do not mind this arrangement or not even notice it. But even more important, these services are designed so intuitively that most users do not even realise about donating the resources, and even when they do, the value these services provide has sufficient incentive.

The success of such services implies that for community cloud as well, the users should be able to join with zero or very little costs. The value proposition of the community cloud services should be strong enough to attract early adopters and keep them committed. The economic mechanisms in place for encouraging reciprocal sharing and ensuring overall system health and stability should be either invisible for non-technical users or very simple to understand and work with.

4 Design of Macroeconomic Policies

We discuss in this section the macroeconomic policies we propose for community clouds, addressing relevant issues of the technical, social, economic and legal aspects of the community cloud system. We approach the problem by having explored some of the mechanisms previously in simulations [10] and also by developing a prototype implementation which is currently deployed in the Guifi community network [11] and which will allow to get users involved and participating in a real world scenario.

4.1 Commons License

The agreement and license to join a community cloud should encourage and help enforce reciprocal sharing for community clouds to work. The Wireless Commons License[4] or Pico Peering Agreement[5] is adopted by many community networks to regulate network sharing. This agreement could serve as a good base for drafting an extension that lays out the rules for community clouds.

[2] http://www.spotify.com
[3] http://www.skype.com
[4] http://guifi.net/es/ProcomunXOLN
[5] http://www.picopeer.net

4.2 Peering Agreements

When different community clouds federate together, agreements should ensure fairness for all the parties. Agreements between different communities should describe the rules for peering between clouds. Within such agreements, local currency exchanges could be extended to address cases of imbalance in contribution across different zones [12].

4.3 Ease of Use

The easier it is for users to join, participate and manage their resources in the community cloud, the more the community cloud model will be adopted. This requires lowering the startup costs and entry barriers for participation. To this end, in terms of an institutional policy, we have developed a Linux-based distribution[6], to be used in the Guifi.net community cloud [11]. It will make the process of joining and consuming cloud services almost automated with little user intervention. This effect will make the community cloud appealing to non-technical users.

4.4 Social Capital

Community clouds need to appeal to the social instincts of the community instead of solely providing economic rewards. This requires maximising both bonding social capital [13] within local community clouds in order to increase the amount of resources and commitment of the users, and bridging social capital in order to ensure strong cooperation between partners in federated community clouds. Research on social cloud computing [14] has already shown how to take advantage of the trust relationships between members of social networks to motivate contribution towards a cloud storage service.

4.5 Transaction Costs

The community cloud, especially in its initial stages, will require strong coordination and collaboration between early adopters as well as developers of cloud applications and services, so we need to lower the transaction costs for group coordination [9]. This can take advantage of existing Guifi.net's mailing list[7], but also of the regular social meetings and other social and software collaboration tools. It also requires finding the right balance between a strong central authority and decentralised and autonomous mode of participating for community members and software developers.

---

[6] `http://repo.clommunity-project.eu`

[7] `http://guifi.net/en/forum`

4.6 Locality

Since the performance and quality of cloud application in community networks can depend a lot on the locality, applications need to be network and location aware, but this also requires that providers of resources should honour their commitment to local community cloud implying that most requests are fulfilled within the local zone instead of being forwarded to other zones. We have explored the implications of this earlier when studying the relationship between federating community clouds [10, 15].

4.7 Overlay Topology

Community networks are an example of scale-free small-world networks [7], and the community cloud that results from joining community networks users is expected to follow the same topology and inherit characteristics similar to scale-free networks. As the overlay between nodes in the community cloud gets created dynamically [16], the community cloud may evolve along different directions as users of the underlying community network join the system. As the applications in community cloud will most likely be location and network aware to make the most efficient use of the limited and variable resources in the network, the overlay steered concentration and distribution of consumers and providers of services direct the state and health of the community cloud.

4.8 Entry Barriers

In order to control the growth of the community cloud and provide a reasonable quality of experience for early adopters and permanent users, different approaches can be considered, for example, a community cloud open to everyone, by invitation only, or one that requires a minimum prior contribution.

4.9 Role of Developers

The developers of the cloud applications are expected to play an important intermediary role between providers of resources and consumers of services, for example adding value to the raw resources and selling them to consumers at a premium. End users could have both the roles of raw resource providers and consumers which find the value of the cloud in the provided applications.

4.10 Service Models

Cloud computing offers different service levels, infrastructure, platform and software-as-a-service (SaaS). Similar to the three economic sectors for provisioning goods, the third level, the SaaS of the cloud reaches the end users. For providing value from the beginning in the community cloud, we propose to prioritize provisioning SaaS at the early stage of the community cloud.

4.11 Value Addition and Differentiation

The community cloud requires services that provide value for users. In addition, these services need to compete and differentiate from the generic cloud services available over the Internet. In this line, FreedomBox[8] services focus on ensuring privacy, and FI-WARE CoudEdge[9] and ownCloud[10] let cloud applications consume resources locally.

5 Conclusion and Future Work

Community clouds take advantage of resources available within community networks for realising cloud-based services and applications tailored to local communities. Being community clouds a case of private provisioning of public goods, economic mechanisms and policies are needed to direct their growth and sustainability. First, we identified the cost and value evolution of the community cloud during its emergence and under permanent operation. A core number of highly motivated contributors is needed at the beginning. Once the community cloud is operational, its value should easily exceed the cost of the minor contribution expected from the users. The socio-economic context of community networks forms the basis for the macroeconomic policies that we proposed for community clouds. We outlined and illustrated these policies that address technical, social, economic and legal aspects of the community cloud system.

Based on the proposed macroeconomic policies, our next step is to design and integrate them in our prototype implementation of the community cloud that we currently deploy in the real-world Guifi.net community network. The resulting empirical studies will help assessing the effect of the proposed economic mechanisms further. Our hope is that community clouds will complement the existing public cloud services paving the way for innovative and interesting applications for local communities.

Acknowledgement. This work was supported by the European Framework Programme 7 FIRE Initiative projects CONFINE, FP7-288535, and CLOMMUNITY, FP7-317879, and by the Universitat Politècnica de Catalunya BarcelonaTech and the Spanish Government through the Delfin project, TIN2010-20140-C03-01.

References

1. Shirky, C.: Here Comes Everybody: The Power of Organizing Without Organizations. Penguin Group (2008)
2. Braem, B., et al.: A case for research with and on community networks. ACM SIGCOMM Computer Communication Review 43(3), 68–73 (2013)

[8] http://freedomboxfoundation.org
[9] http://catalogue.fi-ware.eu/enablers/cloud-edge
[10] http://owncloud.org

3. Bina, M., Giaglis, G.: Unwired Collective Action: Motivations of Wireless Community Participants. In: International Conference on Mobile Business (ICMB 2006), Copenhagen, Denmark, pp. 31–40. IEEE (June 2006)
4. Buyya, R., Broberg, J., Goscinski, A.: Cloud Computing: Principles and Paradigms. John Wiley & Sons, Inc. (2011)
5. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. NIST Special Publication 800(145) (2011)
6. Marinos, A., Briscoe, G.: Cloud Computing
7. Vega, D., Cerda-Alabern, L., Navarro, L., Meseguer, R.: Topology patterns of a community network: Guifi.net. In: 1st International Workshop on Community Networks and Bottom-up-Broadband (CNBuB 2012), within IEEE WiMob, Barcelona, Spain, pp. 612–619 (October 2012)
8. Greiff, M.: Rewards and the private provision of public goods on dynamic networks. Journal of Evolutionary Economics 23(5), 1001–1021 (2013)
9. Coase, R.H.: The Nature of the Firm. Economica 4(16), 386–405 (1937)
10. Khan, A.M., Büyükşahin, Ü.C., Freitag, F.: Towards Incentive-Based Resource Assignment and Regulation in Clouds for Community Networks. In: Altmann, J., Vanmechelen, K., Rana, O.F. (eds.) GECON 2013. LNCS, vol. 8193, pp. 197–211. Springer, Heidelberg (2013)
11. Jiménez, J.: et al.: Deploying PaaS for Accelerating Cloud Uptake in the Guifi.net Community Network. In: IEEE International Workshop on the Future of PaaS 2014, Boston, Massachusetts, USA (March 2014)
12. Punceva, M., Rodero, I., Parashar, M., Rana, O.F., Petri, I.: Incentivising resource sharing in social clouds. Concurrency and Computation: Practice and Experience (March 2013)
13. Coleman, J.S.: Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120 (1988)
14. Chard, K., Bubendorfer, K., Caton, S., Rana, O.F.: Social Cloud Computing: A Vision for Socially Motivated Resource Sharing. IEEE Transactions on Services Computing 5(4), 551–563 (2012)
15. Khan, A.M., Büyükşahin, U.C., Freitag, F.: Prototyping Incentive-based Resource Assignment for Clouds in Community Networks. In: 28th International Conference on Advanced Information Networking and Applications (AINA 2014), Victoria, Canada. IEEE (May 2014)
16. Nakao, A., Wang, Y.: On Cooperative and Efficient Overlay Network Evolution Based on a Group Selection Pattern. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40(2), 493–504 (2010)

Performance and Results of the Triple Buffering Built-In in a Raspberry PI to Optimize the Distribution of Information from a Smart Sensor

Jose-Luis Jimenez-Garcia[1], Jose-Luis Poza-Luján[2], Juan-Luis Posadas-Yagüe[2], David Baselga-Masia[1], and José-Enrique Simó-Ten[2]

[1] School of Engineering in Computer Science (ETSINF), Valencia, Spain
[2] University Institute of Control Systems and Industrial Computing (ai2)
Universitat Politècnica de València (UPV, Camino de vera, s/n. 46022 Valencia, Spain
{jojigar1,dabama1}@inf.upv.es,
{jposadas,jopolu,jsimo}@ai2.upv.es

Abstract. Currently, 3D sensors can be considered an evolution of cameras by providing the image with its depth information. These sensors have a generic function and the programmer has to process the received information in order to be adapted and used in a specific environment. In robots navigation, the 3D information can be useful for basic behaviours such as "obstacles avoidance" or even more complex behaviours such as "maps generation". In this article an image management system provided by the xTion intelligent sensor is presented. The xTion sensor provides a VGA image and a 3D depth, which allows it to be used for several purposes. In order to distribute the data, it is acquired, processed and sent to several clients with a triple buffer system modified to serve the most recent image to the client. The system is programmed in C for Linux and built-in in a Raspberry PI. The article exposes the performance and results from monitoring the frame's delay comparing it with a simple and a double buffer system widely used in this kind of systems.

Keywords: intelligent sensors, buffering, distributing information.

1 Introduction

The use of sensors in robot navigation goes from basic survival behaviours like obstacle avoiding to complex task realization like map generation. In basic survival behaviours, commonly associated to reactive navigation, were commonly employed sensors with a specific task like depth or contact sensors. To complex behaviours, commonly associated to deliberative navigation has been used information from reactive information sensors. Currently, robots need sensors associated to behaviours. For example, distance sensors are utilized to avoid obstacles. However, this information is useful to build an environment map or to locate specific objects during the navigation. So, the information that a sensor produces is interesting it to be distributed to different behavioural processes.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_33,

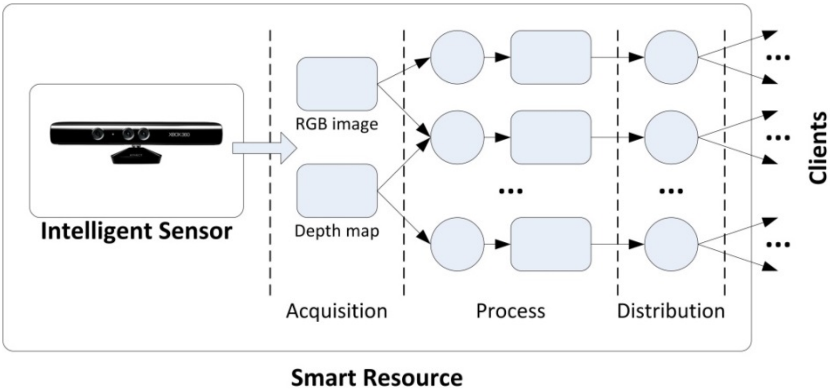


Fig. 1. Smart resource components

The article is organized in this order: next section introduces the related work. Section 3 the implemented system is described, in section 4 the developed triple buffer variation is detailed. In section 5 the experiments and results obtained are described, finally, in section 6 the future working lines and conclusions are exposed.

2 Related Work

An intelligent sensor is one that modifies its internal behaviour to optimize its ability to collect data from the physical world and communicate them in a responsive manner to a host system [1]. Nowadays intelligent sensors with great computation capabilities and large amount or information are offered by the technology, like image and depth of the provided frame.

There are several smart resources employing intelligent sensors [2] generally used on image-based systems. It seems convenient using the intelligent sensors both to reactive navigation to deliberative navigation [3] and [4]. However, using an intelligent sensor with several clients and with different client requirement implies offering the sensor information in the more suitable way to every client and distributes it so that the clients can control enough to perform their behaviours, especially if those behaviours are reactive.

The fact that the information transmitted by a smart sensor can be processed and adapted to the client needs turns it into a smart resource. [5]. Anyway, this process and its later distribution imply a processing cost which can decrease the information quality (figure 1).

In order to optimize the processing several methods to distribute internal information are used. Among them, the use of multiple buffers allows process synchronizing optimizing the speed [6]. The most used methods in graphic processing are the double and the triple buffering [7]. There are triple buffering implementations but they are usually oriented to only one consumer or they are simulations of theoretical models [8].

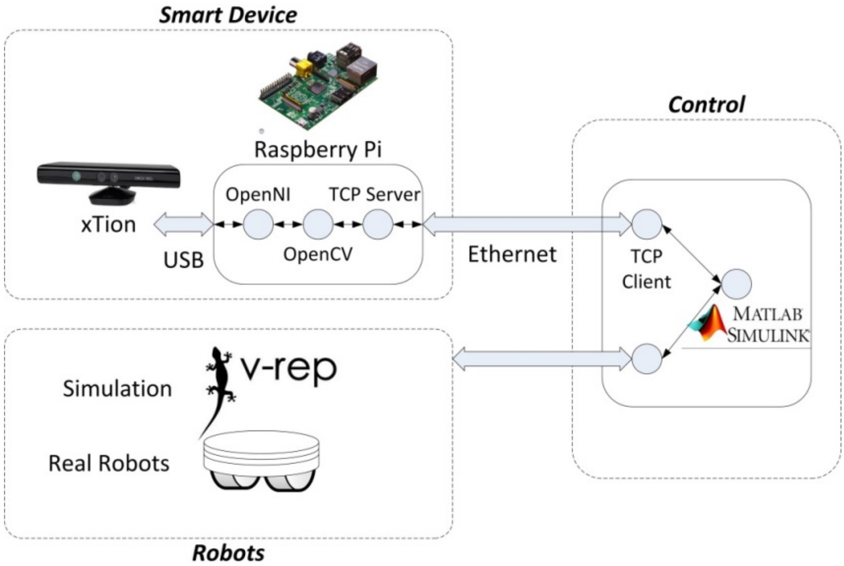


Fig. 2. System developed

In this article it is described a smart device which involves the use of a smart sensor xTion [8] with a synchronized triple buffer with no waiting times. The algorithm's goal is providing frames at the speed (in frames per second or fps) that the client has requested, with minimum delay possible and without the need of synchronizing the data sending. In the article results of delay times measuring and quality of image sending frequency are presented.

3 System Proposed Description

The implemented smart resource acquires images from an xTion sensor to process it with a Raspberry Pi [9] system which distributes it with an Ethernet connexion. These dispositive were chosen because their low weight and ease of programming to make prototypes that can be easily embedded as soon as the system efficiency and strength has been checked. Nowadays the smart resource is been implemented to improve the efficiency in robot navigation control based on the improvement of QoS and QoC [10].

In Figure 2 it is showed the details of the system in which is used the smart resource. The smart device is using the OpenNI library [11] to the xTion sensor connection. The image processing is done using the OpenCV library [12], implementing the connection and distribution with the Linux socket library. The control algorithm are implemented in a server with MatLab [13] which processes the images to the generation of different behaviours that goes from the generation of trajectories to the map creation. Finally it is used the robot simulation environment V-REP [14] with the goal of including the smart resource in real robots.

The robot navigation requires the most recent information, according that, the triple buffer design presented in this article is oriented to sending the most recent frame employing the minimal processing time possible.

4 Method

Because of the robot control system needs the most recent image; the designed triple buffer optimizes the transmitted image making sure that it is the last one without employing more memory than a conventional triple buffer. In this article only has been experimented with a single image processing. Is planned to realize a multiprocessing with multiple clients in future researching. The triple buffer is managed by two threads: the acquisition and process thread (APT), and the sending thread (ST).

The buffer's behaviours are as follows: a buffer is used for the image being acquired. Another contains the image being sent and the third buffer is an auxiliary buffer to avoid bottleneck. The triple buffer is managed by two threads: the acquisition and process thread (APT), and the sending thread (ST). The APT changes between the acquisition buffer and the auxiliary while the ST ends sending its buffer. The Swap function does the buffer's pointers exchange which corresponds. When the exchange is done, the APT sends a signal to ST to pick the last image. The APT behaviour is as follows.

```
I ←0
Adquisition() ← frame
while I < frame.MaxPixel do
   frame.Process()
end while
Swap()
Signal()
```

When the ST has sent an image it waits, using the wait() function, to be communicated by APT that the last image is available. As soon as the ST is unlocked by the APT from the wait function, the ST does the buffers pointer swap to send the last available image.

```
Wait()
Swap()
Send()
```

Because two threads are employed to decouple the acquisition from the distribution, we can find two scenarios: APT serves a frame faster than ST sends it and otherwise. In first case, the frame sequence is lost in exchange for gain immediacy in the information. In figure **3** the sending sequence is shown where it can be appreciated how the frame i+2 is lost because the frame i+3 is acquired before the frame i+1 ends being sent. In figure **4** the opposite case is shown, the ST is faster than the APT so no message in the sequence is lost.

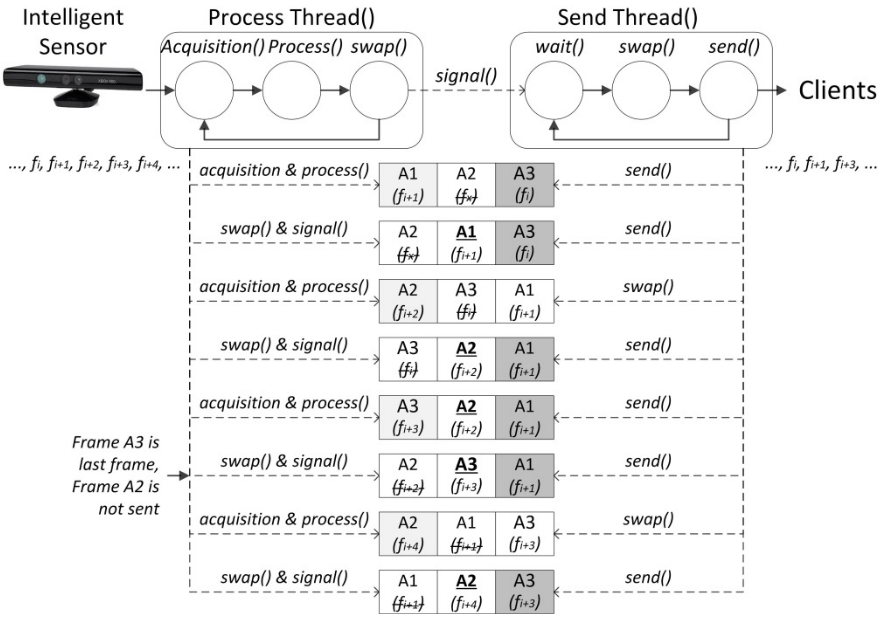


Fig. 3. Sequence of operation of triple buffer with APT faster than the ST

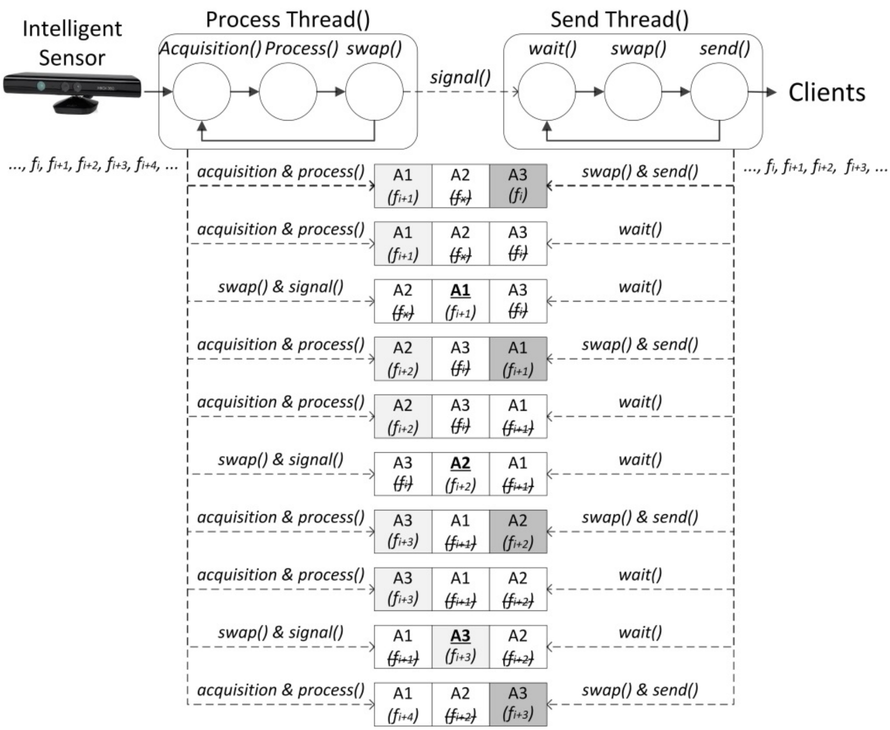


Fig. 4. Sequence of operation of triple buffer with ST faster than the APT

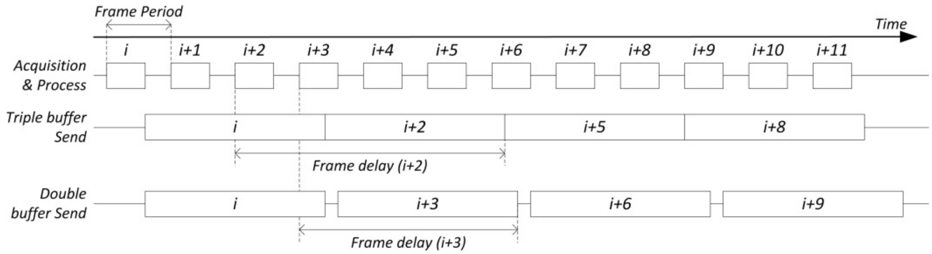


Fig. 5. Variables measured at the experiments

5 Experiments and Results

To check the correct operation of triple buffer implemented, it has been tested using a computer as smart resource to check highest level of efficiency, and a raspberry PI on the final system. Because the system must operate in a distributed closed network, servers and clients work with fixed IP addresses, thereby, the effect of messages noise on network is reduced. All tests were performed with loads of 1000 frames and 30 frames per second (fps) as sampling rate. Previously, it has been verified that the number of frames not affect the operation. Has been measured times with one, two and four clients, comparing the performance of a simple, double and triple buffering based server. The measured variables are the average of the all the frame delays, the frame delay is defined in the equation 1.

$$T_{frame\_delay} = t_{adq} + t_{wp} + t_p + t_{ws} + t_s \tag{1}$$

In equation 1, t_{adq} represents the time inverted to read a frame from the xTion, t_{wp} represents the time that the APT is waiting to obtain the frame and the t_p is the time used by the APT to process the frame, t_{ws} is the time that the ST waits for the new frame, if it is necessary, and t_s is the time inverted to send the frame to all clients. The results for the frame delay, expressed in microseconds, obtained are shown at table 1.

Table 1. Results of the experiments (in microseconds)

	1 client		2 clients		4 clients	
Variables	PC	RPI	PC	RPI	PC	RPI
Simple buffer	67350	3176925	66936	592390	66945	1293636
Double buffer	67350	203086	66936	334824	66945	640506
Triple buffer	66940	199647	67386	332900	66930	611752

The main difference between double and triple buffering is that the triple buffer always has a frame ready to send. However, the triple buffer adds more delay time. These results are predictable but it's necessary that the smart resource can provide to the clients in order to improve the system with quality of service policies (figure 5).

6 Conclusions

In this article a system of acquisition, processing and sending of image frames from an intelligent sensor to several clients interested in a processed image has been presented. To provide the most recent image, a triple buffer that ensures that constraint has been developed.

As can be seen in the table 1, when the image server is a powerful computer (PC) differences between the types of buffer are not significant. However, an embedded system provides less efficient results for a simple buffer, for a double buffering and triple buffering delay times is less in the case of double buffering when we increase the client number. So that, triple buffer offers the same updated information with an added immediacy and more efficiently than double.

In triple buffering there is always a frame available to be sent. So that it is not necessary to wait for the image acquisition to send a new frame. With this method, an intelligent sensor can provide the same performance as the double buffering method, but performing others tasks, as image processing, image segmentation or image resizing.

Currently, we are studying how the buffer used affects to robot navigation. Reactive behaviours require information as quickly as possible, so the triple buffer seems the most appropriate method. However, the immediacy of the double buffer, can improve certain non-critical behaviours as tracking paths, due to they provide latest information, but with less immediacy.

As future work, it is planned to adapt the triple buffer to a system where the obtained data from the intelligent sensor would have several distinct parallel processes, in example, locating free paths to the robot trajectory at the same time that interesting objects to surrounding maps generation are detected like corners, walls or door steps.

Acknowledgments. Coordinated project COBAMI: Mission-based Hierarchical Control. Education and Science Department, Spanish Government. CICYT: MICINN: DPI2011-28507-C02-01/02 and project "Real time distributed control systems" of the Support Program for Research and Development 2012 UPV (PAID-06-12).

References

1. Brignell, J.E.: The future of intelligent sensors: a problem of technology or ethics? Sensors and Actuators 56, 11–15 (1996)
2. Fernandes, J., Laranjeira, J., Novais, P., Marreiros, G., Neves, J.: A Context Aware Architecture to Support People with Partial Visual Impairments. In: Omatu, S., Neves, J., Rodriguez, J.M.C., Paz Santana, J.F., Gonzalez, S.R. (eds.) Distrib. Computing & Artificial Intelligence. AISC, vol. 217, pp. 333–340. Springer, Heidelberg (2013)
3. Chien-Hui, L., Kuan-Wei, L., Ting-Hua, C., Che-Chen, C., Charles, H.-P.: Fall Detection by a SVM-Based Cloud System with Motion Sensors. In: Yueh-Min, H., Han-Chieh, C., Der-Jiunn, D., James J (Jong Hyuk), P. (eds.) Advanced Technologies, Embedded and Multimedia for Human-centric Computing. LNEE, vol. 260, pp. 37–45. Springer, Heidelberg (2014)

4. Grzejszczak, T., Mikulski, M., Szkodny, T., Jędrasiak, K.: Gesture Based Robot Control. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2012. LNCS, vol. 7594, pp. 407–413. Springer, Heidelberg (2012)
5. Lee, C.S., Gyu, M.L., Woo, S.R.: Standardization and challenges of smart ubiquitous networks in ITU-T. IEEE Communications Magazine 51(10), 102–110 (2013)
6. Dias, D.M., Mukherjee, R., Sitaram, D., Tewari, R.: Buffering and Caching in Large-Scale Video Servers. In: Prac. of COMPCON (1995)
7. Tagami, Y., Watanabe, M., Yamaguchi, Y.: Development Environment of 3D Graphics Systems. Fujitsu Scientific & Technical Journal 49(1), 64–70 (2013)
8. Khan, S., Bailey, D., Gupta, G.: Simulation of Triple Buffer Scheme. In: Second International Conference on Computer and Electrical Engineering (2009)
9. ASUS: Xtion Pro Live, `http://www.asus.com`
10. Edwards, C.: Not-so-humble raspberry pi gets big ideas. Engineering & Technology 8(3), 30–33 (2013)
11. Poza-Luján, J.L., Posadas-Yagüe, J.L., Simó-Ten, J.E.: Quality of Control and Quality of Service in Mobile Robot Navigation. International Journal of Imaging and Robotics 8(1) (2014)
12. Norman Villaroman, N., Rowe, D., Swan, B.: Teaching natural user interaction using OpenNI and the Microsoft Kinect sensor. In: Proceedings of the 2011 Conference on Information Technology Education (SIGITE 2011), pp. 227–232. ACM, New York (2011)
13. Bradski, G., Kaehler, A.: Learning OpenCV: Computer vision with the OpenCV library. O'reilly (2008)
14. Ollero, A.: Intelligent mobile robot navigation. STAR, vol. 16. Springer, Heidelberg (2005)
15. Freese, M., Singh, S., Ozaki, F., Matsuhira, N.: Virtual Robot Experimentation Platform V-REP: A Versatile 3D Robot Simulator. In: Ando, N., Balakirsky, S., Hemker, T., Reggiani, M., von Stryk, O. (eds.) SIMPAR 2010. LNCS (LNAI), vol. 6472, pp. 51–62. Springer, Heidelberg (2010)

Mobile Access to Sensor Network: A Use Case on Wildfire Monitoring

Sergio Trilles, Óscar Belmonte, and Joaquín Huerta

INIT, Universitat Jaume I, Spain
{strilles,belfern,huerta}@uji.es

Abstract. These networks provide large volumes of data in many different formats, resolution and scales. The data are of different types and character: from meteorological conditions to air quality and the concentrations of pollutants due to human activity, such as transportation or other industry-related actions. The work in this paper aims at increasing the interoperability and improving accessibility of data provided by sensor networks. To address this problem Geographic Information System (GIS) services, such as the Sensor Observation Service (SOS), in conjunction with Representational State Transfer (RESTFul) architecture are used. A standard-based solution that increases interoperability is presented. It also allows for a better integration of data already published in different semi-structured formats in order to be used by various platforms (web or mobile).Finally, we apply this system in a use case as wildfire monitoring, so it offers determine the state of a wildfire.

Keywords: air quality sensors, meteorological sensors, heterogeneous sensor sources, data interoperability, RESTFul services, wildfire.

1 Introduction

Nowadays there are many types of sensors that measure a diverse variety of environmental physical values. Due to an increasing concern about climate change, natural disasters and global warming, there is an enhancing demand for these sensors to be deployed to monitor our environment towards a sustainable development of cities and an improvement of the quality of human life. We find, for instance, air quality sensors, deployed in urban, industrial and even rural environments. An interoperable access to these environmental data and the integration of these into analytics tools is crucial to extract useful information to assist in the management and reduction of problems that are caused, for example, by pollutants from industries or by traffic [1].

The goal of this project is to provide interoperable access to different data sources, using standard or agile interfaces enabling easy access. To do this, we use Open Geospatial Consortium (OGC) standards following INSPIRE architecture [2]. It should provide environmental data related to 34 themes, including transport networks, land cover and hydrography, INSPIRE provides important parts of the European contribution to a Global Earth Observation System of Systems (GEOSS). For it, we use standard

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_34,

formats such as the Sensor Observation Service (SOS). SOS provides a standardized web service interface which allows clients to access descriptions of associated sensors and their collected observations. We provide other types of interfaces capable of lightening the use of these data sources to be consumed on more restrictive devices such as mobile phones, although without compromising the interoperability standards. We have used an interface Representational State Transfer (RESTFul), following the principles of Internet of Things (IoT) [3] and related to different standards to define sensor data, such as Observations & Measurements (O&M) and Sensor Model Language (SensorML). Finally, we use our system for monitoring wildfire and detect which components are present when a wildfire is declared.

2 Background

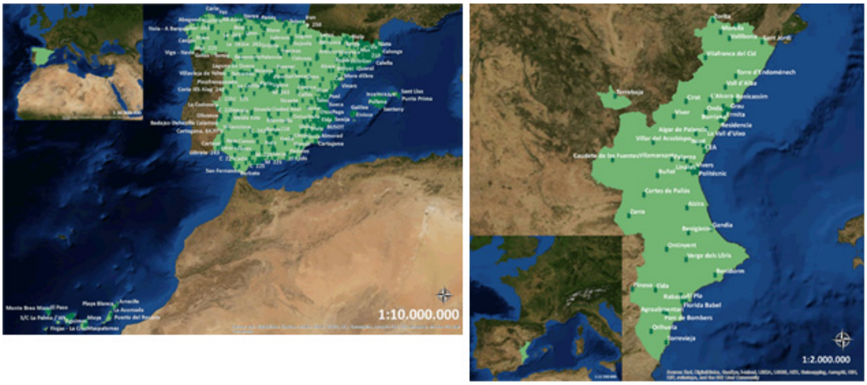
The efforts to standardize interoperability are directed by the OGC. OGC aims to define open standards, interoperable within the GIS and the World Wide Web. The OGC established the Sensor Web Enablement (SWE) as a set of specifications related to sensors, sensor data models, and sensor web services that will enable sensors to be accessible and controllable via the Web [4]. In our work we use three specifications, these are detailed below:

- O&M: standard models and XML Schema for encoding observations and measurements from a sensor, both archived and real-time.
- SensorML: standard models and XML Schema for describing sensors systems and processes; provides information needed for discovery of sensors, location of sensor observations, processing of low-level sensor observations, and listing of task able properties.
- SOS: standard web service interface for requesting, filtering, and retrieving observations and sensor system information.

We use RESTFul with our system. RESTFul [5] is an architectural style for building/creating distributed applications in a server. RESTFul builds on three concepts: representation, state and transfer. RESTFul is an architectural model that can be efficiently implemented as a combination of the Hypertext Transfer Protocol (HTTP) and TCP/IP. Uniform Resource Identifiers (URIs) are used to encode transaction states. With this URI, we can identify a sensor, if we follow the principles of the Internet of Things (IoT) and Web of Things (WoT), this sensor will be a "smart sensor" [6]. IoT describes a concept in which the world of real, physical things is integrated into the virtual world of bits and bytes. This term has first been used in a paper by David Brock in 200 [7]. The term WoT describes the evolution of the IoT [8].

3 Air Quality and Meteorological Data Components and the Study Area

We are working with two data sources, Meteoclimatic and air quality of Valencia network. Meteoclimatic is a huge network of weather stations in real-time and non-professional. The Meteoclimatic network is extended by: Iberian Peninsula, the two archipelagos, southern France and Africa near the Strait of Gibraltar. Currently, there are over 1,000 stations (figure. 1 (a)). Meteoclimatic measured meteorological components such as temperature, humidity, wind, pressure, precipitation or radiation. The refresh rate is irregular and varies from each station, but the most usual is every 10 minutes.



a) b)

Fig. 1. a) Meteoclimatic network. b) Air Quality stations in the Valencia network.

Air quality data source is the network of the air quality of the Valencia Community maintained by local government. It is a network of stations that measures the levels of various pollutants and meteorological elements in urban, rural and industrial areas, extending that control to the whole area of the Valencia Community. This network analyzes contaminants such as, sulfur dioxide (SO2), nitrogen monoxide (NO), nitrogen dioxide (NO2), nitrogen oxides (NOx), carbon monoxide (CO), ozone (O3), benzene (C6H6) and other hydrocarbons such as toluene and xylene; particulate matter with diameter less than 10 microns (PM10), 2.5 microns (PM2.5), and 1 micron (PM1). At present, the air quality network has 61 stations that are operational: 24 in Castelló, 24 in València and 13 in Alacant (figure 1 (b)). These data are published on a website that is validated and updated every hour. Each observation has a delay of 2-3 hours.

4 System Design

In the last years, information technology has moved from desktop and mainframe app to service oriented architectures and distributed computing. This trend applies to current SDIs which offer the possibility to access distributed, heterogeneous geospatial resources. In our system, follows the INSPIRE technical architecture, which contains three layers [2]: content layer, service layer and application layer. At the top layer reside the users and applications (clients). At the middle layer reside all services that provide the required functionality such as accessibility and processing of data.

Figure 2 shows the conceptual architecture of our system. In the left side, it is shown the conceptual view of system's architecture. In the content layer, we have the data, obtained from the Valencia air quality network. In addition we also have scientific models, with which we use to model and process the data. In the services layer, we have added four different services as with an extension of the functionality recommended by INSPIRE; download service, view service, publish service and processing service. Finally, in the application layer, we have applications and portals for displaying information.

On the right side of the figure, there is a more technological view of the proposed system. At the bottom, we can see the pre-processing module, where the system implements the preparation and integration of the information in different structures, such as databases. In this way, it is integrated different data sources in different databases. We have defined different scientific models, which we apply to process the data and derive required information. For example, viewing (heat maps), analysis (clustering), prediction or propagation. We deploy different services, where we publish data and processes based on standard interfaces, such as, SOS, Web Processing Service (WPS), Web Map Service (WMS). We offer other light-weighted interface, such as, RESTFul. In order to develop a scalable and extensible system we follow a pattern design [9]. We use the Façade pattern; it will be able to offer a single point of entry to all the services that we define.

The right side of the figure shows the interaction with the user via the client applications. In our first approach, and due to the increasing demand in mobile devices, we have developed a client application and we have provided our services with lighter interfaces for efficient communications.

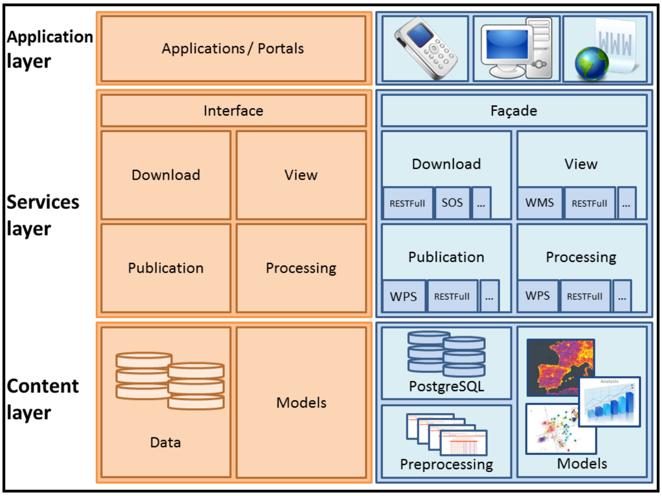


Fig. 2. A proposed System Architecture

5 Implementation

5.1 Services

We have designed a service that implements a RESTFul interface. RESTFul interface offers better access from any device, even from mobile phones, which they have inferior features. We have chosen a web service RESTFul to develop a system that follows the paradigm of IoT and more specifically Web of Things (WoT). WoT is the evolution of IoT [10], which uses web standards such as RESTFul for implementation. In the concept of the WoT, web servers are embedded into everyday objects, which they were turned into “smart things” [11]. The output format is JavaScript Object Notation (JSON). The pattern of the URI-Scheme is defined as:

```
http://<server>/<sensorId>/<method>
```

- server: is the entry point of the sensor platform. It provides a collection of sensors which are attached to the platform.
- sensorId: refers to an identifier for a specific sensor. When accessed this resource should list a collection of all available methods applicable for this sensor.
- method: stands for the method of interaction with the sensor.

For example, “data” method lists observations for the sensorID indicated. The method allows queries between two dates and different components, with parameters "date1", "date2" and "comp". In this case it returned a list.The O&M model is used for representing and exchanging results.

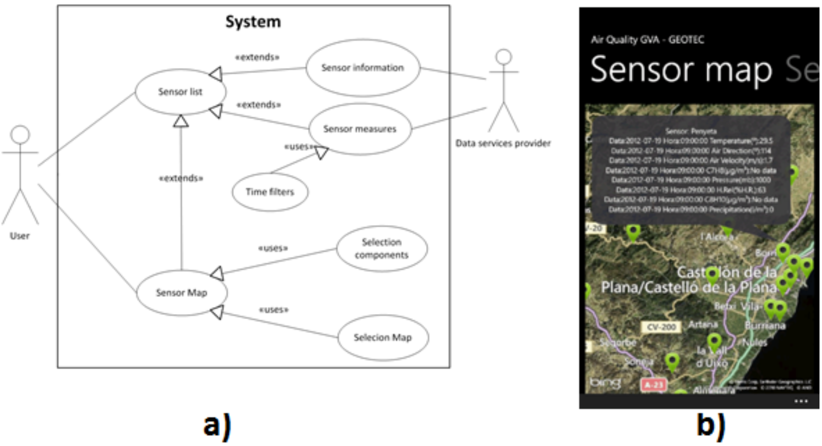


Fig. 3. a) Use case client application. b) Client application (WP7).

5.2 Client Application

For our first prototype, we create an application for the Windows Phone (figure 3 (b)). Figure 3 (a) shows the different use cases of our mobile application. The application has two case uses. The first use is able to list all the sensors and users can consult sensor information, also current and historical measurements. There are time filters for a better search. Furthermore, the application offers other users to locate the sensors on the map and display the latest values for each sensor, where users can select the measured physical values to be displayed and select the type of map.

6 A Use Case: Wildfire Monitoring

During the month of June 2012, two of the fires reported in the Valencia region were especially critical due to their size and aggressiveness. The first one was located in Cortes de Pallars (figure 4). The fire started on June 28th. The second fire was located in Andilla (figure 4) and started on June 29th. The two fires were controlled on July 6th.

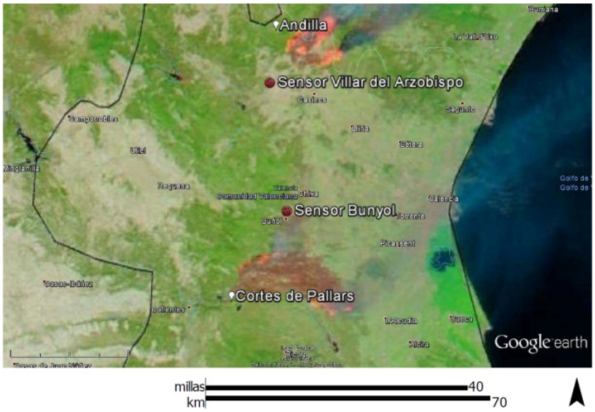


Fig. 4. Fires in the region of Valencia in the summer of 2012 and sensors used

We have taken the two sensors nearest to the affected regions to be able to measure the greater number of contaminants listed above. For the fire in Cortes de Pallars, we took a sensor located in a neighboring town called Bunyol, and visualized the observations for June 27th and 29th and July 7th at 6 a.m. For the Andilla wildfire, we selected a sensor in a neighboring town called Villar del Arzobispo, located a few kilometers away from the burned region. For this wildfire, we conducted observations on the 27th and 30th of June and on July 7th, also at 6 in the morning.

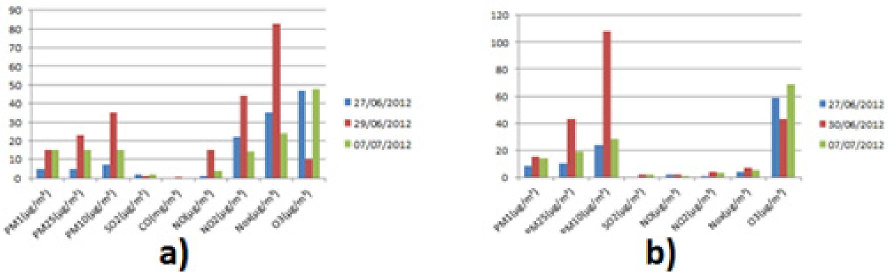


Fig. 5. a) Bunyol observations. b) Villar del Arzobizpo observations.

According to [13], one of the main indications of fire is the increase of the following pollutants: carbon monoxide (CO), oxides of nitrogen (NO, NO2 and NOx) and particles in the air (PMx). In the following graphs (figure 5) we see on June 29th and 30th a significantly increase of the levels of particles (PM10, PM2.5, PM1). In the case of particles of 10 microns and 2.5 microns, we see on July 7th has become normal, but to 1 micron, due to less weight, are still in suspension.

Figure 6 shows different heat maps for PM1 (figure 6(a)), PM2.5 (figure 6(b)), PM10 (figure 6(c)) and NOx (figure 6(d)). As we have seen the levels of the four components increased in the fire area. Heat maps of particles (figure 6(a)(b)(c)) show as particles moved westwards caused by wind. On the other hand, NOx heat map (figure 6(d)) shows an increase in that area of this component.

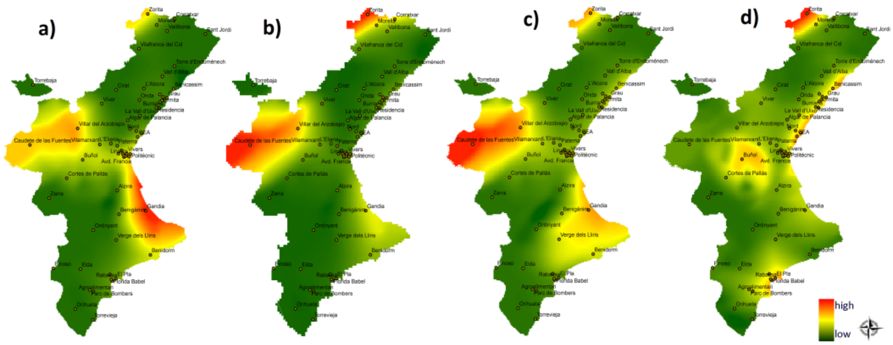


Fig. 6. a) Heat Map PM1 (April 27th), b) Heat Map PM10 (April 27th), c) Heat Map PM2,5 (April 27th), d) Heat Map Nox (April 27th)

We observe a significant increase in oxides of nitrogen in the two sensors during the wildfires, especially in Bunyol's sensor. We see that when the fire has been controlled, the values of the oxides of nitrogen are similar to the first day, as well as the Carbon Monoxide (CO) which is only available in Bunyol's sensor and that increased by one tenth during the wildfire. We have demonstrated that, when two fires were declared, in the two closest sensors, the emission of particles in the air and oxides of nitrogen increased and when fires were controlled the indexes returned to their initial state. The high level of particles and oxides of nitrogen is registered when fires occur, according to [13], so that our system is able to monitor and detect a fire next to any sensor. Furthermore, we can determine and define the hours in which wildfires is active or inactive.

7 Conclusions and Future Work

In this work we have proposed a service-oriented architecture to implement an application, which orchestrates a workflow for the management of a sensor network. We have proposed a system that aims at the integration of data to offer them in a structured and interoperable way. In addition, the service layer in our application is enhanced with a module that implements the design pattern façade, and acts as a "middleware" between the client and the services to increase interoperability by implementing several interfaces and help users to get information easier.

We used different interfaces (SOS, RESTFul ...), thus improving interoperability. RESTFul service is able to represent each sensor by a URI (WoT). But, in addition, we are able to maintain the standard GIS for the sensorial data (SensorML and O&M). We have implemented a mobile client for users to interact with the offered functionality and to be able to access and visualize data from the selected sensor network. To demonstrate the functionality and feasibility of our system we have described an experiment in which we can detect wildfires by the measured observations.

Acknowledgments. This work has been supported by GVA (Grant ACIF/2012/112) and the CITYBENCH project, nº 121_PR3_10_0503, funded by the ESPON program of the EU.

References

1. Gualtieri, G., Tartaglia, M.: Predicting urban traffic air pollution: A gis framework. Transportation Research Part D: Transport and Environment 3(5), 329–336 (1998)
2. INSPIRE EU Directive (2007) Directive 2007/2/ec of the European parliament and of the council of 14 March (2007)
3. Uckelmann, S., Harrison, M., Michahelles, F.: An Architectural Approach Towards the Future Internet of Things. In: Uckelmann, D., Harrison, M., Michahelles, F. (eds.) Architecting the Internet of Things. Springer, Heidelberg (2011)
4. Sheth, A., Henson, C., Sahoo, S.: Semantic Sensor Web. IEEE Internet Computing 12(4), 78–83 (2008)
5. Pautasso, C., Wilde, E.: RESTful web services: principles, patterns, emerging technologies. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 1359–1360. ACM, New York (2010)
6. Guinard, D., Trifa, V., Wilde, E.: A resource oriented architecture for the Web of Things. Internet of Things 1, 8 (2010)
7. Brock, L.: A naming Scheme for Physical Objects. The Electronic Product Code, EPC (2001, 2010)
8. Bröring, A., Remke, A., Lasnia, D.: SenseBox - A Generic Sensor Platform for the Web of Things. In: Puiatti, A., Gu, T. (eds.) MobiQuitous 2011. LNICST, vol. 104, pp. 186–196. Springer, Heidelberg (2012)
9. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Dessign Patterns: Elements of Reusable Object-Oriented Software. Addison Wesley (1995)
10. Duquennoy, S., Grimaud, G., Vandewalle, J.: The Web of Things: Interconnecting Devices with High Usability and Performance. In: Proc. of the International Conference on Embedded Software and Systems, ICESS 2009, Zhejiang, China, pp. 323–330 (2009)
11. Guinard, D., Trifa, V., Mattern, V., Wilde, E.: From the Internet of Things to the Web of Things: Resource-oriented Architecture and Best Practices. Architecting the Internet of Things 97, 129 (2011)
12. Crockford, D.: The application/json Media Type for JavaScript Object Notation (JSON). RFC 4627 (2011)
13. Mejias, E., Manso, R.: Estimación de las emisiones de gases, producidos por incendios detectados por satélite en la Ciénaga de Zapata. Cuba. XIV SimposioInternacionalSelper (2010)

Building Scalable View Module of Object-Oriented Database

Haeng-Kon Kim[1] and Hyun Yeo[2]

[1] School of Information Technology, Catholic University of Daegu, Kyungbuk, 712-702, Korea
hangkon@cu.ac.kr
[2] Dept. of Information and Communication Engineering, Sunchon National University, Suncheon, Jeollanam-do, Republic of Korea
yhyun@sunchon.ac.kr

Abstract. Many researchers and developers have studied object-oriented relational database management systems (ORDBMS) in the past ten or so years, few have published their results that reveal the inside workings of an ORDBMS. Leading database software companies integrated object-oriented features in their DBMS products only recently. These companies do not make the technical core of their products public. Most academic researchers, on the other hand, have worked on ideas, methodologies and analysis of ORDBMS, but few have shown the database engine of a working system. This paper presents a propotype ORDBMS engine that supports objects in databases, including user-defined types, inheritance, and polymorphous method invocaiton. Although a prototype, it is implemented in Java, fully functional and can be extended should additional modules be added in the future. The system is composed of three major components: the query-command module, the view module, and the database module.

Keywords: Object-Oriented Relational Database Management, User Friendly Interface, Inheritance, View Module, Searching Engine, Self-diagnosing, Agent, Ubiquitous computing, CBE (Common Base Event), Crop Production Agent Systems.

1 Introduction

Object-oriented database, as one of the latest developments in the database area, has been widely studied in the past decade [2-16]. One of the primary methods used today for object-oriented databases is the object relational model, which is an extension of the relational model. In object relational model, attributes of objects can be of user-defined types such as chunk/stream, set, nested/reference attributes, as well as the primitive types. In addition, new data types and tables can be defined and created by inheriting from existing ones that also supports polymorphous method invocation. As an example, let's say that we want to create audio and image associated with person objects in a human resource database. Using the object relational model, the person table can be defined as

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_35, © Springer International Publishing Switzerland 2014

```
CREATE TYPE media FROM chunk (length: integer);
  CREATE TYPE audio FROM media;
  CREATE METHOD play FROM
  { /home/media/audio/play.class } INTO audio;
  CREATE TYPE image FROM media;
  CREATE METHOD display FROM
  { /home/media/image/display.class } INTO image;
CREATE TABLE person (ssn: text, name: text, salary: num-
ber, face:image, voice: audio);
```

In this design, chuck is a primitive type (a stream of bytes) and the play/display methods are Java bytecode in the specified directory. Note that both the types audio and image inherit from media type, each having its own method for playing the audio or displaying the image. Now, we can query the database like this:

```
SELECT name, face.display( ) FROM person
WHERE salary < 30000 and voice.length < 5K;
```

This example illustrates some of the important aspects of object relational databases, such as creation of new types, inheritance, and method calls.

In this paper, we investigate problems in the design of database management systems in our research endeavor, object-oriented database management systems (ORDBMS) in particular, and present a database engine that support OO databases. Object-relational database management system is a database system that extends the capability of relational database management system (RDBMS) to provide richer data type system and object orientation. The extension attempts to preserve the relational foundation of RDBMS while providing more modeling power. During the past several years we had built some prototypes using C++ as the implementation language but it appeared that some of the features that were supposed to be supported by ORDBMS were quite difficult to realize using C++. Problems particularly troublesome were the handling of pointers in nested and reference attributes as well as methods applied to multimedia objects in the database. In early 1998, we switched to Java [1] in our implementation. It turns out that Java and ORDBMS fit extremely well. Java provides straightforward support for multimedia object handling, secure object referencing, and elegant inheritance mechanism. We have designed an ORDBMS engine and implement it in Java. A simple user interface was also developed for testing the database engine.

The architecture of the ORDBMS consists of three major components:

- Database module: providing the main concepts and related object classes needed to deal with the manipulation of database schema and low-level data storing and retrieval.
- Query-command module: processing user’s query and other actions, providing mechanism to support commands at the language level (like “Statement” and ”ExecuteQuery” in JDBC).

- View module: providing the main concepts and related object classes used to support the mechanism of attribute referencing, expression composition, projection, filtering, and joining.

The overall architecture is shown in Fig. 1.

We will describe each of the three components in the next sections, along with the Java API class hierarchies. In this paper, class names are in *italic* and start with capital letters.

2 Database Module

This is the lowest level component of the ORDBMS, i.e. it is the module directly interacts with the database. It processes the database schema and stores the metadata in internal tables, which are also known as "catalogs". It also manipulates the storage and retrieval of data in the database. The database module provides the functionality to support various kinds of data types in a uniform way. These data types include reference, set, nested field, chunk (raw stream of bytes for image, audio and video data), and method, in addition to the built-in types such as integer, float, boolean, text, and chunk. The major object classes in this part of the ORDBMS are: *Database, DataType, MethodInfo, TableDef, FieldDef, Record, Attribute, RecordSet, Table, TableRecord*, and *Field*. These Java classes (and some classes in other modules) are organized in several groups whose hierarchies are as shown in Fig. 2 below. These class hierarchies are only a part of the architecture of our system.

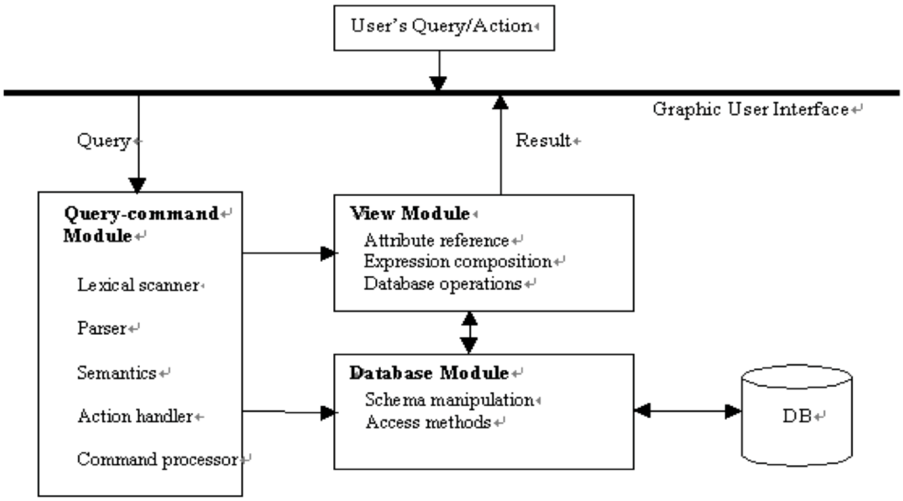


Fig. 1. Overall architecture of the ORDBMS

In the next several subsections, we discuss the major classes of the Database module.

2.1 The *Database* Class

This is the core class that represents the logical concept of a database system. The *Database* object handles the schema; persistently keeps, manipulates, and provides information about the database, and controls the internal works of the system such as file access control. For example, when the user adds a new type to a database, the information of the new type is stored and all of its references are resolved within the *Database* object. The main function of the methods in this class includes:

- Create a new or open an existing database. If an existing database is opened, the *Database* object loads the schema information (definitions of tables, fields, and methods).
- Add and drop table/field/method definitions.
- Close a database.
- Read and write database files.

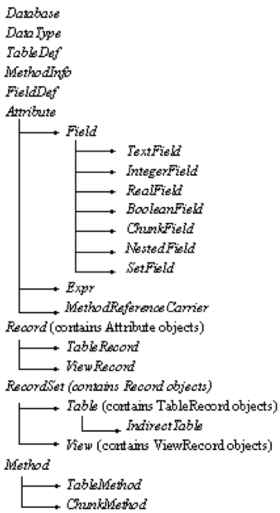


Fig. 2. Class hierarchy

The *Database* object uses *TebleDef* and *FieldDef* objects to maintain the schema information for object values (expressions, constants, operators, etc.), classes for records of database tables, and classes for method attributes. Because a user-defined type may describe objects with multiple attributes, it is treated the same way as tables in the database.

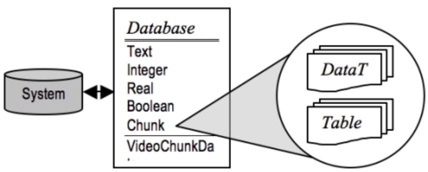


Fig. 3. DataType and TableDef objects inside Database

2.2 The *DataType* Class

In our system, all data types (built-in or user-defined) are represented using DataType objects. A DataType object contains a type name and a reference to the parent type, which may be null. Treating data types uniformly using the same discipline makes the design clean and more manageable.

When a *Database* object is created, it is initially empty. Once the initialization process is done, a set of *DataType* objects are created automatically within the *Database* object to present the built-in data types such as integer, real, text, boolean, chunk, etc. The parent type of each of the built-in types is null. Later, when a user-defined type is created, it is processed and maintained by the *Database* object. There are two kinds of user-defined data types. One is a stream of chunk type. A *DataType* object can directly represent this kind of type because it does not contain more information than that in a *DataType* object. The other kind of data type is structured type, such as a person or an employee. It should be clear that a structured type is simply a table definition in a database. An elegant way to support such structured types is to derive a new class (*TableDef*) from the *DataType* class.

2.3 Table Definition and Tables

Whenever the user defines a type, he/she is defining a table structure. A *TableDef* object contains more information than its parent *DataType*. It represents the definition of a table, which includes the descriptions of all the fields (attributes) of the table. The field definitions are represented by *FieldDef* objects, which will be described in the next section. Additional information stored in a *TableDef* object includes the location on the disk to store the data and a set of field definitions. Conceptually, the user always define a type by creating a *TableDef* object and added it to the *Database* object, which maintains the type like any other *TableDef* objects. Persistence of these types is also guaranteed by *Database*. Fig. 3 shows the *DataType* and *TableDef* objects within the *Database*.

Once a *TableDef* object and the relevant FieldDef objects are created and added to *Database*, actual data can be stored in a Table object according to the definition of the table. A *Table* object acts as an agent between the user and the data storage, as shown in Fig. 2. It handles the overall operations of a table, such as adding or deleting a record. It also controls the movement of the current record. Because the complexity of the kinds of records in tables, storing and retrieval of various kinds of records should be left to other classes specifically designed to handle records and their fields.

2.4 Field Definition and Fields

When we create a table definition, we must define its field definitions first. A *FieldDef* object can be created for this purpose that can be easily added to the *TableDef* object. The important information of a field definition includes the name of the field, its type (reference to a *DataType* object), size, and kind. In our system, there are three kinds of fields: a simple type, a nested or a set type. A nested field is very similar to having a structure variable inside another structure variable in C++ or Java. When a nested field is accessed, we go though each individual element in the field. This will lead to accessing to another nested field. This process continues (recursively) until all elements in the "current working field" are of simple types. For set fields, it is like a table nested inside another table, because what we obtain for accessing a set field is a table (set of records). Fig. 4 shows an example of nested and set fields.

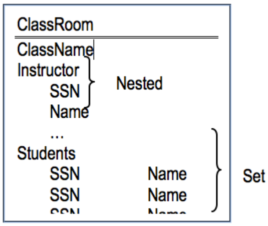


Fig. 4. An example of nested and set fields

To keep track of the data type of the fields is the task of *FieldDef* objects. All built-in types and user-defined stream/chunk types are simple types; all other user-defined types are either nested or set fields.

As stated in the previous section, a *Table* object can be created for storing and retrieval of data after *TableDef* and relevant *FieldDef* objects are added to the schema. The *Table* object only handles the overall operations of a table; it leaves the specific operations to the records in the table to the classes *TableRecord* and *Field*. A *TableRecord* object is given the information about the type and size of the record, and performs the required action to it. Also, it provides a pathway to invoke methods associated with the record.

When a *TableRecord* object performs operations to a record, it also relies on another object (the *Field* object) to deal with individual fields. The *Field* class defines the interface for the *TableRecord* object working with the fields in the table. For example, when a *TableRecord* reads data from storage and brings it to memory, it will repeatedly call the load() method of each *Field* object to retrieve and interpret the data.

Because the fields may be of many different types, we choose to make the *Field* class an interface for *TableRecord* to interact with the fields. The *TableRecord* object does not need to care about the field it is dealing with because all fields have the same interface. For the various kinds of types, we create a set of field derivatives (*TextField*, *ChunkField*, *nestedField*, *SetField*, etc., as shown in Fig. 2) that are derived from the *Field* class. When a new data type is created for a field, the *Field* and *TableRecord* classes need no change.

The *ChunkField* class represents a chunk of data, normally BLOB (Binary large Object) or CLOB (Character large Object), as found in many commercial DBMS. The data of *ChunkField* is variant in size and generally very large. As do many commercial DBMS, we use external files to store the filed data. Each such file is given a unique name by combining file name with the table name. It is also given a unique ID that is stored in the record of the table.

The *NestedField* class represents a field of a structured data type that contains other types. Because the structured data type is also a *TableDef*, the *NestedField* object maintains an internal *Table* object for the database table that is automatically opened for access when the nested field is being accessed. Whenever the value of the table is accessed, the record (*TableRecord* object) of the internal table is returned. From then on, the DBMS programmer can go further to the deeper fields in the same manner. A nested field maintains a pointer pointing to a record by using the OID (Object Identifier) of the record. The record ID maps to the location of the record in the data storage. Fig. 5 illustrates how *NestedField* works with OID.

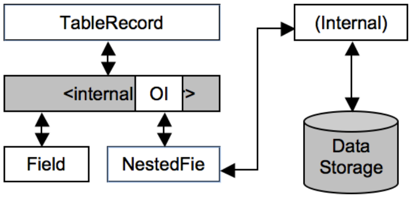


Fig. 5. Access to a nested field using OID

The *SetField* class is designed to handle a field that may contain a table. That is, the value of this field is a table (set of records). For example, we may have a table called Student that contains all students. Then, we can create a table called ClassRoom to keep track of classrooms. This table has a set field called "students" that contains the set of students who attend the class in the particular classroom. This subtable obtained from this field may not contain all the students in the Student table. Moreover, the sub-tables may have duplicates because a student may attend multiple classes. For this reason, we maintain a list of pointers pointing to the original data table. The list of pointers is by itself also a table. So, we create another class called *IndirectTable* that is derived from the *Table* class. An *IndirectTable* object is treated just like any other *Table* object. The only thing special about *IndirectTable* is that it handles pointers.

2.5 Method Definition and Method

ODBMS allows users to create and associate methods with user-defined data types. When a method is created for a type, a method_def object is created, which contains the name of the method, its return type and kind (nested, set, or simple), and the method code itself in the form of a Method object. It is then added to a DataType or a TableDef object the method belongs to. There are two kinds of methods: table method and chunk method. A table method is a method that works on the record basis, and is always associated with a TableDef. When such a method is called, there is always a current table record ready to be accessed for the method. A chunk method works on a byte stream field. It is associated with a user-defined stream/chunk data type. This kind of methods interprets the byte stream as the user defines, mostly for images, audio, or video streams.

Two classes are provided for the two kinds of methods: *Tablemethod* and *ChunkMethod*. They are derived from the *Method* class. The *Method* class, which is an abstract class, defines only the basic interface for a method definition implemented by the DBMS programmer. *TableMethod* class defines more specific interface needed to work on the record basis, whereas *ChunkMethod* defines an interface for working with byte stream. With this design, a method_def object can contain (refer to) a method by having only a reference to a *Method* object, which is a general form of both *TableMethod* and *ChunkMethod*.

The DBMS programmer can create a method by creating a Java class that inherits from with *TableMethod* or *Chunkmethod*, but not from the *Method* class (it is abstract). When the Java class file is created, the DBMS programmer can ask the *db_scheme* object to associate it with a particular table or a data type. Instantiation of an object for such a method (so that the method can be called) is done through Java's dynamic class loading.

3 View Module

The second layer of the ORDBMS is the View module. It provides all the necessary classes so that the ORDBMS programmer can create objects corresponding to the various components in the SQL query issued by the user and constructs an execution plan.

As we have known in relational databases, a view is a result of an SQL query, which is also a table. But may DBMS do not store an actual table for the view; rather, it stores the SQL query itself. When we activate a view, the SQL query is fetched and executed and the result is returned. In this sense, a view is just an SQL query. The data source, from which the data is retrieved for an SQL query, can be either tables (data are actually stored) or other SQL queries (hence nested views).

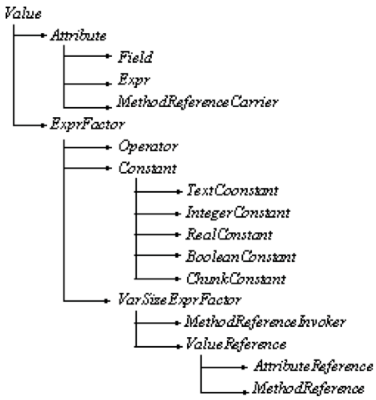


Fig. 6. The *Value* class hierarchy for values in SQL queries

The View module contains classes that handle "values" in an SQL query, including expressions (that in turn contains constants, operators, attribute references, methods, etc. The class hierarchy is shown below in Fig. 6.

3.1 The *View* Class

The *View* class is designed to represent an SQL statement, which in general contains three major parts: an expression list (or field list), a data source list, and a condition. These three parts correspond to the SELECT, FROM, and WHERE clauses of an SQL statement. The condition in the WHERE clause is just an expression but it must be a boolean.

A *View* object does not have its own data; rather, it relies on the data source that may be *Tables* or other *Views*. It creates a temporary table containing the positions (or OIDs) of the records in the *Table/View* data source that satisfy the condition. These positions are generated by applying the condition to each record while the *View* object iterates through the record set. Once the temporary table is generated, the *View* object just uses it as the underlying data source. Each *View* object maintains an internal "current record pointer" that points to a row in the temporary table. This row is by itself another pointer. Hence nested views can be easily handled as a pointer chain. The lowest level of View has pointers pointing to the actual records in the data source. This implementation of *View* using pointer chain is called "record reflection" mechanism that provides some benefits. First, there is no need for large storage space to generate a *View*. Second, it is fast to access the actual data because only positions belong to the *View* are kept. Third, any updates made to the underlying record sets are reflected through the *View* immediately. For example, if a underlying record is deleted, we simply break the reflection chain by null the pointers in the View's temporary table, as shown in Fig. 7.

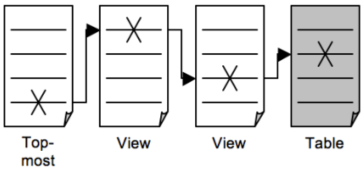


Fig. 7. Deletion of an underlying record in a View via record reflection

3.2 The *ViewRecord* and *Expr* Classes

ViewRecord is designed pretty much similar to *TableRecord*, they both represent records in a table. The difference is that *ViewRecord* objects do not provide data storing capability so that view is always read-only because the result of a view can come from many places (from expressions, join of tables, other views, etc.). The main function of *ViewRecord* is to manage the data obtained from "somewhere" to yield results of the expressions in a list the view is maintaining. In fact, the implementation of

View has an internal *Table* object to work with the temporary table, whereas the *ViewRecord* does not deal with the temporary table at all. The *ViewRecord* is used primarily for simplifying the expression validation process to be consistent with the process handled by *TableRecord*.

An expression can be anything that yields to a value. It can be a constant, a field, a method call that returns a result, or a combination of these. The classes in the Value hierarchy (see Fig. 6) provide all the necessary methods to evaluate an expression in a uniform way. We will discuss several of these classes that are particularly important for the object-oriented DBMS, including AttributeReference, *MethodReference* and *MethodReferenceInvoker*.

3.3 The AttributeReference Class

Recall that fields in an expression may come from different data sources. Each of the data sources has its own internal structure. In order to refer to fields in a uniform way, we use references. One of the commonly used fields is attribute that is represented by *AttributeReference*. Because the object referenced may be some other *Attribute Reference* or *MethodReference*, the getAttribute method of this class recursively calls itself until an *Attribute* object is reached. From the user's point of view, there is no need to go through the reference chain to get the attribute; rather, the user just calls getAttribute and get the target attribute. Using reference also eliminates the need to distinguish field and expression that are encapsulated in the *AttributeReference* object.

3.4 The MethodReference and Method Reference Invoker Classes

MethodReference works in the same ways as *AttributeReference* in the sense that it can also go to several levels due to, for example, inheritance. It refers to a record (Callable record) of which the method is called:

```
public method MethodReference extends Valuereference {
.....
  public Object call(Object[] args) throws Exception {
    try {
      Callable caller = di-
rect ?((RecordSet)source).getRecord():(Callable)((ExprFac
tor)source).getValue();
      return caller.call(MethodName, args);
    } ....
  }
}
```

To certain degree, it is like Class.getMethod() in Java, which gives you a reference to a method. When actually calling a method (may or may not have parameters), we use a *MethodReferenceInvoker* object that encapsulates a *MethodReference* object with a set of parameters (maybe an empty set):

```
public method MethodReferenceInvoker extends Exprfactor {
  private MethodReference mref;
  private Object[] realArgs;
  ......
  public Object getValue() throw Exception {
    return mref.call(getArgs());
  }
  ......
}
```

The following figure shows the relationships of the classes related to the concept of reference.

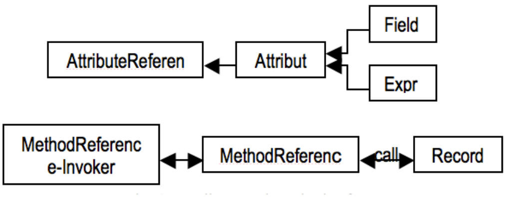


Fig. 8. Attribute and method references

4 Query-Command Module

The highest layer of the database engine is the Query-Command layer that processes SQL-like commands. Three kinds of user input are allowed:

- DDL (Data Definition Language) commands: These commands involve database tables at the schema level. The classes that implement these commands return nothing but will throw an exception if the desired operation fails. These commands are the CERATE command that defines a schema of a table and the DROP command that removes a table from the database.
- DML (Data Manipulation Language) commands: These commands are for accessing the database, including SELECT, INSERT, UPDATE DELETE, etc. Nested queries are supported by the mechanism of view.
- System commands: The user can use these commands to find out the metadata about the database tables. They are the LIST command that displays the types and tables in the database, and the DISPLAY command that displays the internal catalog information about the schema.

This module supports a modified version of the SQL syntax using an LL(1) parser, which is driven by a set of grammar rules defining the syntax. This part of the ORDBMS is responsible for processing user's SQL-like queries and commands. We designed a *CmdProc* class to be a center place where all semantic actions for handling an SQL command are triggered, as shown in Fig. 9.

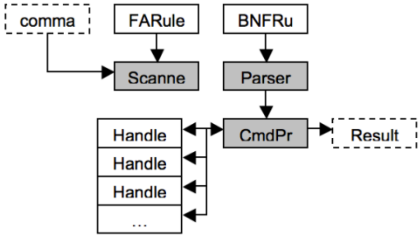


Fig. 9. The architecture of SQL processing

The SQL-like syntax includes all the CREATE, DROP, SELECT, UPDATE, DELETE, etc. statements. In addition, it uses the traditional dot-notation to access or refer to fields and methods of tables. Here are some examples:

```
supervisor.address.city
emp.IncreaseSalary(0.05)
student.getDepartment().chair.name
```

The ODBMS allows a nested field to be treated as a whole. For example, it can be assigned to the field of another object at once. Because we use internal pointers for nested fields, this task is quite straightforward. We use the reserved word ref to be the reference operator in the SQL statement. For example, assume that we have a table called Order that has a field called cust that is of the type Customer. We can write the following SQL statement to insert into the Order table a record that has order_no 0002 by the customer who has placed an order of order_no 0001:

```
INSERT INTO Order(order_no, cust)
VALUES ('0002', ref(SELECT cust FROM Customer WHERE or-
der_no = '0001');
```

Moreover, the ref operator can also be used to reference to an object or a record of a table:

```
UPDATE Order
SET cust = ref(SELECT * FROM Customer WHERE ssn = '123-
45-6789')
WHERE order_no = '0002';
```

The only requirement of the ref operator is that the sub-query to must yield one record.

Nested tables are handled in a similar way, except that it uses a the operator rather than ref operator, as shown in the following example:

```
SELECT *FROM the(SELECT students FROM ClassRoom WHERE in-
structor = 'Dr. Berry');
```

There are some other syntax details for the testing SQL-like query language.

The *CmdProc* class also provides methods to shield the details of processing an SQL statement so that the DBMS programmer can directly pass the SQL statement as a string parameter to the method to carry out the execution, very much like the ExecuteSQLStatement method in Java JDBC.

5 Experimental Examples

To show how the Java ODBMS API works, we creates a "shell" ODBMS with a simple GUI. The ODBMS and all the Java classes in the API are written for JVM 1.2 or JVM 1.1.7 with JFC support. This shell ODBMS simply receives an SQL command from the user and passes it on to the command processor. The result (if any) of the command is then displayed. In the following Fig. 10, we show a few snapshots to illustrate user-defined types, table inheritance, and polymorphous method call. The explanations are given along with the snapshots.

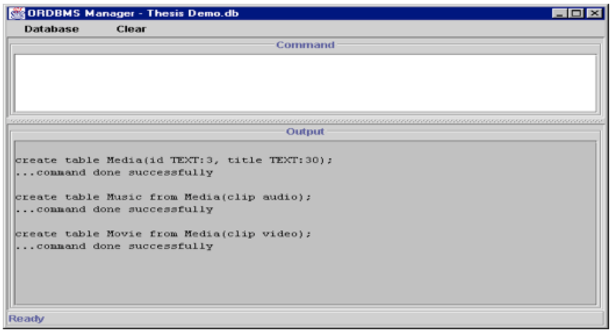


Fig. 10. Three tables: `Media`, `Music`, and `Video` are created. `Media` is the base class of the other two. The derived tables just add only additional field(s) they need.

6 Summary

In this paper, we presented a prototype of an ORDBMS database engine that is implemented in Java. The design is a 3-layer architecture composed of the *Database*, *View*, and *Query-Command* modules. The database engine is totally self-contained, relying on no other tools except the JDK. An SQL-like query language is formally defined and the handling of the data in the database tables is at the physical storage level. The *Database* module handles the schema and access methods of the data, as well as user-defined types and methods. The *View* module is responsible for the analysis and evaluation of expressions/fields/attributes/methods in the user's query. The *Query-Command* module parses and processes user's queries.

The database engine is for object-oriented databases and hence supports objects (chunk, structured, set, ref, etc.) and methods as attributes in a table. It also supports inheritance and polymorphism. The prototype contains over 200 Java classes and interfaces with over 1,000 methods. A media database is used to test the ORDBMS and it works as expected.

We did not discuss much in detail about the implementation of polymorphism in method calls, which is basically a chain of references with OIDs associated with the objects along the chain.

We are currently study some other aspects of ORDBMS, including object indexing and optimization of query execution.

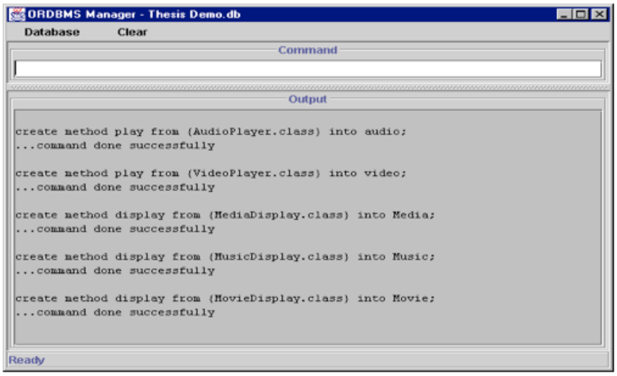


Fig. 11. Methods are associated with data types and tables. In this screen, method `play()` is created for `audio` and `video` data types that were created before (not shown), and method `display()` is created for `Media`, `Music`, and `Movie` table.

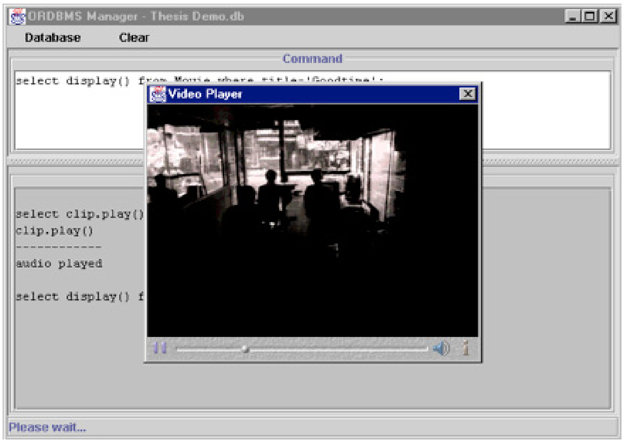


Fig. 12. The method `play()` is called from the field `clip` of `Music` table. And then, the method `display()` is also called on `Movie`. The results are both having music and movie displayed on the screen.

Acknowledgments. "This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the CITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-2008) supervised by the NIPA (National IT Industry Promotion Agency)".

"This research was also supported by the International Research & Development Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning(Grant number: K 2012057499)".

References

1. Corchado, J.M., De Paz, J.F., Rodríguez, S., Bajo, J.: Model of experts for decision support in the diagnosis of leukemia patients. Artificial Intelligence in Medicine 46(3), 179–200 (2009)
2. De Paz, J.F., Bajo, J., López, V.F., Corchado, J.M.: Biomedic Organizations: An intelligent dynamicarchitecture for KDD. Information Sciences 224, 49–61 (2013)
3. Rodríguez, S., de Paz, Y., Bajo, J., Corchado, J.M.: Social-based planning model for multiagent systems. Expert Systems with Applications 38(10), 13005–13023 (2011)
4. Bajo, J., De Paz, J.F., Rodríguez, S., González, A.: Multi-agent system to monitor oceanic environments. Integrated Computer-Aided Engineering 17(2), 131–144 (2010)
5. De Paz, J.F., Rodríguez, S., Bajo, J., Corchado, J.M.: Mathematical model for dynamic case-based planning. International Journal of Computer Mathematics 86(10-11), 1719–1730 (2009)
6. Corchado, J.M., Bajo, J., De Paz, J.F., Rodríguez, S.: An execution time neural-CBR guidance assistant. Neurocomputing 72(13), 2743–2753 (2009)
7. Závodská, A., ŠRamová, V., Aho, A.M.: Knowledge in Value Creation Process for Increasing Competitive Advantage. Advances in Distributed Computing and Artificial Intelligence Journal 1(3), 35–47 (2012)
8. Satoh, I.: Bio-inspired Self-Adaptive Agents in Distributed Systems. Advances in Distributed. Computing and Artificial Intelligence Journal 1(2), 49–56 (2012)
9. Agüero, J., Rebollo, M., Carrascosa, C., Julián, V.: MDD-Approach for developing Pervasive Systems based on Service-Oriented Multi-Agent Systems. Advances in Distributed Computing and Artificial Intelligence Journal 1(6), 55–64 (2013)
10. Khoshafian, S., Dasananda, S., Minassian, N., Ketabchi, M.: The Jasmine Object Database: Multimedia Applications for the Web. Computer Associates International (1998)
11. Kim, W.: Introduction to Object-Oriented Databases. MIT Press, Cambridge (1990)
12. Larson, J.A.: Database directions: from relational to distributed, multimedia, and object-oriented database systems. Prentice Hall, Upper Saddle River (1995)
13. Lejter, M., Meyers, S., Peiss, S.P.: Support for maintaining object-oriented programs. IEEE Transactions on Software Engineering 18, 1045–1052 (1992)
14. Saracco, C.M.: Universal database management: a guide to object/relational technology. Morgan Kaufmann Publishers, San Francisco (1998)
15. Silberschatz, A., Korth, H.F., Sudarshan, S.: Database System Concepts, 3rd edn. McGraw Hill (1997)
16. Tesler, L.G.: Object-oriented approach. Communication of the ACM 34(8), 13–14 (1991)

E-Nose System by Using Neural Networks

Sigeru Omatu and Mitsuaki Yano

Osaka Institute of Technology,
5-16-1 Omiya, Asahi-ku, Osaka, 535-8585, Japan
omatu@rsh.oit.ac.jp, yano@elc.oit.ac.jp
http://www.oit.ac.jp

Abstract. This paper considers a new construction of an electronic nose system based on a neural network. The neural network used here is a competitive neural network by the learning vector quantization. Various odors are measured with an array of many metal oxide gas sensors. After reducing noises from the odor data which are measured under the different concentrations, we take the maximum values among the time series data of odors. They are affected by concentration levels, we use a normalization method to reduce the fluctuation of the data due to the concentration levels. Those data are used to classify the various odors of teas and coffees. The accuracy of the classification is around 96% in case of four kinds of teas and around 89% for five kinds of coffees.

Keywords: E-nose, learning vector quantization, metal oxide gas sensor, odor classification.

1 Introduction

Over the last decade, odor sensing systems have undergone important developments from viewpoints of technology and commerce. James A. Milke [1] has proved that two kinds of metal oxide semi-conductor gas sensors (MOGS) have the ability to classify several sources of fire more precisely than with conventional smoke detector. However, his results achieve only around 85% of correct classification.

This paper proposes a new type of an electronic nose (E-nose) system to classify the various odors under the different densities of tees or coffees using a competitive neural network based on the learning vector quantization (LVQ) method. The sensors used here are MOGSs, which are attached on the seal of the sampling box at grid points in an array. We have used fourteen MOGSs of FIGARO Technology Ltd in Japan. We consider two types of data for classification in the experiment. The first type is four kinds of tees and the second one is five kinds of coffees of similar properties. The classification results of tees and coffees are about 96% and about 89%, respectively, which is much better than the results in [1]–[3].

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_36, © Springer International Publishing Switzerland 2014

2 Princple of MOGS

MOGS used in this paper is the most widely used sensor for making an array of artificial olfactory receptors in the E-nose system. These sensors are commercially available as the chemical sensor for detecting some specific odors. Generally, an MOGS is applied in many kinds of electrical appliances such as a microwave oven to detect the food burning, an alcohol breath checker to check the drunkenness, an air purifier to check the air quality, and so on.

Various kinds of metal oxide, such as SnO_2, ZnO_2, WO_2,TiO_2 are coated on the surface of semi-conductor, but the most widely applied metal oxide is SnO_2. These metal oxides have a chemical reaction with the oxygen in the air and the chemical reaction changes when the adsorbing gas is detected. The scheme of chemical reaction of an MOGS when adsorbing with the CO gas is shown as follows:

$$\frac{1}{2}O_2 + (SnO_{2-x}) \rightarrow O^{-}ad(SnO_{2-x}) \tag{1}$$

$$CO + O^{-}ad(SnO_{2-x}) \rightarrow CO_2 + (SnO_{2-x}) \tag{2}$$

The relationship between sensor resistance and the concentration of deoxidizing gas can be expressed by the following equation over a certain range of gas concentration:

$$R_s = A[C]^{-\alpha} \tag{3}$$

where R_s =electrical resistance of the sensor, A = constant, C =gas concentration, and α =slope of R_s curve.

When the metal oxide element on the surface of the sensor is heated at a certain high temperature, the oxygen is adsorbed on the crystal surface with the negative charge as shown in (1). In this stage the grain boundary area of the metal oxide element forms a high barrier. Then the electrons cannot flow over the boundary and this makes the resistance of the sensor become higher. When the deoxidizing gas, e.g., CO gas, is presented to the sensor, there is a chemical reaction between negative charge of oxygen at the surface of the metal oxide element and the deoxidizing gas as shown in (2). The chemical reaction between adsorbing gas and the negative charge of the oxygen on the surface of MOGS reduces the grain boundary barrier of the metal oxide element. Thus, the electron can flow from one cell to another cell easier. This makes the resistance of MOGS lower by the change of oxygen pressure according to the rule of (3).

Generally, it is designed to detect some specific odor in electrical appliances such as an air purifier, a breath alcohol checker, and so on. Each type of MOGS has its own characteristics in the response to different gases. When combining many MOGS together, the ability to detect the odor is increased. An E-nose system shown in Fig. 1 has been developed, based on the concept of human olfactory system. The combination of MOGS, listed in Table 1, are used as the olfactory receptors in the human nose.

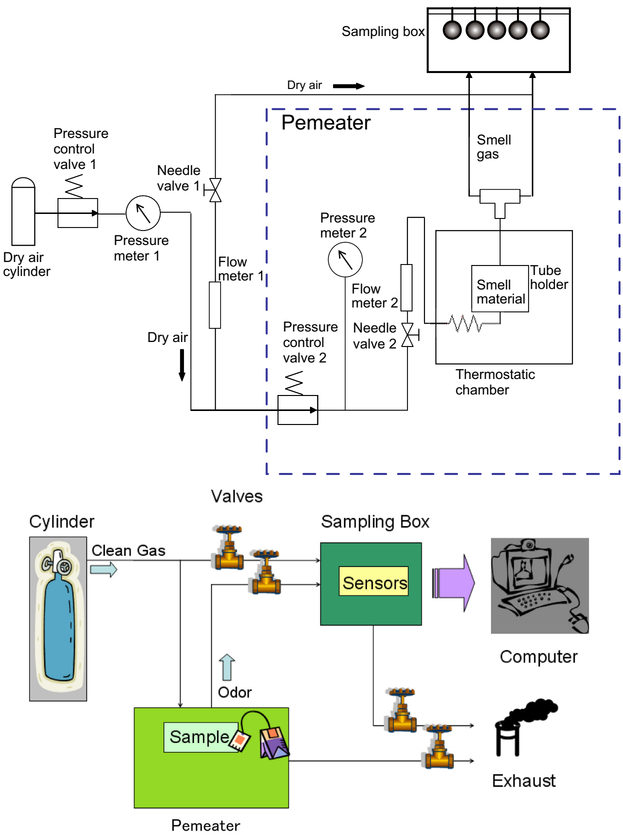


Fig. 1. Structure of the E-nose system

Table 1. List of MOGS used in the Experiment from FIGARO Technology Inc

Sensor No.	Sensor model	Main detecting gas
1	TGS2600	tobacco, cooking odor
2	TGS2602	hydrogen sulfide, VOC, ammonia
3	TGS2610	LP gas, butane, propanel
4	TGS2611	methane
5	TGS2620	alcohol, organic solvent
6,7	TGS826	ammonia, amine compounds
8,9	TGS816	methane, propane, butane(flammable gas)
10	TGS821	hydrogen gas
11	TGS832	chloro fluoro carbon gas
12	TGS825	hydrogen sulfide
13	TGS830	chloro fluoro carbon gas
14	TGS822	alcohol, organic solvent

3 Experimental Data Collection

The odor data used here are shown Tables 2 and 3, which are measured by the E-nose system explained in the previous section and used in the later classification. Note that in Table 3 Mocha coffees of the labels A,B,and C are selected from different companies.

Table 2. odor data of tees used in Experiment I

Label	Materials	No. of samples
A	English tee	20
B	Green tea	20
C	Barley tea	20
D	Oolong tea	20

Table 3. odor data of tees used in Experiment I

Label	Materials	No. of samples
A	Mocha coffee1	35
B	Mocha coffee2	35
C	Mocha coffee3	35
D	Kilimanjaro coffee	35
E	Char-grilled coffee	35

The sample of the raw data of Experiment I is shown in Fig. 2. At the beginning stage of Experiment I, we use clean gas from 2,200s to 2,600s. Then we use the clean gas to erase the odor from 2,600s to 3,800s. After that the same process has been continued. For each data set, the sampling period is 1s and the voltage signal in the clean gas is measured at the beginning of the repetition of experiment. In order to reduce the noise, we apply a smoothing filter to the measurement odor data $v'_s(t)$ such that

$$\bar{v}_s(t) = \frac{1}{3}\sum_{i=0}^{3} v'_s(t-i). \tag{4}$$

Then we take the difference $\Delta\bar{v}_s(t) = \bar{v}_s(t) - \bar{v}_s(t-1)$ and if this $\Delta\bar{v}_s(t) \geq \theta$ for more than N senseors where N is a predetermined number, we assume that the odor has been transported to the odor sonsors. Here, we take $\theta = 0.001$ and $N = 7$. We assume that the standard value $\bar{v}_s^{std}$ to determine the defference of the odor voltage and the clean gas voltage is an average of the clean gas voltages for five seconds before the odor data begins, that is,

$$\bar{v}_s^{std} = \frac{1}{5}\sum_{i=0}^{5} \bar{v}_s(t). \tag{5}$$

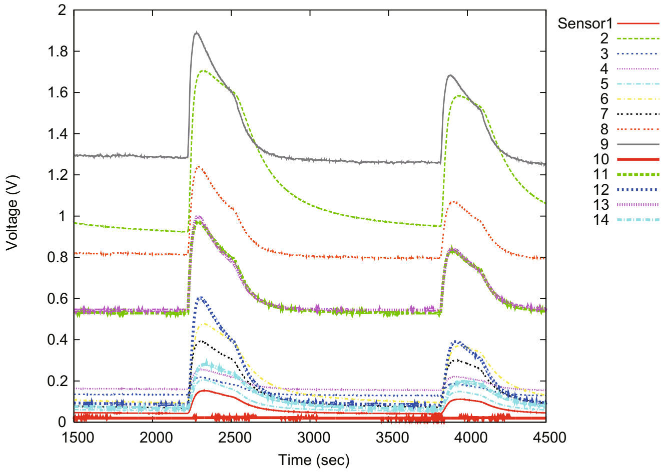


Fig. 2. Full time series data $v'_s(t)$ from a coffee odor in Experiment I

Then the effective voltage of the odor $v_s(t)$ is given by

$$v_s(t) = \bar{v}_s(t) - \bar{v}_s^{std} \tag{6}$$

After testing one odor the MOGS need to be cleaned by removing the tested odor and supplying only the clean gas until the output of the MOGS returns to a stable point. Then a new sample can be tested, repeatedly. This process is just like the human nose which needs to breath the fresh air before being able to recognize a new odor accurately. Some time series data from the experiment show that all odors approach the saturation stages within the measuring periods.

The levels of odor data are different according to the concentration values of the odor. Generally, we take the maximum value of the time series data as the representative characteristic of the odor for the sensor, which means that the maximum value reflects the steady state of the time series due to the mechanism of the sensing devices of MOGSs. Figure 3 shows the relation of the concentration levels and measurement data.

In order to delete the dependence on the concentration of the odors, we arrange those data such that the horizontal axis is the maximum values of a typical sensor (TGS2600) for experimental trials and the vertical axis show the ratio of the measurement values of the different sensors to those of the typical sensor (TGS2600) selected above. To classify the linear regression lines into some groupes, we will define some notations. Let the maximal value of the odor c data for the measurement trial p denote $v_{p,s}^c$. we denote the averageof $v_{p,s}^c$ with respect to s by μ_p^c, that is,

$$\mu_p^c = \frac{1}{S}\sum_{s=1}^{S} v_{p,s}^c, c = 1, 2, ..., C, p = 1, 2, ..., P \tag{7}$$

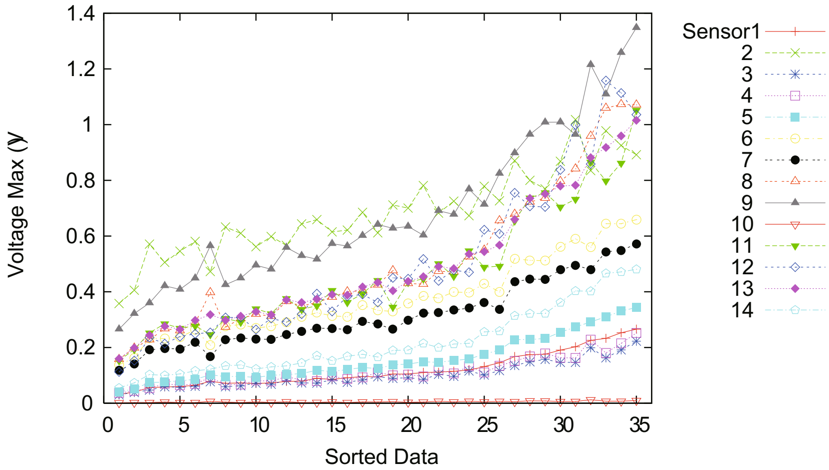


Fig. 3. Sorted data according to the concentration levels from a coffee odor in Experiment II

where C is the total number of odor kinds and P is the total number of experimental trials.

For a fixed odor c, we plot the data $\mu_p^c, v_{p,s}^c$ in the logarithmic scale in the plane. Then we can get a groupe of several kinds of lines given by the following regression linesas shown in Fig. 4:

$$f_s^c(\mu_p^c) = \alpha_s^c \mu_p^c + b_s^c, c = 1, 2, ..., C, p = 1, 2, ..., P. \tag{8}$$

Therefore, we can regard $f_s^c(\mu_p^c)$ or their regression coefficients α_s^c, b_s^c as the representative parameters of the odor data $v_{p,s}^c$ for the odor $c, c = 1, 2, ..., C$, the sensor $s, s = 1, 2, ..., S$,and the experimental trial $p, p = 1, 2, ..., P$.

4 Classification Algorithm

We must train the regression coefficients α_s^c, b_s^c for training samples. We assume that some part P' of total experimental trials P is the training data and the remaining one is the test data. For training phase, we adopt the competitive neural network based on LVQ, which has been used to train the regression coefficients α_s^c, b_s^c with respect to p.

After training, we caluculate the value μ_p by the following equation:

$$\mu_p = \frac{1}{S} \sum_{s=1}^{S} v_{p,s} \forall p \tag{9}$$

where $v_{p,s}, s = 1, 2, ..., S$ are the maximal values of $v_s(t), s = 1, 2, ..., S$ for the measurement trial p. Using μ_p, we get the estimated value $\hat{v}_{p,s}^c$ by the following equation:

$$\hat{v}_{p,s}^c = \hat{f}_s^c(\mu_p^c) = \alpha_s^c \mu_p + b_s^c, s = 1, 2, ..., S, c = 1, 2, ..., C. \tag{10}$$

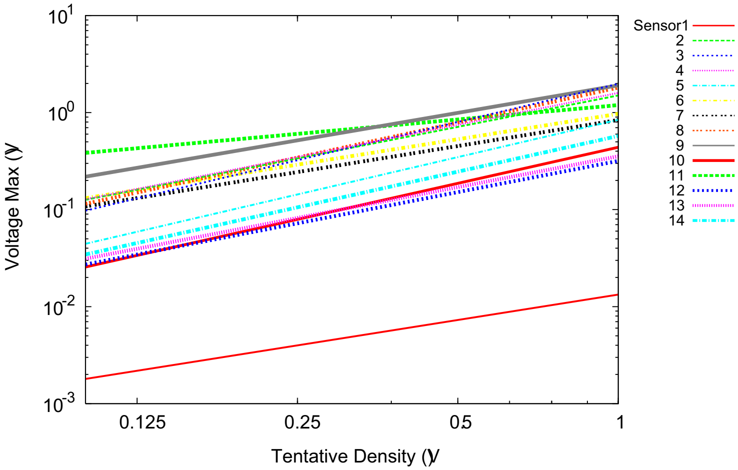


Fig. 4. Linear regression lines in Experiment II

Furthermore, we find c^o such that

$$c^o = arg \min_c d(v_p, \hat{v}_p^c) \tag{11}$$

$$d(v_{p,s}, \hat{v}_{p,s}^c) = \sqrt{\sum_{s=1}^{S} (v_{p,s} - \hat{v}_{p,s}^c)^2} \tag{12}$$

Then the measurement data $v_{p,s}, s = 1, 2, ..., S$ have been classified into the odor c^o which satisfies (11).

5 Classification Results

We have examined two examples, Experiment 1 and Experiment II stated in Section 3. In Experiment I, the training sample number $P' = 15$ and test sample number is five. We exchanged the training data set by 100 times and checked the classification accuracy of the test data sets. Thus, the total number of classification of 500 test samples is checked. The results are summarized in Table 4. Average of the classification is 96.15% and the classification is sufficient for real applications.

Experiment II is to classify the different kinds of coffee. As mentioned in Section 3, odor data A, B, and C are the coffees of Mocha made from different companies. If we regard those coffees as the same group, the classification rates become better. Roughly speaking, the classification of coffee is also very good for real application although those data look similar.

Table 4. Classification results for Experiment I

Odor data	Classification results(96.15%)					
	A	B	C	D	Total	Classification rates
A	500	0	0	0	500	100.0%
B	0	494	6	0	500	98.80%
C	0	71	429	0	500	85.80%
D	0	0	0	500	500	100.0%

Table 5. Classification results for Experiment II

Odor data	Classification results(88.80%)						
	A	B	C	D	E	Total	Classification rates
A	1190	253	29	27	1	1500	79.33%
B	225	1237	9	10	19	1500	82.47%
C	142	7	1325	26	0	1500	88.33%
D	9	14	3	1437	37	1500	95.80%
E	0	18	0	11	1471	1500	98.07%

6 Conclusions

We have presented the reliability of a new E-nose system designed from various kinds of MOGS. The E-nose has the ability to identify various sources with more than 90% of accuracy. It can be concluded that the E-nose is suitable for various applications in real world to classify the difficult odors.

Acknowledgment. This research has been supported by Grant-in-Aid for Scientific Research (B) No. 23360175 of Japan Society for the Promotion of Science and we wish to thank JSPS for their support.

References

1. Milke, J.A.: Application of Neural Networks for discriminating Fire Detectors. In: 10th International Conference on Automatic Fire Detection, AUBE 1995, Duisburg, Germany, pp. 213–222 (1995)
2. Charumporn, B., Yoshioka, M., Fujinaka, T., Omatu, S.: An E-nose System Using Back Propagation Neural Networks with a Centroid Training Data Set. In: Proc. Eighth International Symposium on Artificial Life and Robotics, Japan, pp. 605–608 (2003)
3. Fujinaka, T., Yoshioka, M., Omatu, S., Kosaka, T.: Intelligent Electronic Nose Systems for Fiore Detection Systems Based on Neural Netwoks. In: The Second International Conference on Advanced Engineering Computing and Applications in Sciences, Valencia, Spain, pp. 73–76 (2008)

Modelling an Orientation System Based on Speculative Computation*

João Ramos[1], Ken Satoh[2], Paulo Novais[1], and José Neves[1]

[1] Informatics Department, University of Minho, Portugal
{jramos,pjon,jneves}@di.uminho.pt
[2] National Institute of Informatics, Hitotsubashi, Chyoda-ku, Tokyo, Japan
ksatoh@nii.ac.jp

Abstract. Progress is inherent to a living society, which may occur in several different areas (*e.g.* computation, healthcare) and manners. The present (or now) is the time that is associated with the events perceived directly and in the first time, making, for example, the society to be very inquisitive on assistive technologies and how they may improve the human beings quality of living. This application of scientific knowledge for practical purposes may help the user in his/her diminished capabilities, and, usually, implies a small adaptation on the part of the individual in the use of devices; indeed one of the die down potentials of people with cognitive disabilities is the one of spatial orientation. On the other hand several were the authors that have developed systems to help an human being to travel between two locations. However, once the system is set up the change in the configurations have to be done manually in order to better adjust the system to the user. In order to go round this drawback, in this work it is presented a framework of speculative computation to set up the computation of the next user step using default values. When the information is obtained the computation is revised. Thus, the system may have a faster reaction to the user stimulus or it may start warning the user before he/she takes the wrong direction.

Keywords: Cognitive Disabilities, Mobile Application, Guidance, Person Tracking, Ambient Intelligence, Logic Programming.

1 Introduction

The degree of assistive technology was first done in the United States in 1988 by the Technology-Related Assistance for Individuals with Disabilities Act (also

* This work is part-funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-028980 (PTDC/EEI-SII/1386/2012). The work of João Ramos is supported by a doctoral grant by FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) SFRH/BD/89530/2012.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_37,

named Tech Act), which was later extended in 1998 by the Assistive Technology Act [15,1]. The definition stated that assistive technology is *any item, piece of equipment, or product system . . . that is used to increase, maintain, or improve functional capabilities of individuals with disabilities*. Since 1988 the interest and attention on this kind of technology has grown, leading to the development of a new set of devices to be embedded in the home environment in order to assist the user, monitoring his/her movements and his/her health on a day basis [16].

This new home environment (smarthouse [13]) prevented or minimized the independence loss on the part of the user. Caregivers and physicians may have remote access to the information collected from the environment [3], which may lead paid or not paid helpers to let the user (*e.g.*, elder or person with cognitive disabilities) to stay in his/her home for a long time, and in an independent way.

Despite the life quality improvement that smarthouses brought, the user could only be monitored when he/she was inside the house. If the user is prevented to go outside by the caregiver he/she may be a prisoner of his/her own house, and may not have an active participation in the community.

To improve the user independency outside the premisses, researchers have been developing new ways to use the available technology and assist the user, helping him/her to travel alone among different locations. Since this technology has to be carried out by the user, new devices have to be small, lightweight and resistant [6,5].On the other hand, besides the physical features of the devices, the user interface plays an important role in the utilization and acceptance of the device by the user. Usability challenges are research topics, mainly when the interfaces are developed for people with cognitive disabilities (*e.g.* in [6,5,8] the authors concluded that simple interfaces without distractive elements and simple text are some features that should be considered when developing an application to people with cognitive disabilities).

Considering the lack of spatial orientation of the user (previously described), it was developed a system (described in [12]) that guides the user among different locations using simple interfaces, enabling the caregiver to know his/her current location, which increases either the freedom of the user or the caregiver. The former may travel alone while the latter may be engaged in another work without neglect the care provision.

In Section 2 is presented some related work developed by other authors, considering orientation systems for people with cognitive disabilities. The developed systems are intended to be executed in smartphones. Section 3 brings about the new improvements of the system developed in [12]. In Section 4 it is formulated the framework of speculative computation. Finally, in Section 5 a reflection about this work is outlined.

2 Related Work

The technological development enable almost every one to have smartphones with high processing power, which may run applications that may help the owners to meet their expectations (*e.g.* the incorporation of a GPS module in a smartphone made possible the development of navigation programs). The major problem with these applications is that they have been developed to target the market as a whole, so that may not be used by people with disabilities, *i.e.*, to enable people with cognitive disabilities to travel alone using a navigation system, the applications must be developed according to the client features.

Carmien *et al.* [2] developed a system that enables the consumer to travel among different locations using a public transportation system like the bus; indeed, they conceive an end-user programming tool that enable caregivers to create scripts according to the activities that must be carried out by the person with cognitive disabilities (physical and/or cognitive).

Liu *et al.* [10], picture a prototype for guiding the user through static illustrations with overlaid arrows, audio messages and text messages. The authors studied the applicability of the prototype with several real case situations (*e.g.* it was studied the usability of landmarks in orientation and it was concluded that a near landmark should be used instead of one outside the user's view).

AlzNav from Fraunhover Portugal [7] is an orientation system for elderly people. The orientation process is given by an arrow that acts like a compass and indicates the travel path. This system also provides a localization feature that allows caregivers to know the current user location.

The previous systems/prototypes could successfully guide the user between two locations, improving the user independence since he/she could go out alone without getting lost.

In [12] we presented a functional prototype that works as an orientation and a localization system. The user is guided through an application for Android Operative System using augmented reality, being also given to the caregivers the possibility to know the current user position in real time.

In order to enhance the guiding experience attained so far, it was conceived a Framework for Speculative Computation (presented in Section 4) to be used in these situations. As expected, using speculative computation and default values, the user may use the system to generate possible scenarios for the next step forward before it materializes, *i.e.*, it may also be seen as an alert system. Due to the addition of this new feature, the system initially developed in [12] suffered some modifications. Section 3 describes, in brief, the changes that were done to the system.

3 System Description

In [11] was brought to light an initial description of the system developed in [12]. Using such a prototype the user (*i.e.* a person with mild cognitive disabilities) could travel alone without getting lost. The system topology, *i.e.*, the geometric properties and spatial relations are made up of (Figure 1) the *server*, a mobile application for people with cognitive disabilities (*Cognitive Helper Mobile Solution*), caregivers applications (*Caregiver Applications - Mobile and Web*), and *external services.*

The *server* is the system core and is made up the *Database* module, that stands for itself, and the *agency for integration of the system applications* module, that

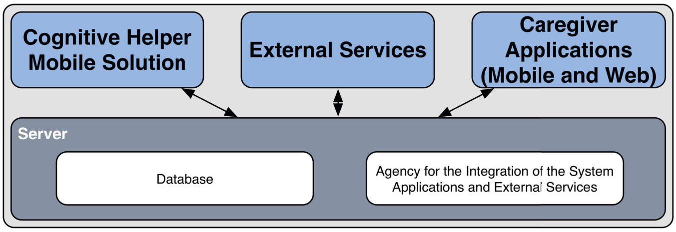


Fig. 1. An intelligible system topology

ensures the communication between the server and the user applications. To extend the features of the system, there is the possibility to connect it to external services through the *agency for integration of the system applications* module, *i.e.*, there is a possibility to interact with other systems like iGenda [4].

To improve the system features it was introduced a new functionality based on *speculative computation* (see Section 4). This practicality is included in the mobile application for people with cognitive disabilities under the *Localization Layer* (Figure 2), since it needs information about the current user location, which is given by the GPS device or by the network. Using this information one may compute the location to where the user should tour.

4 A Speculative System for Users with Cognitive Disabilities

Satoh [14], presented the theory of Speculative Computation and Abduction, which was an extension of the procedure of Kakas *et al.* [9]. This thesis lets the computation continue when one is faced with incomplete information, generating a tentative solution to the problem, as a result of the computational process, *i.e.*,

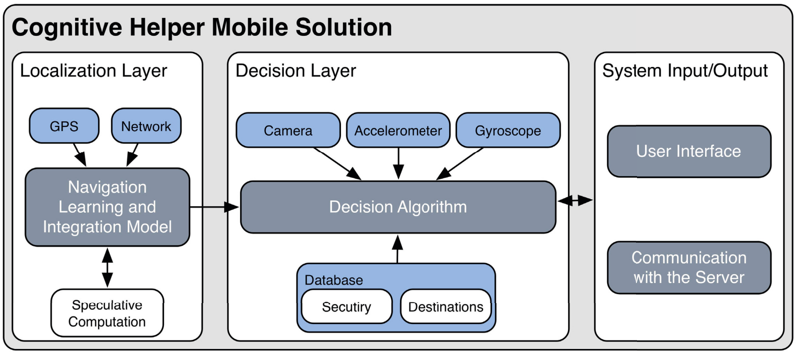


Fig. 2. Detailed framework of the mobile application for people with cognitive disabilities

the missing data is complemented through a default value and, as soon as the real one is obtained, the computational process is re-examined.

A Framework of Speculative Computation in the Orientation Method for people with cognitive disabilities (SF_{OM}) is now defined in terms of the signature $\langle \sum, \epsilon, \Delta, \mathcal{A}, \mathcal{P}, \mathcal{I} \rangle$ [14], where:

- $\sum$ stands for a finite set of constants (an element of $\sum$ is called a system module);
- ϵ denotes a set of predicates called *external predicates*. When Q is a literal belonging to an external predicate and S is the identifier of the information source, $Q@S$ is called an *askable literal*. We define $\sim(Q@S)$ as $(\sim Q)@S$;
- Δ designates the *default answer* set, which is a set of ground askable literals that satisfy the condition: Δ does not contain both $p(t_1, \ldots, t_n)@S$ and $\sim p(t_1, \ldots, t_n)@S$ at once;
- $\mathcal{A}$ is a mark of a set of predicates called *abducible predicates*. Q is called *abducible* when it is a literal with an *abducible predicate*;
- $\mathcal{P}$ signals a set of rules of the form:
  - ▷ $H \leftarrow B_1, B_2, \ldots, B_n$ where H is a positive ordinary literal, where each of $B_1, \ldots, B_n$ is an ordinary literal, an askable literal or an abducible; and
  - ▷ H is the head of rule R and is named as *head(R)* (always non-empty), being R the rule of the form $H \leftarrow B_1, \ldots, B_n$; $B_1, \ldots, B_n$ is the body denoted by *body(R)*, that in some situations is substituted by the boolean value *true*.
- $\mathcal{I}$ is a set of integrity constraints of the form:
  - ▷ $\perp \leftarrow B_1, B_2, \ldots, B_n$, where $\perp$ is a contradiction special symbol and B_1, $B_2, \ldots, B_n$ is an ordinary literal or an *askable literal* or an *abducible*. However at least one of $B_1, B_2, \ldots, B_n$ is an *askable literal* or an *abducible*.

An *askable literal* may have different meanings, namely:

1. An *askable literal* $Q@S$ in a rule $\mathcal{P}$ stands for a question put to a system module S by OM; and
2. An *askable literal* in Δ denotes a default truth value, either *true* or *false*,i.e., $p(t_1, \ldots, t_n)@S \in \Delta$, $p(t_1, \ldots, t_n)@S$ is usually *true* for a question to a system module S, and $\sim p(t_1, \ldots, t_n)@S \in \Delta$, $p(t_1, \ldots, t_n)@S$ is generally *false* for a question to a system module S.

In the logic program given below *path(a,b)* denotes that it is possible to travel between location a and location b; *show_next_point* states that the system must show the next location (which may be an intermediate or the final location) to the user; *show_user_warning* indicates that the system must alert the user, once he/she is going in the wrong direction; and the default values for the travel path of the user are defined in Δ. *user_travel(a,b)*, says that the user will travel from location a to location b; *included(a)* indicates if a location a is part of the route.

To ensure that the user is travelling in the correct path and he/she is alerted if is out the path, a framework to Speculative Computation is given below, in terms of the logic programming suite:

- ▷ $\sum$ = {gps_sensor, recognizer}
- ▷ ϵ = {user_travel, included}
- ▷ Δ = {user_travel(1,2)@gps_sensor, user_travel(2,6)@gps_sensor, user_travel(2,4)@gps_sensor, user_travel(4,5)@gps_sensor, ∼user_travel(2,3)@gps_sensor, ∼user_travel(3,4)@gps_sensor user_travel(6,2)@gps_sensor, included(1)@recognizer included(2)@recognizer, included(3)@recognizer included(4)@recognizer, included(5)@recognizer, ∼included(6)@recognizer}
- ▷ $\mathcal{A}$ = {show_next_point, show_user_warning}
- ▷ $\mathcal{P}$ is a mark of the following set of rules:

$$guide(A, A) \leftarrow .$$

$$guide(A, B) \leftarrow path(A, F),\ show\_next\_point(F),\ user\_travel(A, F)@gps\_sensor,\ guide(F, B).$$

$$guide(A, B) \leftarrow path(A, F),\ user\_travel(A, F)@gps\_sensor,\ show\_user\_warning(F),\ guide(F, B).$$

$$path(1, 2) \leftarrow .$$
$$path(2, 3) \leftarrow .$$
$$path(2, 4) \leftarrow .$$
$$path(2, 6) \leftarrow .$$
$$path(3, 4) \leftarrow .$$
$$path(6, 2) \leftarrow .$$
$$path(4, 5) \leftarrow .$$

- ▷ $\mathcal{I}$ denotes the following set of integrity constraint or invariants:

$$\perp \leftarrow show\_next\_point(F),\ \sim included(F)@recognizer.$$

$$\perp \leftarrow show\_user\_warning(F),\ included(F)@recognizer.$$

To ensure the integrity of the framework there are two invariants stating that the system may not show the next travelling point to the user if it isn't part of the route; and that the system may not alert the user if he/she is travelling in a valid path.

In the setting depicted to above it is assumed that the user will travel among locations 1 and 5 through the intermediate locations 2 and 4, and that he/she

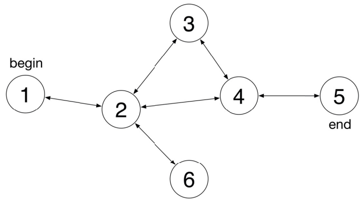


Fig. 3. Possible ways to travel among locations 1 and 5

will take the wrong direction at location 2 to 6. An elucidation of the possible paths that the user may use is presented in Figure 3.

5 Conclusions

The lack of orientation is a serious problem for people with cognitive disabilities. Indeed, caregivers may prevent these people to go out their homes by themselves, since the risk of getting lost may be too high. Thus, some work has been pursuit in orientation methods for people with cognitive disabilities, so they may travel alone while being remotely monitored.

In one' system the user is guided through augmented reality, surpassing the limitations of other ones that use static pictures or the explanation of a compass. On the other hand, caregivers may know in real time the current location of the user on a road map. The system may also connect to external services, thus increasing its effectiveness.

On the other hand, and to increase the system responsiveness, it has been envisage a suite for speculative computation having in mind people with cognitive disabilities.

In the upcoming work we plan to do field tests with users in order to receive their feedback and better adapt the system to their needs. The level of the prompts will be better adjust to each user, according to a speculative program of action.

References

1. Alper, S., Raharinirina, S.: Assisitive Technology for Individuals with Disabilities: A Review and Synthesis of the Literature. Journal of Special Education Technology 21(2), 47–64 (2006)
2. Carmien, S., Dawe, M., Fischer, G., Gorman, A., Kintsch, A., Sullivan, J.F.: Socio-technical environments supporting people with cognitive disabilities using public transportation. ACM Transactions on Computer-Human Interaction 12(2), 233–262 (2005)
3. Carneiro, D., Novais, P., Costa, R., Gomes, P., Neves, J.: EMon: Embodied Monitorization. In: Tscheligi, M., de Ruyter, B., Markopoulus, P., Wichert, R., Mirlacher, T., Meschterjakov, A., Reitberger, W. (eds.) AmI 2009. LNCS, vol. 5859, pp. 133–142. Springer, Heidelberg (2009)

4. Costa, A., Novais, P.: An Intelligent Multi-Agent Memory Assistant. In: Handbook of Digital Homecare, vol. 3, pp. 197–221 (2011)
5. Dawe, M.: Desperately seeking simplicity: how young adults with cognitive disabilities and their families adopt assistive technologies. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2006, pp. 1143–1152. ACM (2006)
6. Dawe, M.: Let Me Show You What I Want: Engaging Individuals with Cognitive Disabilities and their Families in Design. Technology, 2177–2182 (2007)
7. Fraunnhover Portugal: AlzNav (2012), `http://www.fraunhofer.pt/en/fraunhofer_aicos/projects/internal_research/alznav.html`
8. Friedman, M.G., Bryen, D.N.: Web accessibility design recommendations for people with cognitive disabilities. Technology and Disability 19(4), 205–212 (2007)
9. Kakas, A.C., Kowalski, R.A., Toni, F.: The Role of Abduction in Logic Programming. In: Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 5, pp. 235–324 (1998)
10. Liu, A.L., Hile, H., Borriello, G., Kautz, H., Brown, P.A., Harniss, M., Johnson, K.: Informing the Design of an Automated Wayfinding System for Individuals with Cognitive Impairments. In: Proceedings of Pervasive Health 2009, London UK, vol. 9, p. 8 (2009)
11. Ramos, J., Anacleto, R., Costa, Â., Novais, P., Figueiredo, L., Almeida, A.: Orientation System for People with Cognitive Disabilities. In: Novais, P., Hallenborg, K., Tapia, D.I., Rodríguez, J.M.C. (eds.) Ambient Intelligence - Software and Applications. AISC, vol. 153, pp. 43–50. Springer, Heidelberg (2012)
12. Ramos, J., Anacleto, R., Novais, P., Figueiredo, L., Almeida, A., Neves, J.: Geo-localization System for People with Cognitive Disabilities. In: Pérez, J.B., et al. (eds.) Trends in Pract. Appl. of Agents & Multiagent Syst. AISC, vol. 221, pp. 59–66. Springer, Heidelberg (2013)
13. Sadri, F.: Multi-Agent Ambient Intelligence for Elderly Care and Assistance. In: Aip Conference Proceedings, vol. 2007, pp. 117–120. Aip (2007)
14. Satoh, K.: Speculative Computation and Abduction for an Autonomous Agent. IEICE - Trans. Inf. Syst. E88-D(9), 2031–2038 (2005), `http://dx.doi.org/10.1093/ietisy/e88-d.9.2031`
15. Scherer, M.M.J., Hart, T., Kirsch, N., Schulthesis, M.: Assistive Technologies for Cognitive Disabilities. Critical Reviews in Physical and Rehabilitation Medicine 17(3), 195–215 (2005)
16. Stefanov, D.H., Bien, Z., Bang, W.C.: The smart house for older persons and persons with physical disabilities: structure, technology arrangements, and perspectives. IEEE Transactions on Neural and Rehabilitation Systems Engineering 12(2), 228–250 (2004)

Stable Learning for Neural Network Tomography by Using Back Projected Image

Masaru Teranishi, Keita Oka, and Masahiro Aramoto

Hiroshima Institute of Technology, Department of Information Systems and Management, Faculty of Applied Information Science,
2-1-1 Miyake, Saeki-ku, Hiroshima 731-5193, Japan
teranisi@cc.it-hiroshima.ac.jp

Abstract. This paper presents a stable learning method of the neural network tomography, in case of asymmetrical few view projection. The neural network collocation method (NNCM) is one of effective reconstruction tools for symmetrical few view tomography. But in case of asymmetrical few view, the NNCM tends to unstable and fails to reconstruct appropriate tomographic images. We solve the unstable problem by introducing the back projected image in the early learning stage of NNCM. The numerical simulation with an assumed tomographic image show the effectiveness of the proposed method.

Keywords: Few view tomography, neural network, back propagation, collocation method, inverse problem, ill-posed problem, model fitting.

1 Introduction

The Computerized Tomography (CT) is one of effective remote sensed imaging method for medical and industrial non destructive diagnostics[1]. CT is also applied to fusion plasma imaging research area, thanks to its remote sensing facility.

Fusion experimental devices have only few observation port to look inside due to the vacuum vessel structure. Then the viewing angles for tomographic measurement are limited to few and asymmetrical directions. For such asymmetrical few view, standard reconstruction methods used in general CT are difficult to use.

For the few view CT problem, there are mainly two kinds of reconstruction approaches: the regularization of ill-posed equation, and the model fitting reconstruction. Takeda, Ma, et. al proposed a new model fitting CT reconstruction method by using a neural network[2,3] which is called "the Neural Network Collocation Method" (NNCM). The NNCM reconstructs tomographic image by the error back propagation(BP) learning algorithm based on the projection data.

Though NNCM is reported to be useful for few view CT problem[2], view directions are set in symmetric angles. But in actual cases of plasma CT imaging, the view directions are set in "asymmetrical" angles. NNCM has drawbacks

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_38,

for asymmetrical few view CT with instability of learning and often fails to reconstruct appropriate tomographic image.

The paper proposes a stable learning method for NNCM in asymmetrical few view CT. The proposed method stabilizes the reconstruction process by using a back projected image as the teacher data for BP algorithm directly, instead of the projection data, in the early learning stage.

2 Few View CT Problem

This paper deals with two-dimensional tomographic imaging. The goal of CT is to reconstruct spatial distribution of certain physical quantity $f(x, y)$ at spatial position (x, y) of a cross section of the object as an image. We can measure a set of unknown $f(x, y)$ values indirectly by scanning with transparent beams like X-rays and get detector output g_i as form of a line integral, as shown in Fig. 1.

If the reconstruction region is divided into n pixels, the calculation of g_i is approximated by discrete expression as follows:

$$g_i = \int_{L_i} f(x, y)dxdy \approx \sum_{j=1}^{n} a_{ij} f_j \tag{1}$$

where f_j is mean image intensity of jth pixel, a_{ij} is the traversing length of the ith beam through the jth pixel.

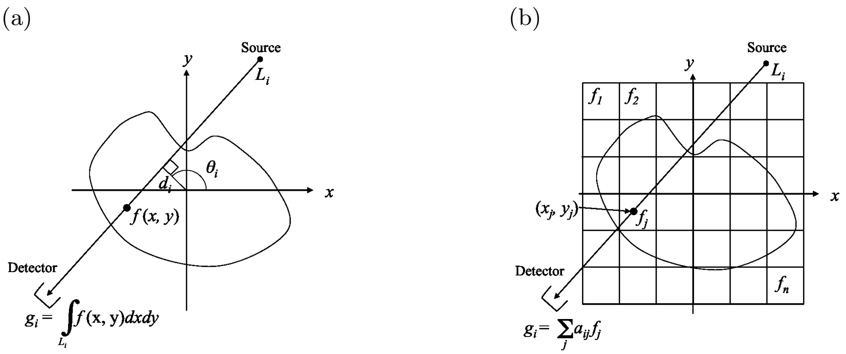


Fig. 1. Projection measurement of g_i along a beam path L_i : (a) continuous form, (b) discrete form

In such discrete projection condition, to reconstruct an unknown tomographic image vector $\boldsymbol{f} = (f_1, f_2, \cdots, f_j, \cdots, f_n)^T$ from the measured projection data vector $\boldsymbol{g}^{meas} = (g_1, g_2, \cdots, g_j, \cdots, g_m)^T$ is equivalent to solve a linear equation as

$$A\boldsymbol{f} = \boldsymbol{g}^{meas} \tag{2}$$

where matrix A is $m \times n$ dimensional and has a_{ij} of Eq.(1) as its element, and is called the "projection matrix". In general CT application, Eq.(2) is solvable because the number of projection data is more enough than the number of unknowns, i.e. the number of pixels of the reconstruction image ($m \geq n$). But in case of few view CT, the number of projection data is extremely less than the number of unknowns ($m \ll n$), so Eq.(2) becomes ill-posed. There are two different approaches to solve such few view CT problems. The one is the model fitting like ART, SIRT[1], and some kind of series expansions[4], the other is the regularization method[1]. The NNCM is a kind of model fitting approach.

3 Neural Network Collocation Method (NNCM)

Fig.2 shows an usage of the NNCM.

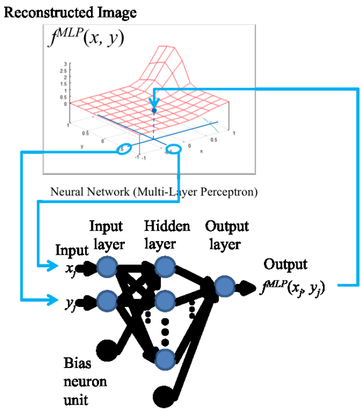


Fig. 2. Usage of the NNCM

The NNCM uses a Multi Layer Perceptron(MLP) type neural network[5]. The MLP consists of three layers: input, hidden, and output layer. The input layer has two neuron units and receives a spatial position (x_j, y_j) of the reconstruction region. The output layer has a neuron unit and outputs the intensity $f^{MLP}(x_j, y_j)$ of the corresponding spatial position. NNCM gets a reconstructed image by feeding the MLP all spatial positions of the reconstruction region sequentially, and getting the corresponding intensities.

The MLP represents the reconstructed image as a synthesized basis functions of spatial positions. Each unit of the hidden layer represents the model basis function of the reconstructed image. In the hidden layer, the kth unit computes its internal state u_k^h as a sum of weighted outputs of the input layer, as

$$u_k^h = w_{k,x}^h x_j + w_{k,y}^h y_j + w_{k,b}^h \tag{3}$$

where $w_{k,x}^h$ is the connection weight from an input layer's unit for position x_j, and $w_{k,y}^h$ denotes the similar way for y_j, and $w_{k,b}^h$ does for the bias unit. The output of kth unit of spatial position (x_j, y_j) is computed as o_k^h by transforming u_k^h with sigmoid function $\sigma(u) = 1/(1+e^{-u})$ as

$$o_k^h = \sigma(u_k^h) = \sigma(w_{k,x}^h x_j + w_{k,y}^h y_j + w_{k,b}^h). \tag{4}$$

The unit of the output layer computes its internal state as the weighted sum of basis functions of the hidden layer's units and transform the state into the output by sigmoid function. The MLP output $f^{MLP}(\boldsymbol{w}, x_j, y_j)$ is determined by current input (x_j, y_j) for the jth pixel and weight vector $\boldsymbol{w}$.

In the NNCM, a MLP reconstructs an appropriate tomographic image for a specific projected data. The reconstruction is achieved by tuning the connection weights using the error back propagation(BP) learning algorithm based on the projection data. The learning result is effective only for one specific projected data. In this meaning, the MLP of the NNCM has no generalized facility for input data. If we want to reconstruct an image for another projection data, then we should reset the MLP connections and let the MLP learn from the beginning. The goal of CT reconstruction is to fit the projection data of the reconstructed image to the measured one, i.e., to reconstruct a tomographic image by minimizing residual of projection $\|A\boldsymbol{f}^{MLP} - \boldsymbol{g}^{meas}\|$. On the other hand, the error function used in the BP algorithm is computed with the MLP output and the corresponding the teacher data. In this way, we have define the error function of BP for NNCM as

$$E = E^{img} = \frac{1}{2}\sum_{j=1}^{N}\left\{f^{MLP}(\boldsymbol{w}, x_j, y_j) - f^{TRUE}(x_j, y_j)\right\}^2 \tag{5}$$

and the BP algorithm modifies the MLP's weights by gradient descent to minimize E as

$$\begin{aligned} \boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} + \Delta\boldsymbol{w}^{(t)} \\ \Delta\boldsymbol{w}^{(t)} &= -\eta\frac{\partial E}{\partial \boldsymbol{w}^{(t)}} + \beta\Delta\boldsymbol{w}^{(t-1)} \end{aligned} \tag{6}$$

where t is an index of the current learning iteration, η is the learning step size, β is the momentum coefficient of the past learning.

But the teacher data $\boldsymbol{f}^{TRUE}$ is not able to use in case of CT because $\boldsymbol{f}^{TRUE}$ is the tomographic image, the solution of the reconstruction problem. So we could not use E^{img} as E of Eq.(6), we have to derive $\Delta\boldsymbol{w}^{(t)}$ from projection residual E^{prj} as

$$E^{prj} = \frac{1}{2}\sum_{i=1}^{m}\left(g_i^{MLP} - g_i^{meas}\right)^2 \tag{7}$$

where g_i^{MLP} is a projection data calculated by using $f^{MLP}(x_j, y_j)$ along the ith beam path. The modification value $\Delta\boldsymbol{w}^{(t)}$ that minimize the projection residual E^{prj} for BP learning along the ith beam projection path is derived as

$$
\begin{aligned}
\Delta \boldsymbol{w}_i^{(t)} &= -\eta \frac{\partial E^{prj}}{\partial \boldsymbol{w}_i^{(t)}} + \beta \Delta \boldsymbol{w}_i^{(t-1)} \\
&= -\eta \sum_{j=1}^{n} \left(g_i^{MLP} - g_i^{meas} \right) \cdot \frac{\partial g_i^{MLP}}{\partial f_j^{MLP}} \frac{\partial f_j^{MLP}}{\partial \boldsymbol{w}_i^{(t)}} + \beta \Delta \boldsymbol{w}_i^{(t-1)} \\
&= -\eta \sum_{j=1}^{n} \left(g_i^{MLP} - g_i^{meas} \right) a_{ij} \frac{\partial f_j^{MLP}}{\partial \boldsymbol{w}_i^{(t)}} + \beta \Delta \boldsymbol{w}_i^{(t-1)}
\end{aligned} \tag{8}
$$

where the partial derivative term in the last row of the above equation is computable in same way of the general BP algorithm. We can now apply BP learning to the CT problem by using Eq.(8) for all projection paths. Eq.(8) enables the NNCM to learn MLP for unknown tomographic image f^{MLP} indirectly via measured projection data g_i^{meas}. We also should note that the weights are updated when the index of projection path i is changed. As the result, the weights are updated m times a learning epoch. The algorithm is called "quasi-on-line learning"[2].

4 Instability of NNCM in Asymmetrical Few View CT

The NNCM is reported to work well in case of the few view tomography[2]. We have applied the NNCM to asymmetrical few view CT applications, then we found that NNCM often tends to fail by increasing projection error in the learning progresses, and finally reconstructs meaningless tomographic image. In the paper[2], the NNCM work well in case of few view projection CT. Because that paper deals only "symmetrical" few view projection case. If directions of the projections are symmetric (Fig.3(a)), NNCM works well even in few view projection. But actually cases of plasma CT take few "asymmetrical" view projection as shown in (Fig.3(b))[4,6]. We have considered that the main reasons of the NNCM instability for asymmetrical few view CT are inappropriate initial weights and over-learning with the projected data.

5 Stable NNCM Learning by Back Projected Image

To improve stability of NNCM learning for asymmetrical few view Projection CT, we introduce an early learning with back-projected image.

The main difficulty of the NNCM for asymmetrical few view CT is considered as the indirect learning. If we can use a direct teacher data as form of image data for learning, the NNCM could be more stable even in asymmetrical few view conditions. But the tomographic image is unknown, hence we cannot use the exact image as the teacher data. An alternative way is to use some kind of quasi-reconstructed image. We can make one of such quasi-reconstructed images by back projecting the project data as

$$
\boldsymbol{f}^{BP} = A^T \boldsymbol{g}^{meas} \tag{9}
$$

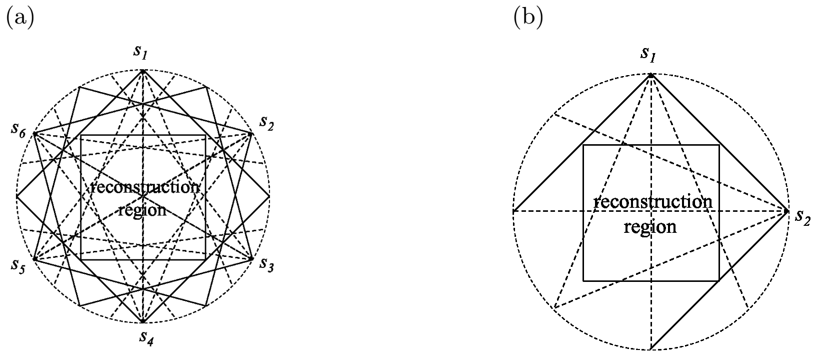


Fig. 3. Few view projection configuration: (a) symmetrical case, (b) asymmetrical case. s_i is source point of the ith beam.

where A^T denotes the transpose matrix of A. Use of the above quasi-reconstructed image as a direct teacher data helps the NNCM to reconstruct appropriate tomographic image more stable.Then the proposed method first learns the MLP by using back projected image at early stage, with small number of learning iterations. After the early learning, the MLP is learned by conventional NNCM method, i.e., by the projection data.

6 Numerical Simulations

We have executed numerical CT reconstruction simulation to validate the effectiveness of the proposed method. To evaluate the reconstruction performance, we use an assumed tomographic image as shown in Fig.4(a) which is typically observed distribution of heat loss at a poloidal cross section of helical plasma[6]. The simulated projection data is calculated by applying the assumed image in two direction few fan beam as shown in Fig.3(b). Each projection view has 50 rays with equal interval angles and are arranged within 90 degree fan. The direction of each fan is crosses each other with 90 degree angle. By using two direction fan beams, we have got the projection data vector $\boldsymbol{g}^{meas}$ as a 100 dimensional vector. The simulation reconstructs the tomographic image at a spatial region in range of $x \in [-1, 1], y \in [-1, 1]$. The reconstruction region has 32×32 pixels. Although the number of pixels seems to be small, the number is suited for the dimension of the projection data.

We have justified the following configurations and learning parameters for the MLP empirically. The MLP has 2 input and 1 output units, and its hidden layer consists of 12 units. The early learning stage of the proposed method takes 100 iterations with learning parameters $\eta = 0.2, \beta = 0.7$, for direct learning with the back projected image as shown in Fig.4(b). After the early learning stage, the reconstruction have continued by the conventional NNCM's projection learning, that takes 2000 iterations with learning parameters $\eta = 0.09, \beta = 0.7$. The conventional NNCM takes only projection data learning for 2000 iterations

(a) (b)

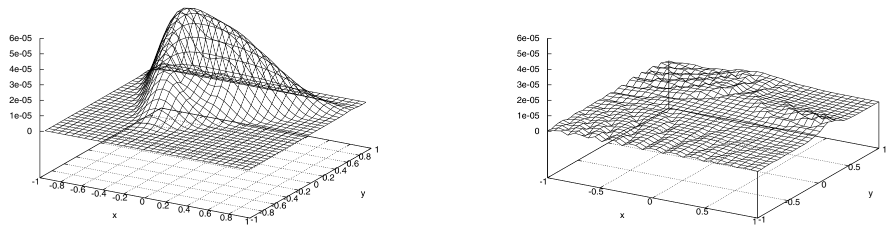


Fig. 4. (a) Assumed image used for the numerical simulation and (b)its back projected image

with learning parameters $\eta = 0.09, \beta = 0.7$. The connection weights $\boldsymbol{w}$ are initialized with small random numbers for every simulation. We have tested same reconstruction simulation by changing initial weights 10 times.

A typical simulation result is shown in Fig.5. As shown in Fig.5(a), the proposed method has continuously reduced the residual of the projection E^{prj} as the learning progressed, in both of the early and sequel learning stages. In contrast, the conventional NNCM failed to reduce the residual in two aspects: the minimum residual is still larger than that of the proposed method, and the residual turned to increase at end of the learning.

The proposed method shows the advantage not only in projection residual, but also in the reconstructed image. The proposed method could reconstruct almost part of the assumed distribution. The conventional NNCM failed to reconstruct. Such advantages of the proposed method are confirmed over all 10 times different initialized trials. The proposed method succeeded to reconstruct 6 times out of 10, and the conventional NNCM had no successful result out of 10 times.

(a) (b) (c)

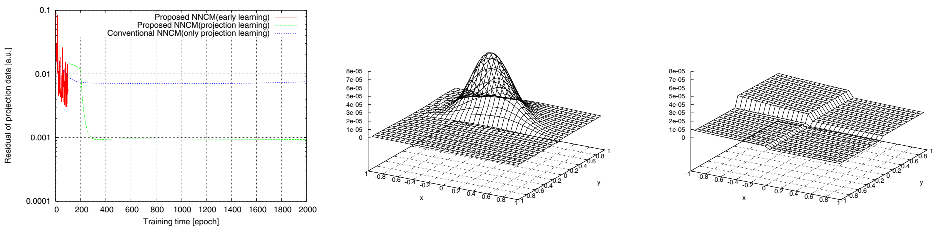


Fig. 5. Reconstruction result:(a) Residual of projection, (b) Reconstructed image by the proposed method, and (c) Reconstructed image by the conventional NNCM

7 Discussion

Though the proposed method could demonstrate the advantages in case of asymmetrical few view CT by the numerical simulation, there are still two problems should be improved:

(1) We had not yet examined enough about most effective length of the early learning. In this paper, we use only a fixed 100 iterations as the early learning. Searching optimal early training length is remained.

(2) We could not say the proposed method reconstructs the tomographic image at satisfied accuracy. As shown in Fig.5(b), the peak distribution of the reconstruction image was shifted in y direction, compared to the assumed distribution.

The major reasons of these problem are initial positions and movements of the basis functions of NNCM. The learning process of the NNCM moves the basis functions of the reconstructed image, by modifying the connection weights of MLP. Since a viewing the learning property of the NNCM in aspect of the basis functions deeply links to the reconstruction performance, we should observe the initial spatial locations of the basis functions and their movement with learning carefully in future works.

8 Conclusion

This paper presents a stable learning method for the neural network tomography. The reconstruction performance of the NNCM for asymmetrical few view CT is stabilized by using the back projected image in early learning stage. Numerical simulation which assumed actual plasma CT imaging showed the effectiveness of the proposed method.

References

1. Natterer, F.: The Mathematics of Computerized Tomography. Wiley/B. G. Teubner, New York (1986)
2. Ma, X.F., Fukuhara, M., Takeda, T.: Neural network CT image reconstruction method for small amount of projection data. Nucl. Instrum. Methods in Phys. Res. A 449, 366–377 (2000)
3. Takeda, T., Ma, X.F.: Ionospheric Tomography by Neural Network Collocation Method. Plasma and Fusion Research 2, S015-1–6 (2007)
4. Nagayama, Y.: Tomography of m = 1 mode in tokamak plasma using least-squares-fitting method and Fourier-Bessel expansions. J. Appl. Phys. 62, 2702 (1987)
5. Rumelhart, D.E., McClelland, J.L.: PDP Research Group: Parallel Distributed Processing. MIT Press, Cambridge (1986)
6. Liu, Y., Peterson, B.J., Iwama, N., Konoshima, S.: LHD Experiment Group, JT-60 team: Application of Tomographic Imaging to Multi-pixel Bolometric Measurements. Plasma and Fusion Research 2, S1124-1–4 (2007)

Security Considerations for Patient Telemonitoring Schemes through Wireless Networks

V. Morales, D. Cortés, N. Gordillo, A. De la Torre, and D. Azpetia

University of Ciudad Juárez, Juárez, Chihuahua, México

Abstract. In this study security aspects are considered for the protection of medical information in patient telemonitoring schemes. In order to propose mechanisms to efficiently protect clinical information, it is presented a case study of a wireless network telemonitoring scheme, the results of a risk assessment to detect security problems in this type of schemes are analyzed.

Keywords: Telemedicine, Telemonitoring, Clinical Information Security, Risk Assesment.

1 Introduction

There are diseases that require constant monitoring of patients that should remain hospitalized so that their biological parameters can be monitored until their health is stabilized or improved. However, in other cases it is not necessary for these patients to remain in the hospital because despite having a chronic illness, their health is not considered to be critical. In the list of the chronic diseases that require constant monitoring or surveillance are diabetes mellitus, cardiovascular diseases, lung diseases, among others. Recently, several studies [1–3] in the area of telemedicine dedicated to the care of patients whose health requires constant monitoring have been proposed. Remote monitoring or telemonitoring is an area of telemedicine in which, through various instruments, diverse biological parameters can be monitored in order to provide timely care in case an emergency situation arises with the patient. Usually these patients are in their own home, so the instruments used to perform remote monitoring should be accessible and easy to use.

Current technologies for patient telemonitoring involve transmission of data through network communications such as Internet, cellular networks, satellite, etc. These data include sensitive medical information that must be protected from tampering. This work presents different schemes and telemonitoring technologies and evaluates, through a risk assessment, the need to employ security mechanisms to guarantee that the information is complete and reliable. In Section 2 telemonitoring schemes proposed for various health areas are described. Section 3 analyzes the security aspects to be considered for the development of telemonitoring systems. In Section 4 a risk assessment process based on current technologies is conducted. Final conclusions are presented in Section 5.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_39,

2 Telemonitoring Schemes

Several biological parameters such as ECG, EEG, insulin level, blood pressure, oxygen saturation, among others can be monitored with biomedical technology available today. However, through a telemonitoring system, related events such as the patient's symptoms or medications can also be recorded and transmitted. Due to the different ways of measuring or capturing clinical information of patients, two areas of remote monitoring can be defined:

- Automated telemonitoring. A biological parameter or event information is acquired through a device that is capable of automatically (or scheduled) transmit such information. This can be done in different ways:
  - Acquisition and transmission in specific periods. According to the need of monitoring the patient a time frame (eg, every 4 hours) is specified so that the clinical information is acquired and immediately sent to a monitoring center through a telecommunications network.
  - Acquisition with no scheduled transmission. A time frame for the acquisition of information is specified, however, it will be transmitted to a monitoring center if one parameter is out of the specified limits or, when a particular event occurs.
  - Acquisition and transmission with alert. As in the first case, information transmission takes place after every acquisition. However, if any value is outside a specified limit or a specific event is present, an alert is generated and transmitted along with clinical information.
- Conducted telemonitoring. Sending clinical information to the monitoring center requires a transmission medium for example through a web application. This type of telemonitoring requires the intervention of the patient or another person, which can reduce its efficiency.

Making appropriate use of information and communication technologies, telemonitoring platforms can help to increase the quality of patient care reducing public expenditure on care and treatment. In recent years several telemonitoring schemes have been proposed as described in [4], [5], [6] and [7]. Having a particular common: the transmission technologies, they vary according to the service provided and the type of patients they treat.

Wireless communication networks offer additional advantages to remote patient monitoring. The patient can be monitored from anywhere with cellular or even satellite coverage. Moreover, not only the information is received and accessed from hospitals or monitoring centers, it can also go directly to relevant medical staff through common mobile devices like cell phones or tablets. A patient telemonitoring system can make use of common wireless communication standards, such as Bluetooth, WiFi or cell phone technologies.

2.1 Case Study: A Wireless Network Telemonitoring Scheme

Figure 1 briefly describes a telemonitoring scheme of diabetic patients developed at the University of Ciudad Juárez. This scheme will serve as a basis for making a risk assessment to detect security problems in telemonitoring systems.

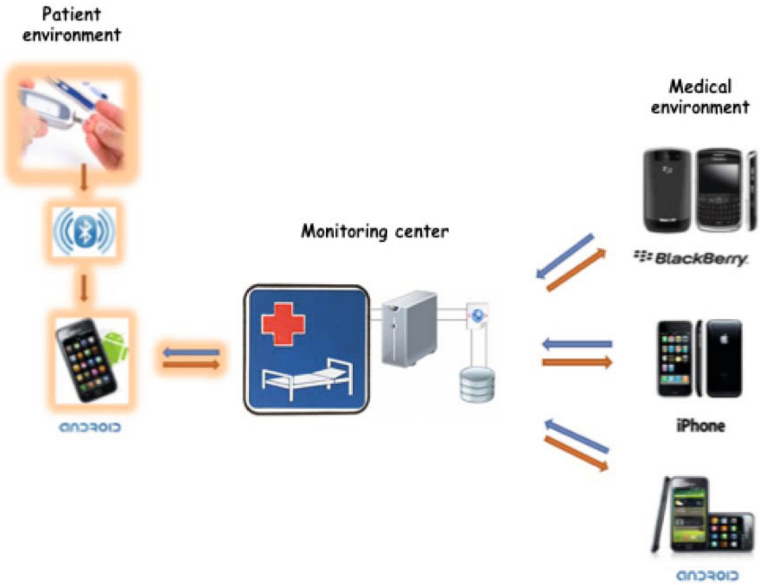


Fig. 1. Components of the case study telemonitoring scheme

The scheme makes use of wireless communication networks, and is composed of the following environments:

- Patient environment. The patient or family performs the tasks of clinical information record from their own home or anywhere else they are. The patient has a commercial glucometer with the ability of transmitting information through Bluetooth. The glucose measurement is transmitted to a mobile phone with an Android native application that receives and retransmits information through a cell network to a database. The application also has the ability to receive information from the monitoring center or the physician in charge.
- Monitoring center. The monitoring center displays patient records through a web application. A list of patients and their most recent measurements is displayed, alerts are also displayed when measurements are considered outside the normal range.
- Medical environment. Consists of mobile applications for multiple platforms (iOS, BB OS, Android) used by physicians who have diabetic patients. Through the application the physician can check the records of glucose levels. The doctor also receives alerts and is able to send recommendations to patients, among other features.

3 Security of Clinical Information

A very important issue is maintaining clinical information safe. We call clinical information to all patient information which is of interest to the physician (e.g. biological parameters, symptoms or events, medical imaging, medication information, lab results). Clinical information should be treated carefully in each of the telemonitoring environments to meet different security objectives [8]. These objectives are: authentication (identifying the client), confidentiality (only who is authorized can access the information), integrity (the information has not been transformed during the processing, transport or storage), availability (to have the information when needed), and audit (to track the accesses made to the information and the operations performed on it).

In most countries clinical information is considered confidential and can only be accessed by authorized clinicians. In telemonitoring schemes, clinical information is generated in the patient environment and is generally transmitted through public means of communication such as Internet, wireless networks, etc., which increases the possibility of unauthorized access [9, 10], and causes the need to include high security checks in the monitoring scheme [8–11]. More important is the data integrity, an alteration in the clinical information could lead to an action that threatens the patient's health. For example, in the scheme described in Fig. 1, suppose that in patient records the glucose value is 130 mg/dL (considered normal) and during the transmission the value is modified and instead, the physician receives a value of 330 mg/dL (considered very high value), the doctor could immediately prescribe the increment of the insulin dose to regulate the glucose level. The patient then follows the instruction and the natural consequence would be that the glucose level decreases considerably, producing severe hypoglycemia. Wireless communications can present serious deficiencies to keep information secure, even if the communication protocols implement their own security mechanisms, they do not prevent some types of attacks.

4 Risk Assessment for Telemonitoring Schemes

In order to accomplish the risk assessment, the methodology proposed by NIST [12] was implemented. This methodology was chosen because of its particular focus on information systems, as well as considering the risk analysis systems in its early phases like the requirements analysis. This methodology is summarized in the following steps: system characterization, threat identification, vulnerability identification, control analysis, likelihood determination, impact analysis, risk determination, control recommendations, and results documentation.

The process is performed separately for each of the three environments, this way the specific risks to each environment could be identified. In the patient environment is assumed that the glucometer has been tested and used for commercial purposes and therefore performs the measurements correctly. As observed in Tab. 1, the highest risk comes from an internal attacker (someone with free access to the devices) that can access the application to cause damage. In Tab. 2, the risks associated with the medical environment are very similar to those of the patient environment. The results of the risk assessment in the monitoring center are presented in Tab. 3. Within this environment we considered the web application, the database that concentrates information from all applications as well as physical infrastructure and telecommunications.

Table 1. Results of the risk assessment in the patient environment

Threat	**Action**	**Motivation**	**Vulnerability**	**Risk**	**Probability**	**Impact**	**Risk level**
External	Intercepting and manipulating glucose value	Fun	Insecure Cellular network	Yes	Medium	High	Medium
	Stolen phone	Economic	Patient neglect	Yes	Medium	Low	Low
Internal	Access the application	Cause damage/delete or manipulate information	Weak or no user access control	Yes	High	High	**High**
	Damaging phone	Cause damage	Sensitive material	Yes	High	Medium	Medium
Patient or user	Lost	None	User distracted	Yes	Medium	Low	Low

Table 2. Results of the risk assessment in the monitoring center

Threat	**Action**	**Motivation**	**Vulnerability**	**Risk**	**Probability**	**Impact**	**Risk level**
External	Remove or manipulate records	Fun / damage	Insecure Cellular network	Yes	Medium	High	Medium
	Stolen phone	Economic	Medical neglect	Yes	Medium	Low	Low
Internal	Access the application	Cause damage/delete or manipulate information	Weak or no user control	Yes	High	High	**High**
	Damaging phone	Cause damage	Sensitive material	Yes	High	Medium	Medium
User	Lost	None	User distracted	Yes	Medium	Low	Low
	Wrong message	None	Neglect Physician / poorly designed application	Yes	Medium	High	Medium

In all cases, special emphasis should be given to mechanisms that mitigate high level risks. Using access control techniques facilitates the intervention of only authorized users and devices in the scheme. Proper implementation of cryptographic protocols will maintain information protected against tampering and unauthorized queries. The selection and implementation of these protocols depend on the communication technology used. The elements used in the monitoring scheme (hardware and software) must have high availability as improper operation or lack of response from any of them could lead to a critical situation. To achieve the desired availability the communication devices and means

Table 3. Results of the risk assessment in the medical environment

Threat	Action	Motivation	Vulnerability	Risk	Probability	Impact	Risk level
External	Remove /manipulate records	Fun /damage /economic	Inadequate access controls	Yes	Medium	High	Medium
Internal	Remove /manipulate records	Damage /economic	Inadequate access controls	Yes	High	High	**High**
	Damaging equipment	Fun /revenge	Sensitive material	Yes	Medium	High	Medium
Technique	Equipment failure	None	Bad maintenance	Yes	Medium	Low	Low
	Power failure	None	Lack of backup power	Yes	Medium	High	Medium
	Lost connection	None	Internet	Yes	Medium	Medium	Medium
Natural	Fire	None	Sensitive material	Yes	Low	High	Low
	Flood	None	Location	Yes	Medium	High	Medium

of transmission used should be carefully assessed. In addition, each of the elements of the scheme shall maintain a unmanipulable record of each of the actions taken, thus tampering can be detected and in some cases prevented.

5 Conclusions

Telemonitoring schemes have significant advantages primarily oriented to the welfare of the patient; they also contribute in reducing costs to health institutions as they avoid some of the major complications in the patients health. The security of clinical information should be paramount part in the design of patient telemonitoring schemes, mainly due to the sensitive nature of this information and the implications arising from unauthorized access or alteration. The main objective of the work presented is to get and analyze the results of a risk assesment for a telemonitoring scheme. A case study was used in order to determine general issues for this kind of telemedicine schemes, so that any scheme with similar characteristics can be designed taking into account the results of this risk assesment. It allows to clearly visualize where to focus efforts to implement security mechanisms in telemonitoring schemes. The case study largely reflects recent telemonitoring proposals, so the results of the risk assessment here presented can be considered for telemonitoring schemes in general, and especially those using wireless communications. However, according to the particularities of each scheme, it is recommended to implement the risk assessment to consider all situations in which patient information is at risk.

References

1. Cleland, J.G., Louis, A.A., Rigby, A.S., Janssens, U., Balk, A.: Noninvasive home telemonitoring for patients with heart failure at high risk of recurrent admission and death. The Trans-European Network-Home-Care Management System (TEN-HMS) Study 45(10), 1654–1664 (2005)
2. Franco, K., Cardenas, A., Pulido, P., et al.: Web-based diabetes control. Journal of Telemedicine and Telecare 10(5), 277–281
3. Nyenwe, E., Ashby, S., Tidewell, J., et al.: Improving diabetes care via telemedicine: lessons from the ADT project. Diabetes Care 34, e34 (2011)
4. Sarrafzadeh, M., Sykes, R.: WANDA: an end-to-end solution for tele-monitoring of chronic conditions. International Journal of Integrated Care 12 (June 15, 2012), URN:NBN:NL:UI:10-1-113097 / ijic 2012-117
5. Neumann, C.L., Menne, J., Rieken, E.M., et al.: Blood pressure telemonitoring is useful to achieve blood pressure control in inadequately treated patients with arterial hypertension. J. Hum. Hypertens (January 13, 2011)
6. Hazenberg, C.E., Bus, S.A., Kottink, A.I., et al.: Telemedical home-monitoring of diabetic foot disease using photographic foot imaging a feasibility study. J. Telemed. Telecare 18(1), 32–36 (2012)
7. Tice, J.: Home Telemonitoring for the Management of Heart Failure. California Technology Assessment Forum, San Francisco (2011)
8. Ureta, J., Garbayo, J.: La seguridad, confidencialidad y disponibilidad de la información clínica, Pamplona, pp. 257–286 (2000)
9. Rash, M.C.: Privacy concerns hinder electronic medical records. The Business Journal of the Greater Triad Area (2005)
10. Huang, Q., Yang, X., Li, S.: Identity Authentication and Context Privacy Preservation in Wireless Health Monitoring System. IJCNIS 3(4), 53–60 (2011)
11. Ray, P.: The Need for Technical Solutions for Maintaining the Privacy of HER. In: 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006, pp. 4686–4689 (2006)
12. Stoneburner, G., Goguen, A., Feringa, A.: NIST Special Publication 800-30: Risk Management Guide for Information Systems. Recommendations of the National Institute of Standards and Technology (2002)

Development of an Ontology for Supporting Diagnosis in Psychiatry

Cátia Silva, Goreti Marreiros, and Nuno Silva

GECAD - Knowledge engineering and Decision Support Research Center
School of Engineering. Polytechnic of Porto,
Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
{1070360,mgt,nps}@isep.ipp.pt
http://www.isep.ipp.pt/

Abstract. Psychiatry is the medical specialty responsible for the diagnosis and treatment of mental disorders, in an attempt to secure mental health in the human beings. However, the human brain is a very complex organ, which makes the diagnosis of this kind of illnesses a hard task. Considering this fact, an ontology was conceived with the task of inferring one or more psychiatric disorders depending on the symptoms presented by a patient.

1 Introduction

The human being possesses a very developed brain when compared to others species. Due to its extreme complexity, this organ can be striken by a large number or mental disorders like mood, personality and eating disorders. Psychiatry is the medical specialty that focuses on the diagnostic and treatment of this kind of problems. Unlike other physical disorders, mental problems can be hard do diagnose, considering some aspects like the absence of specific and easily detectable symptoms as well as the subjectivity of the physician responsible for the diagnosis[2].

While the various technical supporting solutions are promising, they proved to be insufficient due to several factors such as the difficulty in automating the decision-making processes or the identification of the various factors that lead to the onset of psychiatric disorders.

In an attempt to mitigate such situations, we argue that ontologies can have an important role in supporting the diagnosis of this type of problems, particularly when using automated reasoners to, given a set of symptoms for a particular patient, obtain a set of possible disorders affecting that individual. Furthermore, it is our conviction that the ontologies will facilitate the description of the diagnosis rational. These hypotheses arise from the claim that ontologies are the best answer for intelligent systems operating close to the human conceptual level[11].

Considering these facts, the goal of this project consists in creating an ontology-based system capable of supporting the psychiatrists' work by facilitating the diagnosis of mental disorders, specifically disorders of bipolar and depressive

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_40,

aetiology. In particular this paper presents the team's efforts to develop a rich-expressive ontology that will be the base for such system either for capturing knowledge expertise, to perform reasoning over the patient's symptoms and for explaining/describing the diagnosis. The authors decided in a first phase to adopt well-established documents on the domain instead of capturing the knowledge from experts (i.e. clinical). Two standard documents describe the psychiatric problems and diagnosis: ICD-10 (International Classification of Diseases, 10th revision) – from Mental Health Organization – and DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, 4th Edition) – from American Psychiatric Association[3,2]. In this work the DSM-IV was adopted as our core source of conceptual knowledge and diagnostic rules. In fact, contrarily to ICD-10, DSM-IV possesses a group of rules and text-based decision trees that describe the diagnosis process, that we believe is expressible in *OWL DL* ontologies.

This document is structured as follows: initially, we will be presenting some information on the state of the art, going out as well as a brief explanation of the rules used by the DSM-IV in the diagnosis of Mental Disorders. The next step consists of the presentation of the ontology development process adopted as well as the writing of the results obtained in this study.

2 State of the Art

Through time, various authors have developed *AI* tools-based systems capable of helping health care professionals in the diagnosis of one or more pathologies in the various medical specialties.

In 2009 Nunes, Pinheiro and Pequeno[10] used DSM-IV as a reference to create a study about some of the *AI* techniques used to support diagnosis of Obsessive-Compulsive Disorder. In particular, these authors focused their efforts on developing a support system for multi-criteria decision (called *MACBETTH* method),which allows the attribution of a grade to an alternative based on pair comparisons; the program is therefore responsible for analysing both the cardinality and semantics and suggesting a way of solving the problem in cases of inconsistence. This is achieved by suggesting a range of grades and intervals susceptible to changes without making the problem inconsistent. In addition to the *MACBETH* method, Nunes and colleagues were based on the *SINTA* expert system[14], which consists of a visual tool for the development of expert systems based on the input rules of the domain. Although these authors use several *AI* tools in order to achieve their goal[10], they also consider that it is insufficient when seeking for fully automated decision-making processes based on multi-criteria analysis.

Later, in 2008, Kola and peers[11] have proposed the development of an ontology-based system capable of representing and storing information regarding to the various psychotic disorders. However, the existence of different data models regarding neuro-scientific research and clinical practice hinders their interoperability, representing an obstacle when considering the closeness between both fields. The main goal of this ontology is to serve as a connection bond in the domain of psychosis although without going into great details in terms of "dimensionality" and "disorder definition". This ontology should be viewed as a logical approach of some psychiatric concepts due to the lack of consensus inside Psychiatry itself.

That same year, and assuming that most of the information related to mental health is computerized, Hadzic, Chen and Dillon[5] proposed the development of a *Mental Health Ontology (MHO)* capable of modelling concepts related to mental health and relationships that could be used to create a semantic structure appropriate to the storage and retrieval of information; this ontology consists of three sub ontologies, each one representing one of three categories: types of disorders, factors and treatments:

1. By collecting information from both ICD-10 and DSM-IV, Hadzic and peers were able to identify thirteen types of Mental Disorders as well as their subclasses and represent them in the sub ontology related to "Types of Disorders"
2. The second sub ontology is responsible for capturing and representing knowledge related to the factors affecting an individual's Mental Health; the authors considered the existence of five types of factors: Genetic, Physical, Environmental, Personal and Microorganic
3. Finally, the third and last sub ontology of *MHO* is related to Treatments like, for example, psychotherapy, pharmacotherapy, group and familiar therapy as well as electroconvulsive therapy

Later, Lin and colleagues[12] proposed an approach which consists of a distribution network throughout the body whose function is to provide vital energy to all parts of the body, strongly linked to physiological functions and the maintenance of health in humans. This solution interacts with human users via a decision tree algorithm in order to obtain the meridians necessary for diagnosis as well as design of a domain ontology (responsible for storing static knowledge on the meridian system) and a task ontology (used to describe the rules of psychiatric diagnosis based on the indicated meridians, as well as for inferring upon those rules – dynamic knowledge). The goal of this work is to infer the task-ontology rules based on the decision tree, regarding case history data from patients and the training data used. Despite resembling with this work, our approach is based on a reference expert-oriented documents such that the rules are much more complicated than the rules inferred in [12] (mostly based on scalar values).

3 Classification of Mood Disorders According to DSM-IV

In DSM-IV, the section related to Mood Disorders includes a number of diseases that are divided into three distinct parts: Mood Episodes, Mood Disorders and Specifiers relating to disorders of this aetiology that describe either the most recent episode or the course of recurrent episodes.

DMS-IV's Mood Episodes are sets of symptoms for a particular patient and they cannot be considered actual disorders. However, it can be stated that a given patient having all symptoms and characteristics of that Mood Episode can be diagnosed as having that Mood Episode. The DSM-IV defines four types of Mood Episodes: Major Depressive, Manic, Hypomanic and Mixed Episodes[2]. In turn, a Mood Disorder can be seen as a set of Mood Episodes (and other isolated symptoms) while specifiers characterize a particular Mood episode/disorder, such as the type or severity.

Considering the Dysthymic disorder (defined and characterized in DSM-IV), it can be assumed that it is a type of Depressive Disorder where the patient presents with depressed mood for at least two years. Additionally, Dysthymic disorder features by a minimum of two symptoms presented in the following listing:

- Increased or decreased appetite
- Insomnia or hypersomnia
- Fatigue or low energy
- Low self-esteem
- Difficulty concentrating and decision making
- Feelings of hopelessness

Concomitantly, the patient must have the following symptoms:

- Lack of any Major Depressive Episodes in the last two years and of any other Mood Episode throughout the course of the disorder
- Present severe difficulties or impairment in work, social and other daily-life activities
- Lack of association of symptoms with substance abuse, Cyclothymic Disorder or Chronic Psychotic Disorder
- Symptoms are not caused by a Chronic Major Depressive Disorder or a Major Depressive Disorder in partial remission[2]

The definition of the Dysthymic disorder is dependent on the definition of other disorders and vice-versa, eventually, which causes considerable definition and reasoning difficulties.

As mentioned earlier, the aim of the project is to create an ontology for supporting the task of diagnosing mental disorders afflicting the patient, taking into account the information entered by the user (like the physician, the patient or other) on the symptoms (and their characteristics such as the duration, stage, nature, etc.). It is then intended that the episodes and disorders captured in the ontology/KB give rise to the more specific diagnosis.

4 Ontology Development

In Computer Science, ontologies are engineering artefacts capable of formally modelling a system‘s structure, i.e., the most relevant entities and relations considered useful for the modelling and axioms that supply interpretation restrictions and reasoning capabilities to reasoning engines. This section describes the team's effort developing the ontology following the incremental and iterative process *KMP – Knowledge Meta Process*[1], which consists of five basic steps, which will be described in the following subsections as adopted by the team for this ontology.

4.1 System's Economic and Technical Feasibility Study

In this project, the feasibility study consisted in a critical analysis of the rules used by the DSM-IV to make the diagnosis of mood episodes and disorders. Although the analysis of these rules have served as a warning to the impossibility of achieving the initial goals of the project, it played an important role in the developed work since it allowed the identification of several primitive classes which make up the ontology; these classes have served as the basis for the organization of knowledge captured during the reading of the document.

4.2 Kick-off

During the Kick-off phase the team identified and developed the primitive classes and the basic hierarchy of the ontology. Since its main purpose is to diagnose a certain disorder that affects a certain patient (considering the information entered by the user about both the symptoms and their characteristics presented by this same patient), one of the main characteristics of this ontology is the reclassification of mood episodes and disorders considering the input data.

In order to represent knowledge related about each of the Bipolar and Depressive disorders represented in DSM-IV, the ontology conceived in this project followed a basic hierarchy of classes that will be explained below:

- Patient - Represents an individual patient in consultation
- Symptom - Represents all aspects of the symptoms presented by the patient, including the duration of symptoms, the four types of mood episodes identified by the DSM-IV[2] and its aetiology
- Affective Disorder - Identifies and characterizes the various conditions considered by the DSM - IV
- Cause -Identifies the causes of the symptoms, like other medical condition or health problem or the abuse of substances
- Consequence - Identifies the consequences of symptoms, which could consist of social distress or impairment, hospitalization or when symptoms are notorious to others
- Feature - Consist of the specifiers defined by DSM-IV[2]
- Stage - identifies the stage of the disorder (e.g. Chronic or In Remission)
- Severity - Mild, Moderate or Severe (according to DSM-IV[2])

In addition to the above classes, eight properties have been create in order to design rules for defining some of these classes (especially classes related to disorders and episodes). Figure 1 tries to explain the reader a better understanding of these rules.

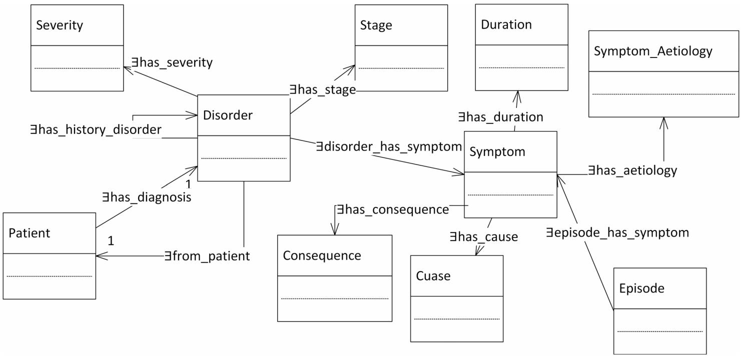


Fig. 1. Ontology primitive classes and their relationships

Unlike what usually happens in the Kick-off phase of the *KMP*, the *ORSD (Ontology Requirements Specification Document)* was not formally created as the goals of the ontology had been previously defined i.e., the capture of decision trees presented in the ontology for future inference of diagnostics.

4.3 Refinement

During the refinement of the ontology we chose to use a top-down approach in the creation of subclasses of primitive classes. Such approach allowed starting from generalized concepts and refine them in order to obtain a class hierarchy increasingly closer to the real needs of the ontology. Figure 2 tries to explain how this refinement was made.

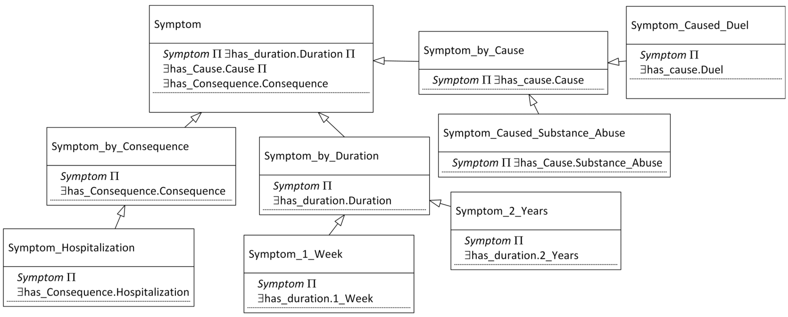


Fig. 2. Example of refinement

Although the vast majority of diagnostic rules explained in DSM-IV were captured and formalized in *OWL DL* considered in early iterations, during the evaluation process the team noticed that some of them were not representable in *OWL*, leading to the adoption of *SWRL*[13]. One of the rules is the following:

Rule 1: Distúrbio(?d), Paciente(?p), de paciente(?d, ?p), paciente tem sintoma(?p, ?s) $\rightarrow$ *distúrbio tem sintoma(?d, ?s)*

This can be read as follows: given a specific disorder d and a patient p, if it is stated that the patient p has symptom s, then it is inferred that the disorder d will also have the symptom s.

5 Evaluation

According to the *KMP* the ontology evaluation is done at three levels: syntactic (related to the linguistic conformity of the ontology properties), semantic (related to the consistency of the ontology, i.e., how it captures the rules represented in DSM-IV's decision trees) and of technological properties (e.g. interoperability or scalability).

We assumed that the testing and debugging of ontology designed in this project fall under the evaluation phase of the *KMP*. In order to test and debug the ontology, several individuals were created in a bottom-up approach. This approach allows the testing of the rules previously defined, allowing early detection of both syntactic and semantic inconsistencies in the ontology. The evaluation was repeated for all the created individuals in order to test all the classes contained in the ontology, which allowed the development of a both semantic and syntactically valid ontology according to the DSM–IV.

6 Conclusions and Future Work

The work described in this paper seeks, in some way to help health professionals in their arduous task of making accurate diagnoses in clinical Psychiatry since it can be useful in suggesting the diagnosis to be made to a particular patient. However, the diagnosis rules adopted in DSM-IV are relatively complex and difficult to formalize in *OWL DL* and *SWRL*. Currently the team doubts about the causes since we are aware of the difficulties in understanding/interpreting certain text-driven ambiguities of the rules, which would have been prevented by a domain-expert participation in the ontology development process.

The evaluation phase allowed concluding that the expressivity capabilities of *OWL* and *SWRL* associated with standard reasoners are sufficient for formalizing and reasoning upon one patient's data. Yet, early efforts for validation and verification of the ontology by domain experts (i.e. psychiatrist) were not successful. This was especially due to the development environment adopted, which inhibits a non-technical stakeholder to full-understand the semantically rich ontology, preventing a full-commitment with the verification and validation of the ontology.

Despite the near future team's efforts will be focused in the domain-expert verification and validation of the ontology, the efforts will not address the development of a domain-expert-oriented user-interface, but the validation and verification with minimal expert intervention. For that, real patient's data will be adopted in the evaluation phase so that the reasoner's inferences (i.e. the diagnosis) can be compared with the psychiatrist' diagnosis. This will allow the revision and refinement of the DSM-IV-based ontology's rules based on the psychiatrist's rules, thus focusing in the core of the system.

Acknowledgments. This work is supported by FEDER Funds through the "Programa Operacional Factores de Competitividade - COMPETE" program and by National Funds through FCT "Fundação para a Ciência e a Tecnologia" under the project: FCOMP-01-0124-FEDER-PEst-OE/EEI/UI0760/2011 and project AAL4ALL (QREN13852).

References

1. Sure, Y., Staab, S., Studer, R.: Methodology for Development and Employment of Ontology based Knowledge Management Applications. ACM SIGMOD Record 31, 18–23 (2002)
2. Diagnostic And Statistidcal Manual of Mental Disorders, 4th edn. American Psychiatric Association, Washington, DC (2000)
3. Mental Health Organization, `http://www.who.int/en/`
4. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. KSL 92-71 (January 1993)
5. Hadzic, M., Chen, M., Dillon, T.: Towards the Mental Health Ontology. IEEE International Conference on Bioinformatics and Biomedicine (2008) 978-0-7695-3452-7/08
6. Lopez, A.D., Murray, C.C.J.L.: The global burden of disease, 1990 2020. Nature America Inc. (November 1998)
7. Orbst, L., Liu, H., Wray, R.: Ontologies for C orporate Web Applications. AI Magazine 24(3) (2003)
8. Russell, S.J., Norvig, P.: Artificial Intelligence A Modern Approach. Prentice Hall, Englewood Clis (1995)
9. Staab, S., Studer, R.: Handbook on Ontologies, 2nd edn. Springer (2009)
10. Nunes, L.C., Pinheiro, P.R., Pequeno, T.C.: An Expert System Applied to the Diagnosis of Psychological Disorders. Av. Washington Soares, 1321 - Bloco J, sala 30, CEP: 60811-905. University of Fortaleza, Fortaleza (2009) 978-1-4244-4738-11/09/
11. Kola, J.S., Harris, J., Laurie, S., Rector, A., Goble, C., Martone, M.: Towards an ontology for Psychosis. Cognitive Systems Research 11, 4252 (2011)
12. Lin, Y.C.: Design and Implementation of an Ontology-Based Psychiatric Disorder Detection System. WSEAS Transactions on Information Science and Applications, 1790–0832 (January 2010)
13. Consortium WWW. SWRL: A Semantic Web Rule Language Combining OWL and RuleML (2004), `http://www.w3.org/Submission/SWRL/`
14. Nogueira, J.H.M., Alcaântara, J.F.L., de Andrade, R.C., de A.e Silva, R.B., Silvestre, R.S.: Grupo SINTA - Sistemas INTeligentes Aplicados. Federal University of Cearaá (UFC) and State University of Cearaá, UECE (1995)
15. Shang, C.: The past, present and future of meridian system research. Clinical Acupuncture and Oriental Medicine 2(1), 115–124 (2000)

Augmented Reality Sign Language Teaching Model for Deaf Children

Jorge Jonathan Cadeñanes Garnica and María Angélica González Arrieta

Department of Computer Science and Automation, Faculty of Science,
University of Salamanca, Spain
{cadenanes,angelica}@usal.es

Abstract. This article describes a Sign Language Teaching Model (SLTM) designed to develop on deaf children different Communication Skills (CS) within a Collaborative Learning Environment with Mixed-Reality (CLEMR). A pilot lesson with the Fingerspelling Alphabet was conducted at the Association of Parents of Deaf Children of Salamanca to determine the Percentage of Development of the Sign Language Communication Skill (SLCS) and others by using a kit of Pedagogical Materials as complementary teaching resources.

Keywords: Augmented Reality, Sign Language Communication Skill, Mixed-Reality Learning Environment, Pedagogical Materials.

1 Introduction

The use of Information and Communication Technologies (ICT) such as the Augmented Reality (AR) has proved to be a successful educational resource for teaching SL topics at primary education [1,17]. In a Teaching/Learning Process it promotes the collaboration between students and teachers to the construction of knowledge within a Collaborative Learning Environment with Mixed-Reality (CLEMR) [14]. To teach SL (in preschool and primary education) it is essential to identify the diverse learning needs of deaf children in order to improve their school achievement and social integration [20]. To do this, we propose a Sign Language Teaching Model (SLTM) and a kit of Sign Language Pedagogical Materials (SLPMs) as complementary teaching resources.

The article is presented as follows: In Section 2, we mention the projects we considered are examples of how AR technology is adapted to educational purposes. In Section 3, we describe the SLTM and SLPMs we are proposing to develop SLCS and other Communication Skills (CS) on deaf children. We also present the instrument used to validate them. In Section 4, we describe the process of a pilot lesson with the Fingerspelling Alphabet at the Association of Parents of Deaf Children of Salamanca (ASPAS) [2]. On section 5, we draw some conclusions and on Section 6 we mention the future research.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_41,

2 Related Works

An AR educational project proposed by Billinghurst et al. [4,5,6,7] called MagicBook covers a CLEMR in a multi-level learning experience. These three levels of collaboration are the Reality, AR and Virtual Reality (VR). On the first level, users can learn with the MagicBook as a physical object that exists in the real world with learners. They can gather around the book and interact together as peers. Also, the readers can explore the book whilst they are sharing their learning experiences in order to comprehend better the topics presented.

On the second level, a desktop application installed on networked PCs allows the visualization of animated avatars and digital environments projected on the PCs screens or AR displays. This level of collaboration allows smooth transitions from Reality to AR or vice versa. The users maintain contact with both realities while they interact with digital objects. On the last level, the users' interactions are no longer in the real world. Everything happens within computer-generated environments. Multiple users wearing AR devices can see each other as avatars. Therefore, their interactions are produced by sharing only digital information while completely immersed within a virtual environment.

With the MagicBook it is possible to learn different topics (previously established by the teachers) by moving between realities. With this technology the students can choose a learning experience from Reality to VR, depending on the skills or information (knowledge) they want to learn or experience.

The interactive desktop tool Realitat3 [17] improves learning and teaching processes by customizing several teaching subjects accordingly with thematic units for primary education. This AR program facilitates the teachers' explanation in classrooms as the same time it promotes the students' interest in learning something new through technology. The interface can be updated with new themes when needed. Their developed AR system provides a basic functionality with menus that allows the tracking of AR markers. Therefore, users can learn by exploring the 3D objects or by playing animations.

3 Sign Language Teaching Model and the Kit of Pedagogical Materials

The SLTM we are proposing is called: Multi-language Cycle for Sign Language Understanding (MuCy) (Fig.1). To design it we are founded on a theoretical background which encompasses the Zone of Proximal Development (ZPD) of Lev's Semionovich Vigotsky [9,12]; the Principles of Social Education for deaf and dumb children [20] and the Reality-Virtuality Continuum proposed by Paul Milgram [15]. The model is also supported by research in neuropsychology which has shown that deaf children could develop good reading and speaking skills by learning them at an early age [13].

The SLTM MuCy is cyclical to the extent that skills gained from knowledge are implemented to communicate with others in different communication systems such as making signs, reading, writing or speaking. Since deaf people develop different levels of Communication Skills they need an outline that helps them to develop a SLCS and other CS mentioned before through interaction with people, technology and information. That is possible when school teachers promote the CLEMR.

We present two psychomotor SL educational levels (Fig.1). The first teaches the proper use of signs with respect to their visual references (words or phrases) and their written words (the action to write down on paper the meaning of those words). The second level enables the development of speech. Hence deaf children can acquire the ability to express themselves with different communication systems when the SL system is not suitable for some social situations. Finally, the concept of CLEMR refers to the use of ICT like AR into social learning interactions. In brief, the CLEMR is similar to Vigotsky's ZPD in the sense that children learn from each other. The MuCy model aims to be a reference point on which deaf children can learn SL in an easy way. Hence, they will be more confident by expressing their ideas socially.

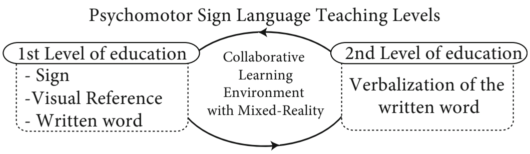


Fig. 1. Multi-language Cycle for Sign Language Understanding (MuCy model)

We are presenting a kit of SLPMs and a desktop AR application created with the Cross-platform Unity3D [18]. These educational materials (Fig.2) were designed in order to offer different SL topics for deaf and non-deaf students. Its design allows deaf students to learn in a logical and sequential way to make signs, read, write and speak.

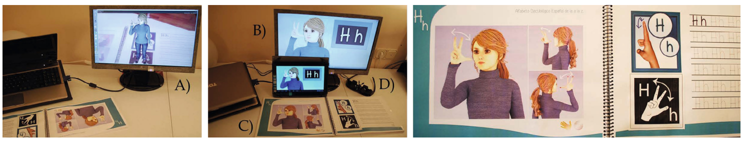


Fig. 2. Pedagogical materials as learning interfaces based on the VR-Continuum. A) An AR avatar making signs in Unity3D, B) Animations displayed on PC screen and Tablet, C) SL Book with sections for reading and writing exercises, D) Vuzix AR eyewear display.

The first SLPM is a *SL Fingerspelling Alphabet book*. It serves as a tangible interface that allows the use of the Vuforia marker-based tracking system for Unity3D [19]. This SL book facilitates the understanding of performing signs as well as the development of reading and writing skills. Words from A to Z are shown in signs printed on the SL book, as well as images of avatars performing signs with their corresponding words. The use of text to teach meanings makes possible to practice reading and writing skills as well.

For the Fingerspelling Alphabet pilot lesson we had created a set of 30 videos that can be watched on either PC screens, Tablets or AR displays. All these, are the intermediate interfaces between Reality and digital contents. With these *videos* as the second SLPMs children can learn to make signs by imitation. The learning experience can also be shared with other students around the screens in a CLEMR [16]. The last SLPM is the *Unity3D AR application*. We used Blender 2.69 [3] to model and animate the avatars. The Cross-platform game engine Unity3D facilitated us the creation of the virtual scenes on which these avatars are displayed. The file format used was the FBX that allows to display 3D objects with high definition. The Unity3D application can be adapted to be used in two modalities: With PC screens or with Head Mounted Displays (HMD) such as Vuzix [21].

To evaluate the SLTM MuCy and the SLPMs we used a Likert's scale survey of five points to measure usability, satisfaction and learning achievement. Two teachers from ASPAS and two parents of the children how participated in this Fingespelling Alphabet pilot lesson answered the survey. To design the questions (Table 1) in accordance with our teaching approach, we took into account the most relevant elements of the Principles of Learning and Teaching P-12 [11] and the Danielson's Group Framework for Teaching [10].

Table 1. Likert's scale survey to validate the MuCy model and the SLPMs

i	Question	Mean	Std.Dev.	%
Q1	The SLPMs help deaf children to remember information through memorization.	4.50	.707	90%
Q2	The two educational levels of the MuCy model help deaf students to cognitively understand relevant information from the SL.	5.00	.000	100%
Q3	Teaching Communication Skills such as reading, writing and speaking help students to create solutions to the socio-cultural problems they face.	5.00	.000	100%
Q4	Learning with a CLEMR helps deaf students to understand a complex situation in parts in order to create diverse learning solutions.	4.50	.707	90%
Q5	Learning with interactive technology helps deaf children increase their learning achievement.	5.00	.000	100%
Q6	I would like to use these pedagogical materials as complementary teaching resources either at home or at school.	5.00	.000	100%
Q7	The MuCy model helps deaf children to organize their learning process according to their educational needs.	4.50	.707	90%
Q8	With these pedagogical materials it is easier to explain the SL positions to the children.	4.50	.707	90%
Q9	Learning with AR avatars increases the interest in speech and makes the children feel more confident that they will learn to speak.	5.00	.000	100%
Q10	The SL book is an adequate tool for teaching the reading and writing for an specific topic.	4.50	.707	90%
Q11	The SLTM MuCy and the SLPMs promote collaborative learning experiences.	5.00	.000	100%
Q12	The SL lessons are more useful and interesting with AR avatars.	5.00	.000	100%

4 Fingerspelling Alphabet Pilot Lesson

The Fingerspelling Alphabet lesson was conducted in order to measure the Percentage of Development of SLCS and other CS reached by two deaf children (Fig.3A). The lesson had a duration of one hour with students located in the same classroom. The two students aged six and seven attended the lesson together with the materials within a CLEMR.

For the lesson we made 30 videos of words from A to Z and the duration of each video was approximately 6 seconds. Hence, for every minute, each student watched and imitated an average of 8 to 10 SL positions. The lesson was divided into four activities, each of them corresponding to a specific SLPMs (Table 2).

Table 2. SL Repetitions, correct answers and total scores in one-hour lesson for the Fingerspelling Alphabet

Fingerspelling Alphabet lesson

Activity	T (mins)	SL Reps. Goal	Session SL Reps.				Xi	Yi	Percent	Score
			MS	RD	WR	SP				
1	20	130	95	0	0	0	95	86	90.53%	9.1
2	10	65	0	58	0	0	58	51	87.93%	8.8
3	10	50	0	0	42	0	42	38	90.48%	9.0
4	20	110	0	0	0	95	95	89	93.68%	9.4
Total	**60**	**355**	**95**	**58**	**42**	**95**	**290**	**264**	**91.03%**	**9.1**
Mean Value	15	88.75	24	15	11	24	72.50	66	**90.65%**	**9.07**
Std. Dev										**0.24**

Activity one (Making Sings: MS) consisted of animated videos having children watch together the avatars performing signs on the PC screen, then they had to imitate the SL positions right after the avatars. For Activities two (Reading: RD) and three (Writing: WR), the children had to use the SL book to practice reading the words of the Alphabet. Next, they had to write those words down on the book, performing corresponding SL positions immediately after writing a word. At the last Activity (Speaking: SP), the children first had to use the

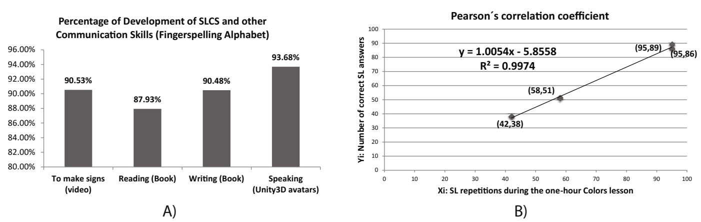


Fig. 3. A) Percentage of Development of SLCS and other Communication Skills, B) Correlation analysis between the SL repetitions during the one-hour Fingerspelling Alphabet lesson and the students' correct SL answers

markers printed on the pages to display the animated avatars on the PC screen, then teachers taught the students to move their lips and tongues to reproduce sounds and to practice speech.

5 Conclusions

In this paper we have presented a SLTM called Multi-language Cycle for Sign Language Understanding (MuCy) and a Kit of Pedagogical Materials to teach SL to deaf children at ASPAS. A Fingerspelling Alphabet pilot lesson within a CLEMR was conducted in order to determine the Percentage of Development of the SLCS and other CS of two deaf children aged six and seven.

It also has been proved that by using ICT technology such as AR avatars increases child interest on speech (Table 1, Q9 and Q12). By using the MuCy model and the SLPMs deaf children developed the SLCS with a 90.53% improvement, for writing skills a 90.48% improvement, for reading skills a 87.93% improvement, and for speaking skills a 93.68% improvement (Fig.3A). The Total Mean Value of the two students' SL Correct Answers was 90.65% development (Table 2).

Mindful of psychomotor relationships between knowledge and communication, it is observed that there is a strong correlation coefficient of 0.99% (Fig.3B) between the SL repetitions from the one-hour Fingerspelling Alphabet lesson and the number of Correct Answers given by the children. It is established that the more they practice SL positions (reading, writing and speaking through the SLPMs) the more they learn to communicate. With cognitive relationships between knowledge and communication, deaf people are able to understand SL concepts through imitation, by watching images or by interacting with technological interfaces (either tangible or digital).

The main contributions of this SLTM are: Firstly, it promotes the development of Communication Skills such as making signs, writing, reading and speaking. Secondly, it establishes the SLPMs which can be used as complementary teachings resources. And finally, it allows deaf children to acquire knowledge within a CLEMR. SL teachers have on their hands a complementary SLTM which ensures the full understanding of concepts, meanings or ideas in accordance to their SL study programs.

6 Future Research

The next step to improve our proposed MuCy model is to keep testing it by conducting another SL pilot lesson with the Rainbow Colors at ASPAS and other deaf people associations and schools. With this, we will be able to make data comparisons between the Alphabet and the Colors lessons in order to have a finalized and tested SLTM and SLPMs with at least two different topics. The core idea is to gradually add more complex SL topics.

Microsoft Research China has demonstrated recognition in real time SL movements performed by a person facing the Kinect camera [8]. Therefore, we also plan to create one more SLPM based on Blender and OpenKinect camera for capture the motion of face, hands and body. So there will be a SL Networked-application that will reproduce and translate sounds, text and images remotely and in real time using Augmented Reality.

Acknowledgments. The authors would like to thank to ASPAS for all the support provided and for the opportunity to facilitate to their students the development of SLCS and other CS with a SLTM and its SLPMs.

References

1. Adamo-villani, N., Carpenter, E., Arns, L.: 3D Sign Language Mathematics in Immersive Environment. In: Proc. of ASM 2006 - 15th International Conference on Applied Simulation and Modeling, Rhodes, Greece (2006)
2. Association of Parents of Deaf Children of Salamanca, ASPAS, Spain (2014), `http://www.aspas-salamanca.es/` (viewed on July 6, 2013)
3. Blender Foundation, Blender 2.69 (2014), `http://www.blender.org/` (viewed on September 1, 2013)
4. Billinghurst, M., Kato, H., Poupyrev, I.: The Magic Book: A Transitional AR Interface. Computers and Graphics 25(5), 745–753 (2001)
5. Billinghurst, M., Kato, H., Poupyrev, I.: The MagicBook. Moving Seamlessly between Reality and Virtuality. Human Interface Technology Laboratory, University of Washington, Hiroshima City University and Sony Computer Science Laboratories (2001)
6. Billinghurst, M., Kato, H., Poupyrev, I.: MagicBook: Transitioning between Reality and Virtuality. In: Proceeding Extended Abstracts on Human Factors in Computing Systems, New York, NY, pp. 25–26 (2001)
7. Billinghurst, M., Kato, H., Poupyrev, I.: Collaboration with tangible Augmented Reality Interfaces. Human Interface Technology Laboratory, University of Washington, Hiroshima City University and Sony Computer Science Laboratories (2002)
8. Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X.: Sign Language Recognition and Translation with Kinect. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS). Institute of Computing Technology, Microsoft Research Asia, Beijing, China (2013)
9. Chaiklin, S.: Vygotsky's educational theory and practice in cultural context. The zone of proximal development in Vygotsky's analysis of learning and instruction. Cambridge University Press (2003)
10. Danielson, C.: The framework for teaching. Evaluation instrument. The Danielson Group (2013), `http://www.danielsongroup.org` (Viewed on December 1, 2013)
11. Department of Education and Early Childhood Development. The Principles of Learning and Teaching P–12 Unpacked (2014), `http://www.education.vic.gov.au` (Viewed on January 10, 2014)
12. Ivic, I.: Lev Semionovich Vygotsky. UNESCO 24(3-4), 773–799 (1994)
13. Mayberry, R.I.: Cognitive development in deaf children: the interface of language and perception in neuropsychology. In: Segalowitz, S.J., Rapin, I. (eds.) Handbook of Neuropsychology, Part. II, 2nd edn., vol. 8 (2002)

14. Mertzani, M.: Considering Sign Language and Deaf Culture in Computer Mediated Communication Environments: Initial Explorations and Concerns. In: 9th Theoretical Issues in Sign Language Research Conference, Florianopolis, Brazil (2008)
15. Milgram, P., Takemura, H., Utsumi, A., Kishino, F.: Augmented Reality: A class of displays on the reality-virtuality continuum. In: SPIE, vol. 2351. ATR Communication Systems Research Laboratories, Kyoto (1994)
16. Pemberton, L., Winter, M.: Collaborative Augmented Reality in Schools. In: Proc. CSCL 2009, vol. 2, pp. 109–111 (2009)
17. Salvador-Herranz, G., Pérez-López, D., Ortega, M., Soto, E., Alcañiz, M., Contero, M.: Manipulating Virtual Objects with your hands: A case study on applying Desktop Augmented Reality at the Primary School. In: 46th Hawaii International Conference on System Sciences (2013)
18. Unity Technologies, Unity3D V4.3 (2014), `http://unity3d.com/unity` (viewed on October 5, 2013)
19. Vuforia Developer, Vuforia™SDK, Unity extension Vuforia–2.8 (2014), `https://developer.vuforia.com/resources/sdk/unity` (viewed on December 22, 2013)
20. Vygotsky, L.: The principles of social education of deaf and dumb children in Russia. In: Proceedings of the International Conference on the Education of the Deaf, London, pp. 227–237 (1925)
21. Vuzix Corporation (2014). Wrap 920AR and 1200DXAR Eyewear (2014), `http://www.vuzix.com/` (viewed on December 22, 2013)

A Multi-agent Simulation: The Case of Physical Activity and Childhood Obesity

Rabia Aziza[1], Amel Borgi[1,2], Hayfa Zgaya[3], and Benjamin Guinhouya[3]

[1] LIPAH research laboratory, Tunis-El Manar University,
El Manar 2092, Tunis, Tunisia
[2] High Institute of Computing/ISI, 2 Rue Abourraihan Al Bayrouni, Ariana, Tunisia
rabia.aziza@gmail.com, amel.borgi@insat.rnu.tn
[3] EA 2994, Public Health: Epidemiology and Healthcare Quality,
Faculty for Health engineering and management/ILIS, University Lille II,
42 rue Ambroise Paré 59120 – LOOS, Lille, France
{hayfa.zgaya,benjamin.guinhouya}@univ-lille2.fr

Abstract. Engaging in a regular physical activity appears to be an important factor in the prevention of childhood obesity, which became one of the major public health challenges worldwide. The literature suggests that the relationship between physical activity and obesity is complex with many intervening factors that come from different aspects of the child's life. Yet, so far, the proposed models do not include all of the identified factors. The main objective of this study is to simulate the child's behavior within his/her social and physical environments in order to understand precisely the relationship between the PA and childhood obesity. This paper proposes a simulation model using the multi-agent paradigm.

Keywords: Complex Systems, Simulation, Multi-Agent Systems, Epidemiology, Childhood Obesity, Physical Activity.

1 Introduction

Multi-Agent Systems (MAS) are widely used in simulating Complex Systems (CS). They create a virtual laboratory that imitates the original context and help understand the complex behaviors and phenomena that occur in a CS, such as emergence, stigmergy, adaptability, etc. [1]. They also help predict for a better decision-making by simulating virtual experiments where scenarios can be tested. Moreover, the MAS paradigm obviates the temporal dimension by simulating long periods in matter of seconds, and helps avoid the costs and effects of these tests on the real system [2]. For example, in the epidemiological field, conducting field surveys can be very expensive, especially if interventions need to be tested. A public health intervention is a set of actions that involves a group of people (in our context of study, it would involve children and their families). It intends to make changes in order to prevent or treat a specific disease [3]. A MAS that models a virtual field (as faithfully as possible to the

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_42, © Springer International Publishing Switzerland 2014

real field) would facilitate understanding the CS's often non linear behavior, and eventually, simulate interventions, predict their results and propose the best strategies to decision makers. Our research lays in the context of childhood epidemiology. It aims to portray the complex relationships between physical activity (PA) and childhood obesity. We use MAS to simulate the daily life of schooled children. This model would take into account the complex interactions between obesity and PA (and related behaviors such as sedentary behavior) while including the large number of factors involved (social, environmental, mood, health, etc.) [4].

The paper is structured as follows. In Section 2, we explain our application context and explain the choice of MAS as paradigm for our simulation. The following section proposes a multi-level MAS architecture and discusses its different components in Section 3. Finally, we conclude with a glance on future work in Section 4.

2 PA and Childhood Obesity in the Literature

2.1 The Relationship between PA and Childhood Obesity

At the beginning of 21st century, awareness of the extent of obesity and its strong negative impact on individual's health and healthcare systems has increased. In fact, it became one of the major public health challenges worldwide, and the WHO even uses the term "*Globesity*" to describe it [5]. Engaging in a regular PA appears to be an important factor in the prevention of obesity and several other chronic non-communicable diseases [Fig.1]. In his book [6], Guinhouya states that: "*the implementation of quality effective interventions first requires a better understanding of the natural history of common forms of obesity in relation to physical inactivity… In the context of children and adolescents, assessing the PA behavior is particularly difficult*". For those reasons, this study is an attempt to model the multifactorial link between PA and obesity for 2 to 18 years old children.

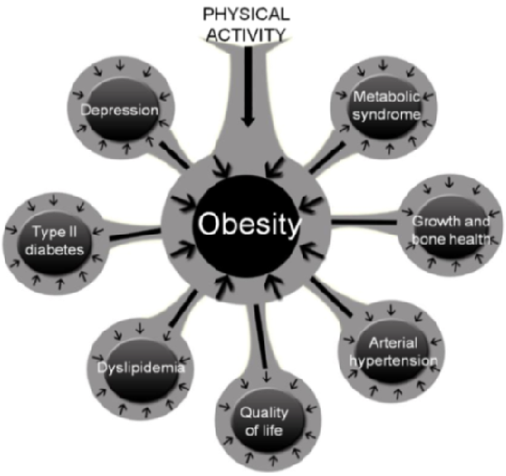


Fig. 1. Obesity as inducer of several diseases and the influence of PA on these diseases through its action on infantile obesity [6]

2.2 Overview of Studies on Modeling PA and Obesity

In the literature, most of the data that fall within the context of our research are epidemiological studies that focus on a subset of the predisposing factors to PA, obesity and/or sedentary behavior (such as depression, self-esteem, neighborhood security, etc.) [7, 8]. The other studies propose mathematical models (mainly based on differential equations and stochastic processes) [9–11], and systemic models (e.g., rule based systems, MAS, etc.) [12, 13]. We are interested in studies that include and examine the relationships between obesity and PA in children. For instance, some authors [9, 11] propose mathematical models that displayed obesity as a social phenomenon, and – among other factors – include PA in their modeling strategy. Dangerfield et al. [10] also proposed a mathematical model for obesity. They have targeted 2-15 year old children, and included both PA and energy intake.

In fact, systems that focus only on either obesity [12] or PA [13] help us understand each one of them, but do not allow an understanding of the complex relationship between them. On the same subject, Guinhouya [6] argued that most interventions aiming to prevent infantile obesity by improving the PA behavior do not take into account all dimensions interfering with the matter, namely, the child's social and physical environments, his/her psychological state, etc. In [14], the authors classified these factors as: demographic, psychosocial, physiological, biological and genetic, factors in the physical and social environments, economic and social status, and motor skills. Each one of these categories is a set of factors. Therefore, we can consider factors and their relationships as a graph with complex interactions [10, 11, 14].

This study proposes a model that allows including all the mentioned factors and their relations from a complexity theory point of view. In fact, a CS's behavior is guided by its details rather than general laws: all agents (in our MAS, agents represent persons) contribute to its functioning through micro movements [15]. Because of that, it is important to model the agents' micro-movements and inner state details.

2.3 MAS for Modeling the Relationship between PA and Childhood Obesity

The complex relationship between PA and childhood obesity depends on the different factors discussed in the previous section. These factors come from different aspects of a person's life: his/her social life, physical environment, feelings, physical health state, etc. Modeling this complex relationship needs to take into account the autonomy of persons, their links with their environments (physical and social), their behavior's heterogeneity, and their cognitive mechanisms. As the MAS paradigm considers all these aspects and levels, we chose it to simulate children's behavior within their environment. In fact, the MAS paradigm represents physical and social environments without limiting the agent's autonomy (his independence of physical and social environments, unlike Cellular Automata [16] for example). Besides that, Bonabeau [17] stated that MAS paradigm is the only one facing the challenge of modeling social systems, because it can take into account, among other things, human behavior, complex reasoning and psychological factors. In addition, agents in a MAS evolve based on a complex set of rules that can differ from one agent to the other.

This allows agents to have heterogeneous behaviors [18]. In order to simulate the agent's micro movements, we focus on the schooled children's environment, and we model their daily activity behaviors.

3 The Proposed Multi-level MAS Architecture

Each individual evolves within a physical environment and communicates with his/her social network while updating his/her inner state. The proposed architecture [Fig. 2] is based on two levels: the environmental level and the intra-agent level.

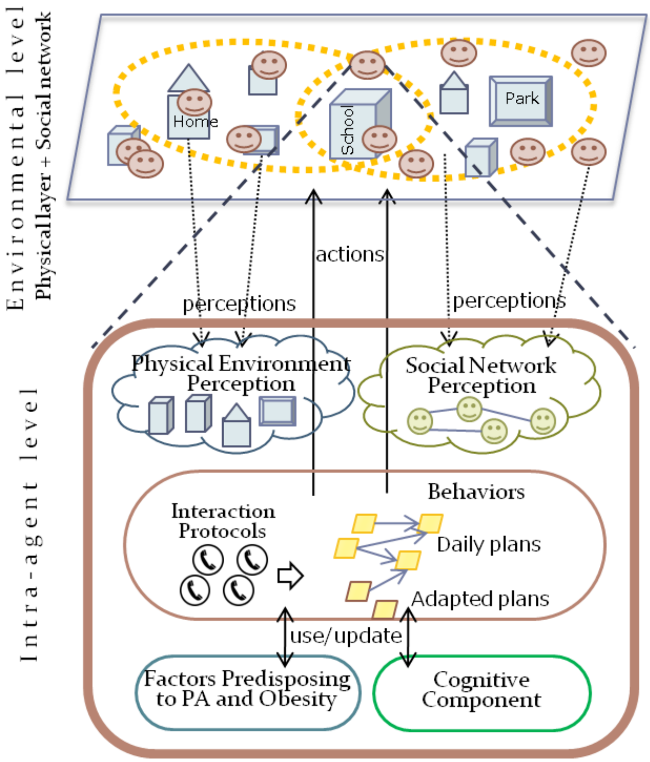


Fig. 2. The proposed multi-level MAS architecture

3.1 Environmental Level

Physical Environment. This layer is composed of neighborhoods. Each neighborhood contains a set of physical components (e.g., home, school, park, path, etc.) simulated by a set of classes. A physical component can offer simple activities (e.g., play, study, eat, sleep, etc) and/or traveling activities (e.g., take the bus, walk, drive by car, etc.). Each activity represents an atomic task an individual can do. It describes:

- a set of conditions for individuals, such as the conditions they need to satisfy in order to practice that activity (age, parental permission required, etc.),
- a list of individuals that are currently practicing this activity,
- the effects that this activity applies on these individuals. For example: altering the individual's preferences, his/her mood and perception of the agents he/she contacted during the activity, and the current physical component. One of the challenges of this research is to accurately specify the effects of each activity.

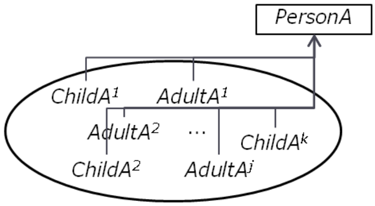


Fig. 3. Agents in the social network layer

Social Network. The social layer contains a set of individuals simulated by person agents (*PersonA*). These agents can be either children (*ChildA*), or adults (*AdultA*). *ChildA* and *AdultA* inherit from *PersonA* [Fig. 3]. The social network reflects relationships between persons and the roles they hold in these relationships. It can be described as a global variable. For instance, this variable would keep a graph of all social relationships in the MAS, and the role each person plays in a relationship. This would suggest that the higher level (system level detaining the variable) mandates rules/laws on the lower lever (*PersonA*). With such an approach, the agent's behavior is boxed and the possible actions he can make would be limited. According to CS modeling and MAS paradigm, relationships between agents should be managed and owned by individuals themselves. Therefore, we dispatch the social network across persons: relationships are represented within every *PersonA*'s perception of his social network. And the macro social network would result from the addition of all micro social perceptions.

3.2 Intra-agent Level

This level focuses on describing the internal components of each individual. In fact, *PersonA* are cognitive agents. They must have the following elements that help them perceive, plan and act [19]:

Perception Components. In a MAS, an agent interacts with the world around him based on his own perspective. He is provided with mechanisms and abilities that allow him to perceive his physical and social environments. These perceptions are updated as he interacts, and they depict knowledge that is embedded within the agent himself [19]. Thus, perception is by definition subjective. It can be updated, and can be correct, false, partially false or incomplete. For example, a person could think that a physical component offers a given activity while the latter is no longer available.

In our MAS, a person needs to perceive both social and physical environments:

- Perceived physical environment: In our model, an agent's perception of his physical environment is information (stored in his memory) about a subset of physical component objects; the ones that the agent knows about. And this knowledge is updated as he grows and interacts with his physical and social environments. For example, update via social interaction can happen when, throughout interactions, persons share some information about what they know (perceive) regarding the physical components and the possible activities they offer.
- Perceived social network: It allows an individual to have social relationships as well as a point of view about the persons he knows. The social perception is therefore a subset of *PersonA* with roles and subjective evaluations assigned to each one of them. The roles allow the description of the relationship's nature (parent, son, friend, instructor, etc.). As for the subjective evaluations, they depict *PersonA's* opinion about others (his feelings towards other agents, how close they are, etc.). Evaluations are expressed using social measures of the cognitive component.

Cognitive Component. This component allows agents to have human reasoning mechanisms. This part is being currently studied. A first analysis of our system's needs led us to modeling preferences and evaluation measures of both perceived physical environment (evaluation of the activities, e.g., what activities does an individual prefer) and perceived social network (evaluation of other individuals, e.g., social influence between two agents).

Factors Predisposing to PA, Sedentary Behavior and Childhood Obesity. The more factors and inter-factor relationships we cover, the more accurate and precise the simulation will be. In fact, it is the dynamic of the factors' network that will help us clarify the complex relationship between PA and childhood obesity. Some of these factors can be calculated based on other intra-agent components. For instance, we can use the agent's social perception and cognitive state to determine some of his psycho-social factors. The factors to be included first are the ones that are mostly encountered in the epidemiological literature. We will also try to inject factors that belong to different categories (the ones stated in [14] and [§2.2]). An analysis of the field studies will allow us to specify how to measure and evaluate these factors, and what influences they have on each other. In fact, the evaluation of a measured factor is expressed semantically. For example, most PA studies in children evaluate this behavior by using the set of ordered linguistic terms {Sedentary Activity, Light PA, Moderate PA, Vigorous PA}. Moreover, the different studies do not propose the same evaluation rules. Such imprecision on one hand, and the use of linguistic variables on the other, lead us to use Fuzzy Logic [20] to represent the considered factors. As for the relationships (such as the influence of the socioeconomic status on PA), they can be described via rules that are also retrieved from epidemiological field studies.

Behavior Component. The agent's behavior is guided by decision-making mechanisms that rely on information from other components. In our model, the life of an individual is simulated by a sequence of activities, allowing him/her to evolve and update his/her social relationships, internal state and perceptions. Therefore, an agent's behavior is modeled as a set of activities, and is organized as daily plans: each plan describes a schedule of activities to be performed during one day.

In order to promote the heterogeneity of agents, we provide them with a mechanism that allows them to follow different plans. In order to do so, different kinds of agents automatically have (more or less) different sets of possible plans, depending on the agent's roles (for *ChildA*: son/daughter, brother/sister, classmate, student, etc. and for *AdultA*: parent, teacher, instructor, etc.). Besides that, every agent is able to modify his plans according to his own will (he relies on his preferences, perceptions, history, etc.). This way, agents will autonomously adjust their plans and create new ones that may be more suitable for them (mentioned as 'Adapted plans' in [Fig. 2]). This ability to adapt one's plan is submitted to the activities' conditions (described in [§3.1]).

The agent's behavior is influenced by the other intra-agent layer components (perceptions, factors and cognitive mechanisms). At the same time, performing the different activities updates the agent's cognitive component and the factors predisposing to PA and obesity. Besides that, the agent's behavior changes the physical environment and the surrounding social relationships.

Interaction Protocols. This component describes the different possible interactions within the perceived social environment. In our MAS, we identify the following communication protocols (*PersonA* – *PersonA*):

- Data Sharing: These protocols allow agents to transfer information to each other. This knowledge can be simple objective data (for example, the parent sends to the child 'today I went jogging for 1 hour') or subjective information (like transferring one's perception, e.g., the parent sends to the child 'I am satisfied with your studies'). This kind of communication will guarantee updating the agents' mutual perceptions. The agents involved are the ones considered to be close enough to perceive each other and share information – close from a social perspective (parent/child) or a spatial perspective (children playing at the same park).
- Other communications: Since we chose to simulate behaviors as a schedules of activities, agents will need to communicate about these activities on different occasions, such as asking permission to perform an activity (like permission to play outside), suggest an activity (offer to go to the park), negotiate an activity (negotiate the time allowed for TV), request to be accompanied to an activity (can be seen as a coordination protocol. Like asking to play a sport together), etc.

4 Discussion and Perspective

In this paper, we have proposed a general MAS architecture for modeling the environment within which evolve the dynamics of the complex relationship between obesity and movement behavior in children. Compared to existing studies, this one considers the model from a CS standpoint, and therefore focuses more on lower level details and relationships between elements. Besides that, the proposed model would allow including most of the factors that are identified in literature as predisposing to PA and obesity in children. The use of learning mechanisms (such as data mining and artificial learning) could help us identify factors and inter-factor relationships that are still unknown in literature. The proposed architecture could be used to study other non communicable diseases where the agent's behavior greatly influences the factors predisposing to that disease, such as type II diabetes, depression, dyslipidemia, etc.

This work must be completed at different levels, particularly in modeling communication protocols, factors and cognitive mechanisms. Furthermore, the next step is the development of a simulator to test the model, and eventually test interventions.

References

1. Mittal, S.: Emergence in stigmergic and complex adaptive systems: A formal discrete event systems perspective. Cogn. Syst. Res. 21, 22–39 (2013)
2. Chen, D., Wang, L., Chen, J.: Large-Scale Simulation: Models, Algorithms, and Applications. Taylor & Francis (2012)
3. Rychetnik, L.: Criteria for evaluating evidence on public health interventions. J. Epidemiol. Community Heal. 56, 119–127 (2002)
4. Aziza, R., Zgaya, H., Guinhouya, B., Borgi, A.: Proposal for an agent-based knowledge model for the study of the relationship between physical activity and childhood obesity. In: Francophone Conference - Management and Engineering of Hospital Systems GISEH., Quebec (2012)
5. Organization, W.H.: Obesity: Preventing and Managing the Global Epidemic. World Health Organization (2000)
6. Guinhouya, B.C.: Physical activity during the development of the child. Medecine Sciences Publications, Paris (2012)
7. Vorwerg, Y., Petroff, D., Kiess, W., Blüher, S.: Physical activity in 3-6 year old children measured by SenseWear Pro®: direct accelerometry in the course of the week and relation to weight status, media consumption, and socioeconomic factors. PLoS One 8, e60619 (2013)
8. Tappe, K.A., Glanz, K., Sallis, J.F., Zhou, C., Saelens, B.E.: Children's physical activity and parents' perception of the neighborhood environment: neighborhood impact on kids study. Int. J. Behav. Nutr. Phys. Act. 10, 39 (2013)
9. Ejima, K., Aihara, K., Nishiura, H.: Modeling the obesity epidemic: social contagion and its implications for control. Theor. Biol. Med. Model. 10, 17 (2013)
10. Dangerfield, B., Abidin, N.Z.: Towards a model-based Tool for Evaluating Population-level Interventions against Childhood Obesity. In: 21st International Conference of the System Dynamics Society, Seoul, Korea (2010)
11. Giabbanelli, P.J., Alimadad, A., Dabbaghian, V., Finegood, D.T.: Modeling the influence of social networks and environment on energy balance and obesity. J. Comput. Sci. 3, 17–27 (2012)
12. Prachai, S.: The Design of Diabetes Simulation System using Multi-Agent. Procedia - Soc. Behav. Sci. 40, 146–151 (2012)

13. Yang, Y., Diez Roux, A.V., Auchincloss, A.H., Rodriguez, D.A., Brown, D.G.: A spatial agent-based model for the simulation of adults' daily walking within a city. Am. J. Prev. Med. 40, 353–361 (2011)
14. Guinhouya, B.C., Apété, G.K., Hubert, H.: The determinants of habitual physical activity in children: update and implications for care and prevention options in pediatric overweight/obesity. Rev. Epidemiol. Sante Publique. 58, 49–58 (2010)
15. Nicolet, J.L.: Risks and complexity. Harmattan Editions (2010)
16. Kari, J.: Theory of cellular automata: A survey. Theor. Comput. Sci. 334, 3–33 (2005)
17. Bonabeau, E.: Agent-based modeling: methods and techniques for simulating human systems. Proc. Natl. Acad. Sci. USA 99, 7280–7287 (2002)
18. Qu, Z., Garfinkel, A., Weiss, J.N., Nivala, M.: Multi-scale modeling in biology: how to bridge the gaps between scales? Prog. Biophys. Mol. Biol. 107, 21–31 (2011)
19. Müller, J.P.: From autonomous systems to multi-agent systems: Interaction, emergence and complex systems (2002)
20. Zadeh, L.A.: Fuzzy sets. Inf. Control. 8, 338–353 (1965)

The Absorptive Capacity-Based View of Training: Enhancing Organizational Performance: An Exploratory Study in Spanish Family Businesses

Felipe Hernández Perlines, María Yolanda Salinero Martín, and Benito Yáñez Araque

Department of Business Administration, University of Castilla-La Mancha,
Cobertizo de San Pedro Mártir, s/n, 45071 Toledo, Spain
{felipe.hperlines,yolanda.salinero,benito.yanez}@uclm.es

Abstract. In recent years academic interest has grown considerably regarding the absorptive dynamic capability, and as such, studying the effect it has on organizational learning, knowledge sharing, innovation, capacity building and performance of the company has greatly increased. Moreover, numerous research papers have studied the effect that training has on organizational performance. However, there are very few studies which link together training, absorptive capacity / learning capability and organizational performance. Investigations of this manner can make interesting contributions to the field of absorptive capacity, learning and the de-velopment of a new approach to more effective human resources training based on absorptive capacity as strategy. Indeed, the aim of this work is to create an approach whereby one can determine whether the connection between training and performance in Spanish family businesses is mediated by absorptive capacity. To this end, we propose the case study method in six Spanish family businesses.

Keywords: Absorptive Capacity, Training, Organizational Learning, Intellectual Capital, Knowledge, Family Business.

1 Introduction

Authors such as [1], as in previous studies, suggest the importance of the absorptive capacity (hereinafter ACAP) to improve performance and the generation of competitive advantages.

In the origins of the concept of ACAP we find the first connection between training and ACAP, which comes from the field of education: "the concept of ACAP can best be developed through an examination of the cognitive structures that underlie learning" [2]. With this in mind, the theory of meaningful learning [3–5] will help us better understand how training and ACAP are intimately linked.

Training is one of the principal intangible elements of intellectual capital items, and is a tool that facilitates learning while helping to smooth this process, "teaching to learn", as it were, and when the ACAP mediates, new knowledge is created and new skills are developed. This may be done individually or in groups, and can be planned or spontaneous, as well as being short or long term [6, 7].

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_43, © Springer International Publishing Switzerland 2014

Our study focuses on the family business (FB, hereafter) that "are the predominant type of company in the world", as is the case in Spain. Several authors have confirmed that in the FB the training process experienced by potential successors influences both their decision to fully integrate into the FB as well as their performance and continuity in it.

Through the case study method, we propose to (a) study the level of development of ACAP in six family companies; (b) essentially describe how to carry out the practice of human resource training development in these companies, (c) assess their strengths and weaknesses as they relate to the components of the ACAP. This proposal intends to finally approximate whether ACAP can be a mediator between the training processes and its conversion into a competitive advantage.

2 Training, Organizational Learning, Absorptive Capacity and Performance

There can be no doubt that in recent years the ACAP is becoming increasingly important to all companies. The first thing that must be done is to contextualize this.

Brettel et al. [1] suggest that the ACAP of a firm is a skill which develops in acumulative form which is path-dependent, and is based on existing knowledge. With respect to the resource-based view (RBV, hereinafter) of the company, the ACAP is a set of business capabilities that can explain differences in competitive advantage. However, because the RBV is a static method that assumes that resources differ in value, rareness, imitability, and substitutability [8–10], it overlooks the underlying dynamics of ACAP. This dynamic is reflected in the vision that emerges from the dynamic capabilities [11, 12], which can be seen as a variation and further development of the RBV. This approach takes into account that the estimated resources are likely to change over time.

The following is a review of chain relationships and the different concepts that concern us.

2.1 Training and Organizational Learning

Several studies directly related training as an element of organizational learning. This is true in the case of Castañeda and Fernández [13], who propose that an instrument consists of six dimensions, three of which are the individual, group and organizational, and three more where the condition for organizational learning occurs: organizational learning culture, training and transmission of information.

2.2 Absorptive Capacity and Organizational Learning

Lane et al. [14] relate directly with learning the first two dimensions of the ACAP consisting of the skills to understand and assimilate new external knowledge. That is, the acquisition of new external knowledge itself does not affect the performance but does so through learning as it will require the effective application of that knowledge. Furthermore, an interesting finding is that the acquisition of prior knowledge only influences learning when combined with high levels of training.

Dimensions of Absorptive Capacity. Zahra and George [15] extend the ACAP of the three original dimensions (identify, assimilate and exploit) to four dimensions (acquisition, assimilation, transformation, exploitation). They propose that ACAP is a multidimensional construct that refers to a dynamic capability relative to the creation and use of knowledge that in turn relates to the ability of the company to gain and sustain a competitive advantage. In their model, these authors distinguish two groups of skills: the potential ACAP (PACAP) and realized ACAP (RACAP). The PACAP includes the acquisition and assimilation capabilities, and includes RACAP transformation and exploitation of knowledge.

Proposed Measure of Absorptive Capacity. Although the work of Cohen and Levinthal [2] highlights the multidimensionality of ACAP, researchers have measured it as a unidimensional construct, often using proxy variables inputs of R&D, as spending on R&D company, the R&D intensity (R&D expenditures divided by sales), and the outputs of R&D, as the number of patents. Flatten et al. [16] propose a scale for measuring the ACAP. In the proposed measure, they evaluate the degree to which a company is engaged in the activities of acquiring knowledge, assimilates the acquired information with existing knowledge, transforms newly adapted knowledge, and commercially exploits their knowledge by transforming it into competitive advantage. This is the scale used to measure the ACAP in our research: we obtained the values of the various dimensions of ACAP in each of the cases studied.

2.3 Absorptive Capacity and Performance in the Company

Several studies have focused on the performance of the company as a result of the ACAP. In fact, empirical results show a significant positive relationship between ACAP and firm performance [14]. The main results of the ACAP model show innovative performance and financial performance in terms of competitive advantage and knowledge transfer between companies.

3 Methodology

We believe it appropriate to develop this study by the case study method, since it is an appropriate method for exploratory research [17]. Also, given the state of the question, despite the immediate evidence of the relationship between training, ACAP and results, the fact is that there are not many studies linking together the three concepts.

With respect to the number of cases studied, we believe that the six selected are enough for two main reasons:

- As opposed to the unit of analysis of a single case, study of multiple cases, like ours, is a powerful tool to create theory because they permit replication and extension among individual cases [18].
- Analysis of six cases has allowed us to reach the "point of redundancy" [19] so the addition of new cases would not significantly enriched sample theoretical saturation [20].

The cases studied correspond to six family businesses, whose location and merchandising range is Spain and the theoretical sample meets the criterion of selection that each has a training department. Briefly, the characteristics of the companies in our study are as follows: company A, auxiliary services sector company, founded in 1962, with more than 80,000 employees; company B, distribution sector company, created in 1977, with more than 70,000 employees; company C, food sector company, established in 1978, with more than 800 workers; company D, pharmaceutical sector company, began trading in 1838, with 550 workers; company E, bottling sector company, established in 1951, with 1,200 workers; company F, food sector company, established in 1991, with 3,800 workers.

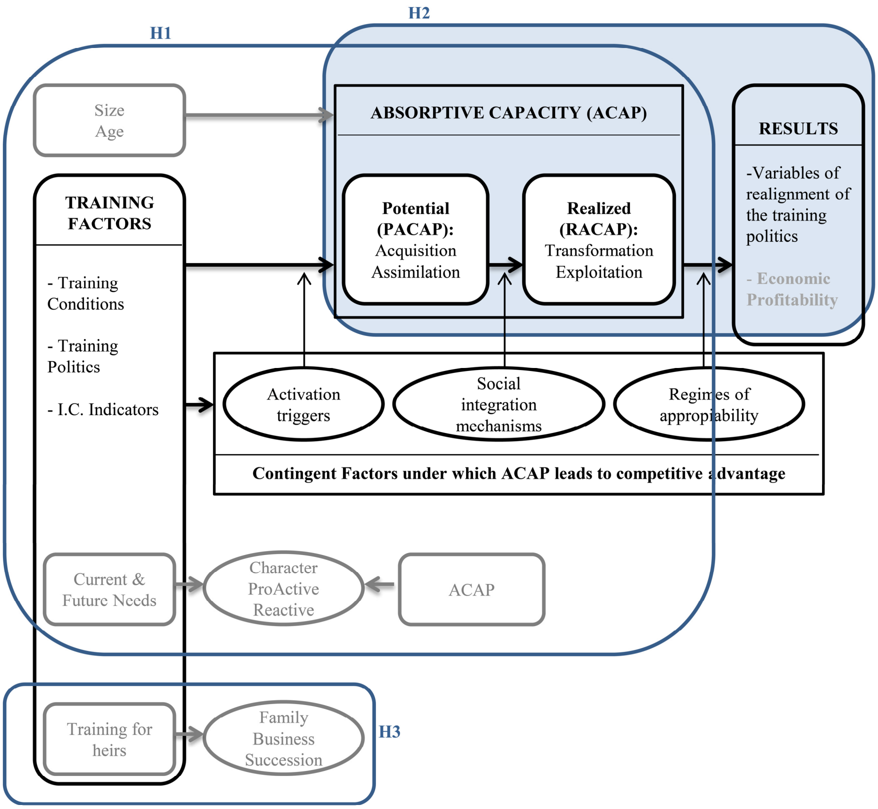


Fig. 1. Tentative preliminary model

Key respondents have been the General Director &/or the Director of Training. The time frame in which data was collected was between June and December 2012.

To gather the data we sent a self-administered questionnaire, using Likert scale questions and other open questions which were more characteristic of the semi-structured interview for qualitative analysis. The questionnaire was posted at: http://www.instituto-csic.es/uclm.grokis/ to make it available to those cases selected. All information has been analyzed with the help of the computer tool ATLAS/ti in order to find analysis codes.

The model is contrasted in two steps through the Proposition 1 (H1) and 2 (H2). A third proposition (H3) completes the tentative model. Figure 1 shows the assumptions made and their interactions in the proposed model.

4 Case Analysis: Findings

To analyze the connection between training, ACAP (as mediator variable) and performance in the Spanish FB, we developed the following propositions.

4.1 Proposition 1 (H1): The Training Has a Positive Effect on ACAP

From the data obtained from the analysis of the cases we observed training factors which are more favorable, both in terms of training politics as intellectual capital indicators, and which have impacted positively the ACAP of the six companies. Likewise, the training improves both the factor and the efficiency ratio of the ACAP: the six companies have good training conditions that correspond to good indicators of efficiency of ACAP. The training is positively related to a better ACAP as much potential as realized.

In turn, those training factors which we studied explain the differences in development between the different dimensions of ACAP. Training factors influenced each of the dimensions of the ACAP, as discussed in the study, and further related the training items rated by the six companies with the dimensions of the ACAP, and other facilitators and inhibitors of ACAP: contingent factors, proactive and reactive character, and size and age of the company.

Contingent Factors of ACAP. Zahra and George [15] identify the activation triggers, social integration mechanisms, regimes of appropriability as key contingencies affecting the ACAP.

Activation Triggers. Triggers moderate the impact that sources of knowledge and experience have on the development of ACAP. These are events that drive companies to respond to certain internal or external stimuli [15]. The current economic crisis is, according to these authors, an internal trigger of the ACAP.

The crisis has negatively impacted the organization of training activities by firms [21]. This is also what we see in this study. However, a crisis, despite being negative, can enhance a company's efforts to achieve and learn new skills and acquire new skills that increase the ACAP [15]. Training is important, if possible, in times of crisis, since the changes caused by the crisis require new skills.

Social integration mechanisms. We found that the six companies, especially the cases B, C and E with greater ACAP made, encourage employees to share knowledge on an informal level, and the training in teamwork is an important training need.

Regimes of appropriability. The appropriatness of human resources training will depend on: a) the durability of the worker in the company, especially because a durable resource favors sustaining competitive advantage [22, 23] and b) of the ACAP.

In our study, both case B and C and F (all other cases, to a lesser extent), career advancement is the incentive used to motivate more workers in conducting training activities.

Current Needs-future. Organizations with higher levels of ACAP tend to be more proactive, exploiting opportunities present in the environment, regardless of the actual results. However, organizations that have a modest ACAP tend to be reactive, looking for new alternatives in response to the lack of a performance criterion [2]. We try to know the reactive nature (training meets current needs) or proactive (training tries to cover future skills requirements in the enterprise) [24]. The companies with greater ACAP (B, C, E, F) respond proactively while in instances of minor ACAP (A, D), the responses are reactive.

Size and Age of the Firm. Here the result is inconclusive because it would require a quantitative analysis to extrapolate it. Nevertheless, our finding is that ACAP is inversely related to the size and age of the cases studied. Liao, Welsch, and Stoica [25] suggest that small firms are better able to respond to changes and are more innovative due to less bureaucracy and more simple hierarchy of their organizational structures.

4.2 Proposition 2 (H2): The ACAP Positively Affects Performance

To measure business performance variables we have used the feedback of the training politics. We took, therefore, a subjective measure for these results: the views of management concerning a number of items of business performance [26]. The use of other more objective outcome variables, such as profitability, would only be feasible from the perspective of quantitative analysis.

We found that the relationship between staff training and business performance is positive, but the intensity of this relationship depends on the level of development of ACAP.

When respondents were asked about what the training was for them, the six companies, but especially company C, responded that they placed high value on associating a business as an investment and also on ways to increase the profitability of the company.

Other items were rated including improving the quality of products or services and reducing the number of accidents.

4.3 Proposition 3 (H3): The Training for Descendants That Is Included in Business Plans Favors the Succession

On the problem of succession, none of the companies include any training on issues of family businesses targeting family descendants, except in the case of company C which includes succession planning in its training. In the words of its CEO: "In this business, the family has always supported people who value loyalty and family. Children are working in the company, all of whom are well prepared academically".

5 Conclusions

The training of employees is a strategic resource that will encourage the development of dynamic capabilities. Traditionally, from the human capital theory [27], training represents an "investment that operators made with the expectation of later profits: in the form of higher revenues from higher productivity" [28]. However, this transfer of training results is not performed directly, but, as we propose, is mediated by the ACAP.

Thus the practice of training is positively related with the ACAP in FBs studied: it represents a direct form of creation and development of ACAP as a source of external knowledge and complementarity. In the same way, training is directly involved in different dimensions of the ACAP: acquisition, assimilation as an indicator of the ability to assimilate new knowledge adopted by the enterprise, transformation and market exploitation. In the cases studied we've also seen the interactions of training with contingent factors of the ACAP such as proactivity-reactivity and size and age of the company.

Therefore, we argue that training improves the ACAP and, in turn, business performance. And we propose training of company heirs as a way of promoting the FB succession.

This means a new approach that has implications for training management and practical HR strategies from the perspective of ACAP. Some of these implications are:

- Establish training plans which are ACAP-oriented; that is, when designing training plans, incorporating in the methodology for training planning, in addition to the training needs, the needs of organizational absorptive capacity. And, as with the ACAP, the training needs of an organization are not only the sum of the training needs of their employees, but must also take into account the dimensions of the ACAP and influence those that demand improvement.
- As we've already described herein, a way of appropriating the returns of the training is to improve the ACAP: ACAP is revealed as a form of appropriating the returns of the training investment in company. This is especially important for the FB, due to the enduring nature of company heirs, and since these are presumed stable, their training is justified in greater measure, if possible, than in non-family firms. To this must be added the importance of training for succession of descendants; therefore, inclusion of training for descendants in the training plans of the company is a good practice for the FB.

- We suggest that the economic crisis, as an activation trigger of the ACAP, may represent an opportunity and an incentive for companies not to give up; rather totally to the contrary, that they emphasize the importance of human resource training. This is especially relevant for the FB, if we take into consideration the claims of some authors such as Lee [29]: the FB tend to lay off fewer employees in periods of economic crisis than their non-family counterparts.

To overcome the limitations of the case study method, would be useful to make a quantitative analysis, as future research way, based on a sample of companies and confirmation of propositions by structural equation model (SEM).

References

1. Brettel, M., Greve, G.I., Flatten, T.C.: Giving up linearity: Absorptive capacity and performance. J. Manag. Issues. 23, 164–189 (2011)
2. Cohen, W.M., Levinthal, D.A.: Absorptive Capacity: A New Perspective on Learning and Innovation. Adm. Sci. Q. 35, 128–152 (1990)
3. Ausubel, D.P.: The psychology of meaningful verbal learning: An introduction to school learning. Grune & Stratton, New York (1963)
4. Ausubel, D.P.: Educational psychology: A cognitive view. Holt, Rinehart and Winston, New York (1968)
5. Ausubel, D.P.: Psicología educativa: un punto de vista cognoscitivo (Roberto Helier D., Trad.). Editorial Trillas, México (1976)
6. Martínez, I.M.: El aprendizaje en las organizaciones. Aplicación al sector agroalimentario (2002)
7. Martínez, I.M., Ruiz, J.: Los procesos de creación del conocimiento: El aprendizaje y la espiral de conversión del conocimiento. XVI Congreso Nacional AEDEM, Alicante (2002)
8. Penrose, E.: The Theory of the Growth of the Firm. Wiley, New York (1959)
9. Barney, J.B.: Firm resources and sustained competitive advantage. J. Manage. 17, 99–120 (1991)
10. Wernerfelt, B.: A Resource-based View of the Firm. Strateg. Manag. J. 5, 171–180 (1984)
11. Easterby-Smith, M., Prieto, I.M.: Dynamic capabilities and knowledge management: An integrative role for learning? Br. J. Manag. 19, 235–249 (2008)
12. Teece, D.J., Pisano, G., Shuen, A.: Dynamic capabilities and strategic management. Strateg. Manag. J. 18, 509–533 (1997)
13. Castañeda, D.I., Fernández, M.: Validación de una escala de niveles y condiciones de aprendizaje organizacional. Univ. Psychol. 6, 245–254 (2007)
14. Lane, P.J., Salk, J.E., Lyles, M.A.: Absorptive capacity, learning, and performance in international joint ventures. Strateg. Manag. J. 22, 1139–1161 (2001)
15. Zahra, S.A., George, G.: Absorptive capacity: a review, reconceptualization and extension. Acad. Manag. Rev. 27, 185–203 (2002)
16. Flatten, T.C., Engelen, A., Zahra, S.A., Brettel, M.: A measure of absorptive capacity: Scale development and validation. Eur. Manag. J. 29, 98–116 (2011)
17. Chetty, S.: The Case Study Method for Research in Small- and Medium-sized Firms. Int. Small Bus. J. 15, 73–85 (1996)
18. Eisenhardt, K.M.: Better stories and better constructs: The case for rigor and comparative logic. Acad. Manag. Rev. 16, 620–627 (1991)
19. Lincoln, Y.S., Guba, E.G.: Naturalistic Inquiry. Sage Publications, Beverly Hills (1985)

20. Eisenhardt, K.M.: Building Theories from Case Study Research. Acad. Manag. Rev. 14, 532–550 (1989)
21. European Foundation for the Improvement of Living and Working Conditions: Preparing for the upswing: training and qualification during the crisis. Publications Office of the European Union, Luxembourg (2011)
22. Grant, R.M.: The resource-based theory of competitive advantage: implications for strategic formulation. Calif. Manage. Rev. 33, 114–135 (1991)
23. Grant, R.M.: Dirección Estratégica. Conceptos, técnicas y aplicaciones. Civitas, Madrid (1995)
24. Leal Millán, A., Alfaro de Prado Sagrera, A., Rodríguez Félix, L., Román Onsalo, M.: El factor humano en las relaciones laborales. Pirámide S.A., Ediciones, Madrid (1999)
25. Liao, J., Welsch, H., Stoica, M.: Organizational Absorptive Capacity and Responsiveness: An Empirical Investigation of Growth-Oriented SMEs. Entrep. Theory Pract. 28, 63–85 (2003)
26. Salinero, M.Y.: La formación como fuente de ventaja competitiva. Análisis de las empresas de Castilla la Mancha (2006)
27. Becker, G.S.: Human Capital. Columbia University Press for the National Bureau of Economic Research, New York (1964)
28. Araujo, A., Barrutia, J., Hoyos, J., Landeta, J., Ibáñez, P.: Comportamiento de las empresas respecto a la formación continua de sus directivos. Cuad. Gestión. 6, 83–98 (2006)
29. Lee, J.: Family Firm Performance: Further Evidence. Fam. Bus. Rev. 19, 103–114 (2006)

LIWC-Based Sentiment Analysis in Spanish Product Reviews

Estanislao López-López[1], María del Pilar Salas-Zárate[1], Ángela Almela[2], Miguel Ángel Rodríguez-García[1], Rafael Valencia-García[1], and Giner Alor-Hernández[3]

[1] Departamento de Informática y Sistemas, Universidad de Murcia. Campus de Espinardo 30100 Murcia, Spain
[2] Department of Modern Languages, Universidad Católica San Antonio de Murcia, Spain
[3] Division of Research and Postgraduate Studies Instituto Tecnológico de Orizaba, México
{estanislao.lopez,mariapilar.salas, miguelangel.rodriguez,valencia}@um.es, aalmela@ucam.edu, galor@itorizaba.edu.mx

Abstract. Opinion mining is the study of opinions and emotions of authors about specific topics on the Web. Opinion mining identifies whether the opinion about a given topic, expressed in a document, is positive or negative. Nowadays, with the exponential growth of social medial i.e. blogs and social networks, organizations and individual persons are increasingly using the number of reviews of these media for decision making about a product or service. This paper investigates technological products reviews mining using the psychological and linguistic features obtained through of text analysis software, LIWC. Furthermore, an analysis of the classification techniques J48, SMO, and BayesNet has been performed by using WEKA (Waikato Environment for Knowledge Analysis). This analysis aims to evaluate the classifying potential of the LIWC (Linguistic Inquiry and Word Count) dimensions on written opinions in Spanish. All in all, findings have revealed that the combination of the four LIWC dimensions provides better results than the other combinations and individual dimensions, and that SMO is the algorithm which has obtained the best results.

Keywords: sentiment analysis, opinion mining, LIWC, machine learning.

1 Introduction

Nowadays, the number of reviews about products or services has increased exponentially on the Web, so that it is becoming impossible for the customers to read reviews expressed in forums and blogs. The information that every user provides is useful, since a customer who buys a product usually search the Web trying to find other customers' opinions to make a decision. On these grounds, last year different technologies to automatically process these reviews have arisen. These technologies are usually known as opinion mining.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_44, © Springer International Publishing Switzerland 2014

Opinion mining (sentiment analysis, sentiment mining, or sentiment classification) deals with the computational treatment of opinion, sentiment, and subjectivity in text. The goal of opinion mining is to determine the attitude or opinion of a presenter or author with regard to a certain topic or target. The attitude could indicate his or her reasoning, opinion, assessment, or affective condition [1]. The opinion mining task can be transformed into a classification task, so different supervised classification algorithms such as Support Vector Machines (SVM), Bayes Networks and Decision Trees can be used to solve this task. On the other hand, LIWC is a text analysis software that allows to extract psychological and linguistic features from natural language text.

The aim of our work is to evaluate how the LIWC features can be used to classify Spanish product reviews into positive or negative opinions using different classifiers. For this purpose, a corpus of Spanish product reviews was first compiled and each opinion was manually tagged as positive or negative. Secondly, the corpus was processed by the LIWC tool to extract the linguistic features. Then, three different classifying algorithms were evaluated with the WEKA tool [2].

This paper is structured as follows: Section 2 presents the state of art on opinion mining and sentiment analysis. Section 3 describes and discusses the text analysis with LIWC dimensions. Section 4 presents the evaluation of the classifiers based on LIWC categories. Finally, Section 5 describes conclusions and future work.

2 Related Work

In recent years, several pieces of research have been conducted in order to improve sentiment classification. Some approaches [3, 4, 5, 6, 7] proposed methods for sentiment classification of English reviews. For example, in [3] three corpora available for scientific research in opinion mining are analyzed. Two of them were used in several studies, and the last one was built ad-hoc from Amazon reviews on digital cameras. Finally, an SVM algorithm with different features is applied, in order to test how the sentiment classification is affected. The study presented in [4] poses an empirical comparison between a neural network approach and an SVM-based method for classifying positive versus negative reviews. The experiments evaluated both methods as regards the function of selected terms in a bag-of-words (unigrams) approach. In [5] a comparative study of the effectiveness of ensemble methods for sentiment classification is presented. The authors consider two schemes of feature sets, three types of ensemble methods, and three ensemble strategies to conduct a range of comparative experiments on five widely-used datasets, with an emphasis on the evaluation of the effects of three ensemble strategies and the comparison of different ensemble methods. Experimental results demonstrate that using an ensemble method is an effective way to combine different feature sets and classification algorithms for better classification performance. Chen, Liu, & Chiu [6] propose a neural-network based approach. It uses semantic orientation indexes as input for the neural networks to determine the sentiments of the bloggers quickly and effectively. Several current blogs are used to evaluate the effectiveness of the approach. The experimental results indicate that the proposed approach outperforms traditional approaches including other neural networks and several semantic orientation indexes. He, & Zhou [7] propose a novel framework where prior knowledge from a generic sentiment lexicon is used to build a classifier. Tagged documents by this classifier are used to automatically acquire domain-specific feature words whose word-class distributions are estimated and are subsequently used to train another classifier by constraining the model's predictions on unlabeled instances. The experiments, the movie-review data and the multi-domain sentiment dataset show that the approach attains comparable or better performance than existing hardly supervised sentiment classification methods despite using no labeled documents.

Furthermore, other proposals [8] introduce methods for sentiment classification of Chinese reviews. Zhai, Xu, & Jia [8] analyze sentiment-word, substring, substring-group, and key-substring-group features, and the commonly-used Ngram features. To explore general language, two authoritative Chinese data sets in different domains were used. The statistical analysis of the experimental results indicate that different types of features possess different discriminative capabilities in Chinese sentiment classification.

Finally, it is worth noting that not many proposals such as the one presented here [9] are focused on sentiment classification of Spanish reviews. In this work, two lexicons are used to classify the opinions using a simple approach based on counting the number of words included in the lexicons. Specifically, an opinion is positive if the number of positive words is greater than or equal to the negative ones, and is negative in the opposite case.

In order to fully analyze the studies described above and compare them with our proposal, we used 4 features: 1) computational learning, 2) linguistic resources, 3) domain, and 4) language.

1. Several machine learning techniques are used, i.e. SVM, Naïve Bayes, among others. Almost all the proposals use computational learning. Specifically, the SVM technique is the most frequently used [3, 4, 5, 8]. Besides, the techniques of Naïve Bayes [5, 7, 9] and neural networks [8] are also used. On the other hand, other pieces of research do not use any machine learning technique [9].
2. The techniques used for detection of polarity in these approaches are n-grams [3, 5, 8], term frequency [3, 5, 8], and semantic orientation indexes [6]. On the other hand, other approaches only use lexical resources [9].
3. Almost all the corpora used in the proposals mentioned above include reviews on products [3, 4, 5, 6, 8, 9]. Other proposals use corpora that include reviews on issues such as: music [3], hotels [3, 8], movies [9], general services [4], DVDs [7] and electronics [7].
4. The English language is the most used in these approaches [3, 4, 5, 6, 7]. However, others languages are used in some proposals, such as Chinese [8] and Spanish [9].

Based on the previous analysis, the present study seeks to study the performance of three different classifying algorithms in the classification of Spanish opinions through the combination of the psychological and linguistic features of LIWC.

3 LIWC

LIWC (Linguistic Inquiry and Word Count) is a software application that provides an effective tool for studying the emotional, cognitive, and structural components contained in language on a word by word basis. Early approaches to psycholinguistic concerns involved almost exclusively qualitative philosophical analyses. More modern research in this field provides empirical evidence on the relation between language and the state of mind of subjects, or even their mental health [10]. In this regard, further studies such as [11] have dealt with the therapeutic effect of verbally expressing emotional experiences and memories. LIWC was developed precisely for providing an efficient method for studying these psycholinguistic concerns, and has been considerably improved since its first version [12]. An updated revision of the original application was presented in [13], namely LIWC2001.

LIWC provides a Spanish dictionary composed by 7,515 words and word stems. Each word can be classified into one or more of the 72 categories included by default in LIWC. Also, the categories are classified into four dimensions: 1) standard linguistic processes, 2) psychological processes, 3) relativity, and 4) personal concerns.

For the present study, a set of reviews in Spanish that include both positive and negative reviews was necessary. Therefore, a corpus with 600 reviews of Technological products such as mobile devices was collected. This corpus contains 300 positives reviews and 300 negative reviews, obtained from online selling websites. Also, each review was examined and classified manually as positive or negative to ensure its quality.

Once the corpus has been obtained, it is analyzed through each of all possible combinations of LIWC dimensions. LIWC searches for target words or word stems from the dictionary, categorizes them into linguistic dimensions, and then converts the raw counts to percentages of total words. The obtained values for the categories were used for the subsequent training of the machine learning classifier.

This analysis aims to evaluate the classifying potential of the dimensions, both individually and collectively.

4 Evaluation and Results

In the present work, WEKA [2] has been used in order to evaluate the classification success of reviews (positives or negatives) based on LIWC categories.

WEKA provides several classifiers, which allows the creation of models according to the data and purpose of analysis. Classifiers are categorized into seven groups: Bayesian (Naïve Bayes, Bayesian nets etc.), functions (linear regression, SMO, logistic, etc.), lazy (IBk, LWL, etc.), meta-classifiers (Bagging, Vote, etc.), miscellaneous (SerializedClassifier and InputMappedClassifier), rules (DecisionTable, OneR, etc.) and trees (J48, RandomTree, etc.). In this case, J48, SMO and BayesNet classification algorithms were applied. These algorithms were selected because they have been used in several experiments obtaining good results in data classification [14], [15].

The classification process involves the building of a model based on the analysis of the instances. This model is represented through classification rules, decision trees, or mathematical formulae. The model is used to generate the classification of unknown data, calculating the percentage of instances which were correctly classified.

In order to evaluate the results of the classifiers, we have used three metrics: precision, recall and F-measure. Recall is the proportion of actual positive cases that were correctly predicted as such. On the other hand, precision represents the proportion of predicted positive cases that are correctly real positives. Finally, F-measure is the harmonic mean of precision and recall.

The experiment was performed by using three different algorithms: the C4.5 decision tree (J48), the Bayes Network learning algorithm (BayesNet) and the SMO algorithm for SVM classifiers [16]. For each classifier, a ten-fold cross-validation has been done. This technique is used to evaluate how the results of a would generalise to an independent data set. Since the aim of this experiment is the prediction of the positive or negative condition of the texts, a cross-validation is applied in order to estimate the accuracy of the predictive models. It involves partitioning a sample of data into complementary subsets, performing an analysis on the training set and validating the analysis on the testing or validation set [17].

Table 1 shows the results of the experiment. The first column indicates which LIWC dimensions are used, i.e. 1) standard linguistic processes, 2) psychological processes, 3) relativity, and 4) personal concerns. In the next columns the results of precision (P), recall (R), and F-measure for each algorithm are presented.

Table 1. Experiment results

	J48			BayesNet			SMO		
	P	R	F1	P	R	F1	P	R	F1
1	0.669	0.671	0.69	0.729	0.727	0.727	0.773	0.773	0.773
2	0.729	0.73	**0.729**	0.763	0.763	**0.763**	0.753	0.752	0.**748**
3	0.663	0.665	0.66	0.711	0.712	0.711	0.726	0.725	0.72
4	0.672	0.671	0.671	0.69	0.691	0.691	0.685	0.685	0.685
1_2	0.733	0.734	0.733	0.815	0.811	**0.812**	0.817	0.817	**0.816**
1_3	0.68	0.682	0.679	0.749	0.748	0.749	0.762	0.763	0.762
1_4	0.701	0.701	0.701	0.743	0.741	0.742	0.762	0.763	0.762
2_3	0.749	0.75	**0.749**	0.809	0.808	0.808	0.793	0.793	0.792
2_4	0.738	0.739	0.739	0.784	0.782	0.783	0.775	0.775	0.774
3_4	0.667	0.667	0.667	0.741	0.741	0.741	0.747	0.748	0.747
1_2_3	0.746	0.746	0.746	0.819	0.815	**0.815**	0.811	0.811	0.811
1_2_4	0.751	0.752	0.75	0.8	0.795	0.796	0.809	0.809	0.809
1_3_4	0.733	0.734	0.732	0.758	0.757	0.758	0.767	0.768	0.767
2_3_4	0.735	0.736	0.734	0.808	0.804	0.805	0.796	0.797	0.797
1_2_3_4	0.76	0.761	**0.759**	0.808	0.804	**0.805**	0.834	0.835	**0.834**

As can be seen in Table 1, the results from the J48 algorithm show that individually, the second dimension, “psychological processes“, provides the best results, with an F-measure of 72.9%. Conversely, the third dimension, “relativity“ provides the worst results, with an F-measure of 66.0%. On the other hand, the combination of all LIWC dimensions provides the best classification result with an F-measure of 75.9%.

The results of the BayesNet algorithm are similar to the ones obtained by the J48 algorithm, although this experiment provides better classification results. Results show that the second dimension, “psychological processes“, provides the best results on its own as well, with an F-measure of 76.3%. Quite the reverse, the fourth dimension, “personal concerns”, provides the worst results with an F-measure of 69.1%. Furthermore, the combination of 1_2_3 LIWC dimensions provides the best classification result, with an F-measure of 81.5%. The results obtained by means of the use of the four dimension are also good, with a total F-measure of 80.5%

The results from the experiment with SMO are better than the ones obtained with the previous algorithms. The results show that, once again, the second dimension provides the best results by itself, with an F-measure of 74.8%. On the contrary, the fourth dimension, “personal concerns”, provides the worst results with a score of 68.5%. Moreover, the combination of all LIWC dimensions provides the best classification result with an F-measure of 83.4%.

General results show that the combination of all LIWC dimensions provides the best classification results with J48 and SMO, and a considerably high score with BayesNet. As regards the performance of individual dimensions, the second one provides the best results, probably due to the fact that written opinions frequently contain words related to the emotional state of the author, which correspond to word stems classified into categories such as: anxiety, sadness, positive and negative emotions, optimism and energy, and discrepancies, among others. All these categories are included in the second dimension, confirming its discriminatory potential in classification experiments.

The fourth dimension provides the worst results, owing to the fact that the topic selected for this study, “technological products”, bears little relation to the vocabulary corresponding to “personal concerns” categories. It can be stated that this dimension is the most content-dependent, and thus the least revealing. Finally, the results show that the combination of different LIWC dimensions provides better results than individual dimensions.

The best results in this experiment were obtained by the SVM models. SVMs have been applied successfully in many text classification tasks due to their main advantages: first, they are robust in high dimensional spaces; second, any feature is relevant; third, they are robust when there is a sparse set of samples; and finally, most text categorization problems are linearly separable [18]. Unlike other classifiers such as decision trees or logistic regressions, SVM assumes no linearity, and can be difficult to interpret outside of its accuracy values [19].

It is difficult to compare the different opinion mining approaches described in the literature, because none of the software applications are available. Indeed, the corpora used for the experiment differ significantly in content and size, concepts and language. A fair comparison of two opinion mining methods would require the usage of the same testing corpus. Taking all this into account, most of the current approaches for opinion mining use the standard measures of precision, recall, and the F-measure. This is the case of the work presented in [9], where the system obtains an F-measure of 63.33% for the Spanish language, which is considerably lower than the one obtained by our approach.

On the other hand, the approaches for the English language obtained better F-measures than the ones obtained here. For example, in [5] and [7] the F-measures obtained are 86.45% and 84.85% respectively. This is probably due to the lower level of complexity of the English language as compared to Spanish, especially as regards grammar, which seems to have a reflection on the final results.

5 Conclusions and Future Work

We have presented a research study based on sentiment classification in order to evaluate the classifying potential of the LIWC dimensions. For our research, the evaluation of the combination of the four LIWC dimensions (standard linguistic processes, psychological processes, relativity, and personal concerns) for sentiment classification has been carried out. Subsequently, in an attempt to evaluate the efficacy of LIWC features, J48, BayesNet and SMO algorithms have been used. Results show that the combination of the four LIWC dimensions provides better results than the other combinations and individual dimensions. The SMO algorithm has obtained the best results.

As further work, the authors are considering to develop new corpora comprising different domains of the Spanish language, since research into sentiment classification in this language is needed. Also, future studies will include the classification of reviews in five categories: highly positive, highly negatives, positives, negatives and neutral, in order to provide a better comprehensive analysis related to other topics and the relation of LIWC features and the reviews classification Furthermore, we will use LIWC features in English language to verify whether this technique can be applied to different languages. Finally, it is also planned to adapt this approach to a feature-based opinion mining guided by ontologies, as in the study presented in [20].

Acknowledgements. This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Commission (FEDER / ERDF) through project SeCloud (TIN2010-18650). María del Pilar Salas-Zárate is supported by the National Council of Science and Technology (CONACYT), the Public Education Secretary (SEP) and the Mexican government.

References

1. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2, 1–135 (2008)
2. Bouckaert, R.R., Frank, E., Hall, M.A., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: WEKA–experiences with a java open–source project. Journal of Machine Learning Research 11, 2533–2541 (2010)

3. Rushdi Saleh, M., Martín Valdivia, M., Montejo Ráez, A., Ureña López, L.: Experiments with SVM to classify opinions in different domains. Expert Systems with Applications 38, 14799–14804 (2011)
4. Moraes, R., Valiati, J.F., Gavião Neto, W.P.: Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications 40, 621–633 (2013)
5. Xia, R., Zong, C., Li, S.: Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences 181, 1138–1152 (2011)
6. Chen, L., Liu, C., Chiu, H.: A neural network based approach for sentiment classification in the blogosphere. Journal of Informetrics 5, 313–322 (2011)
7. He, Y., Zhou, D.: Self-training from labeled features for sentiment analysis. Information Processing and Management 47, 606–616 (2011)
8. Zhai, Z., Xu, H., Kang, B., Jia, P.: Exploiting effective features for chinese sentiment classification. Expert Systems with Applications 38, 9139–9146 (2011)
9. Molina González, M.D., Martínez Cámara, E., Martín Valdivia, M.T., Perea Ortega, J.M.: Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications 40, 7250–7257 (2013)
10. Stiles, W.B.: Describing talk: A taxonomy of verbal response modes. Sage, Newbury Park (1992)
11. Pennebaker, J.W., Mayne, T., Francis, M.E.: Linguistic predictors of adaptive bereavement. Journal of Personality and Social Psychology 72, 863–871 (1997)
12. Francis, M.E., Pennebaker, J.W.: LIWC: Linguistic Inquiry and Word Count. Technical Report. Southern Methodist University, Dallas (1993)
13. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count. Erlbaum Publishers, Mahwah (2001)
14. Nahar, J., Tickle, K., Ali, S., Chen, P.: Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer. Expert Systems with Applications 39, 12371–12377 (2012)
15. Chen, L., Qi, L., Wang, F.: Comparison of feature-level learning methods for mining online consumer reviews. Expert Systems with Applications, 9588–9601 (2012)
16. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13(3), 637–649 (2001)
17. Kohavi, R.: A study of cross–validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2(12), pp. 1137–1143. Morgan Kaufmann, San Mateo (1995)
18. Rushdi Saleh, M., Martín Valdivia, M.T., Montejo, A., Ureña, L.A.: Experiments with SVM to classify opinions in different domains. Expert Systems with Applications 38(12), 14799–14804 (2011)
19. Chen, Y., Lin, C.: Combining SVMs with various feature selection strategies. In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds.) Feature Extraction. STUDFUZZ, vol. 207, pp. 315–324. Springer, Heidelberg (2006)
20. Peñalver-Martínez, I., Valencia-García, R., García-Sánchez, F.: Ontology-guided approach for Feature-Based Opinion Mining. In: Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 193–200. Springer, Heidelberg (2011)

Data Extraction Tool to Analyse, Transform and Store Real Data from Electricity Markets

Ivo F. Pereira[1], Tiago M. Sousa[1], Isabel Praça[1], Ana Freitas[1], Tiago Pinto[1], Zita Vale[1], and Hugo Morais[2]

[1] GECAD – Knowledge Engineering and Decision-Support Research Center
Institute of Engineering – Polytechnic of Porto (ISEP/IPP), Porto, Portugal
{ifdsp,tmsbs,icp,1080598,tmcfp,zav}@isep.ipp.pt
[2] Automation and Control Group, Technical University of Denmark (DTU)
morais@elektro.dtu.dk

Abstract. The study of electricity markets operation has been gaining an increasing importance in the last years, as result of the new challenges that the restructuring process produced. Currently, lots of information concerning electricity markets is available, as market operators provide, after a period of confidentiality, data regarding market proposals and transactions. These data can be used as source of knowledge to define realistic scenarios, which are essential for understanding and forecast electricity markets behavior. The development of tools able to extract, transform, store and dynamically update data, is of great importance to go a step further into the comprehension of electricity markets and of the behaviour of the involved entities. In this paper an adaptable tool capable of downloading, parsing and storing data from market operators' websites is presented, assuring constant updating and reliability of the stored data.

Keywords: Databases, Electricity Markets, Machine Learning, Multi-Agent Simulators, Real Electricity Markets Data.

1 Introduction

Electricity markets are complex environments with very particular characteristics. A critical issue regarding these characteristics concerns the constant changes they are subject to. This is a result of the electricity markets' restructuring, which was performed so that the competitiveness could be increased, but it also had exponential implications in the increase of the complexity and unpredictability in those markets scope [1]. Electricity markets, as competitive environments, require good decision-support mechanisms to assist players in their decisions. Significant development concerning electricity markets players modelling and simulation including decision-support capabilities can be widely found in the literature [2-5].

Liberalized markets provide valuable information generally available to the community. Knowledge about markets grows from these last years and assists the definition of adequate players' profiles and behaviours. This information is also essential to test and validate existing simulation tools, such as MASCEM (Multi-Agent System

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_45,

for Competitive Electricity Markets) [2, 3], making them capable of simulating realistic scenarios.

The realistic modelling of electricity markets, which provides the means for a suitable knowledge extraction from the study of advantageous simulations, requires an extensive search and organization of as much information as possible concerning these markets characteristics, particularities and constraints. Automatic tools, able to gather, store, update and organize data from distinct real electricity markets will be a key issue to improve markets simulators and the modelling of the participating entities, enabling researchers and professionals to extract knowledge and really learn from the last years' experience.

This paper presents a tool that was developed with the purpose of automatically searching for new electricity market data, extracting it from various websites, parsing the information, and storing it in the appropriate database, so that it can be used by the electricity market simulators to model realistic scenarios. This tool is adaptive to the data availability timings; it is capable of dealing with different data formats, and it includes parallel processing capabilities, in order to deal with multiple data sources processing.

This paper is organized in 5 sections. In section 2 an insight on the electricity markets data requirements is presented, both in what concerns the distinct nature of different countries' electricity markets, and the requirements from the currently most important electricity market simulators, namely MASCEM, which is being developed by the authors' research team since 2003 [2]. Section 3 presents the system capable of downloading, analysing and saving information from real electricity markets to provide a database with real historical data. Section 4 illustrates, by means of simple example, the processing of some of the data available at the GME (*Gestore Mercati Energetici* – Italian Energy Market Operator) homepage [6]. Finally, section 5 presents the most relevant conclusions and future implications concerning the presented work.

2 Electricity Markets Data

The liberalization of the electricity sector provides new market rules, the emergence of new market players and new forms of interactions among them [1, 7].

The functioning of liberalized markets over the last years provides valuable information most of the times available to the community trough market operators websites. In fact, market operators such as MIBEL (Iberian Market Operator) [8], Nord Pool [9], EPEXSPOT (European Power Exchange) [10], MISO (central U.S. energy market) [11] and GME [6] provide on their web sites information regarding market proposals and transactions, usually after a period of confidentiality. The available information depends on each different market operator, however, essential information such as market proposals, with quantity and price; accepted proposals and established market prices is usually always available. This information grows in a very dynamic way, as it is put available in the various websites.

The Iberian Market Operator [8], which includes the electricity markets of Portugal and Spain, started on July 2006 with the futures market. One year later, in July 2007, both the day-ahead and the intraday markets started operating.

NordPool [9] is currently the largest energy market in Europe and includes the northern countries of this Continent, namely Norway, Denmark, Sweden, Finland, Estonia, Latvia and Lithuania. There are two different markets in the Nord Pool: the ELSpot (day-ahead market) and the ELbas (intraday negotiations).

The EPEXSPOT [10], covers the central area of Europe, including countries such as France, Germany, Austria and Switzerland. EPEX includes the day-ahead and intraday markets and an established deal with Czech Republic, Slovakia and Hungary for a trilateral market between these countries.

The MISO [11] includes 15 U.S. states and the Canadian province of Manitoba in the day-ahead market.

GME [6], the Italian market operator, includes the day-ahead market (MGP), the intraday market (MI), ancillary services market (MSD) and The Forward Electricity Market.

With the information taken from the operation of different markets, lessons can be learnt from the last years to improve knowledge about markets, to define adequate players' profiles and behaviours, and realistic scenarios.

The need for understanding these market mechanisms and how the involved players' interaction affects the outcomes of the markets, contributed to the growth of usage of simulation tools, with the purpose of taking the best possible results out of each market context for each participating entity. Multi-agent based software is particularly well fitted to analyse dynamic and adaptive systems with complex interactions among its constituents, such as the electricity market. Several modelling tools directed to the study of restructured wholesale power markets have emerged. Some of the most relevant tools in this domain are:

- Electricity Market Complex Adaptive System (EMCAS) [4]: software agents with negotiation competence use strategies based on machine-learning and adaptation to simulate Electricity Markets;
- Agent-based Modelling of Electricity Systems (AMES) [5]: open-source computational laboratory for studying wholesale power markets, restructured in accordance with U.S. Federal Energy Regulatory Commission (FERC);
- Genoa Artificial Power Exchange (GAPEX) [12]: agent-based framework for modeling and simulating power exchanges implemented in MATLAB. GAPEX allows creation of artificial power exchanges reproducing exact market clearing procedures of the most important European power-exchanges.
- Multi-Agent Simulator for Competitive Markets (MASCEM) [2, 3] : a platform based on multi-agent simulation firstly proposed in 2003 [2] that evolved into a complete tool acting in forward, day-ahead, and balancing markets, considering both simple and complex bids, and players' strategic behaviour, using the decision support of ALBidS (Adaptive Learning strategic Bidding System) [13].

These simulators are very good examples of tools able to represent market mechanisms and players' interactions. However, for them to be valuable decision support tools in foreseeing market behaviour, they need to be used in testing adequate and realistic scenarios. Real data analysis by means of a knowledge discovery process will be a crucial forward step to assure that MASCEM agents exhibit adequate profiles and strategies, namely by improving ALBidS strategies.

3 Automatic Tool for Real Markets Information Extraction

After a careful analysis of the available data, since each operator make available different types of information, some of them providing even entities technical characteristics and localization, a database model has been defined. Figure 1 illustrates the Domain Model. As can be seen by Figure 1, the websites provide several files. Each file is related to a day, or may also be referring to a market session. The files usually include several lines, each containing the proposal of an agent, and the corresponding date. The proposals include the volume of traded energy, its price and whether it was a selling or buying offer.

Some requirements for the application are: the needs to assure the treatment of different file types, reliability in storing all the gathered data, as well as the needs to actualize extracted data whenever it is available. Another relevant issue is efficiency regarding the treatment of great amounts of files, which, indeed in the initial use of the application may imply an enormous amount of files to assure gathering historic relevant data.

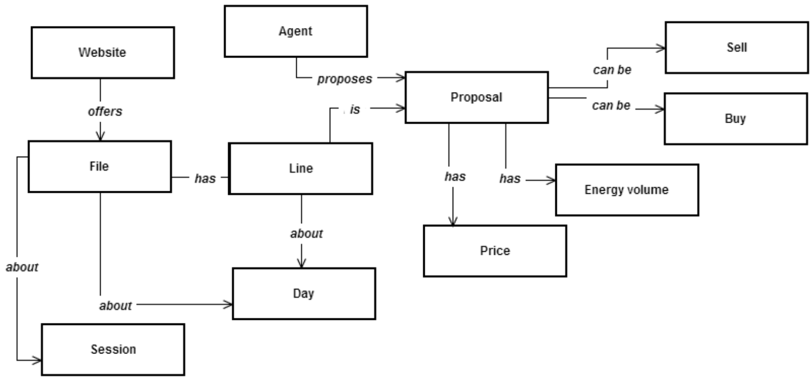


Fig. 1. Domain Model

So that the files management can be performed in the best possible way, the *adapter* software standard is used for each file format. The *adapter* standard is used in circumstances in which a system needs to connect to an external service. In these cases the user should not need to allow such action, as it is intrinsic to the program. For this, it is necessary to create a class that provides the interface expected by the user and that uses the interface of the service provider. This means that it is necessary to create an interface with the signatures of the necessary methods for each *adapter*, and build each *adapter* in an independent fashion. The global class *Fileadapter* was built to provide the required abstraction that enables the system's ability to deal with new and different file formats.

The application was built taking into account that in the future may be necessary to treat new files. The scalability requirement was achieved with success, for both the user interface and the architecture of the application does not require any change to be extended to new types of files, besides integrating a new *FileAdapter*.

Within the developed automatic tool there are four major steps:

- Download data - the download of several files containing the new data. The download depends on the website from which the data is being extracted, and it is performed accordingly to the data type of each file;
- Parse data - the extraction of the stored data from the downloaded files. The parsing of the data includes the analysis of the data fields of each file, from which the information and its associated value are taken;
- Store data - the storage of collected data in the database. The storage of the parsed data takes into account the necessary connections between different sets of data. This enables the data to be stored appropriately, respecting the interconnectivity and dependencies between all data;
- Mechanism for automatic data updates – adaptive technique to automatically define downloads periodicity. The availability timings of each file are analysed so that the developed tool can process all available data as soon as possible.

3.1 Automatic Data Updates

Real electricity markets information grows up in a very dynamic way, which brings the need to check for information updates automatically. For this purpose, a mechanism was developed, which is able to automatically check for information updates for download and data base updating.

For the system to periodically update the information, it uses an adaptive mechanism, which consists of the automatic repetition of the entire process of extracting data from files and setting the new programming period to check for the availability of new files. The system schedules a new check some time after the last time it was checked. In the case of a file containing monthly information, a new file will probably only become available one month later.

The adjustment of the scheduling period by the Adaptive Mechanism for Updating Data is based on the success and failure of foreseen checks. If a check is scheduled but there are no new files, it is considered a failure and so it should made a little adjustment in the period of scheduling. However, if it succeeds in finding a new available file the programming period should be maintained, or reduced to find out if the timing of its availability was sooner than expected.

Figure 2 illustrates the interface of the schedule definition for the mechanism of automatic updates of the automatic data extraction tool.

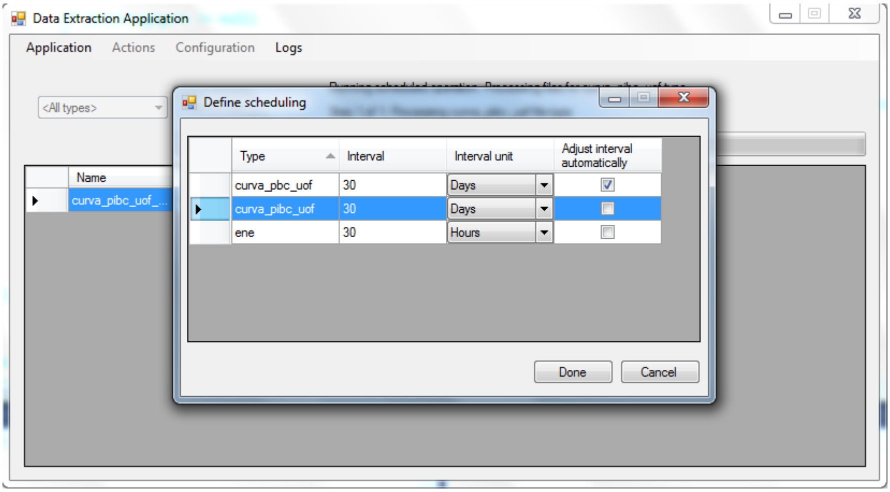


Fig. 2. Automatic Data Updates specification

3.2 Reliability Mechanism

The reliability mechanism assures that one file is considered as successfully completed after passing all treatment phases (Download, Extraction, Save and Parse) without errors. Another feature, specific to the "Save" phase, consists on the use of transactions in inserting the records into the database, ended by a COMMIT transaction ensuring that these records will be stored in the database. However, if an error occurs, it makes the ROLLBACK of the transaction, so that none of the records of the file currently being treated are entered. Thus, it is possible to ensure that all or none of the data contained in the file is stored, avoiding redundancy situations in which only a portion of the file is saved.

3.3 Parallel Processing of Files

After a careful analysis of the websites of market operators, it became clear that there was a huge amount of different files to deal with. In order to assure the efficiency of the tool, the files processing mechanism was implemented with parallelism. Thus, several files can be treated simultaneously, since there is no dependence between them in the treatment process. Then came another concern given that to keep many data in memory due to the processing of multiple files simultaneously, often there were memory errors due to lack of space for allocating new objects.

After different strategies for parallelism were tested, the strategy "Concurrent Queue with foreach loop" was chosen. Table 1 presents the runtimes for each parallelism strategy and the memory errors obtained.

Table 1. Testing strategies for parallelism

ZIP Files			
Strategy	**Number of Files**	**Minutes**	**Number of errors**
Parallel ForEach	3360	40	High
Aggregator:	3360	30	Few
Concurrent Queue	3360	27	None
Txt and Excel files			
Strategy	**Number of Files**	**Minutes**	**Number of errors**
Parallel ForEach	3500	20	Few
Aggregator:	3500	15	None
Concurrent Queue	3500	9	None

4 GME Market Test-Case

GME's webpage provides much information regarding the Italian Market [6]. By accessing GME's website, a list of the different types of files is available. The example that is shown in this test-case refers to the files "*MGP Prezzi*" and "*MGP Quantita*", which contain prices and volumes of the day-ahead spot market, for each day. All files of both types can be accessed through the web page [6], and have to be downloaded separately.

In this case, the presented files are XML files, and their names have the following format: "<year> <month> <day> MGPPrezzi.xml" and "<year> <month> <day> MGPQuantita.xml", where <year> corresponds to the year of the data, <month> is referent to the month and <day> refers to the day (e. g. "20140101MGPPrezzi.xml" corresponds to the prices of January 1st, 2014). The files ending with "MGPPrezzi.xml" contain the information about all the prices for every period and area of that day. The files ending with "MGPQuantita.xml" contain information about the volumes of energy bought and sold for every period and area of that day. In order to ensure the integrity of the data, these two types of files are merged in one, containing all the relevant information of both files for every period and area of each day.

In the first phase of the whole process, Download Data, the files to be analysed and combined are downloaded. The files are placed in a temporary folder, so that at the end of the process they can be removed since they are no longer required.

In the second phase, Parse Data, it is necessary to parse each extracted file and then combine both files into a table. As the files to be read come in XML format it is easy to open them. To make the parse of the files, their structure was studied, and then combined into a table, in which the first row indicates the information in each column, with the order: *Period*, *Area Name1*, *Buy Volume1*, *Sell Volume1*, *... , Area NameN*, *Buy VolumeN*, *Sell VolumeN*. From the second row forward the information is presented. The data type of each field is shown in table 2.

Table 2. Field Types for the combination of the "*MGP Prezzi*" and "*MGP Quantita*" files

Field	*Period*	*Area Name1*	*Buy Volume1*	*Sell Volume1*	...	*Area NameN*	*Buy VolumeN*	*Sell VolumeN*
Data Type	Integer value indicating the period	Numeric value indicating the price for *Area Name1*	Numeric value indicating the volume bought for *Area Name1*	Numeric value indicating the volume sold for *Area Name1*	...	Numeric value indicating the price for *Area NameN*	Numeric value indicating the volume bought for *Area NameN*	Numeric value indicating the volume sold for *Area NameN*

Finally, the period column and each group of the corresponding three columns for each area will result in a database record being further included in the database, corresponding to the third phase (Storage Data) of the process. The information of all files is centred on the same database, in order to be used by MASCEM, but also by other simulation and decision support systems.

5 Conclusions

Electricity markets worldwide suffered profound transformations. With the increase of the competitiveness and consequent decrease of electricity price in sight, the restructuring of electricity markets brought a significant enlargement of the complexity in this sector. In order to deal with this increasing complexity, and the dynamism of the electricity markets environment, simulators and decision support systems became widely used.

Data regarding electricity market players of very distinct nature, with enormous differences in their characteristics; data regarding the market mechanisms of different types from country to country; data regarding the interactions and negotiations between players in different market environments; data regarding the decision support and strategic behaviour of such players; are only a few examples of what is required for electricity market simulators to adequately model the electricity market environment, so that realistic scenarios can be built and an advantageous decision support can be provided.

Even though most of the referred data is available, it is of very difficult access, for diverse reasons; therefore it is not used in the way and extent it should. In order to overcome the problem of the data access and treatment, making it available in an useful way for electricity market studies, in this paper we present our work regarding the development of a tool that provides a database with available information from real electricity markets, ensuring the required actualization mechanisms.

The presented tool has the capacity of collecting, analyzing, processing and storing real markets data available on-line. Additionally, this tool is built in a way that it can deal with different file formats and types, some of them inserted by the user, resulting from information obtained not on-line but based on the possible collaboration with market entities.

This tool includes the capability of managing files using parallel processing, allowing the system to deal with multiple data sources at the same time. The different data files are accessed through an adaptive approach for automatic downloads of new information available on-line. All procedures are secured by a reliability mechanism that prevents from the storage of incomplete or unviable information.

The final result from the continuous execution of the presented tool is the definition and implementation of a database that gathers information from different market sources, even including different market types.

This is a crucial tool to go a step forward in electricity markets simulation, since the integration of this database with a scenarios generation tool, based on knowledge discovery techniques, will provide a framework to study real market scenarios allowing simulators improvement and validation.

Acknowledgements. This work is supported by FEDER Funds through COMPETE program and by National Funds through FCT under the projects FCOMP-01-0124-FEDER: PEst-OE/EEI/UI0760/2011, and PTDC/SEN-ENR/122174/2010.

References

1. Meeus, L., et al.: Development of the Internal Electricity Market in Europe. The Electricity Journal 18(6), 25–35 (2005)
2. Praça, I., Ramos, C., Vale, Z., Cordeiro, M.: MASCEM: A Multi-Agent System that Simulates Competitive Electricity Markets. IEEE Intelligent Systems, Special Issue on Agents and Markets 18(6), 54–60 (2003)
3. Vale, Z., Pinto, T., Praça, I., Morais, H.: MASCEM - Electricity markets simulation with strategically acting players. IEEE Intelligent Systems, Special Issue on AI in Power Systems and Energy Markets 26(2) (2011)
4. Sun, J., Tesfatsion, L.: Dynamic Testing of Wholesale Power Market Designs: An Open-Source Agent-Based Framework. Computational Economics 30(3), 291–327 (2007)
5. Koritarov, V.: Real-World Market Representation with Agents: Modeling the Electricity Market as a Complex Adaptive System with an Agent-Based Approach. IEEE Power & Energy Magazine, 39–46 (2004)
6. GME - Gestore Mercati Energetici – Italian Energy Market Operator, homepage, `http://www.mercatoelettrico.org` (accessed on January 2014)
7. Shahidehpour, M., et al.: Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management, pp. 233–274. Wiley-IEEE Press (2002)
8. OMIE - Operador del Mercado Ibérico de Energia – Iberian Electricity Market Operator, homepage, `http://www.omie.es` (accessed on December 2013)
9. Nord Pool – homepage, `http://www.nordpoolspot.com` (accessed on December 2013)
10. EPEX – European Power Exchange – homepage, `http://www.epexspot.com` (accessed on April 2013)
11. MISO energy – homepage, `http://www.misoenergy.org` (accessed on April 2013)
12. Cincotti, S., Gallo, G.: Genoa Artificial Power-Exchange. Agents Artif. Intell. (2013)
13. Pinto, T., Vale, Z., Rodrigues, F., Morais, H., Praça, I.: Strategic Bidding Methodology for Electricity Markets using Adaptive Learning. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K., Ali, M. (eds.) IEA/AIE 2011, Part II. LNCS (LNAI), vol. 6704, pp. 490–500. Springer, Heidelberg (2011)

Are There Semantic Primes in Formal Languages?

Johannes Fähndrich, Sebastian Ahrndt, and Sahin Albayrak

DAI Lab, Berlin Institute of Technology, Berlin, Germany
{faehndrich,ahrndt,albayrak}@dai-labor.de

Abstract. This paper surveys languages used to enrich contextual information with semantic descriptions. Such descriptions can be e.g. applied to enable reasoning when collecting vast amounts of row data in domains like smart environments. In particular, we focus on the elements of the languages that make up their semantic. To do so, we compare the expressiveness of the well-known languages OWL, PDDL and MOF with a theory from linguistic called the Natural Semantic Metalanguage.

Keywords: Context Description Languages, Contextual Reasoning, OWL, PDDL, NSM.

1 Introduction

Intelligent environments are made up of multiple pervasive or ubiquitous devices that provide a service to the user. One key indicator of such environments is the ability to adapt to changes. The changes are implied by external or internal influences like the introduction or removal of devices or changing application goals [20]. We expect that intelligent environments react to such changes and adapt themselves in a way that the service provided are still available for the users. More than a decade ago *R.J. Sternberg* [22] still emphasises this specifying that intelligence is the ability to adapt to changes in environments (to distinguish between the environment itself we refer to this as context). This point of view implies that an environment becomes more intelligent if it can cope with more or bigger changes in the context. To be able to adapt to contextual changes a cognition is needed to be aware of the actual context and appearing changes. We focus our analysis to environments where such cognition is available. That means, that there exist at least one entity able to perceive the context and able to communicate the actual perception to other entities in the environment.

One can distinguish two types of contextual information: The **defined context** and the **derived context** [13]. In both cases, the devices making up the intelligent environment have to agree on a language to interpret the data collected by the sensing devices. In a defined context (e.g., a specific application) this language can be given to the environment by a domain model. Another approach is to use semantic languages to annotate contextual information and use reasoner that derive knowledge or facts from this annotations. This approach is called derived context. Here every device has its own local model of the environment, without having to agree on a global context model providing information about all devices. Derived context is created by finding patterns in raw data form the sensing devices of an intelligent environment and annotate them with the

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_46,

given semantic language. A reasoner then reasons upon this annotated information to transform the information into a local domain model of the device. This emphasises the requirement to agree upon a semantic language used for the annotation. Furthermore, it underlines why no model of the whole context is needed.

Languages to describe semantic have been subject to research in many research areas. *Bikakis et al.* [1] surveys semantic based approaches and applicable reasoning methods in the domain of ambient intelligence. Two of them are the Semantic Web community and the Agent community. Both have developed a quasi standard language to describe semantics. In the semantic web community the Web Ontology Language[1] [17] (OWL) is been widely used. The agent community uses Planning Domain Definition Language[2] [7] (PDDL) to describe their planning problems. This paper will examine the fundamental concepts making up those two languages. Additionally the study includes the Meta Object Facility [6] (MOF) as a meta-language for artificial languages. We compare these approaches with a theory form linguistics named the Natural Semantic Metalanguage [8] (NSM). NSM states that every naturally developed language is based on 63 semantic concepts.

The paper is structured in the following way: In Section 2 introduces NSM and the basic concept of semantic primes in a nutshell. Furthermore, it describes the three semantic description languages and the difference between the languages. Section 3 takes such insights into account and compares the languages in a more detailed way. Afterwards, Section 4 wraps-up the paper with a discussion of the results.

2 Semantic Primes

The Natural Semantic Metalanguage (NSM) is a linguistic theory originated in the early 1970s [25]. It states that each meaning of an concepts created in a natural language can be represented using a set of atomic terms—so-called universal semantic primes. These primes have an indefinable word-meaning and can be identified in all natural languages [9]. In conjunction with associated grammatical properties NSM presents a decompositional system able to describe all concepts build in the appropriate language. Here, an expression is decomposed into less complex concepts, where the process ends if the expression is decomposed to the atomic level of semantic primes which can not be analyzed further. One can imagine that the decomposition builds a tree, where all leafs are semantic primes [27]. Consequently for each natural language a metalanguage exist that consist of the semantic primes in the specific syntax and their appropriated grammatical properties. About 63 semantic primes exist that can be divided into 16 categories [26].

As well as natural languages, formal defined artificial languages are based on a metalanguage like the Meta Object Facility. This leads to the implication that the concepts

[1] For further information the interested reader is also refereed to: `http://lists.w3.org/Archives/Public/www-webont-wg/2001Dec/0169.html`

[2] For further information the interested reader is also refereed to: `http://www.cs.yale.edu/homes/dvm/`

defined in artificial languages are semantic primes and that such primes can be compared among different languages. Since the bag of semantic primes presented by NSM is empirically well-researched, this work tries to compare three artificial languages utilising this bag of primes. For this comparison, we take the purpose and concepts of the languages into account and match the available primes with each other as foundation to discuss potentially missing primes in the languages.

Table 1. List of semantic primes with no equivalent found in the other languages

Category	Semantic prime
Substantive	SOMEONE, PEOPLE
Determiners	OTHER/ELSE
Quantifiers	TWO, MUCH/MANY
Evaluators	GOOD, BAD
Descriptors	BIG, SMALL
Mental predicates	THINK, KNOW, WANT, FEEL, SEE, HEAR
Speech	SAY, WORDS, TRUE
Actions, events, movement, contact	DO, HAPPEN, MOVE, TOUCH
Space	WHERE/PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE
Intensifier, augmenter	VERY, MORE
Similarity	LIKE
Time	WHEN/TIME, NOW, FOR SOME TIME, MOMENT
Logical concepts	MAYBE

Table 1 lists all semantic primes presented by NSM not matchable with any concept in the examined artificial languages. In the following we will introduce the artificial languages and the list for each language, which semantic primes are used.

The Web Ontology Language (OWL) is a semantic markup language to create structured knowledge representations and enrich them with semantics. OWL is a W3C standard since 2004 and has been continuously developed since [11]. OWL is an extension of the Resource Description Framework [15] and has become one of the most used language to describe knowledge for AI. Since OWL is meant to describe structured knowledge the concepts used are abstract. Table 2 list all equivalents found in comparison with NSM primes.

The Planning Domain Definition Language (PDDL) is a first-order logic based language defined as an extended BNF [7]. Commonly, it is used to provide a standardized way to describe planning problems and the associated domains. The syntax allows to define among others actions, effects, quantifications and constraints and was intended to enable developers to describe the ”physics” of a domain. Given such a description the reasoner uses a goal defined in PDDL to search for a plan that satisfies all constraints, requirements and preconditions. The concepts which are equivalent to semantic primes are listed in Table 3.

The Meta Object Facility (MOF) has been introduced by the Object Management Group and is formally defined e.g. by *Smith et al.* [21]. MOF has been developed to

Table 2. List of semantic primes with and equivalent found in OWL

Category	Semantic prime	OWL
Substantive	I	self.entry
	SOMETHING/ THING	owl:thing
Relational	KIND	owl:SubClassOf
substantives	PART	owl:topObjectProperty
Determiners	THIS	owl:entityURI
	THE SAME	owl:equivalentClass

Table 3. List of semantic primes with and equivalent found in PDDL

Category	Semantic prime	PDDL
Substantive	SOMETHING/ THING	:define
Determiners	THIS	::=
Existence	THERE IS	:exists
Time	BEFORE	:precondition
	AFTER	:effect
	A LONG TIME	:maintain
	A SHORT TIME	:wait-for
Logical concepts	NOT	:not
	CAN	:action
	BECAUSE	:imply
	IF	:when/constrained

model the structure and behaviour of entities in software development. For example, UML[3] implements MOF. Since MOF is quite abstract, the meta language like OWL mostly has structural semantic primes. Table 4 list all equivalents.

3 Comparison of Primes

The compared languages introduce additional concepts that are domain specific and which are not part of the semantic primes (e.g., 'OWL:VERSIONINFO'). Depending on the purpose of the language those additional concepts change. OWL for example was created to describe shared conceptualizations where versioning and backward compatibility is an issue. But from the theory of NSM those concepts could be described using the other semantic primes. Thus they are merely shortcuts. There are multiple extensions to those languages for special cases like the Semantic Web Rule Language (SWERL) [14], which introduces rules to OWL. We now discuss the 16 categories of semantic primes to analyse why such concepts do or do not exist.

Substantives are the first category. In natural language these semantic primes are used to distinguish actors and to separate humans from other things. To describe meaning,

[3] See: `http://www.omg.org/spec/UML/`

Table 4. List of semantic primes with and equivalent found in MOF

Category	Semantic prime	MOF
Substantive	YOU	uri
	SOMETHING/ THING	object
	BODY	instance
Relational	KIND	type, extent
substantives	PART	property
Determiners	THE SAME	element.equals
Quantifiers	ONE	multiplicityElement
Location	BE (SOMEWHERE)	link
Existence	THERE IS	element
	HAVE	classifier
	BE (SOMEONE/SOMETHING)	extend
Life and death	LIVE	create
	DIE	delete
Logical concepts	CAN	operation
	IF	event

humans often reduce description of properties of things to the relation to humans or more precise them self [27]. For example, to describe the concept 'mouse' a semantic, context independent description most likely rely on a degree in Biology. Describing a mouse so that a naïve reader of the description might understand it, the description can refer to the potential context of the reader. In natural languages these readers are other humans, which implies that most description in NSM are in the context of humans and their relation to things. The distinction of 'YOU' and 'I' is needed if roles are described, i.e. in negotiation or contracting. Non of the reviewed language has a concept for 'PEOPLE'. On the one hand semantic description in these languages are thought for artificial reasoners and there a concept of 'PEOPLE' is not needed to describe most concepts. On the other hand, in the area of HCI the distinction of artificial agents and human agents can be of some concern and with that the concept of 'PEOPLE' might be required.

The category of *relational substantives* are well represented in two of the languages, except PDDL as it does not use type hierarchies to define domains. That means that PDDL does not semantically aggregate all instances of one 'KIND'. In semantic descriptions 'KIND' is for example used to describe water: 'something of one kind' [10].

Quantifiers are represented in all tree languages. The exception is the fuzzy representation of 'MUCH/MANY' and 'SOME'. There is a need to enable fuzziness in semantic description languages as motivated by *Stoilos et al.* [23].

The category of *Evaluators* is not represented in any language. PDDL from version 2.1 features numeric fluent to describe e.g., cost for actions, which could be interpreted depending the metric. Here an implicit metric is given to the reasoner, e.g. the *plan minimization metric* [7]. This metric states that the value to be optimized is better the smaller the value. Stating that less is better. We argue that such a metric can be explicitly formalized in the description itself and to define what is 'BAD' or 'GOOD' those concepts need to be part of the description language.

Descriptors are not represented in all three analyzed language. ‘BIG’ and ‘SMALL’ are fuzzy values and are defined in an description. For example, *Wierzbicka* [27] describes mice as small in the following way: ’They are very small. A person can hold one easily in one hand’. Giving a example on what small means and a relation to something every reader of the explanation knows: the size of a hand. In relation to some reference or as constant like the semantic primes ‘A LONG TIME’ and ‘A SHORT TIME’ these can be used to describe relations in e.g. size, intensity, power or time. We can imagine that for example a timeout can be explained by defining ‘A LONG TIME’ as the maximum timeout. We argue that if the semantic description is used by a reasoner to create a heuristic, a metric needs to be defined and with that semantic descriptors are needed which classify the value which is subject to the metric. Further we argue that ‘LONG’ and ‘SHORT’ should be part of the descriptors since they are descriptors which could be used in addition to time with other concepts like spacial distances.

Mental predicates are not represented in all three languages doe to the fact that these predicates are based on the senses of human being. We separate mental predicates in two groups: The first group is based on human cognition, ‘FEEL’, ‘SEE’ and ‘HEAR’. First of all there are two of the senses missing: ‘SMELL’ and ‘TASTE’. Additionally for the domain of intelligent environments and if sensors needs to be described, such predicates can be used. Of cause this is specific for humans, since not every agent involved in an intelligent environment has such cognitive functions. Additionally there could be sensors which extend the human cognitional like a barometer, altimeter or a localization like GPS, which should then be incorporated in the semantic primes as well. Even thou these concept are not used in the analysed modelling languages, they are part of the semantic description of many fundamental concepts like ’HANDS’ [27]. The second group of mental predicates is the mental state of mind: ‘THINK’, ‘KNOW’, ‘WANT’. These are philosophical terms and rarely used in artificial languages. Braubach et al. [2] e.g. describe a Believe, Desire, Intent (BDI) paradigm for agents. Here ‘I believe’ is considered a subset of ‘I think’ [26]. In the BDI paradigm ‘believe’ can be mapped to the semantic prime ‘KNOW’, since it represents the knowledge of the agent and ‘desire’ can be mapped to ‘WANT’ since it describes the internal goals of the agent. But in our analysed languages, all of these concepts are missing.

Speech is - at first - the category which hold one of the basic logical operators ‘TRUE’. All three languages use an implicit representation of the concept ‘TRUE’ since they assume that a reasoner interprets an axiom as fundamentally true. PDDL e.g. ‘define a model to be an interpretation of a domain’s language that makes all its axioms true’ [16]. Thus here again it can be argued to explicitly describe such truth values and with that add a semantic prime to the metalanguage. But we think that ‘TRUE’ should be caped in the category ‘Logical concepts’. Further we use ‘WORDS’ as basic building blocks for our description, and thus need a semantic prime for them. ‘SAY’ has been represented in a formal way as agent communication speech acts [4] and could be directly part of the metalanguage.

Semantic primes in the category *Action* are often defined in an context dependent manner, where the semantic is given by the reasoner of the evaluating the axioms. In PDDL for example the blocks world defines‘MOVE(A,B,T)’ [12]. NSM proposes to add such primes to the metalanguage, to be able to describe events, movement and actions.

The semantic primes in the category *possession* e.g., ‘HAVE’ can be seen as the specialization of composition and aggregation of the semantic prime ‘PART’. The *specification* ‘BE’ denotes a location where something is located and at the same time to be of one type.

Life and death is a category which is not subject to research in formal languages because computer systems rarely need a concept of death or living. Semantically there are many things that can not be described without the concept of ‘LIVE’ and ‘DEATH’. In agent communication e.g.m agents send ‘alive messages’ to other agents, where the interpretation is left to the programmer of the agent.

Time has found its way into almost every formal language. Even a own logic—the Temporal logic—which is a kind of modal logic has been created to model something like: ‘I am hungry until I eat something’. In formal languages time has often been included into the language e.g., PDDL from version 2.1.

The category *space* is subject to research and is formulated in contextual models like the CORBA-ONT [3]. Nevertheless non of the surveyed languages presents primitive elements to describe special properties. The fact that a OWL ontology is required, shows that such semantic primes are necessary for the modelling of contexts. The same can be argued for the semantic primes ‘BE (SOMEWHERE)’ of the category location. Other fuzzy primes like ‘NEAR’ or ‘FAR’ are again hard to grasp in a formal language.

Logical concepts are of cause part of most formal languages. The hurdle primes are again the fuzzy ones: ‘MAYBE’ and ‘CAN’. To describe the meaning of probability, those primes could be part of a language like they are in epistemic logic [24].

Intensifier can be modeled as lexical functions [18] and fuzzy decision systems have been subject to research [19] and thus to make their semantic explicit intensifier should be part of the metalanguage.

Similarities are a huge research area and measures have been studied in depth [5]. The here developed methods like recommender systems try to find entities which are alike. Those methods try do define the prime ’SAME’ for different domains.

4 Conclusion

We have analysed three common semantic description languages and compared their meta languages with the set of semantic primes taken from NSM. We have found that already many of the semantic primes are part of the three formal description languages depending on their focus. The semantic primes that are not yet part of the description languages has been collected in Table 1. Future work will include an in-depth analysis of those primes. Here we want to examine which primes are useful for formal languages and define a set-theoretic semantic for each of them.

Acknowledgement. Research in this paper has been financed by the Schaufenster Elektormobilität http://www.schaufenster-elektromobilitaet.org in the project 16SBB007A.

References

1. Bikakis, A., Patkos, T., Antoniou, G., Plexousakis, D.: A survey of semantics-based approaches for context reasoning in ambient intelligence. In: Mühlhäuser, M., Ferscha, A., Aitenbichler, E. (eds.) AmI 2007 Workshops. CCIS, vol. 11, pp. 14–23. Springer, Heidelberg (2008)
2. Braubach, L., Pokahr, A., Moldt, D., Lamersdorf, W.: Goal representation for bdi agent systems. In: Bordini, R.H., Dastani, M., Dix, J., El Fallah Seghrouchni, A. (eds.) PROMAS 2004. LNCS (LNAI), vol. 3346, pp. 44–65. Springer, Heidelberg (2005)
3. Chen, H., Finin, T., Joshi, A.: An ontology for context-aware pervasive computing environments. The Knowledge Engineering Review 18(03), 197–207 (2003)
4. Colombetti, M., Verdicchio, M.: An analysis of agent speech acts as institutional actions. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 3, pp. 1157–1164. ACM (2002)
5. Duda, R.O., Stork, D.G., Hart, P.E.: Pattern classification and scene analysis. Part 1, Pattern classification. Wiley, Chichester (2000)
6. Favre, J.-M.: Foundations of meta-pyramids: languages vs. metamodels. In: Episode II. Story of Thotus the Baboon, Procs. Dagstuhl Seminar, vol. 4101 (2004)
7. Fox, M., Long, D.: PDDL2.1: An extension to PDDL for expressing temporal planning domains. Journal of Artifical Intelligence Research 20, 61–124 (2003)
8. Goddard, C., Wierzbicka, A.: Semantic and lexical universals: Theory and empirical findings, vol. 25. John Benjamins Publishing Company (1994)
9. Goddard, C.: Natural Semantic Metalanguage: The state of the art. In: Cross-Linguistic Semantics, pp. 1–34. John Benjamins Publishing Company (2008)
10. Goddard, C.: Semantic molecules and semantic complexity(with special reference to "environmental" molecules). Review of Cognitive Linguistics 8(1), 123–155 (2010)
11. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: Owl 2: The next step for owl. In: Web Semantics: Science, Services and Agents on the World Wide Web, vol. 6(4), pp. 309–322 (2008)
12. Gupta, N., Nau, D.S.: On the complexity of blocks-world planning. Artificial Intelligence 56(2), 223–254 (1992)
13. Henricksen, K., Indulska, J.: Modelling and using imperfect context information. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops 2004, pp. 33–37. IEEE (2004)
14. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al.: Swrl: A semantic web rule language combining owl and ruleml. W3C Member Submission 21, 79 (2004)
15. Lassila, O., Swick, R.R.: Resource description framework (rdf) model and syntax specification (1999)
16. McDermott, D.: The formal semantics of processes in pddl. In: Proc. ICAPS Workshop on PDDL, pp. 101–155. Citeseer (2003)
17. McGuinness, D.L., Van Harmelen, F., et al.: Owl web ontology language overview. W3C Recommendation 10(2004-03), 10 (2004)
18. Melčuk, I., Wanner, L.: Lexical functions and lexical inheritance for emotion lexemes in german. In: Benjamin, J. (ed.) Lexical Functions in Lexicography and Natural Language Processing, Amsterdam, Philadelphia, pp. 209–278 (1996)
19. Grundlagen der Theorie Unscharfer Mengen. Fuzzy decision support-systeme (1994)
20. Salehie, M., Tahvildari, L.: Self-adaptive software: Landscape and research challenges. ACM Transactions on Autonomous and Adaptive Systems 4(2), 1–42 (2009)

21. Smith, J., DeLoach, S., Kokar, M., Baclawski, K.: Category theoretic approaches of representing precise uml semantics. In: ECOOP Workshop on Defining Precise Semantics for UML (2000)
22. Sternberg, R.J.: Beyond IQ: A triarchic theory of human intelligence (1985)
23. Stoilos, G., Simou, N., Stamou, G., Kollias, S.: Uncertainty and the semantic web. IEEE Intelligent Systems 21(5), 84–87 (2006)
24. van der Hoek, W.: Epistemic logic for AI and computer science, vol. 41. Cambridge University Press (2004)
25. Wierzbicka, A.: Semantics: Primes and Universals. Oxford University Press (1996)
26. Wierzbicka, A.: English: meaning and culture (2006)
27. Wierzbicka, A.: Mental lexicon. In: The Slavic Languages: An International Handbook of Their History, Their Structure and Their Investigation. Mouton de Gruyter, Berlin (2009)

The Age of Confidentiality: A Review of the Security in Social Networks and Internet

Antonio Juan Sánchez[1] and Yves Demazeau[2]

[1] Universidad de Salamanca, Salamanca, Spain
anto@usal.es
[2] Laboratoire d'Informatique de Grenoble - CNRS, Grenoble, France
Yves.Demazeau@imag.fr

Abstract. Security based on content analysis in social networks has become a hot spot as a result of the recent problems of violations of privacy by governments to international security agencies. This article is an approach to the implementation of programs for extraction and analysis of the web information, this includes the technical difficulties and the legal consequences involved.

Keywords: Social networks analysis, security, data meaning, information fusion, wrappers, protection data laws, organizations, agents.

1 Introduction

Today, security has become a fundamental point in the information era. Just as all eyes are on this critical point and it affects all levels from the individual human being to international corporations. The cyber-crime has grown exponentially since the massive use of Internet and mobile devices, always protected by the anonymity of the criminals. The new technologies and their ability to analyse large amounts of information (Big Data) have made a breakthrough, the anonymity of aggressors is no longer such. Every step we take is reflected on the Internet in different ways, through different databases scattered around the globe, is what has been called digital fingerprint.

Every adverse comment we left on Facebook, every tweet that may contain traces of a past or future crime, every comment in a blog that may damage the self-image, all of them can be now analysed and linked with our personal data. These data can be extracted from the same sources where can be found documents with sensitive content (photos, videos) or information about our life and work (LinkedIn, Google+, telephone records on-line yellow pages extracted as, listings in public entities, etc.). In most cases, user data from mass use websites are data that can be accessed freely. Throughout this article the different strategies and tools that can be used to get personal content on Internet will be discussed as well as the legal issues involved, and finally a platform that supported all this strategies will be proposed.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_47, © Springer International Publishing Switzerland 2014

2 Extracting Information from the Web

The semantic web has tried to simplify the process of extracting information, but the fact is that today the semantic capacity has not yet been implemented in most of the websites.

Scraping tools (download web content most used) today are Wget[1], cURL[2] or Pillow[3], although there are numerous libraries for most languages as Scrapy[4] and Beautiful Shop[5] (both for Phyton), Html Agility Pack (.NET)[6], JSoup (Java)[7], htmlSQL (PHP)[8]. Beautiful Shop is the most complete of all bookstores because it supports even malformed HTML.

The crude extract of the contents of the web is the most developed stage of the analysis and involves minor problem. The difficulty increases when we talk about information processing to get useful information obtained.

There is no doubt that the nature of the information obtained depends on the type of paper used, but also we must consider that in each document can be found associated information does not correspond to the nature of this: for example, a photograph is mainly graphical information, but the metadata of the image file may also contain information about where it was made (many GPS mobile devices leave a mark) or the date and time it was made. In addition, passing an OCR to the picture, we can find texts, which could then be processable by a computer.

There are many different types of tools to extract information from the obtained data. To process the text got on web pages, first of all it's necessary to select and then label content that may be useful to us. Wrappers are used to select [2][3] (data mining programs specialized in information extraction). There are five different approaches to generate wrappers [6]:

- **Wrapper induction** need specific rules. These rules can be determined by a user or by a training-induced. In both cases supervised algorithms [5]. Within this category are the NLP (Natural Language Processing) [7].
- **Model based Wrappers** operate inversely to wrapper induction. They try to find underlying structures within the web content to serve as a model to extract data [8].

---

[1] GNU Wget. `http://www.gnu.org/software/wget/`
[2] cURL. Command line tool for transferring data with URL syntax. `http://curl.haxx.se`
[3] Pillow. Tool for Phyton for transferring data with URL syntax `https://pypi.python.org/pypi/Pillow/`
[4] Scrapy. Screen scraping and web crawling framework `http://scrapy.org/`
[5] Beautiful Soup. Scraping framework for Python. `http://www.crummy.com/software/BeautifulSoup/`
[6] Html Agility Pack L. Scraping framework for PHP. `https://github.com/hxseven/htmlSQL`
[7] JSopup. Scraping framework for Java. `https://github.com/hxseven/htmlSQL`
[8] htmlSQL. Scraping framework for PHP. `https://github.com/hxseven/htmlSQL`

- **The self-extracting wrappers** select relevant content using regular expressions and dictionaries [4]. Among the most comprehensive tools in this section we find Camaleon [20] that supports HTML, Javascript and cookies.
- **The ontology based** seek constant within the content and build objects around them. The most important work in this aspect has been made in the Brigham Young University [23] and it's based on a RDF extraction system [21] and DAML ontology[9].
- **Automatic design languages** are definition languages wrappers, generally allow the definition of grammars through EBNF or regular expressions with help of the typical exceptions to procedural languages. Notable in this kind of languages: Minerva, Web-OQL and WebL.
- **The HTML-based generators** are strictly based on the knowledge of how this language tag to select the important parts. Among the most important tools are W4F, XWRAP, Road Runner or LIXTO.

The optimum solution for generate wrappers would be a mixed technique of Wrappers based in Model and Ontology, generating data structures based in the objects found in the ontology.

The most widely model used to classify information properly is an ontology. Ontologies can generate a vocabulary with the data, specify the meaning of terms and formalize the relationships between them using a suitable logic [9]. Some wrappers facilitate the process of step to ontology, by providing the ontology formed.

The fundamental difficulty of step to ontology comes from the scalability, and it is essential if we think that the extraction and storage of the web information is taken continuously once the process is started. The higher an ontology becomes, the greater the difficulty in applying an algorithm for reasoning about it, is a problem of type 2NEXPTIME-complete [9]. The proposed solutions to this problem agree on the division of the ontology groups, through a clustering algorithm [25], which always grow at much less speed than the set [10].

3 Is It Legal to Extract Data from Internet for Purposes Related to Security?

The data protection laws and copyright is an important point to deal since they mark the limit of the data that may be extracted from Internet. In Spain the legislative rules that regulate are “Ley 15/1999 de Protección de Datos Personales” or LOPD, “Ley 34/2002 de Servicios de la Sociedad de la Información y el Comercio Electrónico” (LSSICE), and “Ley 56/2007 de Medidas de Impulso de la Sociedad de la Información”.

Many websites, such as Twitter, Facebook o Google+, require to leave public information from the information extractor (usually through a registration in the web developer) and require that the information obtained can’t be used outside of the social network. On the other hand, the screen scraping is approved in Spain as legal

[9] DAWL-OIL. OWL. Web Ontology Language `http://www.w3.org/2001/sw/wiki/OWL`

technique if it is done under certain assumptions [11], ie, the extraction of sources such as blogs, newspapers or news pages is only protected by copy-laws of intellectual character.

In Spain, Ley 25.326 protects the data of citizens, and their possibility to freely disclose their personal data. Transfers of data to companies to obtain services usually involve the protection of these data. Furthermore, according to the data protection law 15/1999, the personal data collected -with consent- of a certain website may be used only for purposes directly related to the functions of the transferor and transferee and must be a connection between the activities and purposes of the communication of the transferor and transferee companies.

Thus, initially, the extraction of data from social network users for studies or statistics outside the network is prohibited by law, but according to the Spanish Law of Criminal Procedure 579, a judge may grant the interception of communications in cases of national security, avoiding also Article 8 of the European Convention for the protection of Human Rights and Fundamental Freedoms affecting privacy.

4 Data Analysis

4.1 How to Analyze the Data Obtained

Each of the sources available on the Internet gives us 5 different perspectives fingerprint left by web users. For your personal data is necessary to merge information from different sources, such as social networking, government listings, company records, state news bulletins or telephone directories, and delete those erroneous data.

From the comments in conventional social networks can be drawn about the personality characteristics of the subjects: hobbies, interests, and political, sexual and moral preferences.

Of employment-oriented social networks like LinkedIn or Infojobs can get your professional profile, professional curriculum and academic data.

Merging data from their networks of professional and personal contact can get your closest contacts. If we consider these networks of relationships between people are graphs, you can apply all the related theory and small-world networks to establish communities and circles of trust. There are many practical examples of this, to highlight Anheier, Gerhards and Romo [12] Wouter De Nooy [13] and Burgert Senekal [14], based on the theories of Ronald S. Burt [15].

Through photographs and multimedia documents can plot profile of the subject and get a map and a calendar of your visited websites. The retrieval of this information is not too problematic, there are specialized tools and techniques in it. To extract information automatically from the metadata multimedia files can be used as Tika libraries[10], MediaInfo[11] or Exif Tags Collection[12]. To extract information from sound

[10] Tika Project. `http://tika.apache.org/`
[11] MediaInfo Project. `http://mediainfo.sourceforge.net`
[12] Exif Tags Collection Project. `https://www.codeproject.com/script/Members`

content type and pass it to text would require other libraries like Microsoft Speech API for .NET[13], Cmu Sphinx[14] for Java o Julius[15] , Kalidi[16], Simon[17], iATROS[18] for C. To extract textual information from videos and photos there is a wide range of OCR type solutions both free and paid license.

Crowd of on-line applications are dedicated today to make this information collection and create comprehensive profiles of users of social networks, so far these profiles have always responded to public properties and / or objective of subjects. The greatest interest in security and privacy do not provide these features, but those more personal that allow make a profile more focused on the psychological and personal.

4.2 Data Groupings

As discussed above managing the wealth of information that involves extracting data from the web takes to divide it to make it operatively computable.

The partition can be accomplished by different techniques at different times of the process. The nature of the data makes these may be divided at least when perform design ontologies and networking. For the division of space relations of the individual, detection techniques of communities are commonly used [16], while for the ontologies, besides the graph analysis procedures, cluster analysis (either hierarchical or partition) can be applied.

4.3 Extracting Results

After the whole process of clustering of terms in the ontology and the distinction of communities only would analyse the "psychological" data of individuals, ie verifying that the texts employees doing their comments or reviews not show signs of offense (whether physical or virtual).

In this case, special ontologies will be required very tight to the terminologies searched cybercrimes, since the semantics used is substantially different from the ordinary natural language.

These ontologies can be created from the application of various supervised algorithms on texts.

The self-learning system for detecting future crimes or from changes in ontologies, is based on anomaly detection in texts. In David Guthire studies [18] about significant deviations from the context different detection methods can be found.

---

[13] Microsoft Speech API `http://www.microsoft.com/downloads/details.aspx?FamilyID=5e86ec97-40a7-453f-b0ee-6583171b4530`

[14] Cmu Sphinx. `http://cmusphinx.sourceforge.net/sphinx4/`

[15] Julius in Sourceforge. `http://julius.sourceforge.jp`

[16] Kalidi in Sourceforge `http://kaldi.sourceforge.net`

[17] Simon. `http://simon-listens.blogspot.com/`

[18] iATROS. `http://prhlt.iti.es/software/iatros/doc/speech/`

Moreover, another problem is the detection of fake profiles. Studies of the National Institute of Technology of Roukela [19] prove that by techniques of SVM (Support Vector Machine) 95% of the fake profiles can be detected taking into account the following attributes of each profile: number of friends, education / current job, amount of text on own information, marital status, number of images, number of comments in other people's profiles, percentage among friends of the same gender and friends in total, percentage between applications of "friendship" sent and received, number of groups to which it belongs and the number of "I like you" (or similar).

5 Conclusions

5.1 Case Studies of Implementation of Cybersecurity Systems

Both companies and public entities have begun in recent years to promote cybersecurity projects. Large intelligence agencies internationally as CIA, Interpol, DEA or FBI have their own software but for safety information they are inaccessible.

Other projects due to its link with research teams are a little more open to analysis.

Riot is a project of the U.S. defense company Raytheon, for monitoring individuals from storage and fusion of data on social networks like Facebook, Twitter or FourSquare. Riot uses the contents of comments and the information contained in the metadata of media relations for plotting graphs, maps the movement of individuals, and predict future movements.

In August 2012, the University of Sydney released version 2.0 of the tool GEOMI (maximum penetration geometry) for data visualization. GEOMI allows police and security agencies to visualize and analyze complex relationships in social networks, email and phone records [22].

5.2 Proposal for a System with Agents

Automated search of crime through the content analysis in online social environments still involves technical difficulties that will soon be overcome:

- Detect false profiles
- Find a optimal rime dictionary
- Automated the information extraction from different systems
- Simulated future scenarios of crime

As has been shown, some government cybersecurity programs in different countries already operate with reduced automation. The main problem of application software for the cyberdetection of crimes is the maintenance required, since it is very difficult to ensure effective probability of success in generating new dictionaries criminal terms.

In any case, the results should never be understood as significant evidence of a crime but cyberdetection software should be better understood as an example of decision support system.

As the process of collection, storage and analysis of information is complicated, to work in the future an approach with a society of distributed agents [24] is proposed (see Figure 1).

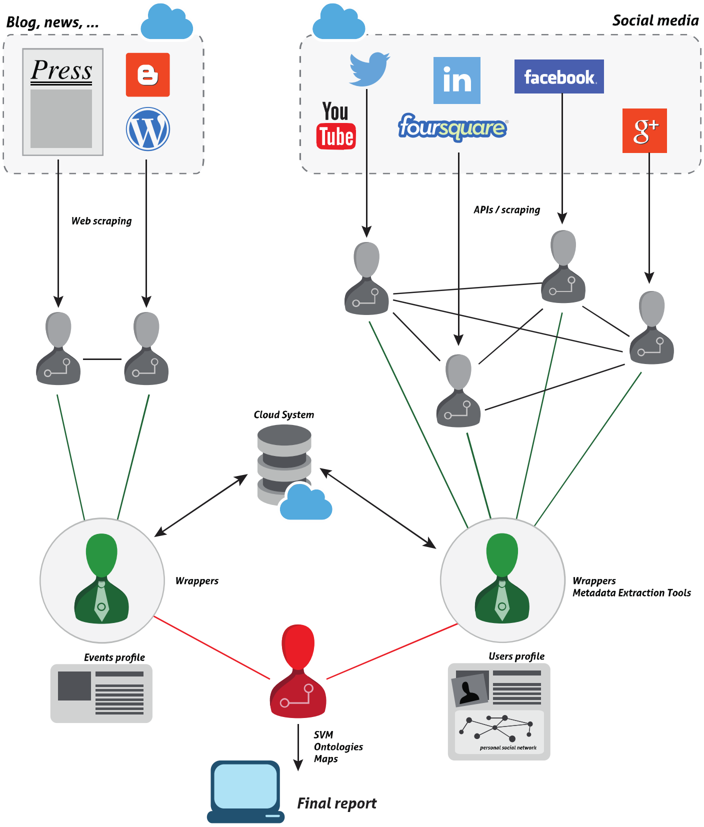


Fig. 1. Extraction and analysis process with agents

Acknowledgement. This work has been partially supported by the MINETUR project PIAR, TSI-100201-2013-20.

References

1. W3C Semantic Web Activity. World Wide Web Consortium (W3C) (November 7, 2011) (retrieved November 26, 2011)
2. Kushmerick, N., Weld, D.S., Doorenbos, R.: Wrapper Induction for Information Extraction. In: Proceedings of the International Joint Conference on Artificial Intelligence (1997)

3. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer (2007)
4. Dalvi, N., Kumar, R., Soliman, M.: Automatic wrappers for large scale web extraction. Proceedings of the VLDB Endowment 4(4), 219–230 (2011)
5. Kushmerick, N., Weld, D.S., Doorenbos, R.: Wrapper Induction for Information Extraction. In: Proceedings of the International Joint Conference on Artificial Intelligence (1997)
6. Laender, A., Ribeiro-Neto, B., Silva, A., Teixeira, J.: A Brief Survey of Web Data Extraction Tools. SIGMOD Record 31(2) (June 2002)
7. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009) ISBN 978-0-596-51649-9
8. Alderberg, B.: NoDoSe – A Tool for semi-automatically extracting structured and semistructured data from text-documents. In: Proceddings of ACM SIGMON International Conference on Management of Data, Seattle
9. Glimm, B., Horrocks, I., Motik, B., Shearer, R., Stoilos, G.: A Novel Approach to Ontology Classification. J. of Web Semantics 14, 84–101 (2012)
10. Stuckenschmidt, H., Klein, M.: Structure-based partitioning of large concept hierarchies. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 289–303. Springer, Heidelberg (2004)
11. Sentencia del Tribunal Supremo de España 9 de Octubre de, número 572/2012 (2012)
12. Anheier, H.K., Gerhards, J., Romo, F.P.: Forms of capital and social structure of fields: Examining Bourdieu's social topography. American Journal of Sociology 100, 859–903 (1995)
13. De Nooy, W.: Fields and networks: Correspondence analysis and social network analysis in the framework of Field Theory. Poetics 31, 305–327 (2003)
14. Senekal, B.A.: Die Afrikaanse literêre sisteem: 'n Eksperimentele benadeing met behulp van Sosiale-netwerk-analise (SNA), LitNet Akademies (2012)
15. Burt, S.R.: Structural Holes: The Social Structure of Competition. Havard University Press (1992)
16. Coskun, G., Rothe, M., Teymourian, K., Paschke, A.: Applying Community Detection Algorithms on Ontologies for Identifying Concept Groups. In: Proc. of the Fifth International Workshop on Modular Ontologies (WoMO 2011), pp. 12–24 (2011)
17. Leskovec, J., Lang, K.J., Mahoney, M.W.: Empirical Comparison of Algorithms for Network Community Detection. In: WWW 2010: ACM WWW International Conference on World Wide Web (2010)
18. Guthrie, D.: Unsupervised Detection of Anomalous Text. Department of Computer Science University of Sheeld (July 2008), http://nlp.shef.ac.uk/Completed_PhD_Projects/guthrie.pdf
19. Barrientos, F., Ríos, S.A.: Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones. Revista de Ingeniería de Sistemas 27 (Septiembre 2013)
20. The Camaleon Web Wrapper Engine, http://chameleon.readthedocs.org/en/latest
21. RDF. Semantic Web Standards, http://www.w3.org/RDF/
22. Ahmed, A., et al.: GEOMI: GEOmetry for Maximum Insight. In: Healy, P., Nikolov, N.S. (eds.) GD 2005. LNCS, vol. 3843, pp. 468–479. Springer, Heidelberg (2006), http://sydney.edu.au/engineering/it/~cmurray/geomi.pdf

23. Liddle, S.W., Embley, D.W., Scott, D.T., Yau, S.H.: Extracting Data Behind Web Form. In: Olivé, À., Yoshikawa, M., Yu, E.S.K. (eds.) ER 2003. LNCS, vol. 2784, pp. 402–413. Springer, Heidelberg (2003)
24. Garijo, F., Gómes-Sanz, J.J., Pavón, J., Massonet, P.: Multi-agent system organization: An engineering perspective. In: Pre-Proceeding of the 10th European Workshop on Modeling Autonomous Agents in a Multi-Agent World, MAAMAW 2001 (2001)
25. Castellanos-Garzón, J., García, C., Novais, P., Díaz, F.: A visual analytics framework for cluster analysis of DNA microarray data. Expert Systems With Applications 40(2), 758–774 (2013) ISSN: 0957-4174

Extracting Sentences Describing Biomolecular Events from the Biomedical Literature

Tiago Nunes, Sérgio Matos, and José Luís Oliveira

DETI/IEETA, University of Aveiro,
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
{tiago.nunes,aleixomatos,jlo}@ua.pt
http://bioinformatics.ua.pt/

Abstract. The scientific literature is one of the main sources of information for researchers. However, due to the rapid increase of the number of scientific articles, satisfying a specific information need has become a very demanding task, and researchers often have to scan through a large number of publications in search of a specific nugget of information. In this work we propose the use of supervised machine learning techniques to retrieve and rank sentences describing different types of biomolecular events. The objective is to classify and rank sentences that match any general query according to the likelihood of mentioning events involving one or more biomolecular entities. These ranked results should provide a condensed, or summarized, view of the knowledge present in the literature and related to the user's information need.

Keywords: Sentence-based Information Retrieval, Biomedical Literature, Biomolecular Events.

1 Introduction

Due to the rapid increase of the number of scientific articles, finding articles satisfying a specific information need has become a very demanding task, and researchers often have to scan through a large number of publications in search of specific nuggets of information. In the biomedical field, the most popular information retrieval system, PubMed, combines the MeSH based indexing provided by the MEDLINE literature database with boolean and vector space models for document retrieval [1], giving researchers access to over 22 million citations.

Over the past years, a range of systems have been developed to help users quickly and efficiently search and retrieve relevant publications from the MEDLINE collection. This was facilitated by the free availability of MEDLINE data and Entrez Programming Utilities [2], which make it possible for external entities – from either academia or industry – to create web-based tools that provide alternative ways to search over PubMed. Lu presented a comprehensive review of tools complementary to PubMed [3]. Most of these take advantage of the domain knowledge available in databases and resources such as Entrez Gene, UniProt, Gene Ontology (GO) or the Unified Medical Language System

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_48,

(UMLS) to process the titles and abstracts of texts and present the extracted information in different forms, such as relevant sentences describing biological processes, relationships extracted between various biological entities, or in terms of co-occurrence statistics between domain terms.

In this work, we studied the problem of retrieving informative sentences describing biomolecular events, such as binding, localization or regulation, for example. Classifying an arbitrary sentence from a given scientific paper as being informative or not is an abstract task, highly dependent on the information need of the reader. Classifiers relying on large sets of complex, handcrafted rules have been shown to yield good results in the identification of key sentences from specific types of text, like web pages, news articles, clinical records, narrative admission reports, and radiology reports [4–7]. However, creating such classification systems requires human labour, is extremely time consuming and demands both linguistic and domain-specific knowledge of the text being classified. For this reason, an automatic approach for the identification and extraction of features that make a sentence informative in a given domain presents multiple benefits. First, being an automatic system, once developed it can be easily and quickly applied to different types of informative sentences. Second, it can be constantly improved whenever new, larger sets of annotated data are available.

We propose the use of supervised machine learning techniques to classify and rank sentences describing different types of biomolecular events. These classifiers could be used within an information retrieval system to rank sentences from scientific documents that match a given query, according to the likelihood of mentioning events involving one or more biomolecular entities. These ranked results should provide a condensed, or summarized, view of the knowledge present in the literature and related to the user's information need. Additionally, this could also be used as a starting point for extracting and curating bio-pathway knowledge from the literature.

2 Data and Methods

2.1 Corpora

In order to train classification models for our purpose, we adapted data from different corpora aimed at the identification and extraction of biomolecular events and corresponding arguments. This is a complex text mining task that has attracted large interest recently but that is far from being successfully tackled. Although the task we propose is much less complex, we can adapt and use the data used for event extraction. These datasets are described next.

BioNLP-ST 2011 Datasets. The main tasks of BioNLP-ST 2011 were GENIA (GE), Epigenetics and Post-translational Modifications (EPI) and Infectious Diseases (ID). The GE task focused on extraction of bio-events related to genes or gene products, without distinguishing between genes and proteins [8].

Table 1. Datasets statistics after pre-processing and tagging of sentences

Dataset	Sentences	Informative		Non-Informative	
GE	11579	4785	41.32%	6794	58.68%
EPI	7688	1394	18.13%	6294	81.87%
ID	3184	1226	38.51%	1958	61.49%
Angiogenesis	2784	1260	45.30%	1524	54.70%
Total	**25235**	**8665**	**34.34%**	**16570**	**65.66%**

The EPI task was centred on events relating to epigenetic change, including DNA methylation and histone modification, as well as other common post-translational protein modifications [9]. The ID task was an event extraction task focusing on the bio-molecular mechanisms of infectious diseases [9].

Angiogenesis Corpus. The angiogenesis gold standard corpus was constructed from MEDLINE abstracts randomly selected from a pool of documents mentioning angiogenesis events. Documents were then manually annotated by domain experts with terms and bio-events related to angiogenesis [10].

2.2 Data Preparation

The four datasets were pre-processed using Neji [11], a Java framework for document processing and concept recognition and annotation. After applying sentence splitting to extract individual sentences from the 3,256 abstracts and 87 full-text documents, a total of 25,235 sentences were identified. In order to train the sentence classification models, we used as positive instances the sentences containing gold-standard annotations for a bio-event, while all the remaining sentences were used as negative instances. In some cases, the bio-event annotations may span over more than one sentence, i.e. the event trigger appears in one sentence while at least one of the event participants appears in another sentence. In such cases, those sentences were marked as negative.

Pre-processing and identification of informative sentences where bio-events are mentioned resulted in the tagging of 8,665 sentences, corresponding to 34.34% of the complete sentence collection. Detailed statistics about the sentences marked in each dataset are presented in Table 1.

For assessing the classification and retrieval performance, we considered the common evaluation measures $Precision = TP/(TP+FP)$, $Recall = TP/(TP+FN)$ and $F1 = 2 \times Precision \times Recall/(Precision + Recall)$, were TP refers to true positives, FP to false positives, and FN refers to false negatives.

Another important way of comparing the performance of binary classifiers is by evaluating the receiver operating characteristic (ROC), and namely by calculating the respective area under the curve (AUC), which measures the probability that the classifier will assign an higher rank to a random positive instance than to a random negative instance. Additionally, we also measured accuracy, defined

Table 2. Classification performance results for the baseline (simple bag-of-words)

Dataset	F-score	Prec.	Recall	Av.Pr.	AUC	Accuracy
All	0.703	0.708	0.698	0.755	0.774	0.797
GE	0.745	0.748	0.742	0.798	0.783	0.790
EPI	0.688	0.712	0.667	0.720	0.803	0.890
ID	0.773	0.789	0.758	0.820	0.816	0.829
Angio	0.738	0.753	0.725	0.801	0.764	0.768

as the ratio of true classification $(TP + TN)$ in relation to the total number of instances. Moreover, since IR systems usually return a ranked list of documents, it is desirable to measure performance while considering the order in which the documents are presented. This can be obtained through the average precision metric, which measures the average of the precision over different recall values.

3 Results and Discussion

3.1 Baseline

In order to set a starting point for feature exploration and reference values to try to improve the classification model, we established a baseline consisting of one of the simplest possible models for text classification, the bag-of-words. Classification results for a Linear SVM trained on bag-of-words features are presented in Table 2. Cross-validation results on the whole dataset showed reasonable performance, achieving accuracy of almost 80%, f-measure of 70%, average precision of 75.5% and AUC score of 77.4%. Separate cross-validation evaluations on each dataset suggest the classification method is relatively stable, with performance variances bellow 10%. Best results were attained on the ID dataset, with accuracy, average precision and AUC around 82%, f-measure of 77.3% and precision and recall of 78.9% and 75.8%, respectively.

3.2 Features

N-Grams. To improve our feature set we started by extending the basic bag-of-words model with combinations of bigrams, trigrams, and 4-grams.

TF-IDF Vectors of Trigrams. TF-IDF (term frequency - inverse document frequency) is a form of normalization of absolute term counts such as the ones used in bag-of-words or n-grams models, which takes into account the relative occurrence of certain terms across all documents of a corpus. If is commonly used in IR and text classification applications to improve both retrieval and classification results. We tested normalizing our trigrams classification model using TF-IDF values to try to improve prediction results.

TF-IDF Vectors of Trigrams, POS and Chunk Tags Combinations. In addition to TF-IDF vectors of word trigrams, we tested the value of also using TF-IDF scaled vectors of trigrams of POS and chunk tags. These types of linguistic features model sentence structure at the grammatical level without tying the model to specific lexicons and should improve generalization to different kinds of sentences.

Verbs, Concepts and Syntactic Arcs. We extended the TF-IDF vectors model with semantic features related to the domain, such as the presence of biologically relevant verbs in sentences, semantic features related to biological concepts and syntactic features related to the number of arcs of a certain distance (in number of hops) between recognized concepts and biologically relevant verbs.

3.3 Overall Features Comparison

Figure 1 compares the performance of different classifiers built with different combinations of the features described above, using stratified 10-fold cross-validation. As can be seen, four of the proposed models achieved higher performance results than the baseline model (BoW model, Table 2).

When using n-grams, evaluation results displayed improvements across all measures, with the exception of recall reductions between 1 and 3% for the EPI and ID datasets. Trigrams (plus words and bigrams) proved to be the best n-gram combination, showing results consistently better than the baseline and other tested n-gram combinations. The use of trigrams provided a 6% precision boost, 2% AUC increase and improvement of 2.5% on both accuracy and f-measure.

The use of TF-IDF scores, both for the simple BoW and bag-of-trigrams models, significantly improved precision of the estimators, with increases over the BoW baseline of 7% . Cross-validation results showed consistent improvements in all evaluation metrics, with increases over the previous trigram model in term of precision and AUC (+1.2%), recall (+1.8%) and f-score (1.5%).

Models trained with POS and with POS-Chunk pairs performed considerably worst than the TF-IDF trigrams model, suggesting that when training and testing on datasets dominated by the same lexicon, a trigram model of words can better classify text. The addition of these grammatical features to the TF-IDF trigram model also caused a slight decrease in performance, since these features augment dimensionality considerably and can add redundancy and noise to the lexical feature set. We can also conclude that pairing POS tags with chunk tags yields a better classification model than using only POS tags. Combining trigrams of POS tags with trigrams of POS-Chunk tags pairs seems to generate less precise classification models but with higher recall.

In summary, the model trained with TF-IDF normalized trigrams achieved the best performance across all metrics. Classification results obtained with this model (Table 3) represent improvements over the baseline of 7% in precision and 2% in recall, reflected in the f-measure by an increase of 4.4%. AUC was improved

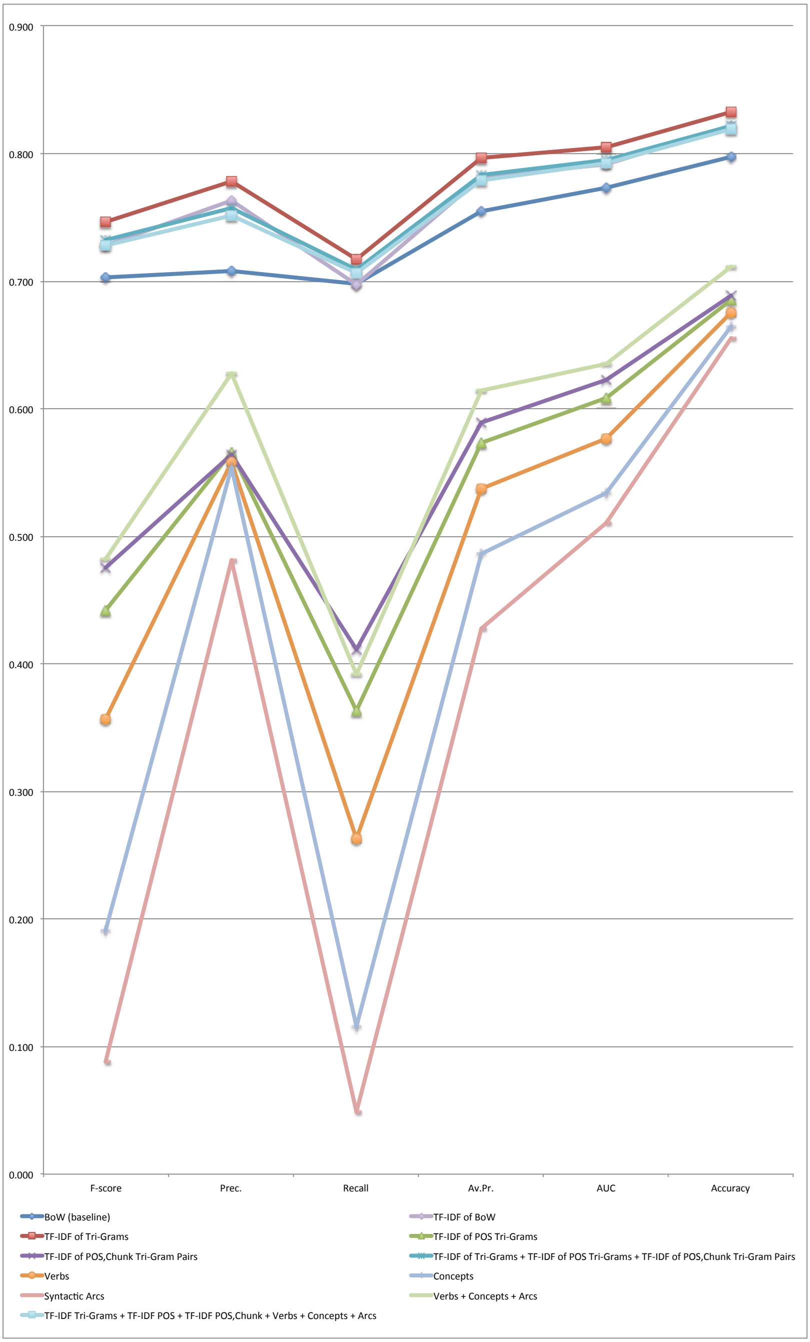


Fig. 1. Classification performance results for different types of features using stratified 10-fold cross-validation on the aggregated dataset

Table 3. Classification performance results for the TF-IDF normalized trigrams model

Dataset	F-score	Prec.	Recall	Av.Pr.	AUC	Accuracy
All	0.746	0.779	0.717	0.796	0.805	0.833
GE	0.787	0.783	0.792	0.831	0.819	0.823
EPI	0.728	0.782	0.681	0.761	0.819	0.907
ID	0.819	0.845	0.796	0.860	0.852	0.865
Angio	0.768	0.781	0.757	0.824	0.790	0.793

by 3.2%, average precision by 4.2% and overall accuracy by 3.5%. The next best performing models were trained with TF-IDF bag-of-words features, and with a combination of TF-IDF trigrams, TF-IDF vectors of POS tags trigrams and TF-IDF vectors of POS-Chunk tag pairs trigrams. Performance differences between these top performing models were marginal, with variance under 3%.

4 Conclusions

The objective of this work was to evaluate the use of classification methods for identifying and ranking sentences describing biomolecular events.

We conducted exploratory work for the identification and evaluation of various lexical, syntactic, grammatical and semantic features of text that could potentially be used for discrimination of these sentences. We concluded that one of the simplest models for text classification – the bag-of-words – can be used with satisfactory results. Since this model is based on lexical features of sentences (words), it works best when training data has a vocabulary aligned with the target data for classification. We also found that trigrams produced the best results for our data and that the use of TF-IDF weighting techniques on top of bag-of-words and n-grams improved classification precision, with a negative effect on recall. Other types of features considered, such as part-of-speech (POS) and chunk tags, the presence of biologically relevant verbs, presence of biological concepts and concept diversity, and syntactic features related to existence of dependencies between verbs and concepts, failed to produce satisfactory results when used in isolation. However we expect that these features could allow creating models with better generalization than using bags-of-words alone. This is a research direction that we are currently exploring, together with the application of feature selection methods.

Acknowledgements. This work was funded by FEDER through the Operational Program Competitiveness Factors - COMPETE and from National Funds through FCT - Foundation for Science and Technology, in the context of the projects FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011) and Incentivo/EEI/UI0127/2013. S. Matos is supported by FCT, under the FCT Investigator Programme.

References

1. Jensen, L.J., Saric, J., Bork, P.: Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics 7(2), 119–129 (2006)
2. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., et al.: Database resources of the national center for biotechnology information. Nucleic Acids Research 35(suppl. 1), D5–D12 (2007)
3. Lu, Z.: PubMed and beyond: a survey of web tools for searching biomedical literature. Database: the Journal of Biological Databases and Curation 2011 (2011)
4. Cafarella, M.J., Downey, D., Soderland, S., Etzioni, O.: Knowitnow: fast, scalable information extraction from the web. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 563–570. Association for Computational Linguistics (2005)
5. Uzuner, Ö., Zhang, X., Sibanda, T.: Machine learning and rule-based approaches to assertion classification. Journal of the American Medical Informatics Association 16(1), 109–115 (2009)
6. Hripcsak, G., Friedman, C., Alderson, P.O., DuMouchel, W., Johnson, S.B., Clayton, P.D.: Unlocking clinical data from narrative reports: a study of natural language processing. Annals of Internal Medicine 122(9), 681–688 (1995)
7. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., Johnson, S.B.: A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association 1(2), 161–174 (1994)
8. Kim, J.-D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T., Yonezawa, A.: The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. BMC Bioinformatics 13(suppl. 11), S1 (2012)
9. Pyysalo, S., Ohta, T., Rak, R., Sullivan, D., Mao, C., Wang, C., Sobral, B., Tsujii, J., Ananiadou, S.: Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. BMC Bioinformatics 13(suppl. 11), S2 (2012)
10. Wang, X., McKendrick, I., Barrett, I., Dix, I., French, T., Tsujii, J., Ananiadou, S.: Automatic extraction of angiogenesis bioprocess from text. Bioinformatics 27(19), 2730–2737 (2011)
11. Campos, D., Matos, S., Oliveira, J.L.: A modular framework for biomedical concept recognition. BMC Bioinformatics 14(1), 281 (2013)

TKG: A Graph-Based Approach to Extract Keywords from Tweets

Willyan Daniel Abilhoa and Leandro Nunes de Castro

Natural Computing Laboratory, Mackenzie Presbyterian University, São Paulo, Brazil
abilhoa.willyan@gmail.com, lnunes@mackenzie.br

Abstract. Twitter is a microblog service that generates a huge amount of textual content daily. All this content needs to be explored by means of text mining, natural language processing, information retrieval, and other techniques. In this context, automatic keyword extraction is a task of great usefulness. A fundamental step in text mining techniques consists of building a model for text representation. This paper proposes a keyword extraction method for tweet collections that represents texts as graphs and applies centrality measures for finding the relevant vertices (keywords). The proposal is applied to two tweet collections of Brazilian TV shows and its results are compared to those of TFIDF and KEA.

Keywords: Knowledge Discovery, Text Mining, Keyword Extraction, Graph Theory, Centrality Measures.

1 Introduction

The application of data analysis techniques in social media data [1] can be used to extract patterns, helping individuals and organizations to discover the acceptance level of products [2], to detect disasters and anomalies [3], to monitor events, and to forecast the performance of politicians in election campaigns [4], among many other potential applications. Text data requires, among others, the use of text mining [5], natural language processing [6], and information retrieval [7] for analysis. In the specific case of text mining methods, documents are represented using the well-known vector space model [8], which results in sparse matrices to be dealt with computationally. Besides, when the target application involves Twitter messages, this problem becomes worse due to the short texts, informality, grammatical errors, buzzwords, slangs, and the speed with which real-time content is generated.

Keyword extraction is the task of finding the words that best describe the subject of a text. Its applications include indexing, summarization, topic detection and tracking, among others [9]. This paper proposes a technique to extract keywords from collections of Twitter messages based on the representation of texts by means of a graph structure, from which it is assigned relevance values to the vertices based on graph centrality measures. The proposed method, named TKG, is applied to one real-world datasets related to Brazilian TV shows and compared with the TFIDF method and the

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_49, © Springer International Publishing Switzerland 2014

KEA algorithm, having human-based keyword extraction as benchmark. Several variations of TKG are proposed, including different forms of building the graph connections and weighting them, plus different centrality measures to extract keywords from the graphs. The experiments performed showed that some variations of TKG are invariably superior to others and the other algorithms for the problem tested.

The paper is organized as follows. Section 2 gives a brief explanation about text representation and keyword extraction. Section 3 introduces the proposed technique, and Section 4 covers the experiments performed. Section 5 brings a general discussion about the work and perspectives for future research.

2 Graph-Based Text Representation

2.1 General Idea

In graph-based text representation, a text can be seen as a collection of interconnected words [10]. Thus, a *textual graph* is a graph in which the set of vertices corresponds to the collection of terms, and the set of edges provides the relationships among terms obeying some criteria, such as the proximity between the terms or the existence of a given relationship of these terms in the same sentence, such as a co-occurrence relationship [11, 12, 13]. Some of main graph-based models to represent texts include the works of [11], [13], [14], and [15]. The textual graph proposed by [11] captures information about the order, frequency and context of terms. To capture the notion of order, lists of pre-processed terms in the same order with which they appear in the original document are obtained. In the textual graph formed, each vertex corresponds to a term or concept, and is represented in a unique way. Each edge is constructed based on the proximity and co-occurrence relationships between concepts within a given window, which is moved along the list.

2.2 Centrality Measure

Centrality measures are discriminative properties of the importance of a vertex in a graph, and are directly related to the structure of the graph. The identification of central vertices is a major task in graph analysis [5]. In this context, the notion of importance or centrality of a vertex can be related to the *degree*, *closeness*, and many other features of a given vertex [5]. The *degree centrality* (C_i^D) defines the importance of a vertex i according to the number of edges incident upon it [16]. This number quantifies the interaction level of an element with others in the network. The *closeness centrality* (C_i^C) of a vertex i is defined as the inverse of its *farness*, where the fairness of a vertex i is defined as the sum of its distance to all other vertices [17]. Therefore, the more central a vertex, the lower is its total distance to all other vertices [18].

Since a text is represented by a graph and centrality measures are a possibility for highlighting important nodes (terms) according to the structure of the graph, these are used here as weighting criteria for words in a document.

2.3 Keyword Extraction

A *keyword* can be understood as the smallest unit of one or more terms which summarizes and identifies the subject of a text [19]. One of the main tasks in the analysis of large textual databases is the discovery of keywords within a document collection, which is a process known as *keyword extraction*. This task is a core technology for many text mining applications, such as indexing, summarization, classification, clustering, topic detection and tracking, among others [19]. The simplest, and probably the most effective, solution for this task is the assignment of keywords by a person that reads the full document. However, when a massive number of documents is continuously received, for example in Twitter tracking applications, humanly reading tweets in search for keywords becomes an unfeasible and error prone task [19].

In this context, it becomes necessary to design and implement techniques capable of automatically extracting keywords. Such techniques might be divided into [20]: 1) *corpus-oriented*, requiring a collection of documents; or 2) *single document-oriented*, performed on individual documents. Despite their differences, these techniques follow two basic steps [21]: 1) define heuristics, such as a measure of similarity, term frequency or predetermined relationships between words; and 2) locate and define a set of one or more words that describe the topic of a text with precision.

The method to be proposed in this paper is a graph-based and corpus-oriented approach, because the keywords will be extracted from a set of tweets represented as a graph of messages posted in Twitter. This text representation model allows the capture of statistical information, for example, term co-occurrence frequency, centrality measures, and aggregation of linguistic features.

3 TKG: A Graph-Based Approach to Extract Keywords from Tweets

The technique proposed in this paper, named TKG (standing for *Twitter Keyword Graph*), consists of three sequential steps: 1) document pre-processing; 2) textual graph building from preprocessed tweets; and 3) keyword extraction, which involves the calculation of the centrality measures for each vertex (token), the ranking of these vertices, and the selection of keywords based on their rank.

In the document pre-processing step, each tweet collected is seen as a document $\mathbf{d} \in D$, where $D = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N\}$ is the set (corpora) of N documents collected. In the *preprocessing* step, a document $\mathbf{d}_i \in D$ is tokenized; each *token* or *term* is a basic unit of the document. Then, the stopwords are removed and the document $\mathbf{d}_i$ is represented by its attribute vector $\mathbf{d}_i = \mathbf{t}^i = \{t^i{}_1, t^i{}_2, \ldots, t^i{}_K\}$, of K terms, $t_j \in T$, $j = 1, \ldots, K$. Note that K is the number of tokens extracted from the whole collection of documents. In the *textual graph building* step a token corresponds to a vertex and an edge defines the co-occurrence relationship between tokens. Given the token set T, it is possible to define the vertices and edges of the graph $G = (V, E)$, where $V = \{t_1, t_2, \ldots, t_K\}$ is the set of vertices and $E = \{e_{1,1}, e_{1,2}, \ldots, e_{I,J}\}$ the set of edges. In the *vertex assignment* step, a complete reading of the attribute vector is performed and one vertex is created for each token. In the *edging* step, the definition of the edges can be performed according to two heuristics, inspired by the model proposed in [11], based on co-occurrence: 1) *nearest neighbor edging* (TKG_1); and 2) *all neighbors edging* (TKG_2). The two heuristics for edge creation include the *weighting* information of an edge given by the frequency, $f_{i,j}$, with which tokens co-occur. The weight w_{ij} of an edge is assigned according to the three following ways: 1) *same weight edges* (W^1); 2) *weight as co-occurrence frequency* (W^F); and 3) *weight as inverse co-occurrence frequency* ($W^{1/F}$). After the graph is built, the centrality measures described previously (degree centrality C_i^D, closeness centrality C_i^C, and eccentricity C_i^E) are calculated for each vertex $v_i \in V$ in G, and a rank for each measure is generated. With this, the n best ranked vertices are selected as keywords.

4 Performance Assessment

The proposed technique, TKG, is evaluated according to its properties: edging heuristics (*nearest-neighbor edging*, named TKG_1; *all neighbors edging*, named TKG_2); edge-weighting possibilities (*same weight assignment*, W^1; *weight as co-occurrence frequency*, W^F; *weight as inverse co-occurrence frequency*, $W^{1/F}$); and centrality measures to be assigned to vertices (C_i^D, C_i^E and C_i^C).

Given the TKG properties, the goal of the experiment is to preliminary validate the TKG performance and compare it with other approaches, KEA [22] and TFIDF [8], from the literature, and human keyword extraction. In this analysis, all algorithms performed the task of extracting keywords from two Brazilian TV shows, called "Trofeu Imprensa" and "*A Fazenda*", with 100 tweets each. The Portuguese stopword list was inserted in the extracting candidates step, and no thesaurus was used.

In principle, there is no "correct" set of keywords for a given document, not even humans may agree on the keywords they extract from documents. Therefore, to assess the performance of the proposed algorithm and the others from the literature, the following methodology was adopted:

1. Three persons (humans) were invited to suggest an unspecified number of keywords from the documents. After that, the intersection of these sets for each show is determined: these new sets contain the relevant documents, {*Relevant*}, for each show.
2. The TKG variations are compared with the TFIDF and KEA methods, having the human-defined sets as references. Here, the TKG configurations were denoted as follows:

$$TKG_{\text{“EDGING”}}|W^{\text{“WEIGHTING”}}|C^{\text{“CENTRALITY”}} \tag{1}$$

   where "EDGING" may assume the TKG_1 or TKG_2 heuristics, "WEIGHTING" may assume the W^1, W^F or $W^{1/F}$ heuristics, and "CENTRALITY" may assume the C^D, C^E or C^C centrality measures. The comparisons are performed according to the Top-10 rankings from the TKG configurations, TFIDF and KEA. The centrality measures adopted in this experiment for the TKG rank creation were C^C and C^E.

3. The well-known *Precision*, *Recall*, and *F-measure* from information retrieval were used as metrics [23]:

$$\mathrm{Pr} = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \quad (2)$$

$$\mathrm{Re} = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|} \quad (3)$$

$$F\text{-}measure = F = 2 * \frac{\mathrm{Pr} * \mathrm{Re}}{(\mathrm{Pr} + \mathrm{Re})} \quad (4)$$

As the number of keywords retrieved by each method was defined as 10, to compute Pr the number of relevant documents retrieved was equal to the number of keywords proposed by the method that appear in at least one of the human lists. By contrast, to compute Re the number of relevant documents was considered to be the number of keywords in the intersection set of the three human lists.

Table 1 presents the sets of keywords suggested by each reader for the tweets. The common keywords (intersection) among these sets are highlighted in bold.

Table 1. Keyword sets suggested by each human reader for each TV show. In bold are those keywords proposed by all readers.

Reader 1	Show 1	***mel***, ***fronckowiak***, ***micael***, ***borges***, ***trofeu***, ***imprensa***, ***sbt***, ***emissoras***, *trofeuimprensa, paula, fernandes,* ***premio***, *restarnotrofeuimprensa, lombardi,* ***premiacao***, *rebeldes, silvio, santos*
	Show 2	***fazenda***, ***record***, ***xepa***, *andressa, reality, show, afazenda,* ***barbara***, *evans, rede, foraxepa, denise, rocha, novela, juliana, silveira, monique, trevisol*
Reader 2	Show 1	***mel***, ***fronckowiak***, *chay, suede,* ***micael***, ***borges***, ***sbt***, ***emissoras***, ***premiacao***, ***trofeu***, ***imprensa***, *edicao, rebelde, juntas,* ***premio***, *compartilhe*
	Show 2	***fazenda***, *afazenda,* ***record***, ***barbara***, *assistindo, assistir, vendo, roca, banho, mateus,* ***xepa***
Reader 3	Show 1	***trofeu***, ***imprensa***, ***sbt***, ***mel***, ***fronckowiak***, *chay, suede,* ***micael***, ***borges***, *hoje,* ***emissoras***, *trofeuimprensa,* ***premio***, ***premiacao***, *restarnotrofeuimprensa*
	Show 2	***fazenda***, ***record***, ***barbara***, *assistir, evans, dona,* ***xepa***, *reality, show, votacao, rede, foraxepa, factor, Denise*

Table 2 shows the results obtained by the TKG variations, TFIDF, and KEA. Due to space constraints, only some of the variations are shown in this table. Keywords over a gray background are those that match the set of relevant keywords, and those highlighted in bold type appear in at least one of the human sets (in addition to the other detached ones). The former are used to compute recall (Re), and the latter to compute precision (Pr). The results presented in this table show that the all neighbors edging scheme (TKG_2) when combined with the inverse co-occurrence frequency was superior to the nearest neighbor edging (TKG_1), giving the best results when using the C^C centrality measure. It is also observed that the TFIDF method found slightly better results than KEA. At the lines regarding to the performance measures, the superior results, from TKG, were highlighted in bold.

Table 2. Summary of the results for all methods applied to each set of tweets. Keywords over a gray background match the set of relevant keywords, and those printed in gray appear in at least of the sets proposed by humans. Precision, Recall and F-measure are calculated as in Eqs. (2), (3) and (4), respectively. In this case, the TKG configurations have used closeness (C^C) as centrality measure.

Show 1: Trofeu Imprensa					
	$TKG_1 \mid W^{1/F} \mid C^C$	$TKG_2 \mid W^1 \mid C^C$	$TKG_2 \mid W^{1/F} \mid C^C$	*TFIDF*	*KEA*
1	imprensa	trofeu	Imprensa	**hoje**	sbt
2	trofeu	imprensa	Trofeu	**trofeuimprensa**	trofeu
3	sbt	sbt	Sbt	daqui	imprensa
4	**trofeuimprensa**	**hoje**	**hoje**	**juntas**	**trofeuimprensa**
5	**juntas**	**trofeuimprensa**	**trofeuimprensa**	emissoras	**hoje**
6	**hoje**	percam	**juntas**	percam	daqui
7	emissoras	mel	Emissoras	mel	emissoras
8	**silvio**	fronckowiak	Daqui	micael	**juntas**
9	participacao	micael	Percam	**chay**	participacao
10	daqui	**chay**	Mel	imprensa	percam
Pr	80%	**90%**	**80%**	80%	70%
Re	40%	**60%**	**50%**	40%	40%
F	53.33%	**72.00%**	**61.54%**	53.33%	50.91%
Show 2: A Fazenda					
1	fazenda	Record	Record	record	fazenda
2	rede	fazenda	Fazenda	fazenda	record
3	record	afazenda	**Rede**	**afazenda**	**rede**
4	dormir	barbara	**Afazenda**	**rede**	comecar
5	ofuro	hoje	**denise**	**denise**	**afazenda**
6	**mateus**	mulher	barbara	barbara	comeca
7	explica	gente	dormir	**dormir**	barbara
8	motivos	falando	**rocha**	hoje	foto
9	**factor**	**denise**	peoa	mulher	**assistir**
10	papo	voc	ofuro	falando	hoje
Pr	40%	40%	**70%**	70%	60%
Re	50%	50%	**75%**	75%	75%
F	44.44%	44.44%	**72.41%**	72.41%	66.67%

5 Discussions and Future Trends

This paper proposed a keyword extraction technique of Twitter messages in which tweets are represented as graphs. The proposed approach, called TKG, relies on three phases: text pre-processing; graph building; and keyword extraction. The text pre-processing methods used are standard ones from the text-mining literature, such as tokenization, stemming, and stopwords removal. Graph building takes each token as a vertex and edging can be performed in one of two ways: using a nearest-neighbor approach, or using an all neighbors approach. Finally, keywords are extracted from the graph by applying, in cascade, some graph centrality measures.

The results obtained in the experiments showed that building the graph using an all neighbors edging (TKG_2) scheme provided superior performance, and assigning weights to the edges valued by 1 (W^1) or based on the weight as the inverse co-occurrence frequency ($W^{1/f}$) were superior for both cases. Closeness centrality was the keyword selection method chosen.

Some advantages of TKG when compared with the other approaches might be highlighted in terms of its simplicity and scalability. The simplicity of TKG is given because it does not require any external knowledge, e.g. a model provided by training. The scalability is achieved due to the way in which the edging processes receive a vector of tokens and assigns vertices and edges in order to build the graph, which requires basically local updates in the structure.

Possible future works include performing extensive sets of experiments to validate TKG and its variations, and using other centrality measures not discussed here, individually or in combination. Further studies might also be directed to refining the structure of the graph using some heuristics, such as the elimination of noisy edges or vertices. A clear difficulty of the proposed technique is to find the best configuration of TKG properties. Studies in this direction might be also explored in the future.

Acknowledgments. The authors thank Mackenzie University, Mackpesquisa, CNPq, Capes (Proc. n. 9315/13-6) and FAPESP for the financial support.

References

1. Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S.: Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons 54, 241–251 (2011)
2. Yoshida, M., Matsushima, S., Ono, S., Sato, I., Nakagawa, H.: ITC-UT: Tweet Categorization by Query Categorization of On-line Reputation Management. In: Conference on Multilingual and Multimodal Information Access Evaluation (2010)
3. Prabowo, R., Thelwall, M.: Sentiment analysis: A combined approach. Journal of Informetrics 3, 143–157 (2009)
4. Bermingham, A., Smeaton, A.: On Using Twitter to Monitor Political Sentiment and Predict Election Results. Sentiment Analysis Where AI Meets Psychology, 2–10 (2011)
5. Feldman, R., Sanger, J.: The Text Mining Handbook Advanced Approaches in Analysing Unstructured Data, Cambridge (2007)
6. Hirschman, L., Thompson, H.S.: Overview of evaluation in speech and natural language processing. In: Survey of the State of the Art in Human Language Technology, pp. 409–414. Cambridge University Press and Giardini Editori, Pisa (1997)
7. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press (1999)
8. Salton, G., Yang, C.S., Yu, C.T.: A Theory of Term Importance in Automatic Text Analysis. Journal of the American society for Information Science 26, 33–44 (1975)
9. Zhang, C., Wang, H., Liu, Y., Wu, Y., Liao, Y., Wang, B.: Automatic Keyword Extraction from Documents Using Conditional Random Fields. Journal of Computational Information Systems, 1169–1180 (2008)
10. Gross, J.L., Yellen, J.: Graph Theory and Its Applications, 2nd edn. Chapman & Hall/CRC (2006)
11. Jin, W., Srihari, R.K.: Graph-based text representation and knowledge discovery. In: Proceedings of the 2007 ACM Symposium on Applied Computing, vol. 7, pp. 807–811 (2007)
12. Palshikar, G.K.: Keyword Extraction from a Single Document Using Centrality Measures. In: Ghosh, A., De, R.K., Pal, S.K. (eds.) PReMI 2007. LNCS, vol. 4815, pp. 503–510. Springer, Heidelberg (2007)

13. Zhou, F., Zhang, F., Yang, B.: Graph-based text representation model and its realization. Natural Language Processing and Knowledge Engineering (NLP-KE) 8(1), 21–23 (2010)
14. Schenker, A., Last, M., Bunke, H.: Classification of Web documents using a graph model. Document Analysis and Recognition 1, 240–244 (2003)
15. Hensman, S.: Construction of conceptual graph representation of texts. In: Proceedings of Student Research Workshop at HLT-NAACL, Boston, pp. 49–54 (2004)
16. Nieminen, J.: On the centrality in a graph. Scand. J. Psychol. 15, 332–336 (1974)
17. Wasserman, S., Faust, K., Iacobucci, D.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1995)
18. Hage, P., Harary, F.: Eccentricity and centrality in networks. Social Networks 17, 57–63 (1995)
19. Zhang, K., Xu, H., Tang, J., Li, J.: Keyword Extraction Using Support Vector Machine. In: Yu, J.X., Kitsuregawa, M., Leong, H.-V. (eds.) WAIM 2006. LNCS, vol. 4016, pp. 85–96. Springer, Heidelberg (2006)
20. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic Keyword Extraction from Individual Documents. Text Mining: Applications and Theory, 1–20 (2010)
21. Lott, B.: Survey of Keyword Extraction Techniques. UNM Education (2012)
22. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA Practical Automatic Keyphrase Action. In: Proceedings of the 4th ACM Conference on Digital Library (DL 1999), Berkeley, CA, USA, pp. 254–226 (1999)
23. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Academic Press (2001)

Outdoor Robotic Companion Based on a Google Android™ Smartphone and GPS Guidance

Eduard Clotet, Dani Martínez, Javier Moreno, Marcel Tresanchez, Tomàs Pallejà, Davinia Font, Mercè Teixidó, and Jordi Palacín

Department of Computer Science and Industrial Engineering,
University of Lleida, 25001 Lleida, Spain
{eclotet,dmartinez,jmoreno,mtresanchez,tpalleja,dfont,
mteixido,palacin}@diei.udl.cat

Abstract. This paper presents an outdoor mobile robotic companion that is guided by GPS. The robotic companion is a wheeled mobile robot that uses a Google Android smartphone as a central processing unit. The mobile robot takes advantage of the integrated smartphone features such as wireless connectivity, USB interface for external electronic boards, and the GPS sensor. Moreover, the mobile robot companion can be controlled by another smartphone in order to define companion behaviors such as "go to location", "follow me", and "keep close" all of them based on the use of GPS information. The mobile robot has been tested as in long walks in order to evaluate the effectiveness of the artificial companion effect.

Keywords: Companion, mobile robot, smartphone based robot, GPS.

1 Introduction

The smartphone application market is growing and offering an increasing number of new functionalities to the users. The increasing capabilities of such mobile devices allow the creation of applications with huge computational and peripherals requirements as well as the direct control of the sensors and actuators of a mobile robot. For example, Google Android smartphones have several embedded sensors, actuators, connectivity modules, and multi-core processors which can be freely used to develop new applications. This paper proposes a new robotic application in which a first smartphone is used as the main control unit of a mobile robot and a second smartphone (the user personal phone) is used to establish a wireless link with the mobile robot and control its behavior.

This paper proposes the development of an outdoor robotic companion in order to create an emotional link with a person. The mobile robot will use the GPS from its smartphone as the main external localization sensor and will follow the other GPS information provided by the user smartphone. The mobile robot will be able to use artificial vision or other sensor information for obstacle avoidance; however, the companion effect will be only based on the GPS information provided by a personal smartphone.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_50, © Springer International Publishing Switzerland 2014

The objective of this paper is to assess the positioning error and the companion effect through live tests in outdoor conditions. Such tests will validate the accuracy and performances of the low cost GPS sensors integrated with the smartphones. Finally, the viability of using these devices for positioning applications will be discussed.

2 Related Work

There are many works that proposes the use of GPS sensors in robotics for localization and guidance [1, 3]. For example, in [2] a mobile robot which uses a notebook with a GPS/GPRS module was used to receive displacement orders anywhere from an enabled server. In [4], the computation of the path accuracy was improved by applying a Kalman filter to the GPS location information. Moreover, there are few research works about robotics which are based on smartphone devices. For example, in [5] a practical programming methodology was proposed in order to develop an autonomous soccer mobile robot controlled by an Android smartphone. In a similar direction, in [6] was presented a teleoperation implementation between a Google Android smartphone and a mobile robot which is capable of transmitting images and motion orders for a real-time communication.

3 Materials and Methods

3.1 Mobile Robot

The mobile robot (Figure 1) is based on a modified electric remote control crawler. A new case for the crawler that simulates a dog shape was designed and made through a fast prototyping method using a 3-D ABS printer taking into account the new requirements of the mobile robot. The control circuits were removed and replaced by a Sony Ericsson Xperia U smartphone connected to a small USB interface board. The smartphone uses the USB connection to send motion orders to an electronic board which controls directly all the motion system of the robot: two DC motors for forward motion and a servo motor for direction control. The board is based on a PIC24F MCU by Microchip. In addition, the crawler is powered by a 4600mAh Ni-MH battery that operates at 7.2V.

Fig. 1. Mobile robot companion

The Xperia U smartphone integrates a dual-core 1GHz processor with 512MB of RAM. There are several embedded sensors in the smartphone such as an accelerometer, a gyroscope, a GPS, two cameras, and a digital compass. As communication methods, it has a Wi-Fi 802.11 b/g/n module, Bluetooth 2.1 connectivity, and a USB 2.0 port. The installed operating system in the mobile phone is the Android version 4.0.4. Finally, the mobile robot smartphone is configured to work as a Wi-Fi access point to establish point to point connections with other personal smartphone devices.

3.2 User Smartphone

The smartphone used for this paper as the user personal phone was a Sony Ericsson Xperia P. This device has a dual-core 1GHz processor with 1GB of RAM. It has a 4” touch screen for user interaction and also integrates some embedded sensors such as gyroscope, accelerometer, digital compass, cameras, and GPS. The smartphone includes the following modules wireless communications modules: Wi-Fi 802.11 b/g/n, Bluetooth 2.1, Near Field Communication (NFC), and USB 2.0.

3.3 Android SDK

Google provides a complete set of tools and methods in Java language for the implementation of Android applications. The Android Software Development Kit (SDK) [7] is composed mainly by high level methods and other specialized tools to simplify the functionality and the use of many features from Android devices such as sensors, wireless connectivity, I/O, and native computation. The SDK is freely available together with an extensive repository of tutorials and developer documentation. Moreover, the Integrated Development Environment (IDE) used in for programming, compiling and debugging the main smartphone applications was Eclipse. In this work, the motor control routines and the robot behavior process are based on the methodology described in [5].

4 Methodology

The application proposed in this paper requires a connectivity protocol to engage the communication between the mobile robot and the user in order to develop the companion effect. On the one hand, the guidance of the mobile robot uses a route estimation procedure derived directly from the onboard GPS and the GPS positioning information provided by the user smartphone. On the other hand, it has been developed several services and methods in both smartphones in order to perform different mobile robot actions such as “go to location”, “follow me”, and “keep close”.

4.1 Connectivity

The mobile robot establishes a wireless local communication with the user personal phone. The communication is performed over Wi-Fi protocol using the mobile robot smartphone as an access point. Additionally, a wireless local area network can be used in order to an extended the operation range though access point roaming. The data streams implemented between both applications use TCP sockets with a custom application-level communication protocol. The implementation of the communication methodologies in both Android applications is performed in separate execution threads in order to avoid the blocking of the main applications. On the other hand, the applications have direct access to wireless connection parameters using Android SDK routines in order to handle and automatize all the connection procedures.

4.2 GPS Sensor

Most Google Android devices integrate a GPS sensor which is accessible by the applications. So, the application is able to get location data from the smartphone GPS using a sensor listener implementation from Android SDK. Regarding the application framework, the developer is able to choice among 5 abstract types of sensor delay or the exact rate in microseconds for newer Android versions. However, Android is not a real-time operating system and in practice, the sensor listener does not provide new data in an established constant frequency. In this case, the fastest option, *SENSOR_DELAY_FASTEST*, has been selected in order to obtain GPS positioning at the maximum speed. Regarding the Android implementation, there is a method from the listener implementation which is called every time a new location point is generated in order to store it or use it as the new provided positioning information.

4.3 Path Planning

This methodology is currently applied for limited network infrastructures or limited areas within the Wi-Fi range of the mobile robot access point. All the services offered by the mobile robot are natively based on position routing between two two-dimensional geo points which have the terrestrial latitude and longitude as spatial components. The basic behavior of the mobile robot is to drive from the origin geo point (provided by the GPS of the mobile robot) to the destination geo point (provided by the GPS of the user smartphone). Thus, the companion mobile robot updates a trajectory path using trigonometry and vector operations by estimating the minimal rotation angle and the distance to the next geo point.

The mobile robot first calibrates its orientation by estimating its direction vector from few geo points captured at its first displacements. So, the direction vector is obtained by simply calculating the vector between two geo points: the previous position and the current position. Once an approximation of the direction vector is given, the robot is able to calculate the optimal route to the next target point. This result will be the minimal rotating angle that the mobile robot will have to perform in order to face the target point. On the other hand, the estimated linear distance between the two geo points is used as a relative reference for the motor powering. In order to increase the effectiveness of the companion effect the path planned by the mobile robot is updated when the current GPS location point is updated.

4.4 Companion Application

The Android application running on the mobile robot defines the companion effect. The companion application is currently designed to interact only with a single user client but in the future this feature can be increased to manage a swarm of companions. The implementation of this application requires several services such as the USB manager, the TCP communications, and the GPS sensor listener. The USB interface is prepared for sending motion commands from the smartphone to the motor control board which processes the commands and powers the motors. For the communications and data exchanging, the companion application is listening for incoming client TCP connections. When a client is connected to the mobile robot, the data exchanging begins and, once the service is finished, it is disconnected and returns to listen for a new client connection state. The GPS service is capturing location values continuously and performing the path planning procedure with the current target location. The target locations are managed and updated depending on the service requested by the client.

4.5 User Application

Personal smartphones are commonly used as a location tool. The user application developed in this work takes advantage of this feature and exchanges its location with the robot companion. Figure 2 shows an outdoor scenario where the robot companion follows the user in real-time. The user is then able to choice among different companion modes implemented in the mobile robot. Currently the behaviors implemented are:

- "go to location". In this mode the user can send a current or stored GPS location to the mobile robot and the mobile robot goes to this location. The current implementation computes the shorted path and cannot handle obstacles in this path; the robot stops in case of detecting obstacles.
- "go to location and return". This mode is similar to playing with a dog with a ball. The user can select a relative orientation and radius and then the mobile robot goes to this location and returns to the current user location.
- "keep close". This mode is a random implementation of the "go to location and return" in order to simulate complex exploratory behaviors in the companion mobile robot.
- "follow a trajectory". In this mode the user can select a previously stored trajectory and then the mobile robot follows this trajectory.
- "follow me". In this mode the GPS location of the user is continuously submitted to the mobile robot and the mobile robot updates continuously a trajectory path in order to stay close to the user (close to the smartphone of the user). Please note that the GPS location provided by a static smartphone changes almost randomly and that two GPS at the same position generates different positioning information so, as a consequence of this mode, the mobile robot can be continuously performing small erratic displacements.

5 Results

This section assesses the companion mobile robot behavior and accuracy. The experiments performed are focused on evaluating the accuracy of the positioning information provided by the two smartphones used in this paper. The mode evaluated is “follow me”; in this mode the user continuously sends positioning information to the mobile robot and thus, the companion is continuously computing a path in order to keep itself close to the user. In these experiments the minimum distance between the mobile robot and the user was defined as 3 meters in order to increase the battery operation; that is, the mobile robot will stop when its position is in a radius of 3 meters around the position of the user (obtained with the user smartphone).

In this experiment, one user starts walking a circular trajectory in a large square and his personal smartphone starts to submit its GPS positions to the mobile robot. The mobile robot starts to move when it detects that the onboard GPS position is at a distance higher than the GPS position of the smartphone of the user. The maximum turn of the driving wheels is limited in order to reduce the trajectory changes of the mobile robot originated by sudden variances in the GPS positioning information. Additionally, the velocity of the mobile robot is lower than the user normal velocity and this effect originates a cumulative delay in the path followed by the mobile robot.

Figure 3 shows the satellite map of the experimentation place in which is represented the GPS lectures from both smartphones. The red line depicts the path followed by the user and the blue line the path followed by the mobile robot. The information shown by both figures are only informative as the information provided by the GPS which can have large absolute positioning errors. According to this positioning information the largest distance between the mobile robot and the user was 7.5 m and the average distance was 5.1 m. The main drawback of this companion implementation is the accumulated error generated in both GPS positioning devices that are not correlated (two GPS at the same position does not provide the same location information). Moreover, the visual impression during the experiments was that the mobile robot keeps a constant distance with the user. On the other hand, the emotional impression is that the mobile robot follows the user rather than its trajectory and the positioning errors are interpreted as a kind of natural exploratory curiosity.

Fig. 2. Robot companion following the user

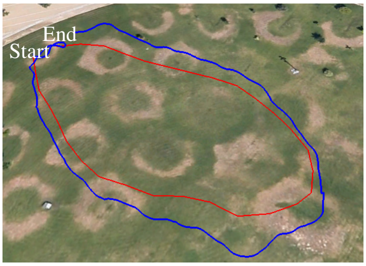


Fig. 3. Trajectories registered in one experiment projected over a satellite map: user (red line) and robot (blue line)

6 Conclusions

This paper presents an outdoor wheeled mobile robotic companion implemented with a Google Android smartphone as the main processing and control unit. The behavior of the mobile robot is controlled remotely by means of another Google Android smartphone. The companion link between the mobile robot and the remote smartphone is accomplished by means of the GPS information provided by these two devices; that is, the remote control sends its GPS position to the mobile robot and the mobile robot compares its own GPS position with the reference received. Nevertheless, one of the main problems of this approach is that two standard GPS receivers placed in the same physical place provide different location information with differences expressed in ranges of meters.

The experiments performed in large squares show that the mobile robot follows the user and that there is a connection between the mobile robot and the user that effectively creates an emotional companion link. So, the developed system is able to be implemented in any final application which requires this kind of dynamics using Android devices. In this direction, future works will be focused on improving the path planning generated from the GPS positioning information.

References

1. Willgoss, R., Rosenfeld, V., Billingsley, J.: High precision GPS guidance of mobile robots. In: Proceedings of the Australasian Conference in Robotics and Automation, Brisbane (2003)
2. Velagic, J., Osmic, N., Hodzic, F., Siljak, H.: Outdoor navigation of a mobile robot using GPS and GPRS communication system. In: Proceedings of ELMAR, Zadar, pp. 173–177 (2011)
3. Choi, H.S., Park, O.D., Kim, H.S.: Autonomous mobile robot using GPS. In: International Conference on Control and Automation, Budapest, vol. 2, pp. 858–862 (2005)
4. Hou, K., Sun, H., Jia, Q., Zhang, Y.: The design of the spherical robot autonomous GPS-based positioning and navigation system. In: International Conference on Intelligent System Design and Engineering Application, Changsha, vol. 2, pp. 626–630 (2010)

5. Martínez, D., Moreno, J., Font, D., Tresanchez, M., Pallejà, T., Teixidó, M., Palacín, J.: A Practical Mobile Robot Agent Implementation Based on a Google Android Smartphone. In: Omatu, S., Neves, J., Rodriguez, J.M.C., Paz Santana, J.F., Gonzalez, S.R. (eds.) Distrib. Computing & Artificial Intelligence. AISC, vol. 217, pp. 625–632. Springer, Heidelberg (2013)
6. Moon, S.W., Kim, Y.J., Myeong, H.J., Kim, C.S., Cha, N.J., Kim, D.H.: Implementation of smartphone environment remote control and monitoring system for Android operating system-based robot platform. In: Proceedings of the International Conference on Ubiquitous Robots and Ambient Intelligence, Daejeon, South Korea, pp. 211–214 (2012)
7. Android Developers website, http://developer.android.com (accessed January 2014)

A Threshold Scheme for 3D Objects Based on Cellular Automata

Angel Martín del Rey

Department of Applied Mathematics
Institute of Fundamental Physics and Mathematics
University of Salamanca, Spain
delrey@usal.es

Abstract. In this work a novel (k, n)-threshold scheme to share a secret 3D object is proposed. It is based on the use of a particular type of finite state machine called cellular automata. More specifically, a reversible 3D cellular automaton with memory is used since the secret to be shared can be interpreted as a configuration of such cellular automaton. This protocol is shown to be perfect and ideal.

Keywords: Threshold Schemes. 3D objects. Cellular Automata. Reversibility.

1 Introduction and Preliminaries

Over the last years there has been an increasing interest in the design and management of 3D objects, which is due, among other things, to the appearance and use of 3D printers. As a consequence, it is reasonable to suppose that in the next years there will be a great proliferation of 3D digital files and 3D multimedia processing tools. This scenario will create an ideal medium for malicious users to compromise the security and privacy of such objects. Then, a major challenge will be the protection of confidentiality, integrity and authenticity of 3D information.

Unfortunately, too few security protocols for 3D objects have been appeared in the scientific literature. The great majority are watermarking protocols ([1,5,6,11,13]), although hash schemes have been also proposed (see [2,10]). Nevertheless, any encryption algorithm or secret sharing scheme has been appeared. Particularly, this paper deals with the design of a type of secret sharing schemes called (k, n)-threshold scheme.

Secret sharing schemes are multi-party cryptographic protocols that share a secret among a set of participants such that only qualified subsets of these participants can recover the original secret, and any information about it can be derived if non-qualified subsets of participants try to recover the secret.

A special and very interesting type of secret sharing schemes is called *threshold scheme*, which can be defined as follows: Let k and n be positive integers such that $k \leq n$; a (k, n)-threshold scheme is a method of sharing a secret K among a set of n participants in such a way that any subset of k participants can compute the value of the secret, but no group of $k - 1$ or fewer can do so.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_51,

Set $\mathcal{P} = (P_1, P_2, \ldots, P_n)$ the set of participants. Let D be the *dealer*: a third trusted party who is not in $\mathcal{P}$. When D wants to share a secret K among the participants belonging to $\mathcal{P}$, D securely distributes some partial information S_i (called share o shadow) to each participant $P_i \in \mathcal{P}$ (so any participant does not know any other participant's share). Some qualified subsets of participants, called *access structure*, $\mathcal{B} \subseteq \mathcal{P}$, can recover the secret K by pooling their shadows together.

A threshold secret sharing scheme is called *perfect* if any non-qualified subset of participants gets no information about the secret, and has no advantage over an outsider in guessing the secret K. An *ideal* threshold secret sharing scheme is a perfect threshold scheme where the sizes of the shadows $S_1, S_2, \ldots, S_n$ are equal or smaller than the size of the secret K.

As is mentioned above, the main goal of this paper is the design of a (k, n)-threshold scheme for 3D objects. Traditional secret sharing schemes are not suitable for 3D objects. This is mainly due to the trade-off between data and slow speed (in 3D objects there is a large amount of data and the speed of traditional protocols is not sufficiently fast) and to the intractable high redundancy (in 3D objects there exists high redundancy which make traditional methods fail to conceal all visible information in some shadows). Consequently, new paradigms are required. The novel method proposed in this work is based on the use of reversible memory 3D cellular automata.

Cellular automata are simple models of computation capable to simulate complex phenomena (see [9,12]). Roughly speaking, a cellular automaton consists of a discrete spatial lattice of sites called cells, each one endowed at each time with a state belonging to a finite state set (usually $\mathbb{F}_2 = 0, 1$). The state of each cell is updated in discrete time steps according to a local transition function which depends on the states of the cells in some neighborhood around it. As the lattice is finite some type of boundary conditions must be imposed: usually null and periodic boundary conditions are considered. 3D cellular automata are characterized by the three-dimensional arrangement of their cells which allows them to simulate different types of phenomena such as spray freeze drying process ([3]), HIV infection ([7]), cracking process of rocks ([8]), dendritic solidification ([14]), or alloy solidification ([15]).

The rest of the paper is organized as follows: In section 2 the basic theory about 3D cellular automata is introduced; the threshold secret sharing scheme for 3D objects is proposed in section 3; in section 4 the security analysis is performed, and the conclusions are presented in section 5.

2 3D Cellular Automata

A 3*D-cellular automaton* (3D-CA for short) is a particular type of discrete dynamical system formed by $p \times q \times r$ memory units called *cells* that are arranged uniformly into a three-dimensional space called its *cellular space* ([4]).

Each cell is endowed with a state from the finite state set $\mathbb{F}_2 = \{0, 1\}$ at every step of time; in this sense $s_{ijk}^t \in \mathbb{F}_2$ stands for the state of the (i, j, k)-th cell at time t. The states of all cells change synchronously in discrete steps of time according to a local transition rule f, where its variables are the states of the m neighbor cells at the previous step of time: $s_{i_1j_1k_1}^{t-1}, s_{i_2j_2k_2}^{t-1}, \ldots, s_{i_mj_mk_m}^{t-1}$. As a consequence:

$$\begin{aligned} f\colon \mathbb{F}_2^m &\to \mathbb{F}_2 \\ \left(s_{i_1j_1k_1}^{t-1}, \ldots, s_{i_mj_mk_m}^{t-1}\right) &\mapsto s_{ijk}^t = f\left(s_{i_1j_1k_1}^{t-1}, \ldots, s_{i_mj_mk_m}^{t-1}\right) \end{aligned} \tag{1}$$

for every $1 \leq i \leq p, 1 \leq j \leq q, 1 \leq k \leq r$.

Usually, Moore neighborhoods are considered (see Figure 1), that is, the neighborhood of the cell (i, j, k) is formed by the 27 nearest cells around it, including itself:

$$\begin{aligned} V = \{&(i-1, j-1, k-1), (i-1, j, k-1), (i-1, j+1, k-1), (i, j-1, k-1), \\ &(i, j, k-1), (i, j+1, k-1), (i+1, j-1, k-1), (i+1, j, k-1), \\ &(i+1, j+1, k-1), (i-1, j-1, k), (i-1, j, k), (i-1, j+1, k), \\ &(i, j-1, k), (i, j, k), (i, j+1, k), (i+1, j-1, k), (i+1, j, k), \\ &(i+1, j+1, k), (i-1, j-1, k+1), (i-1, j, k+1), (i-1, j+1, k+1), \\ &(i, j-1, k+1), (i, j, k+1), (i, j+1, k+1), (i+1, j-1, k+1), \\ &(i+1, j, k+1), (i+1, j+1, k+1)\}. \end{aligned} \tag{2}$$

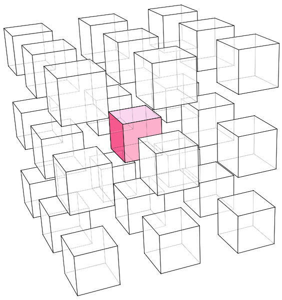


Fig. 1. Moore neighborhood in a three-dimensional cellular space

As the cellular space is finite, null boundary conditions can be established in order to assure a well-defined evolution of the cellular automaton: $s_{ijk}^t = 0$ if the cell (i, j, k) is not in the cellular space. The *configuration at time t* of the 3D-CA is the boolean vector whose coordinates are the states of the cells of the 3D-CA at time t:

$$C^t = \left(s_{111}^t, s_{112}^t, \ldots, s_{pqr}^t\right) \in \mathbb{F}_2^{pqr}. \tag{3}$$

Note that the coefficients are ordered following the lexicographical ordering of their subindices. The *global transition function* of the 3D-CA is a transformation that yields the configuration at the next time step during its evolution, that is:

$$\begin{aligned} \Phi\colon \mathbb{F}_2^{pqr} &\to \mathbb{F}_2^{pqr} \\ C^{t-1} &\mapsto C^t = \Phi\left(C^{t-1}\right) \end{aligned} \tag{4}$$

A 3D-CA is *reversible* if it is possible to compute its inverse evolution by means of another cellular automaton called its *inverse*. If Φ is the global transition function of the main CA, Φ^{-1} is the global transition function of the inverse CA.

A 3D-CA is said to be *m-th order memory* cellular automaton if the state of each cell at t not only depends on the state of its neighbor cells at the previous step of time $t-1$, but also on the states of these neighbor cells at $t-2, t-3, \ldots, t-m$. In this case the global transition function is defined as follows:

$$\begin{aligned} \Phi \colon \mathbb{F}_2^{pqr} \times \ldots \times \mathbb{F}_2^{pqr} &\to \mathfrak{C} \\ \left(C^{t-1}, \ldots, C^{t-m}\right) &\mapsto C^t = \Phi\left(C^{t-1}, \ldots, C^{t-m}\right) \end{aligned} \tag{5}$$

The configurations $C^0, C^1, \ldots, C^{m-1}$ are called *initial configurations* of the memory 3D-CA.

Finally, note that a 3D object of dimension $p \times q \times r$ can be interpreted as a configuration of a 3D-CA where the state of the cell at position (i, j, k) is 1 if this position is occupied in the 3D object, and 0 if the position is free.

3 The Threshold Scheme

In this section the (k, n)-threshold scheme to share a secret 3D object K of dimension $p \times q \times r$ is introduced. It is based on the use of a reversible k-th order memory 3D-CA where one of its initial configurations stands for the secret K, and the computed shadows will be n consecutive configurations of the 3D-CA. Specifically, the method consists of three phases: the setup phase, the sharing phase and the recovery phase.

3.1 The Setup Phase

In this phase the following steps are performed by the dealer:

(i) The dealer chooses randomly an integer number $T \gg \max\{p, q, r\}$.

(ii) The dealer chooses the cellular automaton to compute the shadows: it is a reversible k-th order memory 3D-CA defined by the Moore neighborhood V and the following local transition function:

$$s_{ijk}^{t} = \bigoplus_{\substack{1 \leq l \leq n-1 \\ (\alpha,\beta,\gamma) \in V}} s_{\alpha\beta\gamma}^{t-l} \oplus s_{ijk}^{t-n}, \quad 1 \leq i \leq p, 1 \leq j \leq q, 1 \leq k \leq r. \tag{6}$$

(iii) Finally, D defines the initial configurations of the 3D-CA as follows:

$$C^0 = K, C^1 = Q_1, \ldots, C^{n-1} = Q_{k-1}, \tag{7}$$

where $Q_1, \ldots, Q_{k-1}$ are random 3D objects.

Note that the inverse cellular automaton of the reversible 3D-CA defined in (6) is given by the following local transition function:

$$\widetilde{s}_{ijk}^{t} = \bigoplus_{\substack{1 \leq l \leq n-1 \\ (\alpha,\beta,\gamma) \in V}} \widetilde{s}_{\alpha\beta\gamma}^{t-l} \oplus \widetilde{s}_{ijk}^{t-n}, \quad 1 \leq i \leq p, 1 \leq j \leq q, 1 \leq k \leq r. \tag{8}$$

3.2 The Sharing Phase

This phase consists of two steps:

(1) The dealer computes the configurations $C^T, C^{T+1}, \ldots, C^{T+n-1}$ of the 3D-CA defined in (6).
(2) The dealer D securely distributes the parameter T and the n configurations computed in the last step among the n participants:

$$P_1 \leftrightsquigarrow C^T, P_2 \leftrightsquigarrow C^{T+1}, \ldots P_n \leftrightsquigarrow C^{T+n-1}. \tag{9}$$

Note that the access structure is defined as follows:

$$\mathcal{B} = \{(P_i, P_{i+1}, \ldots, P_{i+k-1}), \, 1 \leq i \leq n-k\}. \tag{10}$$

Consequently, there are $n - k$ qualified subsets that can recover the original secret.

3.3 The Recovery Phase

In this phase the qualified subsets of participants can recover the secret K by simply computing the evolution of the inverse 3D-CA using their shadows as the initial configurations.

Specifically, the qualified subset $(P_i, P_{i+1}, \ldots, P_{i+k-1})$ recovers K by computing the configuration $\widetilde{C}^{T+k-1}$ of the inverse 3D-CA defined by (8) using

$$\widetilde{C}^0 = C^{T+i+k-2}, \widetilde{C}^1 = C^{t+i+k-3}, \ldots, \widetilde{C}^{k-1} = C^{T+i-1} \tag{11}$$

as the initial configurations of such cellular automaton.

3.4 An Illustrative Example

In what follows the protocol proposed above is illustrated by means of a simple example. In this case $k = 3$ and $n = 5$ and, for the sake of simplicity, we will consider a 3D secret object of dimension $4 \times 4 \times 10$. In Figure 2, the secret and the other two initial configurations are shown.

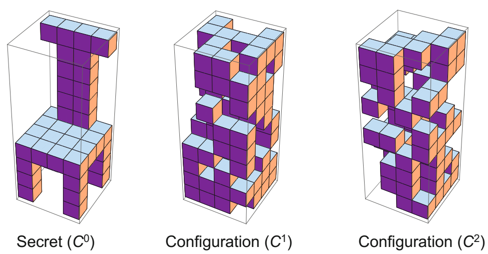


Fig. 2. Initial configurations of the 3D-CA: the secret to be shared (C^0) –left– and the two random tridimensional arrays (C^1, C^2) –middle and right–

If we set $T = 20$, then the shadows computed are the configurations $C^{20}, C^{21}, C^{22}, C^{23}$ and C^{24} of the 3D-CA (see Figure 3).

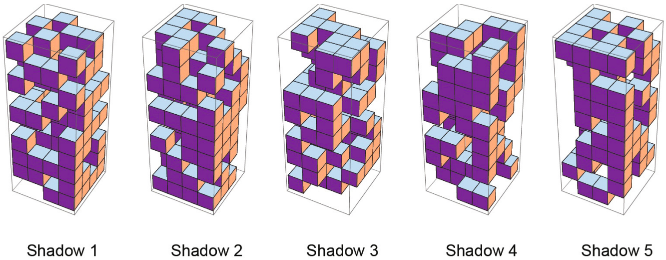


Fig. 3. Shadows computed using the $(3,5)$-threshold scheme

Note that in this case the access structure is the following:

$$\mathcal{B} = \{(P_1, P_2, P_3), (P_2, P_3, P_4), (P_3, P_4, P_5), (P_1, P_2, P_3, P_4), \\ (P_1, P_2, P_3, P_5), (P_1, P_2, P_3, P_4, P_5)\} \qquad (12)$$

4 Analysis of the Protocol

The threshold multi-secret sharing scheme presented in this work is perfect since it is not possible to recover the original secret from $k-1$ or less shadows. Note that the inverse 3D-CA needs k initial configurations $\widetilde{C}^0 = C^{T+i+k-2}, \widetilde{C}^1 = C^{T+i+k-3}, \ldots, \widetilde{C}^{k-1} = C^{T+i-1}$ to compute its evolution in order to obtain the secret: $\widetilde{C}^{T+k-1} = C^0$. If any shadow is lacked during the recovery phase, the inverse cellular automaton cannot be able to compute its evolution.

An opponent could try to substitute the missing share by another (different) one but no information will be obtained about the original secret. For example, taking into account the example shown in the last section, if we modify only one position in one of the shadows the secret computed is far different from the original one (see Figure 4-(a)). Moreover, if the shadows are interchanged in the recovery phase, the secret recovered is also unrecognizable (see Figure 4-(b)).

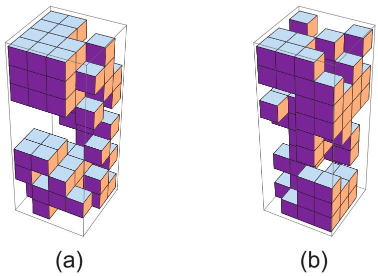


Fig. 4. (a) Secret recovered by the qualified subset (P_1, P_2, P_3) when the position $(2,2,5)$ of the third shadow C^{22} is modified. (b) Secret recovered when the recovery phase is performed using $\widetilde{C}^0 = C^{21}, \widetilde{C}^1 = C^{22}$, and $\widetilde{C}^2 = C^{20}$

Furthermore, the proposed scheme is also ideal since the shadows and the original secret have the same dimensions: $p \times q \times r$.

To measure the influence of the change of one position on the shadows we will compute the parameter λ which stands for the percentage of different positions values between the shadows obtained from original 3D secret which differ in only one position. Suppose that $S_1, S_2, \ldots, S_n$ and $S'_1, S'_2, \ldots, S'_n$ are the shadows obtained from the following two 3D secrets: K_1 and K'_1, which differs in only one position. For each pair of shadows $\{S_h, S'_h\}$, where

$$S_h = \left(w_{ijk}^h\right)_{1 \le i \le p, 1 \le j \le q, 1 \le k \le r}, S'_h = \left(w'^h_{ijk}\right)_{1 \le i \le p, 1 \le j \le q, 1 \le k \le r}, \tag{13}$$

we define a tridimensional array

$$D_h = \left(d_{ijk}^h\right)_{1 \le i \le p, 1 \le j \le q, 1 \le k \le r}, \tag{14}$$

such that:

$$d_{ijk}^h = \begin{cases} 0, \text{ if } w_{ijk}^h = w'^h_{ijk} \\ 1, \text{ if } w_{ijk}^h \neq w'^h_{ijk} \end{cases} \tag{15}$$

Then:

$$\lambda\left(S_h, S'_h\right) = \frac{\sum_{1 \le i,j,k \le n} d_{ijk}^h}{pqr}. \tag{16}$$

Several tests have been performed on the proposed scheme about the influence of only one position change and similar results are obtained. For example, in the case introduced in subsection 3.4, the coefficients obtained are the following:

$$\begin{aligned} \lambda\left(S_1, S'_1\right) &= 48.125, \\ \lambda\left(S_2, S'_2\right) &= 50.000, \\ \lambda\left(S_3, S'_3\right) &= 46.250, \\ \lambda\left(S_4, S'_4\right) &= 43.750, \\ \lambda\left(S_4, S'_4\right) &= 48.125. \end{aligned}$$

As it is shown, the percentage of different positions values between the shadows S_h and S'_h is about 50%, and as each position can assume only two values, 0 or 1, these are good results.

5 Conclusions

In this work, a (k, n)-threshold scheme has been proposed to securely share 3D objects. It is based on the use of a boolean reversible 3D-cellular automata with memory, where Moore neighborhoods are considered. This is a novel protocol since, as far as we know, this is the first secret sharing scheme for 3D objects appeared in the scientific literature.

It is shown that this scheme is perfect and ideal. Moreover, any information about the original secrets can be derived if we permute the correct shadows or the parameters of the 3D-CA are slightly modified. The protocol exhibits good statistical properties and resists the differential attack.

Acknowledgements. This work has been supported by Ministry of de Economy and Competitiveness (Spain) and European FEDER Fund under project TUERI (TIN2011-25452).

References

1. Chou, C.M., Tseng, D.C.: A public fragile watermarking scheme for 3D model authentication. Comput. Aided Des. 38(11), 1154–1165 (2006)
2. Fernandes, E., Delaigle, J.F.: Geometric soft hash functions for 2D and 3D objects. In: Proc. SPIE 5306, Security, Steganography, and Watermarking of Multimedia Contents VI, vol. 784 (2004)
3. Ivanov, S., Troyankin, A., Gurikov, P., Kolnoochenko, A., Menshutina, N.: 3D Cellular automata for modeling of spray freeze drying process. Comput. Aided Chem. Eng. 29, 136–140 (2011)
4. Kari, J.: Theory of cellular automata: A survey. Theor. Comput. Sci. 334, 3–33 (2005)
5. Kishk, S., Javidi, B.: Optical watermarking of 3D objects for authentication in transmission and storage. In: Proc. of the 16th Annual Meeting of the IEEE Lasers and Electro-Optics Society, pp. 732–733 (2003)
6. Montañola Sales, M., Rondão Alface, P., Macq, B.: 3D objects watermarking and tracking of their visual representations. In: Proc. of the Third International Conferences on Advances in Multimedia (MMEDIA 2011), pp. 111–113 (2011)
7. Mo, Y., Ren, B., Yang, W., Shuai, J.: The 3-dimensional cellular automata for HIV infection. Physica A 399, 31–39 (2014)
8. Pan, P.Z., Yan, F., Feng, X.T.: Modeling the cracking process of rocks from continuity to discontinuity using a cellular automaton. Comput. Geosci.-UK 42, 87–99 (2012)
9. Sarkar, P.: A brief history of cellular automata. ACM Comput. Surv. 32, 80–107 (2000)
10. Tarmissi, K., Ben Hamza, A.: Information-theoretic hashing of 3D objects using spectral graph theory. Expert Syst. Appl. 36, 9409–9414 (2009)
11. Wang, K., Lavoué, G., Denis, F., Baskurt, A.: Three-dimensional meshes watermarking: Review and attack-centric investigation. In: Furon, T., Cayre, F., Doërr, G., Bas, P. (eds.) IH 2007. LNCS, vol. 4567, pp. 50–64. Springer, Heidelberg (2007)
12. Wolfram, S.: A New Kind of Science. Wolfram Media Inc., Champaign (2002)
13. Yeo, B.L., Yeung, M.M.: Watermarking 3D objects for verification. IEEE Comput. Graph. 19(1), 36–45 (1999)
14. Zaeem, M.A., Yin, H., Felicelli, S.D.: Modeling dendritic solidification of AlCu using cellular automaton and phase-field methods. Appl. Math. Model. 37(5), 3495–3503 (2013)
15. Zhao, Y., Qin, R.S., Chen, D.F.: A three-dimensional cellular automata model coupled with finite element method and thermodynamic database for alloy solidification. Journal of Crystal Growth 377, 72–77 (2013)

Generation Method of the Trigger Signal for the Automatic Capture System to the Harmful Animals with Intelligent Image Processing

Fumiaki Takeda*

Kochi University of Technology, Japan
takeda.fumiaki@kochi-tech.ac.jp

Abstract. Up to now, some damage by the harmful animals such as deer and wild boars are increased abruptly. Their damage such as crash between cars and harmful animals, eating all domestic vegetables are occurred at not only local area but also urban area. Under these backgrounds, an efficient and reliable capturing system for the harmful animals is requested, extremely. We aim at to develop an efficient automatic capturing system for harmful animals using image as a final product. Especially, we focus its generation method of the trigger signal for the capture devices. For the all days use without human's monitor and operation, we propose a new generation method for the trigger signal using an intelligent image processing and a new camera device Kinect. We consider activation of the trap according to the animal's shape and size furthermore motion by the optical flow analysis, automatically. The Kinect is equipped infrared radiation projector and its receiver camera, which can get the depth image. We construct an experimental device, which is simple and miniature of the real size and investigate detection ability. We discuss its effectiveness of the generation method for the realization of the capture system.

Keywords: Image Processing, Harmful Animals, Kinect, Depth image, Optical flow.

1 Introduction

Until today, some damage by the harmful animals such as deer and wild boars are increased, day by day. Their damage such as crash between cars and harmful animals, eating all domestic vegetables, and so on are occurred at not only domestic area but also urban area. Its total damage costs are more than $600 millions for one prefecture per one year in Japan. Conventionally, various trap systems [1]-[6] were developed and released to the field as shown in Fig.1. However, they were almost all operated by the manual. Furthermore, its capturing ability is not so enough because of human's obscure operation [1]-[3]. Under these backgrounds, an efficient and reliable automatic capturing system for the harmful animals is requested, extremely.

* 185, Miyanokuchi, Tosayamada-cho, Kami-city, Kochi, 782-8502, Japan.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_52,

Therefore, we aim at to develop an efficient automatic capturing system to the harmful animals using image as a final product in the field as shown in Fig 2 without human's operation. This shows ideal system that can capture the harmful animals according to their spices and number using plural visible and infrared cameras, automatically. It can communicate another surveillance network commuter by itself after the capturing them. Especially, for all day's use of this system, we adopt new image acquisition devices such as a Kinect for the construction of the automatic capturing system. This camera device is developed and came onto the market as game equipment. We use this device for generation of the trigger signal of the trap activation according to the analysis result of the animal's shape and size, furthermore, motion detection by the optical flow method. In this paper, we focus its generation method of the trigger signal for the trap device. About the confirmation of the effectiveness for our generation method of the trigger signal, we also construct an experimental device which is simple and miniature of the real size. We layout that infrared camera device Kinect and the trap which is constructed by 1/10 scale model compared with real one as the experimental device in the laboratory.

First, in the experiment for the reappearance and objectivity, we use a toy, which is moved by itself with the spring as the target object. Second, we adopt the two hamsters, which move each other randomly, in addition. Finally, we discuss its feasibility and possibility of our generation method of the trigger signal for the real trap system, which can be used for all day use based on these experimental results.

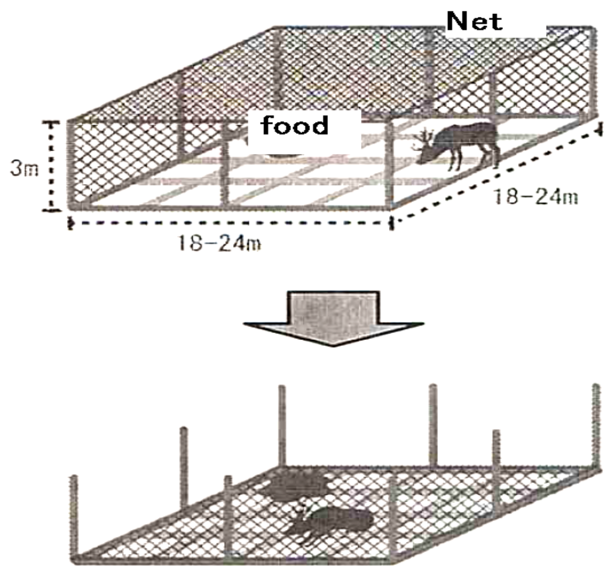


Fig. 1. The conventional trap system

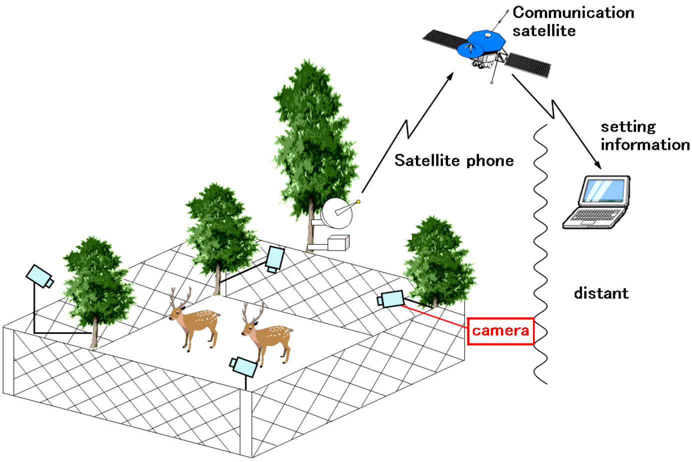


Fig. 2. Final proposed capturing system overview

2 System Construction

We denote here an experimental system constructed by the infrared device Kinect and the trap device, which is constructed by 1/10 scale model compared with real one in the laboratory as shown in Fig.3. We confirm and discuss detection ability using our generation method of the trigger signal for the trap.

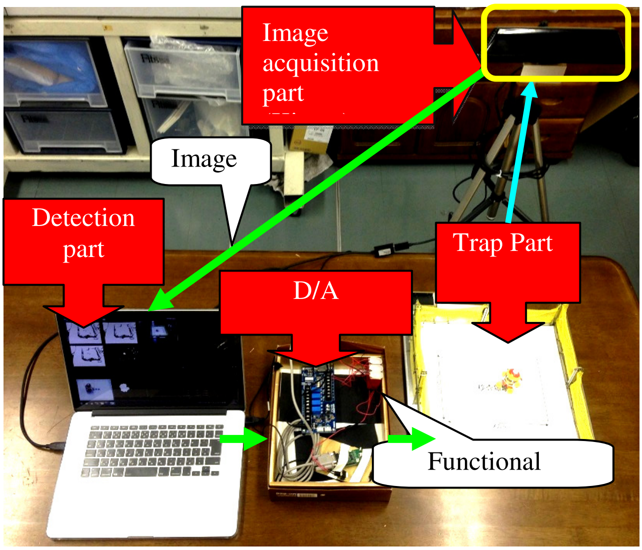


Fig. 3. Experimental trap system overview

2.1 Image Acquisition Part

We adopt a new image acquisition device such as a Kinect as shown in Fig.4 and Table 1 for the construction of the new capturing system. This camera device is developed and came onto the market as game equipment. Especially, the Kinect is equipped infrared projector and its receiver camera, which can get the length between the camera and a target object. After this, we use this length information as the depth image of the target object. Figure 5 shows an example image by this one.

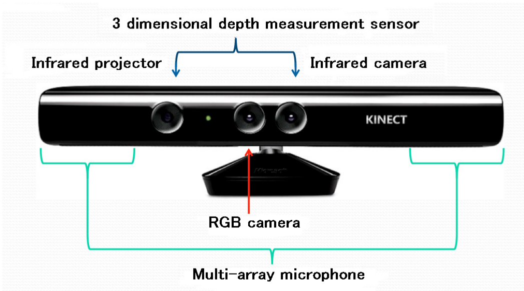


Fig. 4. Construction of the Kinect

Table 1. Specification of the Kinect

RGB camera	resolution VGA(640×480)
camera	resolution VGA(640×480)
frame rate	30fps
recognition length	0.4～6m
vertical view	43.5degree
horizontal	57.5degree
tilt range	±30degree
audio format	16-kHz, 16-bit monaural PCM

Fig. 5. Example of the depth image using infrared radiation

2.2 Detection Part

We use this depth image from the Kinect for operation of recognition of the target object. First of all, we should recognize harmful animals by their shape and size with blob analysis [7], which is the domain constructed by several neighboring pixels. Still more, we should recognize them either true animals or not. Because there is sometimes miss recognition owing to the similar shape of the rocks or threes with using only shape and size of the animals [3]. Therefore, we need motion detection from the extracted blob in the depth image. The optical flow method can make possible to find motion of the animals easily. It is given by equation (1).

$$\left.\begin{array}{l} AE(x, y, dx, dy)=\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} \mid P_n(x+i, y+j) - \\ P_{n-1}(x+i+dx, y+j+dy) \mid \end{array}\right\} \quad (1)$$

Here AE is block matching value on block (N×M), P is frame, dx, dy denotes motion vectors.

Each procedure is described bellows.

1. First of all, we execute blob analysis for the depth image.
2. We account number of the pixel in the outline of the labeled blob as the size of the animal.
3. We compare the number of the pixel with some threshold value for the size judge.
4. We set the rectangle to the labeled blob and measure its long and short sides.
5. We get an aspect ratio of long and short sides as the shape of the animal.
6. We compare its aspect ratio with some threshold value for the shape judge.
7. We analyze depth image by the optical flow method for x and y direction.
8. We compare the analysis results of the optical flow method with some threshold value for the motion judge.

Figure 6 shows these procedures.

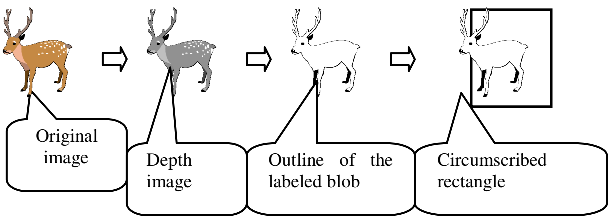


Fig. 6. Setting procedure of the circumscribed rectangle

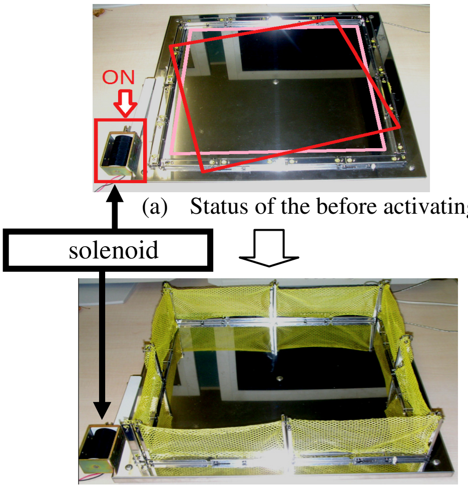
2.3 Trap Part

We describe the capturing part which we construct 1/10 scale model compared with real one [1]-[3], [5], [6] as the experimental device in the laboratory for the confirmation of the effectiveness of generation method of the trigger signal. The trap is constructed by several net units. The trap is activated by the spring and solenoid switch. Figure 7 shows the status of the before activation (a) and the one of the after activation (b) on the trap.

3 Experiments

To confirm the effectiveness of the proposed method, first of all we should consider reappearance and objectivity of the experiments. So we use here a toy which is moved by itself with the spring as the target object as shown in Fig. 8. Furthermore, we use real small animals such as two hamsters in addition as a trial. Toy's specification and its judged value is denoted bellows.

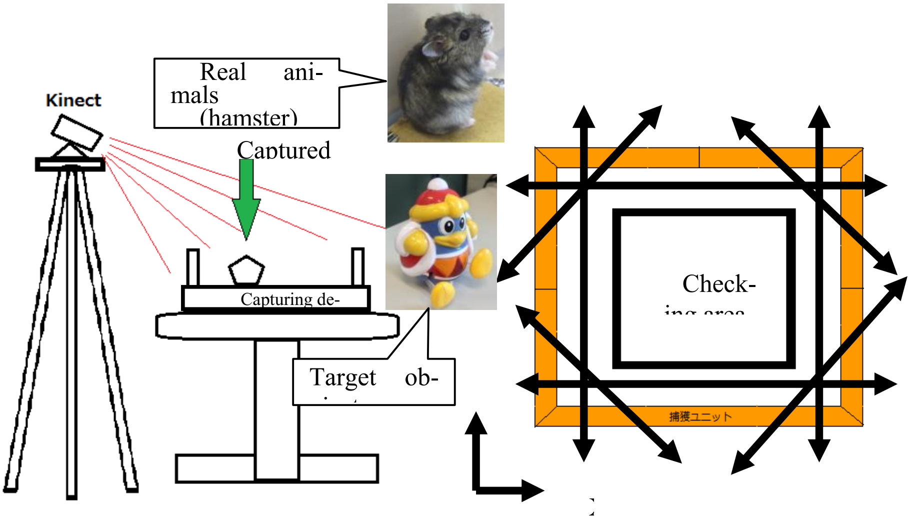
Size is 780 pixels (judged value is more than 780), aspect ratio is 1.0 (judged value is 0.7 ~ 1.3), moving speed is 10cm/s (judged value is more than 20 for x or y direction). Furthermore, the judge values of the hamster are also same. Here, we confirm the recognition of the target object and detection with the display of the PC and the trap. Finally we discuss basic feasibility and possibility for the real trap system based on the experimental results with our proposed generation method for the trigger signal.



(a) Status of the before activating on the trap

(b) Status of the before activating on the trap

Fig. 7. Construction of the trap system



(a) Layout of the experimental de-

(b) Checking area for the top

Fig. 8. Setting feature of the experimental device

Table 2. Experimental result under the day light (305 lux)

	The check of outside			The check of inner		
	Total number	Number of success	Success ratio	Total number	number of miss detection	Success ratio
Xdrection	10	10	100%	10	0	100%
Ydrection	10	10	100%	10	0	100%
Slanting direction	10	10	100%	10	0	100%

Table 3. Experimental result under the night (52 lux)

	The check of outside			The check of inner		
	Total number	Number of success	Success ratio	Total number	number of miss detection	Success ratio
Xdrection	20	20	100%	20	0	100%
Ydrection	20	20	100%	20	0	100%
Slanting direction	20	20	100%	20	0	100%

3.1 Experimental Conditions

Figure 8(a) shows setting feature of the experimental device. First, we use a toy which moves one way. We make experiments under the day light (305 lux: Room size: 10m×8m×3m) and night (52 lux: Room size: 10m×8m×3m). Moving direction, which is denoted by several arrows, of the toy in the experimental device is x and y, and slanting as shown in Fig. 8(b). We make experiments check in the checking area and outside of it.

3.2 Experimental Results

We got good detective ability with the proposed generation method as shown in Table 2 and 3. We could confirm the detection with the display of the PC and the trap behavior. Furthermore, we made additional experiments by using real small animals such as hamster as shown in Fig. 8, as a tentative trial. They are two and behave each other randomly. Its detective ability was 100% with ten times trials. From both experimental results, we confirmed our proposed generation method of the trigger signal is effective in the laboratory environment, basically.

4 Conclusion

We proposed a concept of the automatic capture system for harmful animals with intelligent image processing and the depth image. Especially, we focused on the generation method of the trigger signal for the trap. We constructed 1/10 scale model compared with real one as the experimental device in the laboratory. We got good detective ability in the regulated conditions in the laboratory using a toy and real small animals such as hamsters. We basically confirmed the feasibility and possibility of our proposed generation method for the trigger signal to the conventional capture devices. It is supposed that our proposed method makes possible to autonomous capture devices without human direct operation. Still more, this can be extended to application of another kind of animals such as wild boars, wolves and ownerless dogs according to its shape and size. However, it is still needed that we should implement our method of the generation for the trigger signal to the conventional capture devices [1]-[3], [6] and make experiments in the real fields.

References

1. Conner, M.C., Soutiere, E.C., Lancia, R.A.: Drop-netting deer: Costs and incidence of capture myopathy. Wildl. Soc. Bull. 15, 434–438 (1987)
2. Ramsay, C.W.: A drop-net deer trap. J. Wildl. Manage. 32, 187–190 (1968)
3. Goderey, M.E.R., Browne, R.W.: A review of some mule deer capture operations. ASTM Spec. Tech. Publ (Am. Soc. Test. Mater.) 974, 170–176 (1988)

4. Webb, S.L.: Assessing the Helicopter and Net Gun as a Capture Technique for White-Tailed Deer. J. Wildl Manag. 72(1), 310–314 (2008)
5. Barrett, M.A.: Translating Bait Preference to Capture Success of Northern White-Tailed Deer. J. Wildl. Manag. 72(2), 555–560 (2008)
6. Webb, S.L., Gee, K.L.: Survival and fidelity of an enclosed white-tailed deer population using capture-recapture-reporting data. Popul. Ecol. 52(1), 81–88 (2010)
7. Wilson, R., Meulemans, P., Calwaya: Image Sequence Analysis and Segmentation Using Gblobs. In: Proc. Int. Conf. Image Process., vol. 2, pp. 483–487 (1998)

2-Scene Comic Creating System Based on the Distribution of Picture State Transition

Miki Ueno, Naoki Mori, and Keinosuke Matsumoto

Graduate School of Engineering
Osaka Prefecture University
1-1 Gakuencho, Nakaku, Sakai, Osaka 599-8531, Japan
ueno@ss.cs.osakafu-u.ac.jp

Abstract. Understanding picture by computer has become one of the most important topics in computer science. However, there are few researches have been reported about human like picture understanding by computer. The main reason of difficulty is that lots of picture expressions contain more lyric aspect than natural language expressions. Comic is one of the best target of picture understanding researches because pictures in comics express simply and clearly story, therefore we can presume that pictures in comics have strong universality. Picture understanding is defined as understanding situations and estimating transition between current scene and next scene. In this paper, The novel method which generates pictures using prepared picture parts and image objects databases is proposed. We also show the 2-scene comics creating system using user inputs picture and propose the representation of picture state transition.

Keywords: Automatic Drawing Picture, 2-scene Comic Creating System, Picture State Transition, Picture Model, Introducing Semantics to Picture, Comic Engineering.

1 Introduction

Creating products of human intellectual activities such as novels, comics and musics is ultimate purpose of artificial intelligence. The picture is one of the most important representations of such activities, and representation using pictures is a strong way to communicate because pictures in some cases include more information rather than text[1] and are understandable for various people regardless of age and national origin.

In the field of picture research, despite current advances of image recognition[2,3,4] and analyzing brain activities for images[5], there are few research about semantics of picture[6,7]. To solve this problem, we have proposed a novel dialogue system called *Picture Information Shared Conversation Agent*(**Pictgent**)[8] that shares conversational background knowledge by showing prepared pictures with model to user. Though we proposed the picture model to represent semantics of picture in Pictgent, we have not obtained complete one because of difficulty of defining what picture is. In order to solve the problem, we have analyzed the structure of existing comics whether there are certain relation between the story of 4-scene comics and transition of object. However, unexpectedly, we failed to obtain clear constitution of pictures in comics. We found that

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_53,

lots of stories only use objects of human and balloon objects, hence the stories depend on natural language in balloon and it is difficult to create semantics of picture by existing comics. There have been reported another researches[9] of semantic representation of picture, but no research has proposed complete picture model.

In this paper, we tried to propose system which can create comics like pictures flexibly based on objects in the prepared dataset. We focused on 2-scene comics first because this is the minimum set of pictures which contains some kinds of object transition. After user inputing for 1st scene, the proposed method generates 2nd pictures based on 1st picture automatically with utilizing picture objects transition database.

To achieve computer understanding pictures, 3 issues are required as follow:

1. Estimate situations in pictures.
2. Represent feel or emotion of picture numerically.
3. Expect next scene after current pictures.

We propose the method of creating 2-scene comics by user' inputs in section 2. We show the constitution of application in section 3. Finally, in section 4, we present the conclusions of this study.

2 Proposed Method

In this section, we describe the proposed 2-scene comic creating system.

2.1 Representation of Picture in This Paper

In order to represent semantics of pictures for computer, we propose the "picture model" in which we define each of elements in pictures including background images and balloons as certain object.

It is difficult to create reasonable picture model for all pictures. Therefore, as previously mentioned, we focused on 2-scene comics because 2 sequential pictures have enough information rather than single picture. In this method, we should prepare dataset including images of objects, information of objects, types of areas, drawing operator, and transition. Each picture has above information and history, and those are representation of picture in this study.

2.2 Picture Model

In this study, initial picture p_0 is generated by system. p_0 is a blank picture or a background picture and has several areas determined by system. User can operate picture as user input. User operation set in time t is represented by $\boldsymbol{u_t}$, and $\boldsymbol{u_t}$ contains all kinds of user's operation for picture p_t. p_{t+1} is created applying $\boldsymbol{u_t}$ to p_t. We describe this as follows:

$$p_{t+1} = \prod_{u_i \in \boldsymbol{u_t}} u_i p_t \tag{1}$$

where u_i represents each user's operations in time t.

Fig. 1. cup (original image)

Fig. 2. cup (apply $\hat{\phi}_{\mathrm{divide}}$)

Fig. 3. cup (apply $\hat{\phi}_{\mathrm{rotate}}\hat{\phi}_{\mathrm{divide}}$)

The relation among p_t, o_t, $\boldsymbol{u_t}$, $\{O_t\}$ and $\{U_t\}$ is following:

$$\{O_t\} = \{o_t \in p_t\} \tag{2}$$

$$\{U_t\} = \{\boldsymbol{u_1}, \boldsymbol{u_2}, ..., \boldsymbol{u_t}\} \tag{3}$$

$$p_{t+1} = \prod_{u_i \in \boldsymbol{u_t}} u_i p_t = \prod_{u_i \in \boldsymbol{u_t}} u_i \prod_{u_i \in \boldsymbol{u_{t-1}}} u_i p_{t-1} \tag{4}$$

$$\dots \tag{5}$$

$$= \prod_{j=0}^{t} \prod_{u_i \in \boldsymbol{u_j}} u_i p_0 \tag{6}$$

Definite operation of $u_i \in \boldsymbol{u_t}$ is represented by drawing operator ϕ and operand object in p_t. Each ϕ also has specific augments such as color, angle or other objects. Object o_t is able to represent object transition. Object o_n in n-th scene has Markov model as bellow:

$$o_n = \prod_{i=0}^{N_n^{\max}} \phi_i o_0 \tag{7}$$

where $N^{\max}$ is total number of ϕ to applied to o. $N_n^{\max} \geq n - 1$ because each u_i contains at least one ϕ.

2.3 Base Drawing Operator $\hat{\phi}$

Drawing operator ϕ has the function of transform for pictures. Product of several drawing operators is also drawing operator and the set of base operator is defined as $\{\hat{\phi}\}$. Increasing the number of base operators, the ability of representation of the system also increases more and more because system can create lots of new images utilizing current objects.

2.4 Implementation of Drawing Operator ϕ

In this paper, the system is written by Java and the operator is implemented by command pattern in design pattern[10]. Base drawing operator corresponds to command

and product of drawing operator corresponds to macro-command. In Java implementation, product of operator is represented by using queue.

For example, ϕ_{crack} which represents cracked transition is composed of base drawing operator $\hat{\phi}_{\text{divide}}$ $\hat{\phi}_{\text{rotate}}$.

ϕ_{crack} is applied to object o from right side, 3 base drawing operators are applied to o as follows:

$$\phi_{\text{crack}} o = \hat{\phi}_{\text{rotate}} \hat{\phi}_{\text{divide}} o \tag{8}$$

Figure 1-3 show the variation of image by applying base drawing operator of ϕ_{crack}.

2.5 Flow of Proposed System

The basic flow of proposed system is described as follows[1]:

1. Select the base image as background for initial states. We define this picture as p_0.
2. System waits for user inputs. User can decide several actions to p_0. In this paper, user action is only setting an object in objects set shown by system GUI. User can set their interested object on one area of p_0 with drawing operator and emotion information. Any number of objects are added to p_0 under condition of existing only one object in each area. When user put an object to p'_0, new areas decided by set object are added to picture.
3. After applying all user inputs $\boldsymbol{u}$ to p_0, p'_0 is obtained. This p'_0 become 1st scene.
4. System tries to change objects in p'_0 based on user decided drawing operator and emotion information to target object. New image of changed objects is obtained by transition DB.
5. Applying all changes to p'_0, p_1 is obtained. p_1 represents 2nd scene.

   2-scene comic : p_1 is final output.

   ***n*-scene comic :** p'_1 becomes p_2. System waits for user inputs again and repeats above flow until p'_{n-1} becomes p_n.

2.6 User Input $\boldsymbol{u}$

As mentioned above, user action is putting some objects on one area of the picture. User decides the object and the area to set shown by system. Set objects in $\{O_{\text{set}}\}$ and picture objects in p_t are different as though set objects become picture objects after user setting, where $\{O_{\text{set}}\}$ represents a group which is applicable to picture. When user sets the object to picture, user also sets the "drawing operator" and "emotion" information to decide transition pattern of setting object. Finally, user can decide following information by one user input.

Object o in $\{O_{\text{set}}\}$. The object information of user setting object.
Area a. Set area of o.

[1] p'_t represents the latest picture after user decides several actions for p_t in eq. 1. The operator applied to p'_t is decided uniquely based on user action and p_{t+1} is obtained after it is executed.

Table 1. Object DB

ID	Object Name	Class Name	Transition:Frequency	Image File Path	Included Area ID
1	cup	container	drink:4,crack:2,drop:1	/fig/cup.png	null
2	cat	animal	bite:5,attach oneself to:3,run away:2	/fig/cat.png	null
3	girl	human	stand:3,sit:1, $\cdots$	/fig/girl.png	2, 3, 4

Some objects have several areas with IDs where user can set a new object.
Class name is important information to use external resources such as Wordnet.

Table 2. Transition DB

ID	Object ID	Area ID	Operator Name	Emotion Name	Drawing Operator
1	*	*	divide	sad	crack(Object ID)
2	*	2	*	*	drop(Object ID)

*: Wildcard
In database, we describe $\phi_{\mathrm{crack}} \rightarrow \mathrm{crack}(x)$.

Operator m**.** Operator type of setting object o.
Emotion e**.** User can set the emotion e of setting object. This information effects transition of o.

Using those information, system decide the transition of each objects in the current picture. All transitions are represented by change of image. Therefore, system have to prepare new images of all objects in picture after transition. Creating new images has the difficulty of huge range of variation. To solve this problem, we prepared only basic image sets and created various images to apply drawing operator to an images. For example, we can create various image pattern of cup such as cracked cup, rotated cup or different colored cup from only one cup image.

2.7 Databases in the Proposed System

Our final purpose is to establish a database which can represent the transition of objects in 2.2. However, to make this, we have to decide the complete distribution model of each object. Since this is a difficult task, we simplified the object transition that all objects have only 1 transition result.

The transition pattern of each object is decided by type of object, area of object, operator for objects and emotion of objects uniquely. We created those patterns based on existing pattern of comics and common sense, e.g. cup is cracked or we use the umbrella in the rain. Although we set only 1 typical pattern in this study, transition patterns will be extended based on distribution of existing data in future work. We created 4 databases (DBs) for managing those information as follows.

Object DB. Object DB manages the information of all objects. In object DB, each object has the attributes of object name, class name, what kinds of transition and the rarity in existed comics and common sense, image file path of image, area IDs. Table 1 shows the example of object DB.

Table 3. AreaName DB

ID	Area Name
1	whole
2	ground
3	hand

Table 4. Area DB

ID	Area ID	Object ID	Start point of area	End point of area
1	1	*	(*, 170)	(*, 190)
2	2	3	(20, 40)	(40, 60)

*: Wildcard

Transition DB. Transition DB manages the distribution of object transition. This database contains user input $\boldsymbol{u}$ as key and drawing operator ϕ as value. User Input $\boldsymbol{u}$ is represented by following equation:

$$\boldsymbol{u} = (\text{object } o, \text{area } a, \text{operatorname } m, \text{emotion } e) \tag{9}$$

Operator name m is name string corresponding to each base drawing operator name. Table 2 shows the example of transition DB.

In current transition DB, only 1 operator is considered for each $\boldsymbol{u}$ although $\boldsymbol{u}$ has the distribution of several operators.

Area Name DB and Area DB. Area Name DB and area DB manage the area information in picture. Though area is not appeared visually in the picture, this is important to control the location of set objects. When user sets the object by mouse operation, user can recognize the area by specific color.

Table 3 shows area name DB which has the relation between area ID and area name. Table 4 shows area database which has the start point and the end point of area. The start and end point of same area name are different among objects.

3 Application

We developed 2-scene comic creating system based on the proposed method. Figure 4 shows the outline of system. Figure 5 shows picture model of 2nd scene.

Followings are explanations of each part of system corresponding to the number in figure 4.

1. **Selectable background image.** Select button of the base image for 1st scene. The blank picture or prepared picture is able to be selected.
2. **1st scene image.** The display area of 1st scene image. After applying user input, p_0 becomes p'_0.
3. **Selectable areas.** Areas where user can set object in 1st scene.
4. **2nd scene image.** The display area of 2nd scene image which is generated by applying user input of 1st scene.
5. **Operator & emotion.** The display area of operator and emotion for set objects.
6. **Selectable object.** The display area of $\{O_{\text{set}}\}$ which is set of user selectable objects.
7. **Picture model.** Picture model XML of 2nd scene.

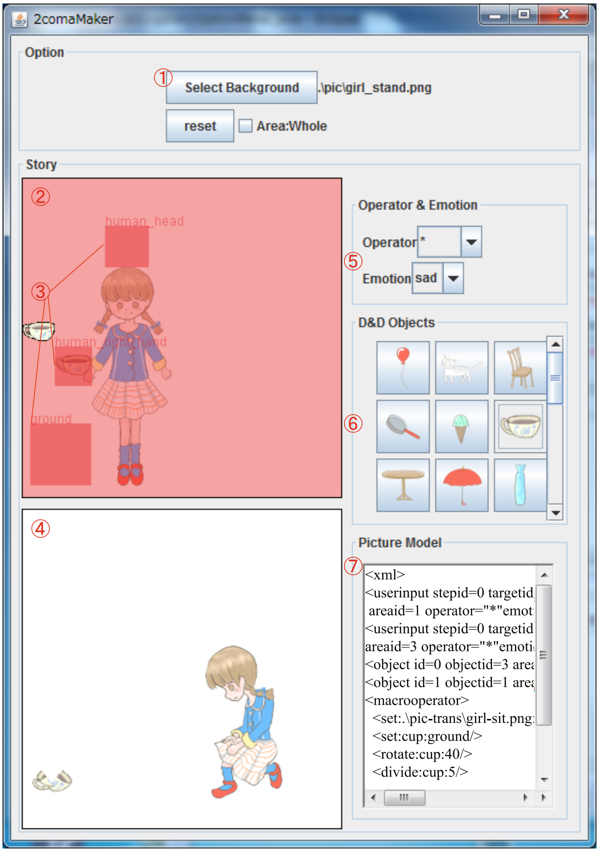


Fig. 4. Outline of 2-scene comic creating system

```
<xml>
<userinput stepid=0 targetid=3 areaid=1 operator="*"emotion="*"/>
<userinput stepid=0 targetid=1 areaid=3 operator="*"emotion="sad"/>
<object id=0 objectid=3 areaid=1 />
<object id=1 objectid=1 areaid=2 />
<macrooperator>
  <set:.\pic-trans\girl-sit.png:whole/>
  <set:cup:ground/>
  <rotate:cup:40/>
  <divide:cup:5/>
  <zoom:cup:50@40/>
</macrooperator>
</object>
</xml>
```

Fig. 5. XML of 2nd scene

Table 5. XML Tag of Picture Model

Tag Name	Explanation
<object>	Information of object. The attributes of this tag are object ID in picture, ID of object DB and area ID.
<macrooperator>	Information of applied drawing operator to the object.
<userinput>	Information of user's inputs. The attributes of this tag are picture step, target object ID, area ID, operator name and emotion name.

4 Conclusion

In this paper, we proposed the method of generating pictures automatically by applying drawing operators to prepared objects.

Furthermore, we developed creating 2-scene comic system utilizing the proposed method. Followings are important future work: (1) What kinds of information object has. (2) How to relate user input and transition. (3) How much drawing operators are required. If we solve those problems completely, we will create conversational system between humans and computers, interactive picture book and drawing application which displays comments about user's drawing.

Acknowledgment. This work was supported by Grant-in-Aid for JSPS Fellows(Grant Number:25·10978).

References

1. Kurlander, D., Skelly, T., Salesin, D.: Comic Chat. In: SIGGRAPH 1996 Conf. Proc., pp. 225–236 (1996)
2. Herot, C.F.: Graphical Input through Machine Recognition of Sketches. Computer Graphics (Proc. SIGGRAPH) 10(2), 97–102 (1976)
3. Eitz, M., Hays, J., Alexa, M.: How Do Humans Sketch Objects? ACM Trans. Graph. (Proc. SIGGRAPH) 31, 44:1–44:10 (2012)
4. Fu, H., Zhou, S., Liu, L., Mitra, N.J.: Animated Construction of Line Drawings. ACM Trans. Graph. (Proc. SIGGRAPH ASIA) 30(6) (2011)
5. Walther, D.B., Chai, B., Caddigan, E., Beck, D.M., Fei-Fei, L.: Simple line drawings suffice for functional MRI decoding of natural scene categories. Proc. National Academy of Sciences 108(23), 9661–9666 (2011)
6. Ito, K., Matsuda, M., Durst, M.J., Hasida, K.: SVG Pictograms with Natural Language Based and Semantic Information. SVG Open (2007)
7. Gross, M.D., Do, E.Y.: Drawing on the Back of an Envelope: A Framework for Interacting with Application Programs by Freehand Drawing. Computers & Graphics 24(6), 835–849 (2000)

8. Ueno, M., Mori, N., Matsumoto, K.: Picture Information Shared Conversation Agent: Pictgent. In: Omatu, S., Paz Santana, J.F., González, S.R., Molina, J.M., Bernardos, A.M., Rodríguez, J.M.C. (eds.) Distributed Computing and Artificial Intelligence. AISC, vol. 151, pp. 91–94. Springer, Heidelberg (2012)
9. Arai, N., Matsuzaki, T.: Can a Robot Get into the University of Tokyo?: The Artificial Intelligence Project at the National Institute of Informatics. Journal of the Japanese Society for Artificial Intelligence 27(5), 463–469 (2012)
10. Gamma, E., Helm, R., Johnson, R., Vlisside, J.: Design Patterns: Elements of Reusable Object- Oriented Software. Addison-Wesley (1995)

A Brief Approach to the Ear Recognition Process

Pedro Luis Galdámez, María Angélica González Arrieta, and Miguel Ramón Ramón

University of Salamanca, Plaza de los Caídos, 37008 Salamanca, Spain
{peter.galdamez,angelica}@usal.es, miguel.ramon@dgp.mir.es

Abstract. This paper offers an approach to biometric analysis using ears for recognition. The ear has all the assets that a biometric trait should possess. Because it is a study field in potential growth, this study offers an approach using SURF features as an input of a neural network with the purpose to detect and recognize a person by the patterns of its ear, also includes, the development of an application with .net to show experimental results of the theory applied. Ear characteristics, which are a unchanging biometric approach that does not vary with age, have been used for several years in the forensic science of recognition, thats why the research gets important value in the present. To perform this study, we worked with the help of Police School of Ávila, Province of Spain, we have built a database with approximately 300 ears.

Keywords: Neural Network, SURF Algorithm, Ear Recognition.

1 Introduction

The ear has been used as a means of human recognition in forensic activities for a long time. During the investigation of various crime scenes, earprints commonly been used to identify a suspect especially when there is no information of fingerprints. A recognition system based on images of the ears is very similar to a typical face recognition system, however, the ear has some advantages over the face, for example, their appearance does not change due to expression and is little affected by the ageing process, its color is usually uniform and the background is predictable.

Although the use of information from ear identification of individuals has been studied, is still an open question by specifying and determining whether or not the ear can be considered unique or unique enough to be used as a biometric. Accordingly, any physical or behavioural trait can be used as biometric identification mechanism provided which is universal, that every human being possesses the identifier, being distinctive and unique to each individual, invariant in time, finally measurable automatically or manually, the ear accomplish all these characteristics.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_54,

2 State of Art

The first technique known to detect the ears is raised by Burge and Burger [17] who have made the process of detection using deformable contours with the observation that initialization contour requires user interaction. Therefore, the location of the ear is not fully automatic. Meanwhile Hurley et al. [7] used the technique of force field, this process ensures that it is not required to know the location of the ear to perform recognition. However, only applies when he has the image specifies the ear without too much noise. In [19], Yan and Bowyer have used manual technique based on two previous lines for detection, where you take a line along the border between the ear and face while another line crosses up and down the ear.

The strategy proposed by Alvarez et al. [15] uses ovoid contour of the ear where the limit is estimated by fitting the shape of an ear in the image using a combination of snake line and ovoid models, like the proposal of Burge and Burger this technique requires an initialization. Ansari and Gupta [20] presented a process based on the outer ear helices edges, they use 700 samples collected at IIT Kanpur, the strategy only relies on the outer helix curves. Yuan and Mu [16] have proposed a skin-color and contour information technique, they perform the ear detection considering ear shape elliptical and fitting an ellipse to the edges to get the accurate ear position. Attarchi et al. [21] have shown an ear detection process based on the edge map. It relies on the hypothesis that the longest path in edge image is the ear outer boundary.

A. Cummings [2] shows a strategy using the image ray transform which is capable of highlighting the ear tubular structures. The technique exploits the helix elliptical shape to calculate the localization. Kumar et al [1], have introduced a proposal where uses skin segmentation and edge map detection to find the ear, once they find the ear region apply an active contour technique [22] to get exact location of ear contours, the technique has been tested on 700 ear images. Like these techniques there are many other significant proposals.

The most used technique for face recognition [18], principal component analysis (PCA), is also suitable for use in ear recognition. The first application of PCA to ear recognition was by Victor et al. [3] they used PCA to perform an comparative analysis between face and ear recognition concluding that the face performs better than the ear. However, Chang et al. [14] also have accomplished a comparison using PCA and found that ears provided similar performance to faces, they concluded that ears are essentially just as good as faces for biometric recognition. There are many proposals to solve the problem, in this paper only has done a small review from some of them, the next section introduce an attempt to solve the ear recognition problem in a practical way, applying the concepts studied for 2D images to develop an application which allow to perform the ear recognition in real-time video.

3 Ear Recognition System

Most of ear biometric articles have centred their attention on recognition using manually cropped ear images. However, for a robust ear recognition system is desired to detect the ear from the profile face image in an automatic way.

4 Detecting and Tracking the Ear

There exist some techniques which could be used to detect ear automatically. However, These techniques usually can detect the ear only when a profile face image do not contain a noisy or big background around the ear. This section proposes an useful ear localization technique which attempts to solve these issues.

4.1 Ear Localization

OpenCV and its wrapper for .Net framework EmguCV includes different object detectors based on the Viola-Jones framework. Modesto Castellón-Santana et al. [5] have developed a haarcascade classifier to be used with OpenCV to detect left and right ears. This classifier represents a first step to create a robust ear detection and tracking system. The Application is developed in C#.

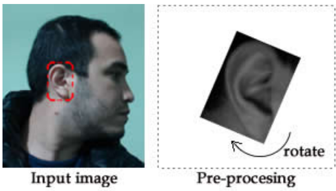


Fig. 1. Ear detection

With the ear identified we proceed to perform the pre-processing task, converting the image to gray scale and we begin the normalization process, first we perform the segmentation of the image applying a mask to extract only the ear, then the image is converted to an edge map using the canny edge filter. If w is the width of the image in pixel and h is the height of the image in pixel, the canny edge detector takes as input an array $w \times h$ of gray values and sigma. The output is a binary image with a value 1 for edge pixels, i.e., the pixel which constitute an edge and a value 0 for all other pixels. We calculate a line between major and minor y value in the edge image to rotate and normalize each image, trying to put the lobule of the ear in the centre. This process is trying to get all the images whose shape is similar to the image to identify. We identify some points on the external shape of the ear and the angle created by the center of the line drawn before and the section in the ear's tragus with the major x value.

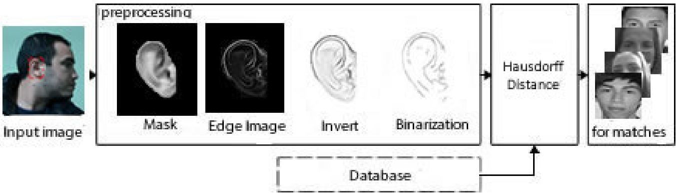


Fig. 2. Image pre-processing

Once the pre-processing is complete we proceed to compute a Match within the database using the contours of the ear form, with this we are trying to reduce the number of candidates. The Hausdorff distance measure used in this document is based on the assumption that the ear regions have different degrees of importance, where characteristics such as tragus, antitragus and helix contour play the most important role in ear form. The algorithm applied is based on what is stated in [13]. It operates basically in the comparison of edge maps. The advantage of using edges to match two objects, is that this representation is robust to illumination change. In this stage we also compute the SURF features to track the ear in the video.

4.2 Tracking the Ear

Speeded Up Robust Features (SURF)[9] is a scale and rotation invariant interest point detector and descriptor. It has been designed for extracting highly distinctive and invariant feature points from images. One of the basic reasons to use SURF for the feature representation is to analyse how the distinctive characteristics works in images, and at the same time is to found more robust with respect to change, taking into account the point of view, rotation, scale, illumination and occlusion [9] as compared to other scale and rotation invariant shape descriptors such as SIFT [6] and GLOH [12]. The result for the feature vectors SURF is the relative measured to the dominant orientation to generate each vector that represent an invariant with respect to rotation of the image.

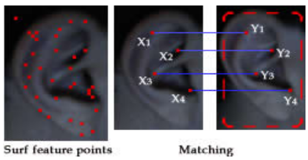


Fig. 3. Example of SURF Features

The way SURF process pairing is using the most proximate neighbour ratio pairing. To get the greatest pairing match for a key-point of a picture inside in another picture is elucidated by detecting the most proximate neighbour in the other key-points from a second picture where the most proximate neighbour is defined as the key-point with the least euclidean distance from the known key-point of the first picture between their characteristic unidirectional matrices. Due to the fact that these SURF vectors are invariant to the image rotation, the process of ear detection combining the previous viola-jones approach with the SURF vectors becomes robust and efficient.

Fig. 4. Tracking Ear using SURF Features

The approach to isolate the ear in the image, the prototype we used for the ear identification should reveal the characteristics of scale and rotation immutability. To calculate such prototype in a suggested method, an invariant shape characteristic to rotation and scale was used. Among numerous scale and rotation invariant shape characteristics, SURF [11] offers respectable distinctive features and at the same time it is robust to variations in viewing circumstances, rotations and scales. SURF denotes a picture by first detecting some exclusive feature points in it and then by describing them with the support of a unidirectional feature descriptor matrix.

5 Ear Recognition Using a Neural Network

The ear image is recreated through the SURF algorithm as a set of salient points, where each on is associated with a vector descriptor. Each can be of 64 or 128 dimensions. The 128 dimensional descriptor vector is considered the more exacting feature based in the knowledge that is always best to represent the image with the most powerful discriminative features possible. A method to obtain a unique characteristic fusion of one sole individual is proposed by combining characteristics acquired from various training instances of the individual. If we have n ear images of an individual for training, a fused prototype is gained by fusing the feature descriptor array of all training images collected, considering the redundant descriptor array only once.

Having all the images processed, a collection was made with tags indicating to whom each image and fusion vector, belongs. These vectors are used as inputs to train the network. In the training algorithm, the unidirectional matrices or vectors of values belonging to an individual, are taken as positive returning 1 as the neuron output assigned to that user and 0 to other neurons.

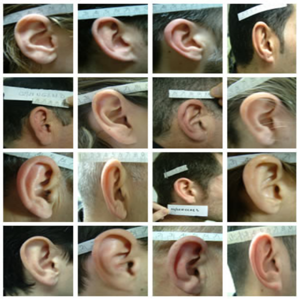


Fig. 5. Avila's Police School Database

6 Experimental Results

The results obtained in the process of detection and recognition of the ear are presented in this section, Table 1 shows the percentages of accuracy when only using the Viola-Jones classifier included in OpenCV vs the potentiation accomplished by adding the tracking with SURF features. It can be seen that in 2D images or photographs the difference is not so evident, however when the process is done on video, the difference is almost 10 percentage points, and is only done when considering the location of the ear in the video in different pose and lighting conditions. If you take into consideration the time it succeeds in maintaining the object identified, the algorithm combined with SURF tracking is much more accurate because these features allow you to place the image even if it has a 180 degrees event that does not happen with ear.

Table 1. Ear Detection (Haar-Cascade and adding SURF Tracking)

	#Images	*EarLocalization(%)*	
		Haar − Cascade	*With SURF tracking*
2D Images	308	92.53	98.70
Real Time Video	308	86.69	95.13

In Table 2 we can observe the results of the recognition process in normal conditions with controlled lighting. At this stage we have compared the results obtained with traditional algorithms such as PCA and Fisher to check the validity of our work. In this sense the results are encouraging, using SURF features as input of a neural network with different test subjects, we get a recognition percentage higher than traditional algorithms in video, however, the precision decreases to 14% being always greater in a variable range from 10 to 20% compared to the PCA and Fisher when we change the normal conditions.

Summarizing with perspective and illumination in normal conditions, we get 86% of succeed in recognition with PCA, 87% with fisher algorithm, using the neural network with SURF descriptors, the percentage increased to 92%, over more than 300 attempts of different individuals.

Table 2. Normal Conditions

	PCA		*Fisher*		*NeuralNetwork*	
	Positive	*Negative*	*Positive*	*Negative*	*Positive*	*Negative*
Positive	131	49	197	37	269	11
Negative	41	118	38	41	23	107

The parameter settings of the neural network used in this method are dynamic, the output neurons depends on Hausdorff Distance filter stage where the algorithm selects some possible answers to the recognition problem in order to reduce the amount of candidates to solve the problem. The hidden layer is created dynamically, respecting that the number of hidden neurons should be between the size of the input layer and the size of the output layer, should be 2/3 the size of the input layer, plus the size of the output layer; and less than twice the size of the input layer based on the research of Jeff Heaton [8].

7 Conclusion and Future Work

The algorithms perform a good ear recognition process if the video captures an image very similar with one in the training set. The Neural network using SURF Descriptor as Input appears to be better over variation in lighting. The neural network makes a better performance than the PCA traditional method over changes on illumination and perspective. Changes in pre-processing process allows better results if all images have the same angle and illumination, other techniques of pre-processing images may improve the ear recognition process. If these techniques allow recognize a person through the ear, exist other methods like Ray Image Transform, Histograms of Categorized Shapes, Edge orientation pattern that can obtain better results.

As future work, the most interesting and useful for the police is to achieve the development of an application not only able to propose candidates from the image of an ear but to achieve the identification and recognition of a criminal using an ear otogram found at a crime scene. The ear otogram is nothing more than a photo print of the ear identical to that obtained from fingerprints. Criminals sometimes place their ear on the door and from there they the police can get the ear photo print. The results of this research are pointing towards that goal, although preliminary, they show a significant progress to approach the final purpose, recognition based on these otograms.

References

1. Kumar, A., Hanmandlu, M., Kuldeep, M., Gupta, H.M.: Automatic ear detection for online biometric applications, pp. 146–149 (2011)
2. Cummings, A., Nixon, M., Carter, J.: A novel ray analogy for enrolment of ear biometrics. In: BTAS 2010, pp. 1–6 (2010)
3. Victor, B., Bowyer, K., Sarkar, S.: An evaluation of face and ear biometrics. In: ICPR 2002, vol. 1, pp. 429–432 (2002)
4. Carreira-Perpinan, M.A.: Compression neural networks for feature extraction: Application to human recognition from ear images. Master's thesis, Faculty of Informatics, Technical University of Madrid, Spain (1995)
5. Castrillón-Santana, M., Lorenzo-Navarro, J., Hernández-Sosa, D.: An Study on Ear Detection and Its Applications to Face Detection, pp. 313–322 (2011)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
7. Hurley, D.J., Nixon, M.S., Carter, J.N.: Force field feature extraction for ear biometrics. Computer Vision and Image Understanding, 491–512 (2005)
8. Heaton, J.: Introduction to Neural Networks for C#, 2nd edn. (2010), `http://www.heatonresearch.com`
9. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Computer Vision and Image Understanding, 346–359 (2008)
10. Jolliffe, I.T.: Principal Components Analysis, 2nd edn., New York (2002)
11. Bustard, J., Nixon, M.: 3D morphable model construction for robust ear and face recognition. In: CVPR 2010, pp. 2582–2589 (2010)
12. Krystian, M., Cordelia, S.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1615–1630 (2005)
13. Lin, K.-H., Lam, K.-M., Siu, W.-C.: Spatially eigen-weighted Hausdorff distances for human face recognition. Polytechnic University, Hong Kong (2002)
14. Kyong, C., Kevin, W.B., Sudeep, S., Barnabas, V.: Comparison and combination of ear and face images in appearance-based biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1160–1165 (2003)
15. Alvarez, L., Gonzalez, E., Mazorra, L.: Fitting ear contour using an ovoid model. In: Proceedings of IEEE International Carnahan Conference on Security Technology (ICCST 2005), pp. 145–148 (2005)
16. Yuan, L., Mu, Z.-C.: Ear detection based on skin-color and contour information. In: ICMLC 2007, vol. 4, pp. 2213–2217 (2007)
17. Burge, M., Burger, W.: Ear biometrics in computer vision. In: ICPR 2000, vol. 2, pp. 822–826 (2000)
18. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
19. Yan, P., Bowyer, K.W.: Empirical evaluation of advanced ear bio-metrics. In: Conference on Computer Vision and Pattern Recognition, vol. 3, pp. 41–48 (2005)
20. Ansari, S., Gupta, P.: Localization of ear using outer helix curve of the ear. In: Proceedings of the International Conference on Computing: Theory and Applications (ICCTA 2007), pp. 688–692 (2007)
21. Attarchi, S., Faez, K., Rafiei, A.: A new segmentation approach for ear recognition. In: Blanc-Talon, J., Bourennane, S., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2008. LNCS, vol. 5259, pp. 1030–1037. Springer, Heidelberg (2008)
22. Lankton, S., Tannenbaum, A.: Localizing region-based active contours. IEEE Transactions on Image Processing 17(11), 2029–2039 (2008)
23. USTB Database, University of Science and Technology Beijing

Integration of Mobile Robot Navigation on a Control Kernel Middleware Based System

Eduardo Munera Sánchez, Manuel Muñoz Alcobendas,
Juan Luis Posadas Yagüe, Jose-Luis Poza-Luján,
and J. Francisco Blanes Noguera

Institute of Control Systems and Industrial Computing
Polytechnic City of Innovation
Polytechnic University of Valencia, Spain
{emunera,mmunoz,pblanes}@ai2.upv.es, {jposadas,jopolu}@disca.upv.es
www.ai2.upv.es

Abstract. This paper introduces how mobile robots can perform navigation tasks by implementing a system based on the control kernel middleware (CKM), and how can take benefit of this. Smart resources are also included into the topology of the system, improving the distribution of the computational load required by the system tasks. The CKM and the smart resources are both highly reconfigurable, even on execution time, and they also implement fault detection mechanisms and Quality of Service (QoS) policies. By combining of these capabilities, the system can be dynamically adapted to the requirements of its tasks. Furthermore, this solution is designed to be implemented by almost every type of robot. The distribution of load make this system suitable even for those configurations which are provided with a low computational power. All these benefits are improved by exploiting the smart resources capabilities, and the dynamic performance of the system.

Keywords: Distributed Control Systems, Control Kernel, Robot Navigation, Limited Resources Management, Embedded Systems.

1 Introduction

A navigation system is a must for every kind of robot which has to perform in an autonomous way and deal with an uncertain dynamic environments [14]. Although navigation is a well known topic, it is always associated to a high computational load in comparison with other tasks. Thus many researches have been focused on how to deal with this load, or the way to reduce it. In every case, computational capabilities of the robot have to be designed for being able to face its execution in a proper way.

Besides, navigation system also implies a strong requirement of data acquisition. Even more, these data may be particularly complex in those cases in which visual information is used[4]. Therefore, the type of sensor, the reliability of the provided data, and its supplying rate will affect on the performance of the navigation system. So, should be considered a proper acquisition and management of the perceptual data, which is required for nourishing the navigation system.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_55,

Finally, a middleware-based implementation will improve the performance and the reliability of the system. It also offers the possibility of working with high level abstraction, and produce portable and reusable code. Therefore, this implementations provide a great support for developing robot architectures [6].

1.1 Related Works

There are many middleware solutions focused on how to deal with sensor management(data acquisition) and navigation system. One of the more used framework in robotics is Robot Operating System (ROS) [13], which offers high level capabilities. ROS is known to work properly with collaborative robot networks, and for improving in many aspects of communications between robots and data management. But shows a lack of generality on low level robot configuration, and does not provide a real-time core.

Another example is Open RObot COntrol Software (OROCOS) [3] where main features are compiled in two libraries: one for kinematics and dynamics, and other for bayesian filtering methods. It is distinguished for offering hard real-time control, data exchange tools and event-driven services. However it has a lack of capabilities on behavior management for mobile robot operations. Is usually extended by frameworks like Robot Constrution Toolkit (Rock) [1].

It also can be introduced the middleware Yet Another Robotic Platfform (YARP) [5]. YARP offers a set of libraries and tools for establishing a decoupled configuration of the robotic platform. For this purpose devices are isolated in a similar way that is done in the architecture proposed in this work. But YARP excludes the control system management, which relies on an underlying operating system.

Some robotic-specific frameworks offers more concrete capabilities, such as CArnegie MEllon Navigation (CARMEN) [7], which is focused on navigation. It offers a full support for navigation tasks like sensor control, obstacle avoidance, localization or path planning. Despite of this, it disregards low level control, behaviour, or real-time management. Instead, there can be found behaviour-specific framework for robotic platforms just as Integrated Behaviour-Based Control (IB2C) [12], used for generation, fusion, and management of behaviours.

As a conclusion, there is no framework with full support for the navigation process, ranging from the lowest level real-time system, to the highest behaviour management. That support is need in order to adapt of the requirements of the navigation process and the behaviour tasks. This adaptability is bounded, in every case, by the reconfiguration capability offered by the system devices.

1.2 Outline

This paper is structured as follows: Section 2 shows a brief description of the structure of the used control kernel middleware (CKM) and the integration of smart resources into the CKM topology. The main contribution is introduced along section 3, introducing the advantages of using the CKM and the smart resources as the support of the navigation method. The paper ends with some conclusions about the work in section 4, and the future lines are collected in section.

2 Framework

In this section is depicted the current implementation of the CKM evolved form the proposal described in [2]. The CKM is responsible of core tasks, and offers mechanisms to support the navigation process.

2.1 Control Kernel Middleware

This topology, as is is shown on Fig. 1, is characterized as a distributed control system such as is defined in [8]. Main elements of this system are:

- Full Middleware (FMW): A full version of the CKM. Implements full services support.
- Tiny Middleware (TMW): This reduced version of the CKM implements basic control and communication services. A detailed description of tiny and full midddleware can be found on [8].
- Physical sensors and actuator: Physical elements connected with a low profile TMW implementation for data management and communication.
- TCP & RS-485 connections: Provide communication capabilities to the system. RS-485 is used for control data, while the TCP is employed for system configuration and communication between RS-485 subnetworks. Both connections must offer Quality of Service (QoS) capabilities.

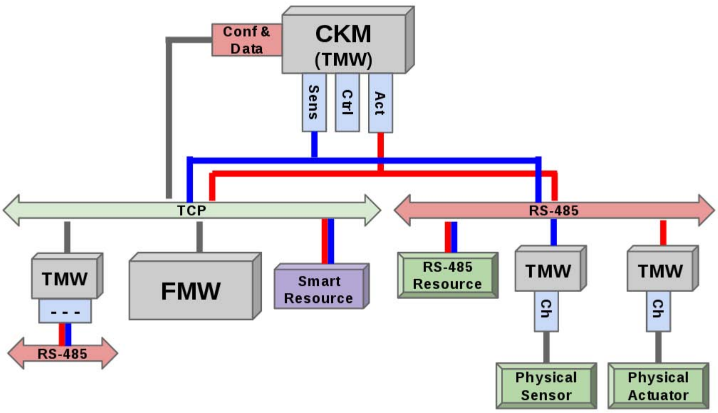


Fig. 1. Topology of a CKM based system

- RS-485 resources: Devices which can be communicated by RS-485 without implementing CKM. More devices are usually sensors or actuators.
- Smart resources: Are introduced in next section.

2.2 Smart Resources

Smart resources are devices with specific computational capabilities that offer a TCP interface in order to access to provide services. This services are usually related to sensorization or actuation tasks that works with big amount of data and requires advance processing.

In the case of navigation tasks, smart sensors will be only considered. Navigation implies the acquisition and management of several data, usually provided by different kind of sensors. The processing of all this information is a highly resource consuming task in both, memory and computational power. An smart sensor can reduce this situation by a simple TCP interface which offers preprocessed information about the environment leaving only to the CKM the data fusion step.This sensors will take profit of the QoS advantages.

3 Robot Navigation

Once the framework has been introduced, the main contribution of this work is detailed: the Integration of a CKM-based support for mobile robot navigation. In order to validate this proposal is presented a use case. In this example is detailed how this integration can improve the performance of the system by its implementation on a certain robot.

3.1 Middleware Support

The CKM is a highly suitable solution for robotic platforms. As is detailed in [9], the CKM can be used for establishing a mission-based control on several robotic platforms. Mission oriented tasks are in most cases extremely related with the navigation process, which is defined as a "non-goal oriented tasks". This paradigm allow the robot behaviours to be influenced by the navigation task needs during the fusion process.

This implementation is oriented to offer a future support for a localization method derived from the one presented in [10]. This method is characterized as a reliability-based particle filter, where a reliability factor provides a statistical computation of how accurate is the position estimation. A proper configuration of the middleware, and the use of smart sensors leads to a distribution of the computational load, and consequently a better performance of the navigation tasks.

For improving execution of the navigation algorithm, the reliability factor (R) is designed to affect the operation mode of the smart sensors. That way, sensor is dynamically reconfigured to suit the localization requirements. In eq. 1 is computed the coefficient factor (f_{mode}) according to he reliability R.

Both values must be normalized between 0 and 1, and and weighted according to the values of (w_R) and (w_{mode}). The obtained f_{mode} is used to select the operation mode which is bounded by a threshold such as is shown on eq. 2.

$$f_{mode}(t) = \frac{R.w_R + f_{mode}(t-1).w_{mode}}{w_R + w_{mode}} \tag{1}$$

$$\begin{cases} 0 \leq f_{mode}(t) < thres_1 & \rightarrow MODE = 0 \\ \vdots & \\ thres_x < f_{mode}(t) < thres_{x+1} & \rightarrow MODE = X \\ \vdots & \\ thres_n < f_{mode}(t) \leq 1 & \rightarrow MODE = N \end{cases} \tag{2}$$

3.2 Use Case

For this case of study it has been chosen a wheeled mobile robot called KertrolBot, which main characteristics can be reviewed in [15]. This platform is improved by the addition of a depth camera (a common device for robot navigation) integrated as a Smart Sensor. Smart Sensor is composed by an Asus Xtion camera is connected to a Raspberry Pi, which provides a TCP interface. According to this interface, it can be applied for some concrete information about the environment, just as information about the closest object, or distance to a certain colour object. Consequently is avoided to process raw camera data in the main CKM device, running on the core of the KertrolBot.

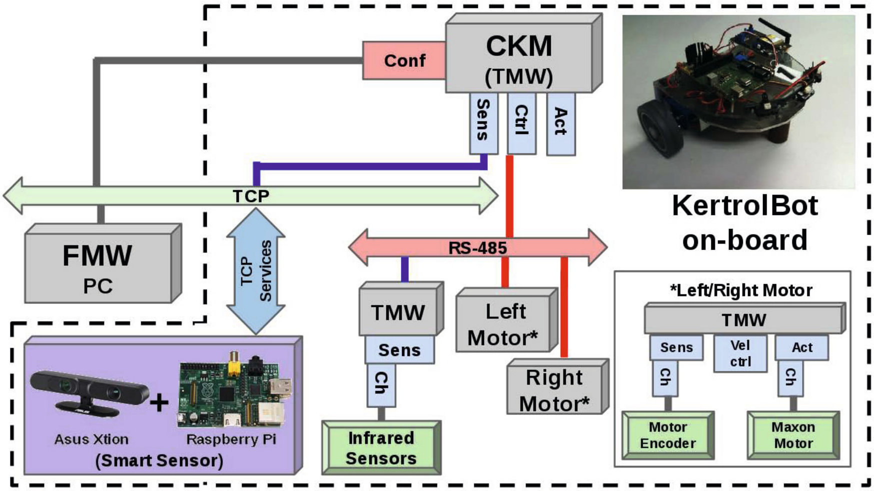


Fig. 2. Diagram of the use case

The proposed configuration is illustrated in Fig 2 where the following elements can be distinguished:

- KertrolBot on Board:
  - Core: This main unit implements a CKM verison for behaviours control.
  - Infrared Sensors (IR): Reduced CKM implementation which acquires raw data form IR sensors and offer the core unit a processed value of it.
  - Motors: Reduced CKM which interprets control signals from the core unit and executes a low level control on each wheel motor.
  - Smart Sensor: Offers high-level services about sensorial information concerning the depth camera.
- FMW: Full middleware implementation running on an external PC that manages the configuration of the system.

Main objective of this implementation is to bring the robot the capability of being localized in the environment, and performing an optimum management of the resources. This goal is achieved by the integration of the Smart Resources into the CKM topology. In this case, data acquisition is optimized by being adapted to the dynamic of the localization algorithm and its requirements. This adaptation is feasible thanks to the implementation of a CKM support. It manages the reconfiguration of system services, like the data acquisition tasks. Furthermore it offers a proper execution support for the localization process, being defined as a "non-goal oriented task".

Therefore the Smart Sensor (depth camera), as the main provider of environmental data, must be adapted to the localization performance. According to its different requirements, the camera can switch between these operation modes:

- Mode 1: Used on lost robot situation. Aims to obtain as much environmental information as is possible. For that purpose maximum resolution (VGA - 640x480) is used for a better landmarks detection, and 10 frames per second (FPS) for a longer processing between frames.
- Mode 2: Localization is not fully reliable. Offers same resolution, but improves the frame rate up to 20 FPS.
- Mode 3: Localization is reliable (most common mode). It deals with a smaller resolution (QVGA - 320x240) for improving the processing time, and 20 FPS.
- Mode 4: System reactivity is required regardless of the localization. QVGA resolution remains, but frame rate is increased to 33 FPS.

Mode switching is triggered by the value of the mode factor (f_{mode}) described on the equation 2 and according to the threshold value for each mode 1. As far as f_{mode} depends on the localization reliability factor, and its previous values, it reflects the quality of the system and its requirements.The relation between the reliability and the active mode is detailed in Fig 3.

The integration of QoS mechanisms will improve the reliability in this system. For a localization method, QoS helps to detects unexpected situations on data acquisition on Smart Sensor. This will affect to the computed reliability factor, and be reflected in the dynamic of the localization method. This variation will trigger mode switching on Smart Sensor trying to solve the acquisition problem.

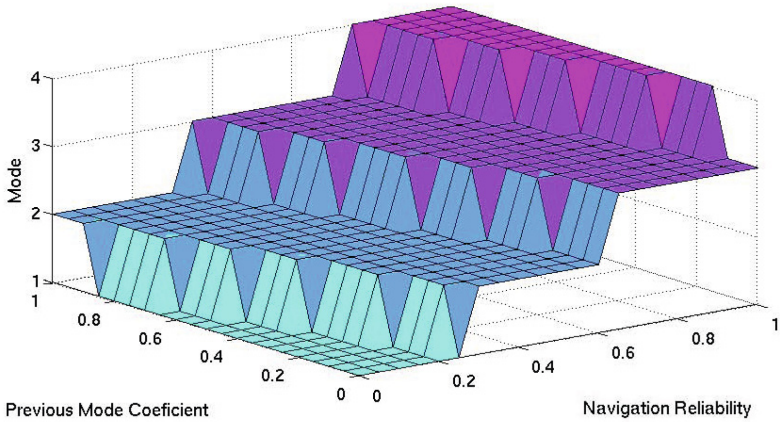


Fig. 3. Mode selection on Xtion smart sensor

4 Conclusions

According to the work previously exposed, the use of Smart Resources as a part of a CKM system helps to improve the optimization level of the localization process. A proper management of mode switching on the Smart Sensor based on the localization reliability provides a sensorial adaptation to the requirements of the localization algorithm. This integration also allow to distribute the computational load increasing the capacity of the system. Furthermore it offers a solid support for future QoS integration that will improve system reliability.

5 Future Lines of Work

As future work, it must be implemented a reliability-based particle filter, as an evolution of the one presented in[10]. This process will take profit of the support architecture here described. One of the main goals to achieve, is to proper management of the relation between the reliability factor and the mode switching on the Smart Sensor. It also must be characterized how this relation will affect the dynamic of the system, and the navigation needs. Finally, will be studied how QoS may help to detect malfunctions, and the way it can be managed in the localization [11].

Acknowledgments. This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under the CICYT project Mission Based Control (COBAMI): DPI2011-28507-C02-02.

References

1. Rock (Robot Constrution Toolkit), http://www.rock-robotics.org/
2. Albertos, P., Crespo, A., Simó, J.: Control kernel: A key concept in embedded control systems. In: 4th IFAC Symposium on Mechatronic Systems (2006)
3. Bruyninckx, H., Soetens, P., Koninckx, B.: The Real-Time Motion Control Core of the Orocos Project. In: IEEE International Conference on Robotics and Automation, pp. 2766–2771 (2003)
4. De Souza, G.N., Kak, A.C.: Vision for mobile robot navigation: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(2), 237–267 (2002)
5. Fitzpatrick, P., Metta, G., Natale, L.: Towards long-lived robot genes. Robotics and Autonomous Systems (2008)
6. Mohamed, N., Al-Jaroodi, J., Jawhar, I.: Middleware for robotics: A survey. In: 2008 IEEE Conference on Robotics, Automation and Mechatronics, pp. 736–742. IEEE (2008)
7. Montemerlo, M., Roy, N., Thrun, S.: Perspectives on standardization in mobile robot programming: The carnegie mellon navigation (carmen) toolkit. In: Proceedings of 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), vol. 3, pp. 2436–2441. IEEE (2003)
8. Muñoz, M., Munera, E., Blanes, J.F., Simo, J.E., Benet, G.: Event driven middleware for distributed system control. XXXIV Jornadas de Automatica, 8 (2013)
9. Muñoz, M., Munera, E., Blanes, J.F., Simó, J.E.: A hierarchical hybrid architecture for mission-oriented robot control. In: Armada, M.A., Sanfeliu, A., Ferre, M. (eds.) First Iberian Robotics Conference of ROBOT 2013. AISC, vol. 252, pp. 363–380. Springer, Heidelberg (2014)
10. Sánchez, E.M., Alcobendas, M.M., Noguera, J.F.B., Gilabert, G.B., Ten., J.E.S.: A reliability-based particle filter for humanoid robot self-localization in RoboCup Standard Platform League. Sensors (Basel, Switzerland) 13(11), 14954–14983 (2013)
11. Poza-Luján, J.-L., Posadas-Yagüe, J.-L., Simó-Ten, J.-E.: Relationship between Quality of Control and Quality of Service in Mobile Robot Navigation. In: Omatu, S., De Paz Santana, J.F., González, S.R., Molina, J.M., Bernardos, A.M., Rodríguez, J.M.C. (eds.) Distributed Computing and Artificial Intelligence. AISC, vol. 151, pp. 557–564. Springer, Heidelberg (2012)
12. Proetzsch, M., Luksch, T., Berns, K.: Development of complex robotic systems using the behavior-based control architecture iB2C. Robotics and Autonomous Systems 58(1), 46–67 (2010)
13. Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: Ros: An open-source robot operating system. In: ICRA Workshop on Open Source Software, vol. 3 (2009)
14. Roy, N., Burgard, W., Fox, D., Thrun, S.: Coastal navigation-mobile robot navigation with uncertainty in dynamic environments. In: Proceedings of the 1999 IEEE International Conference on Robotics and Automation, vol. 1, pp. 35–40. IEEE (1999)
15. Nicolau, V., Muñoz, M., Simó, J.: KertrolBot Platform: SiDiReLi: Distributed System with Limited Resources. Technical report, Institute of Control Systems and Industrial Computing - Polytechnic University of Valencia, Valencia, Spain (2011)

Shared Map Convolutional Neural Networks for Real-Time Mobile Image Recognition

William Raveane and María Angélica González Arrieta

Universidad de Salamanca, Salamanca, Spain

Abstract. We present a technique for improving the speed of a convolutional neural network applied to large input images through the optimization of the sliding window approach. Meaningful performance gains and memory bandwidth reduction can be obtained by processing images in this manner, factors which play a crucial role in the deployment of deep neural networks within mobile devices.

1 Introduction

The Convolutional Neural Network (CNN) [1] has become a general solution for image recognition with variable input data. CNNs consist of two stages – one for automated feature learning, and another for classification – both of which can be successfully trained in tandem through gradient descent of the error surface [9]. Its results have consistently outclassed other machine learning approaches in large scale image recognition tasks [6], outperforming even human inspection of extensive datasets [2].

Compared to other feature-based computer vision methods such as SIFT [8] or HOG [4], CNNs are much more robust and tolerant to shape and visual variations of the images or objects intended to be recognized. However, contrary to such methods, an execution of a CNN will only recognize features on a single image block of size equal to the input dimensions of the network. As CNNs are usually trained with small image patches, this recognition area is likewise small. As a result, to run image recognition over a larger image size, it is necessary to repeatedly apply the same network over multiple regions. This is a very common technique named sliding windows, albeit a time consuming one as the execution time naturally grows in proportion to the number of sampled blocks.

With the increasing use of embedded hardware, it has naturally become a priority to endow mobile devices with computer vision capabilities. The use of CNNs, when applied through a sliding window methodology, allows a large range of important image recognition tasks to be carried out, many of which would have a great impact on the everyday usage of mobile hardware by end users. Some examples of this are text recognition [10] for visual language translators, human action [5] and face [7] recognition for greater user interactivity with social applications, or even traffic sign recognition [3] for embedded automotive applications. The unique task of logo recognition is taken as a sample usage of mobile implemented CNNs in this work, something which would have large

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_56,

opportunities for commercial applications to increase company brand loyalty, perception and awareness among consumers, depending on the context it is used in. However, the same methods and network architecture described here would be equally applicable to solving similar problems, such as those described above.

Due to the high computational requirements of a CNN, the need for mobile computer vision has traditionally been met by outsourcing image analysis to a remote server in communication with the device over an internet connection. This approach, while effective, introduces large delays and is hardly an appropriate solution when user interactivity and real-time responsiveness are paramount. As embedded hardware capacity continues to grow with each new generation of low energy processors, this trend has gradually shifted towards implementing image recognition algorithms on the device itself with all computations carried out locally. Regardless, these devices continue to display performance limitations, as well as having intrinsic architecture constraints which result in slow memory access. It is therefore important to find new possible optimizations, so as to better utilize the computational power of the device.

We introduce one such improvement applicable to CNNs in particular. Our process contrasts with the traditional sliding window method where overlapping patches of a large image are sequentially analyzed by the network over individual executions. Instead, we apply per-layer computations over the entire image space, thereby producing a continuous flow of information from input to output in a single execution of the CNN. This results in a substantial boost in the throughput of the algorithm, especially when executed in an embedded mobile environment. Although applicable to most other platforms, we discuss the deployment of this algorithm on a mobile device as its memory access architecture makes it particularly sensitive to the improvements of the proposed algorithm.

This paper addresses these issues in particular. In Section 2, some background work is reviewed detailing the functionality of CNNs and sliding windows in general. We then introduce in Section 3 an optimized approach for the techniques discussed herein, including the architecture constraints that must be made to implement the proposed system. Finally, Section 4 concludes with the results obtained from the optimized method when compared to the traditional sequential algorithm and visits possible future enhancements to the presented work.

2 Background

The network on which our system is based upon is a standard CNN composed of alternating convolutional and max-pooling layers for the feature extraction stage, and one or more linear layers for the final classification stage. Fig. 1 depicts the layer structure of such a network, and it is the reference architecture used here to describe the concepts of the framework presented.

The first layer in the network consists of one or more neurons containing the image data to be analyzed, usually composed of the three color channels of the incoming image.

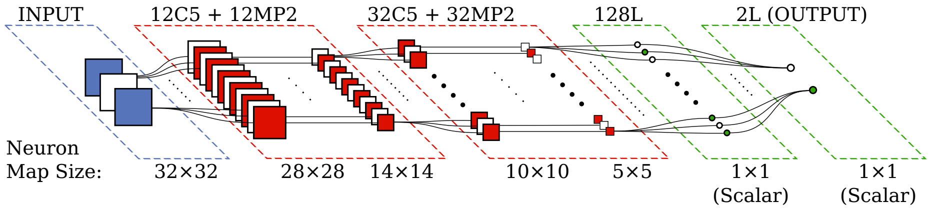


Fig. 1. A typical convolutional neural network architecture, with three input neurons for each color channel of an analyzed image patch, two feature extraction stages of convolutional and max-pooling layers, and two linear layers to produce a final one-vs-all classification output

The notation N_jXK_j is used to describe the following layers, where N_j is the neuron map count of layer j, $X \in \{C, MP, L\}$ denotes the layer type, and K_j is the primary parameter value for the neurons in that layer.

The first part of every feature extractor stage is a convolutional layer. Each neuron linearly combines the convolution of one or more maps from the preceding layer. The map of a neuron in this layer will be slightly smaller than the incoming maps by an amount referred to as the kernel padding. This padding arises from the boundary conditions of the convolution algorithm and is defined as ${}^{K_j}/_2 - 1$, where K_j is the size of the convolutional kernels of layer j. Therefore, the layer's map size will be given by $M_j = M_{j-1} - {}^{K_j}/_2 - 1$, where M_{j-1} is the the preceding layer's map size.

Every convolutional layer is paired to a max-pooling layer which primarily reduces the dimensionality of the data. A neuron in this layer acts on a single map from a corresponding convolutional neuron in the previous layer. Its task is simply to pool as many adjacent values in the map as stated by the pool size parameter, to then determine the maximum value among them, and finally to pass this value as a subsample of the pooled region. The result is a map size that is inversely proportional to said parameter as given by $M_j = {}^{M_{j-1}}/_{P_j}$, where P_j is the pooling size factor of this layer.

The linear layers classify the final feature maps extracted on the previous layers by means of a linear combination operation identical to that of a regular multi layer perceptron, working with single pixel maps for both their input and output, such that $M_j = 1$ at every layer.

Ultimately, the output of the final classification layer decides the best matching class describing the input image, according to preselected training targets. Fig. 2 shows the information flow leading to this classification for a given image patch, where the CNN has been trained to identify a particular brand logo. The output of this execution is composed of two scalar values, each one representing the likelihood that the analyzed input image belongs to that neuron's corresponding class. In this case the logo has been successfully recognized as is dictated by the higher valued output neuron for the class "Logo".

Image recognition of images larger than the input size of a CNN is implemented by the sliding window approach, and its performance is intrinsically

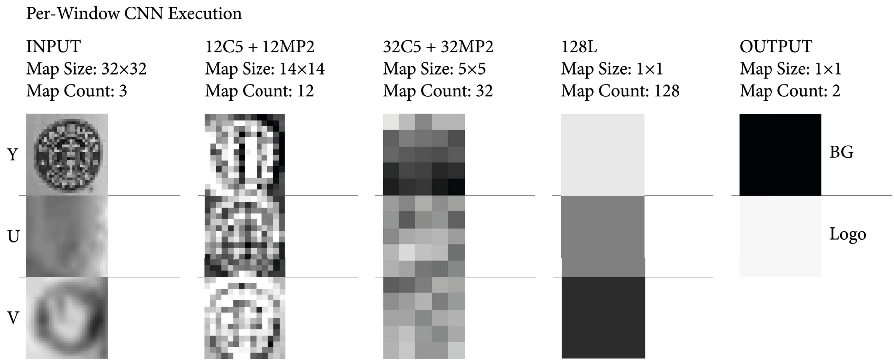


Fig. 2. A visualization of the data flow through the network, showing the first three neuron maps of each stage of the CNN. Note the data size reduction induced at each stage.

dependent on the details of this method. This algorithm is defined by two quantities, the window size S, usually fixed to match the network's designed input size; and the window stride T, which specifies the distance at which consecutive windows are spaced apart. The stride distance, therefore, inherently establishes the total number of windows analyzed for a given input. Therefore, it is important to choose the stride wisely, as this distance is inversely proportional to the resolution of the classifier, but also to the computing power invested in analyzing the number of resulting windows, W. For an input image of size $I_w \times I_h$, the total window count is given by:

$$W = \left(\frac{I_w - S}{T} + 1\right)\left(\frac{I_h - S}{T} + 1\right) \quad \Longrightarrow \quad W \propto \frac{I_w I_h}{T^2}$$

Figure 3 depicts the operation of this method applied on an input image downsampled to 144×92, extracting individual windows with $S = 32$ for the simplified case where $T = {}^{S}\!/_{2}$. A network analyzing this image would therefore require 40 executions to fully analyze the extracted window. The computational requirement is further compounded when a smaller stride is selected – an action necessary to improve the resolving power of the classifier. At $T = {}^{S}\!/_{8}$, for example, 464 separate CNN executions would be required.

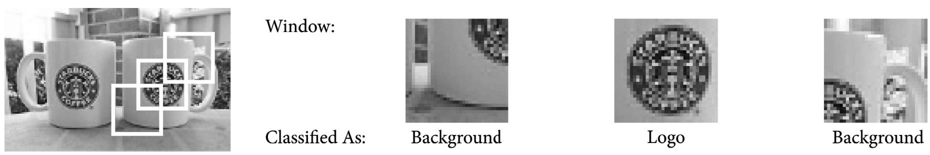


Fig. 3. An overview of the sliding window method, where an input image is subdivided into smaller overlapping image patches, each being individually analyzed by a CNN

3 Optimized Network Execution

The method we propose introduces a framework wherein the chosen stride has no significant impact at all on the execution time of the feature extraction stages of a CNN – the most computationally demanding section of the network – as long as the selected stride is among a constrained set of possible values. This is made possible by two considerations. (i) Allowing each layer of the CNN to process the entire image at once as a single shared map instead of individually processing single windows. (ii) Guiding the image data through the network so as to perfectly align individual pixels over computational columns spanning the various network layers, by virtue of the imposed stride constraints.

By their nature, convolutional neural networks are designed with built-in positional tolerance. This is in part achieved by the reuse of the same convolutional kernel over the whole neuron map. Similarly, a max-pooling neuron performs the same singular algorithmic action at any point within its map. As a result of this behavior, the output of these layers is independent of the pixel offset within the map, such that overlapping windows will share the same convolved values. This is demonstrated in Fig. 4.

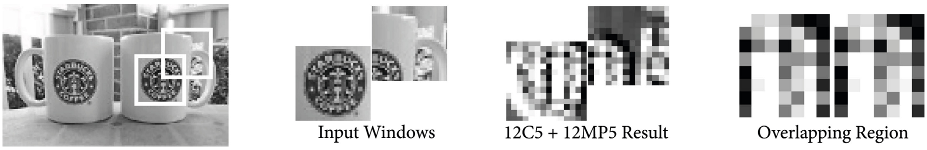


Fig. 4. Two adjacent windows extracted from an input image, passed through the 12C5 + 12MP5 feature extractor. A detailed view of the convolved maps in the overlapping top-right and bottom-left quarters of each window show that these areas match exactly.

This leads to the possibility of streamlining the feature extractors by running their algorithms over the full input image at once. Hence, each C + MP neuron will output a single map shared among all windows, where subdivisions of this map would normally match the outputs of the corresponding windows, had they been executed separately as in the traditional method. This greatly reduces the expense of re-calculating convolutions on overlapping regions of each window.

Figure 5 shows an overview of the shared map process, which passes the input image in its entirety through each stage of the network. By doing this, the output layer now produces a continuous and localized class distribution over the image space, a result which contrasts greatly to that of a single classification value as was previously seen in Fig. 2. The output of this execution consists of image maps where each pixel yields the relative position of all simultaneously classified windows. Similar to the per-window execution method, the intensity value of a pixel in the output map represents the classification likelihood of the corresponding window. Note how the relative position of both logos in the input image has been discovered after only one shared map execution of the network.

An account of the window size and stride is also displayed, illustrating how it evolves after each layer, while the total window count remains the same. Here, the correspondence of each 32×32 window in the input image can be traced to each 1×1 pixel in the output maps.

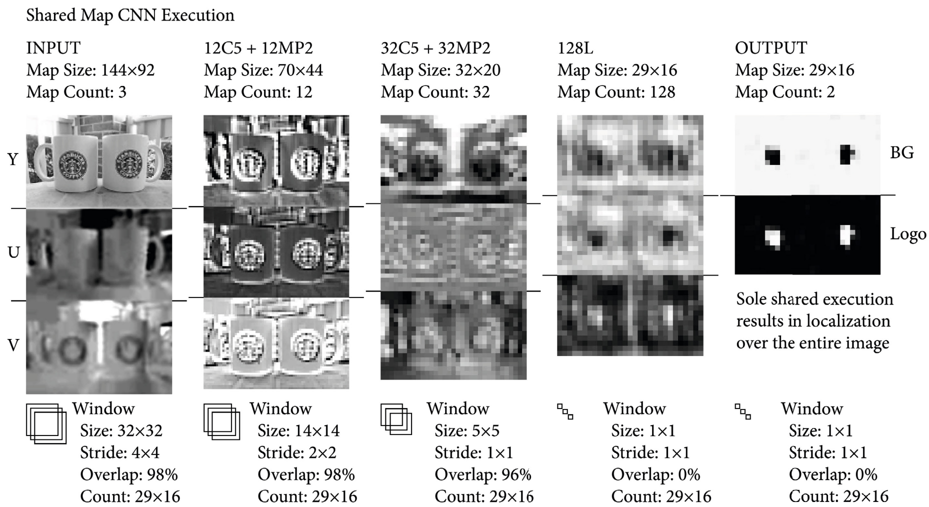


Fig. 5. The shared map execution method for a convolutional neural network, where each layer processes an entire image in a single pass, where each neuron is now able to process maps with dimensions that far exceed the layer's designed input size.

The operation of the shared map process relies heavily on the details of the dimensionality reduction occurring at each layer within the network. For this reason, it is necessary to lay certain constraints that must be enforced when choosing the optimum sliding window stride. The window size and stride at each layer is affected by the parameters of the layer in a well-defined manner:

$$S_j = \begin{cases} S_{j-1} - K_j - 1 & if\ j \in Convolutional\ Layers \\ S_{j-1}/P_j & if\ j \in Max\ Pooling\ Layers \\ S_{j-1} & if j \in Linear\ Layers \end{cases}$$

$$T_j = \begin{cases} T_{j-1} & if\ j \in Convolutional\ Layers \\ T_{j-1}/P_j & if\ j \in Max\ Pooling\ Layers \\ T_{j-1} & if\ j \in Linear\ Layers \end{cases}$$

Where the size S_j and stride T_j of a window at layer j depends on the parameters of the layer and the size and stride values at the preceding $j-1$ layer.

This equation set can be applied over the total number of layers L of the network, while keeping as the target constraint that the final size and stride must remain whole integer values. By regressing these calculations back to the input layer $j = 0$, one can find that the single constraint at that layer is given by:

$$T_0 \quad | \quad \prod_{j=1}^{j=L} \begin{cases} P_j & if\ j \in Max\ Pooling\ Layers \\ 1 & otherwise \end{cases}$$

In other words, the input window stride must be divisible by the product of the pooling size of all max-pooling layers in the network.

Choosing the initial window stride in this manner, will ensure that every pixel in the final output map corresponds to exactly one input window.

4 Results and Conclusions

Table 1 gives a summary of the results of this technique. The tests were carried out on a mobile device equipped with a quad core 1.3 GHz Cortex-A9 CPU and a 12-core Tegra 3 GPU, running a parallel optimized GPU implementation of the CNN architecture in Fig. 1 over a large 512×512 input image.

Table 1. Results of tests with several input layer stride T_0 configurations, from the closest packed 4×4 to the non-overlapping 32×32 layouts. A window count W and the overlap coverage O is shown for each window stride selection. An average over 20 test runs for each of these configurations was taken as the execution time for each of the methods described herein – the traditional per-window execution method, and our shared map technique.

$\mathbf{T_0}$	**W**	**O**	**Execution Time (ms)**		**Speedup**
			Per-Window	Shared Map	
4×4	14,400	98.4%	115,212	4,087	28.1
8×8	3,600	93.8%	28,847	1,548	18.6
12×12	1,600	85.9%	12,789	1,025	12.5
16×16	900	75.0%	7,234	843	8.5
20×20	576	60.9%	4,608	757	6.1
24×24	400	43.8%	3,189	686	4.7
28×28	289	23.4%	2,312	621	3.7
32×32	225	0.00%	1,801	599	3.0

It is of great interest to note the final 32×32 configuration. Regardless of the fact that there is no overlap at this stride, a 3.0 speedup is observed over running the windows individually. This is due to our method introducing inherent memory bandwidth reductions through its pipelined execution approach, where the entire image needs to be loaded once per execution. This contrasts the traditional approach where loading separate windows into memory at different times requires each to be individually sliced from the original memory block – a very expensive operation in the limited memory throughput of mobile devices.

Acknowledgements. This work has been carried out by the project Sociedades Humano-Agente: Inmersión, Adaptación y Simulación. TIN2012-36586-C03-03. Ministerio de Economa y Competitividad (Spain). Project co-financed with FEDER funds.

References

1. Ciresan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J.: Flexible, High Performance Convolutional Neural Networks for Image Classication. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, pp. 1237–1242 (2011)
2. Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Convolutional Neural Network Committees For Handwritten Character Classification. In: 11th International Conference on Document Analysis and Recognition, ICDAR (2011)
3. Ciresan, D.C., Meier, U., Masci, J., Schmidhuber, J.: Multi-Column Deep Neural Network for Traffic Sign Classification. Neural Networks (32), 333–338 (2012)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893 (June 2005)
5. Ji, S., Xu, W., Yang, M., Yu, K.: 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 221–231 (2013)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks, vol. 25, pp. 1106–1114 (2012)
7. Lawrence, S., Giles, C., Tsoi, A.C., Back, A.: Face recognition: a convolutional neural-network approach. IEEE Transactions on Neural Networks 8(1), 98–113 (1997)
8. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2006, vol. 2, pp. 2169–2178 (2006)
9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE (November 1998)
10. Wang, T., Wu, D.J., Coates, A., Ng, A.Y.: End-to-end text recognition with convolutional neural networks. In: ICPR, pp. 3304–3308. IEEE

Using Multi-Objective Optimization to Design Parameters in Electro-Discharge Machining by Wire

Alberto Ochoa[1], Lourdes Margain[2], Julio Arreola[1], Guadalupe Gutiérrez[2], Geovani García[1], and Fernando Maldonado[1]

[1] Juarez City University, México
[2] Universdad Politécnica de Aguascalientes, México

Abstract. The following paper describes the main objective to follow the methodology used and proposed to obtain the optimal values of WEDM process operation on the machine Robofil 310 by robust parameter design (RPD) of Dr. G. Taguichi [1], through controllable factors which result in more inferences regarding the problem to noise signal (S / N), which for this study is the variability of the hardness of samples from 6061, also studied the behavior of the output parameters as the material removal rate (MRR) and surface roughness (Ra), subsequently took the RPD orthogonal array and characterized the individuals in the population, each optimal value is a gene and each possible solution is a chromosome, used multi-objective optimization using Non-dominated Sorting Genetic Algorithm to cross and mutate this population to generate better results MRR and Ra.

Keywords: Multi-objective optimization, Genetic Algorithms and Robust Design.

1 Introduction

WDEM metallurgical process, which erodes the material of the workpiece by a series of sparks area made between the workpiece and the wire (electrode) separated by a thin film of fluid or liquid dielectric supply. The wire cuts like a band saw, but instead of teeth sparks made the cut, creating a path for each download, which ionizes the fluid in the slot area rises to extremely high temperatures, so casting material is removed. The particles which are generated by removal of material are segregated in the dielectric supply and its flow through are filtered and remote from the system.

2 Design of Experiments (DOE)

In the early experimentation man's need to control the areas that directly affect their lives routed efforts toward deeper understanding of the DOE where they took great minds such was the case and the legacy he left Sir Ronald Fisher was a great researcher using the basis of statistical methods , innovated design and development experiments. The development of testing methods is strengthened along with the

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_57,

technologies in environments determined by the conditions and limitations of companies or countries, this leads to complex methods of experimentation. Today modern methods of experimental design used in all areas including industrial, commercial and financial. You can understand how the experiment designed where deliberate changes are induced in the input variables of a process, so that may be possible to identify the causes of changes in output responses. One of the tools is the full factorial DOE design that uses a series of experiments investigating one factor at a time, has two advantages. The first is that the factorial experiments are much more efficient to estimate the main effects and the second reason is that the interaction between the factors can be evaluated in a factorial experiment. Fractional Factorial is defined as a fraction of the full factorial design , in other words , are performed only part of the full factorial experiments , this follows from the need for companies to minimize costs of research and time that you spend on this. Ensuring a good job that supports the quality the RPD is the part of the methodology of Taguchi [1] involving designs of fractional treatments where the choice of factors and levels of each factor are crucial aspects of the design of the experiment and will be dictated by the knowledge matter and time constraints or cost experiment. In the terminology of Taguchi [1] all factors involved in the RPD. The types of factors in the study, beginning with the signal factor and the quality characteristic or factors often named output Moreover the controllable factors are called parameters, which are not controlled are called noise factors. The noise variables or factors are most sensitive to changes in environmental conditions during production and, therefore, transmits responses to the variability in the process of interest. The interaction of factors is important to determine how the findings of the experiment can be applied generally. Factor systems are often complex, especially if a number of factors to be analyzed is large. The concept of robust design experimentation identification represents the shape of the noise factors that are not controllable, considered explicitly or implicitly, where its effect is to minimize disappear or indirectly, i.e. without directly control levels factor can be controlled, to find the right combination of these factors effect of noise is minimized [4]. An elementary lens of robust design is the combination of levels of the controllable factors, in which no noise factors affect the process, although these are not controlled.

3 Methodology

The methodology proposed by Dr. Taguchi [1] for robust parameter design (RPD) is a series of steps in which are contained the phases of the design of experiments, in order to strengthen the WEDM process [2] to the noise generated by the variability of toughness throughout the entire material. In this research is analyzed unconventional WEDM process in cutting 6061 specification aluminum specimens in a machine Robofil 310. Controllable factors have been identified and noise conditions, which affect Ra documentary research and MRR. It strengthens the process by applying Taguchi [1] methodology. In a second stage using genetic algorithms, which is characterized according to the orthogonal array of individuals RPD, population, genes and chromosomes, based on the fact that two objectives are analyzed becomes a multiobjective optimization study. Which aims to maximize the MRR and minimize Ra.

3.1 Parameters and Factors de WEDM

The machine Robofil 310 works by technology, in other words, the ideal values assigned according to the characteristics of the material to be processed, so that when programming a new technology for 6061 is thrown ideal values, which are composed of two forms of fixed parameters remain constant throughout the experiment and the controllable parameters which research suggests three principles: 1. Pulse ON, 2. Pulse Off, and 3. Wire feed, which will be subsequently chromosome genes. Another important factor is the noise, so that a material meets the specifications and rigorously including hardness, certification is required to ensure that the material actually meets what is stated in the certificate. The following tables (1 and 2) are observed values of each parameter and operating levels used for the RPD, and the noise factor and Rockwell Hardness values "B". [5]

Table 1. Levels of the Parameters

Parameter	**Level 1 (Low)**	**Level 2 (high)**	**Optimal**	**Unit**
Pulse ON-time A	**0.6**	**1.2**	1.0	μs
Pulse Off-time B	**6.0**	**10.0**	8.0	μs
Wire speed WS	**6.0**	**12.0**	9.0	m/min

Table 2. Levels Noise Factor

Noise factor (Hardness)	**Value (HRB)**	**Confidence Interval (HRB)**
Level 1	52.96	52.54, 53.37
Level 2	55.21	54.71, 55.71

The initial experiment established each factor in only two levels, so that significant main effects and interactions can be quickly identified and explored further.

Pulse ON-time (A). This parameter is the duration of the spark erosion rates which are primarily affected by the pulse parameters. The spark is a bridge, the current is generated and removal work is accomplished. The longer the spark is maintained, the greater the removal of material. Therefore, the resultant craters are larger and deeper, and the surface is just rougher.

Pulse Off-time (B). Against part of the pulse On-time is the duration of this pulse and resting place when ionized again the die Electric can affect the speed of operation in general, the longer the off-time is longer the time machining. But this is an integral part of the EDM process and must exist. The Off-time also regulates the stability of the process.

Wire Feed Speed (WS). This parameter enters the study due to the strong interaction we have with the two pulses, the parameter itself regulates the feed rate of the thread through a nozzle disposed servos.

Orthogonal Array. The orthogonal arrangement of two noise levels for Robust Parameter Design is of the following form for consideration (see table 3).

Orthogonal Array $\boldsymbol{L_4 = 2^3}$

Table 3. Orthogonal Array

Block	A	B	WS	Noise	Noise
1	1	1	1	Low level	High level
2	1	2	2		
3	2	1	2		
4	2	2	1		

3.2 Implement Results RPD

Output Variables MRR y Ra

After performing the experimental results are analyzed, the following table 4 shows the data obtained from each block, taking the MRR and Ra as output variables or characteristics that we intend to optimize quality, we used the statistical tool Minitab 16 in order to study their behavior.

Table 4. Results of MRR and Ra

Block	Noise Level	MRR $(\frac{in^2}{hr})$	Machining Time(min)	Roughness (μin)
1	1	3.765	15.26	124.16
2	1	7.294	9.50	135.33
3	1	4.308	13.56	128.66
4	1	9.220	6.40	157.33
5	2	3.609	15.42	124.16
6	2	6.254	10.00	135.33
7	2	4.072	14.20	128.66
8	2	8.923	6.52	157.33

The behavior of the quality characteristics or output variables MRR and Ra, is determined that seeks to maximize MRR and Ra seeks to minimize. The following table 5 identifies the type of property and signal / noise ratio for each response variable.

Table 5. Noise Signal

Response Variable	Characteristics	Signal/Noise (S/N)
Ra	While smaller is better	$-10\ \log\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^2\right]$
MRR	While bigger is better	$-10\ \log\left[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{Y_i^2}\right]$

Select Values to Strengthen

After choosing the factors and their levels that most impacted the quality characteristics is obtained optimal values. Implementing predicted levels and performing test runs the following gains are obtained.

• In the case that the process needs a roughing treatment maximizing material removal rate, the expected gain in this design is 19.1501 S/N ratio.

• In the case that the process need a fine finish treatment, the surface roughness is minimized by applying this robust design, and you get a gain of 41.8796 S/N ratio. The optimal levels of operation selected for MRR are: Pulse ON-time (A)=2, Pulse Off-time (B)=2 and Wire speed (Ws)=1. The optimal levels of operation selected for Ra are: Pulse ON-time (A)=1,2 Pulse Off-time (B)=1 and Wire speed (Ws)=1.

The research sets the parameters at two levels of experimentation, this to identify important effects that are conducted on the response variables interact to simultaneously control the three parameters, the studied further observed:

Strong interactions are taken into controlling factors are: Length of drive (click on time) "A" and wire feed speed "Ws". Time between pulses (off press time) "B" and wire feed speed "Ws".

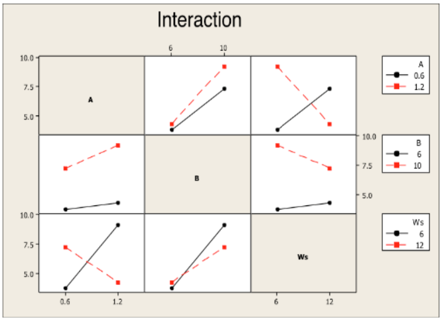


Fig. 1. Interactions for MRR

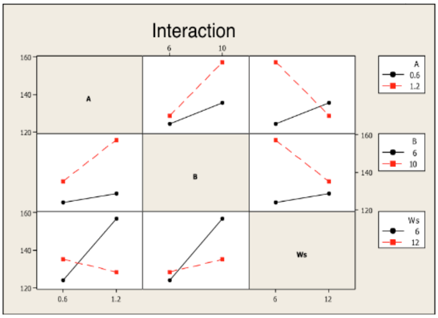


Fig. 2. Interactions for Ra

Genetic Algorithms (GA)

GA are based on a study of computer systems trying to understand and copy adaptation of processes and mechanisms in the evolution towards the improvement and monitoring of the effects of these processes and mechanisms. In other words, evolutionary computation algorithms are trying to exploit the properties of the fields related to population genetics. GA share similar to the early stages of DOE where adaptation through a process learns about beneficial change through trial and error are possible solutions of a population properties, the environment is defined by a specific objective function for each problem. In each case, the evolutionary process addresses the process of adaptation of the population, usually addressing the mechanisms of evolution are the cross, mutation and migration. Reproduction is accomplished by copying the best individuals from one generation to the next, the best solution will be improving from one generation to the next cyclically to comply with the proposed number of generations in the algorithm. The crosses were carried out with a rate of probability assigned. The genetic mutation takes into account the diversity in the population by random replacement of individuals in the population, and likewise that the crossing is done by a certain percentage of probability in the programming of the algorithm. In Addition migration of Individuals Between subpopulations is used if the island model is employed. Standard genetic programming algorithm is expressed as shown below.

The Standard GA

{
Generate initial population. Sa
Assess population Sa
While stopping criteria not satisfied repeat
{
Select elements of Sa Sa put +1
Elements crossing in Sa Sa and put +1
Mutate elements in Sa Sa and put +1
Evaluate new population Sa1
Sa = Sa +1}
}

Multi-Objective Optimization (MOO) for WEDM

In this study, the MOO is a tool that quantitatively determines the relationship between quality characteristics MRR and Ra with the optimal combination of machining parameters. Confirmation is the final step in any process of design of experiments, the purpose is to validate the conclusions drawn during the analysis in this case validate the results of the RPD. Non-dominated Sorting Genetic Algorithm (NSGA) was used to confirm the optimal machining parameters for multiple output targets by using combinations proposed by the RPD , this particular case there are two objectives MRR and Ra where the former seeks to maximize and minimize the second , ie the greater the improvement in machining performance outputs in the MRR and the minimum value is the best outlet for Ra. NSGA provides a pseudocode listing of the classification to minimize a (MRR) and maximize another function (Ra) . The two objective functions were determined by applying a linear regression on the values of the proposed orthogonal array in the RPD and the results of the output variables. The following figure shows the equations describing the two objective functions within the code of NSGA -II, the objective "0 " refers to the MRR and aims to maximize , therefore the objective "1" is the Ra and seeks to minimize.

```
#ifdef MRR Ra

void test_problem (double *xreal, double *xbin,
int **gene, double *obj, double *constr)

{
obj[0] = 7.55 + 2.33*xreal[0] + 0.996*xreal[1]
- 0.150*xreal[2] - 0.191*xreal[3];


obj[1] = (0-1)*(89.8 + 22.1*xreal[0]
+ 4.98*xreal[1] -1.46*xreal[2] - 0*xreal[3]);
    return;
}

#endif
```

Fig. 3. Objective functions

The code has been written for posix compliant operating systems and uses standard piping method provided by GNU C library. The routines should work on any unix and unix like OS having gnuplot installed and which are posix compliant.

NSGA-II: Non-dominated Sorting Genetic Algorithm – II

The NSGA-II is an evolutionary algorithm belong to the field of evolutionary computation inspired by the study process and mechanisms of biological evolution. The process of Darwinian evolution by natural selection that served to explain the variety of species and their adaptation to the environment. The mechanisms described how the evolution took place through the alteration, modification and reproduction of genes or genetic material. *Authors: Dr. Kalyanmoy [3] Deb, Sameer Agrawal, Amrit Pratap, T Meyarivan Article Title: A fast and elitist multiobjective genetic algorithm: NSGA-II. Journal: IEEE Transactions on Evolutionary Computation (IEEE-TEC) Year: 2002.* The objective of the NSGA-II algorithm to improve the fit is WEDM adaptive solutions of the population limited to a Pareto front by a set of objective functions in this case two objective functions, maximize and minimize Ra MRR. The algorithm uses an evolutionary process to substitute for evolutionary operators including selection, genetic crosses, and genetic mutation. The population is classified into a hierarchy of groups based on the population by Pareto dominance. Similarity between the members of each subgroup is evaluated in the Pareto front, and the resulting groups and similarity measures are used to promote a diverse of solutions.

Parameters of WEDM in NSGA-II

The parameters characterizing the problem are used in the RPD, with 4 real variables, three factors control for DPR are: A, B and Ws and add the factor noise hardness of the material more as a real variable. This is each individual has 4 genes that vary in ranges previously estimated by the RPD. The range or boundaries of real variables or genes on chromosome; units they are measured in the RPD settled in this part only quantifiable value in real numbers is taken. The limits on real variables of level 1 (low) are: Pulse ON-time A=0.6, Pulse Off-time B=6.0, Wire speed WS=6.0 and Wire speed WS=6.0. The limits on real variables of level 2 (high) are: Pulse ON-time A=1.2, Pulse Off-time B=10.0, Wire speed WS=12.0 and Wire speed WS=66.21. The population size should be in multiples of four because there are four real variables studied, being thus a population of 400 individuals, the number of generations is estimated by analyzing NSGA applications in different samples was estimated and found to optimal 100 ensures generation where the number of generations with no changes in any variable. In Figure 4, all relevant parameters are observed to characterize the problem in this case the WEDM process using NSGA.

The probability of cross enables iterating percentages ranging from 60% to 100% of the population may be high such that thus generate new individuals with better characteristics of solution, for this study, a 70% value for the variable cross took real. Taking into consideration the lack of diversity in the population and the random replacement of individuals the probability of mutation of real variables was estimated at 30% of the population. The value of the index for cross distribution and mutation of 15 individuals was determined for the first and 30 for the second, there is no binary variables as actual numbers are estimated in the RPD previously.

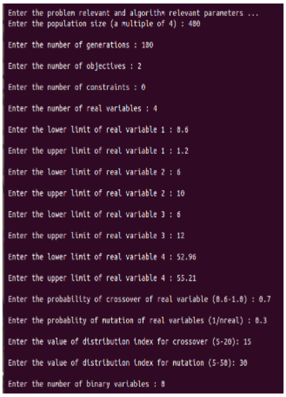


Fig. 4. Parameters of WEDM in NSGA

Implement Results NSGA-II: Objective Functions MRR and Ra

After running the NSGA-II algorithm, the best values for each variable are collected after 100 generations met the proposed algorithm, running as optimal values shown in the following table 6.

Table 6. Values of the objective functions

Variables	Objetive functions 1	Objetive functions 2
MRR ($\frac{in^2}{hr}$)	8.86	2.57
Ra (µin)	157.3	115.4
A (µs)	1.19	0.6
B (µs)	9.99	6.00
Ws (m/min)	6.00	11.99
Hardness (HRB)	55.20	55.20

It is noteworthy that the results can be read as follows: for each objective function the maximized or minimized values of each of the two quality characteristics (MRR and Ra) as a side effect are determined magnitudes are also obtained the another feature that is not objective; Also the four variables are optimized value. The first column is the quality characteristic of MRR which meets the objective to be maximized and the second column contains the minimum value of Ra feature also serving its purpose. After completing the study NSGA implemented to optimize the results, we can draw the thrown values and convert them into optimal levels of operation, the results using GA for "Levels selected for MRR" are: Pulse ON-time (A)=2, Pulse Off-time (B)=2, Wire speed (Ws)=1, Noise factor (Hardness)=2. The results for "Levels selected for Ra" are: Pulse ON-time (A)=1, Pulse Off-time (B)=1, Wire speed (Ws)=2, Noise factor (Hardness)=2.

This means that even a maximized value of MRR of magnitude 8.86 **(in^2/hr)** should be chosen operating levels in column 1, to a minimized value of Ra magnitude 115.4 (μin) is will choose the operating levels of the second column.

Experimental Confirmation

Experimental confirmation is the final step in the process of design of experiments in this study in order to validate the results developed during the analysis phase of implementing RPD. Validation of the results obtained after calculating the signal / noise generated operating levels.

Calculation of S/N

The ratio S/N is calculated for each design, starting with the original design that sheds the WEDM, then the Design of Experiments based on RPD and finally the design using multi-objective optimization to the implementation of NSGA-II. The S/N ratio is measured in decibels (db) to express variability. The higher the value of the S/N ratio, the lower the variability. The following table 7 represents the gain S/N against reducing variability.

Table 7. Gain S/N against reducing variability

Gain S/N (db)	Reducing Variability	Gain S/N (db)	Reducing Variability
0.1	1.1 %	0.2	2.3%
0.5	5.6%	1.0	10.9%
1.5	15.9%	2.0	20.6%
2.5	25.1%	3.0	29.3%
4.0	37.0%	5.0	43.9%
6.0	50.0%	9.0	64.6%
12	75.0%	15	82.3%
18	87.5%	21	91.2%
24	93.8%		

Confirmation Test

Confirmatory testing was performed in 310 Robofil machine which data quality characteristics of the material removal rate (MRR) and surface roughness (Ra) when cutting controlled implementation levels were collected optimal RPD obtained by NSGA-II and the experimental verification and making the S/N were obtained with the statistical tool of Minitab 16 and described below, it is worth mentioning that there are two different tests since a confirmation meets the objective of maximize material removal rate (MRR) and the other objective is to minimize the surface roughness (Ra). This reflects the fact that each quality characteristic has its priority as the target. The results for MRR confirmation are: Original Design (WEDM)=11.3276, Robust Design (RPD)=19.1501, Multi-objective Optimization (NSGA-II)=19.1500. The results for Ra are: Original Desing (WEDM)=-43.9362, Robust Design (RPD)=-41.8796 and Multi-objective Optimization (NSGA-II)=-41.0954.

The S/R for Robust Design (RPD) and multi-objective optimization using NSGA-II, to maximize MRR differs in magnitude, this is because as the two experiments were optimal thereof levels, so the same gain is obtained for both experiments.

The Robust Design (RPD) and Multi-Objective Optimization using NSGA-II, to minimize Ra if you have a marked difference, this will be reflected in the profit that has everyone on the original design.

Conclusions

The selection of the design of experiments (DOE) provided optimal levels WEDM machining with the correct selection and excellent methodology as is Dr. Taguchi [1] RPD generating product quality, and impact on the total cost. Were optimized and confirmed the results with Multi-objective optimization using NSGA. With the interpolation of the values of S/N and Figure 7. The variability is reduced: Robust Design (RPD)=57% and Multi-objective Optimization (NSGA-II) = 57%.

For the case, that the process needs a roughing regime maximizing material removal rate, the expected gain with this rugged design and confirmed by NSGA-II is 57%. This results in more than half of the reduced machining time yielding optimal values of operation and minimizes the cost of inputs. Robust Design (RPD) = 21% and Multi-objective Optimization (NSGA-II) = 29%.

In a Multi-objective optimization the process needs a system of fine finish, the surface roughness is minimized by applying robust design (RPD) gain on the original design by machine yielded 21% is obtained a gain of 29% on Original Design is obtained, this means that the quality characteristic Ra is further optimized using GA and NSGA-II specific. These values reflect a decreased tool wear deviations from specifications and will have a greater control of the process. Also by the RPD it was found that the whips interactions are taken into control factors are:

- Duration of discharge (pulse on time) "A" and wire feed speed "Ws".
- Time between pulses (pulse off time) "B" and wire feed speed "Ws".

GA has the ability to find the global optimal parameters, while traditional optimization techniques such as RPD normally tend to be found in local optima, this check was achieved in the study. Performance measures with the same importance to the quality characteristics and their weights, where the high value and low value MRR for Ra were equally important objectives in the WEDM process were evaluated.

References

1. Taguchi, G., Taguchi, G., Taguchi, G.: Taguchi on robust technol-ogy development: Bringing quality engineering upstream. ASME Press, Nueva York (1993)
2. Mahapatra, S.S., Patnaik, A.: Optimization of wire electrical dis-charge machining (WEDM) process parameters using Taguchi method. Journal of Advanced Manufacturing Technology (2006)
3. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A Fast and Elitist multi-objective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, IEEE-TEC (2002)
4. Kuehl, R.O.: Diseño de Experimentos: Principios Estadísticos de Diseño y análisis de investigación. International Thomson, México (2001)
5. Liao, Y.S., Huang, J.T., Su, H.C.: A study on the machining-parameters optimization of wire electrical discharge machining. Journal of Materials Processing Technology (1997)

Learning Bayesian Networks Using Probability Vectors

Sho Fukuda and Takuya Yoshihiro

Faculty of System Engineering, Wakayama University, Wakayama, Japan

Abstract. Bayesian networks are regarded as one of the essential tools to analyze causal relationship between events from data. To learn the structure of good Bayesian networks from data as quickly as possible is one of the important problems that several studies have been tried to achieve. In this paper, we propose a new algorithm to learn Bayesian network structure based on a genetic algorithm that evolves probability vectors. Through performance evaluation, we found that this probability-based approach is effective to learn better Bayesian network structure with less time.

1 Introduction

The Bayesian network is a well-known probabilistic model that represents causal relationships among events, which has been applied to so many areas such as Bioinformatics, medical analyses, document classifications, information searches, decision support, etc. Recently, due to several useful tools to construct Bayesian networks, and also due to rapid growth of computer powers, Bayesian networks became regarded as one of the promissing analytic tools that helps detailed analyses of large data in variety of important study areas.

To learn a near-optimal bayesian network structure from a target data, efficient optimization algorithm is required that searches an exponentially large solution space for near-optimal bayesian network structure, as this problem was proved to be NP-hard [1]. To find better Bayesian network structures with less time, several efficient search algorithms have been proposed so far. Cooper et al., proposed a well-known deterministic algorithm called K2 [2] that searches for near-optimal solutions by applying a constraint of the order of events. As for the general cases without the order constraint, although several approaches have been proposed so far, many of which uses genetic algorithms (GAs), which find good Bayesian network structures within a reasonable time [3][4][5]. However, because we are facing on large data, more efficient algorithms to find better Bayesian network models are expected.

In this paper, we propose a new approach to learn Bayesian networks, that is based on genetic algorithm that evolves a probability vector. Through the evaluation, we show that our new approach is efficient to find good Bayesian network structures.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_58,

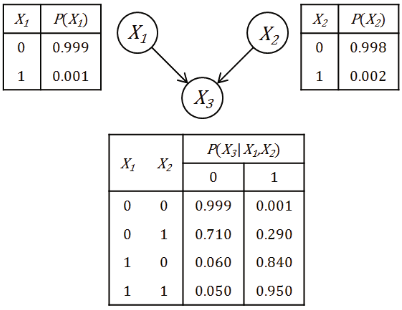


Fig. 1. A Bayesian Network Model

2 Constructing Bayesian Networks

2.1 Bayesian Network Models

A Bayesian network models and visualizes the causal relationship among events through graph representation. In a bayesian network model, events are represented by nodes while causal relationships are represented by edges. See Figure 1 for example. Nodes X_1, X_2, and X_3 represent distinct events where they take 1 if the corresponding events occur, and take 0 if the events do not occur. Edges $X_1 \to X_3$ and $X_2 \to X_3$ represent causal relationships, which mean that the probability of $X_3 = 1$ depends on events X_1 and X_2. If edge $X_1 \to X_3$ exists, we call that X_1 is a parent of X_3 and X_3 is a child of X_1. Because nodes X_1 and X_2 do not have their parents, they have own prior probabilities $P(X_1)$ and $P(X_2)$. On the other hand, because node X_3 has two parents X_1 and X_2, it has a conditional probability $P(X_3|X_1, X_2)$. In this example, the probability that X_3 occurs is 0.950 under the assumption that both X_1 and X_2 occur. Note that, from this model, Bayesian inference is possible: if X_3 is known, then the posterior probability of X_1 and X_2 can be determined, which enables us to infer events that causes the child event.

The Bayesian networks can be learned from the data obtained through the observation of events. Let $O = \{o_j\}, 1 \leq j \leq S$ be a set of observations, where S is the number of obervations. Let $o_j = (x_{j1}, x_{j2}, \ldots, x_{jN})$ be a j-th observation, which is a set of observed values x_{ji} on event X_i for all $i (1 \leq i \leq N)$, where N is the number of events. We try to learn a good Bayesian network model θ from the given set of observations. Note that the model θ should be able to explain the observation O, i.e., O should be likely to be observed under θ. As an evaluation criterion to measure the level of fitting between θ and O, we use AIC (Akaike's Information Criterion) [6], which is one of the best known criterion used in Bayesian networks. Formally, the problem of learning Bayesian networks that we consider in this paper is defined as follows:

Problem 1. From the given set of observations O, compute an Bayesian network model θ that has the lowest AIC criterion value.

2.2 K2 Algorithm

K2 [2] is one of the best used traditional algorithms to learn Bayesian network models. Note that searching good Bayesian network models is time consuming because the problem to learn Bayesian networks is NP-hard in general [1]. K2 avoids the problem of running time by limiting the search space through the constraint of totally order of events. Namely, for a given order of events $X_1 < X_2 < \ldots < X_N$, causal relationship $X_k \to X_l$, where $k > l$ is not allowed. Note that this constraint is suitable for some cases: if events have their time of occurrence, an event X_k that occurred later than X_l cannot be a cause of X_l. Several practical scenes would be the case.

The process of K2 algorithm applied to a set of events $X_1, X_2, \ldots, X_N$ with the constraint $X_1 < X_2 < \ldots < X_N$ is described as follows:

(1) Select the best structure using two events X_N and X_{N-1}. Here, the two structures, i.e., $X_{N-1} \to X_N$ and the independent case, can be the candidates, and the one with better criterion value is selected.
(2) Add X_{N-2} to the structure. Namely, select the best structure from every possible cases where X_{N-2} has edges connected to X_{N-1} and X_N. Namely, from the cases (i) $X_{N-2} \to X_{N-1}$ and $X_{N-2} \to X_N$, (ii) $X_{N-2} \to X_{N-1}$ only, (iii) $X_{N-2} \to X_N$ only, and (iv) where X_{N-2} has no edge.
(3) Repeat step (2) to add events to the structure in the order $X_{N-3}, \ldots, X_2, X_1$.
(4) Output the final structure composed of all events.

Though K2 requires low computational time due to the constraint of event order, many problems do not allow the constraint. In such cases, we require to tackle the NP-hard problem using a heuristic algorithm for aproximate solutions.

2.3 Related Work for Non-ordered Bayesian Network Models

Even for the cases where the constraint of order is not allowed, several approaches to learn Bayesian network models has been proposed. One of the most basic method is to use K2 with random order, where randomly generated orders are applied repeatedly to K2 to search for good Bayesian network models.

As more sophisticated approaches, several ideas have been proposed so far. Hsu, et al. proposed a method to use K2 algorithm to which the orders evolved by genetic algorithms are applied [3]. Barrière, et al. proposed an algotirhm to evolve Bayesian network models based on a variation of genetic algorithms called co-evolving processes [4]. Tonda, et al. proposed another variation of genetic algorithms that applies a graph-based evolution process [5].

This paper propose a new efficient approach that evolves a probability vector from which Bayesian network models are generated.

3 Proposed Method

3.1 Population-Based Incremental Learning (PBIL)

We propose a new efficient algorithm to learn good Bayesian network models based on genetic algorithms with probability vector, to show that the probability based learning is effective for finding good models. Our idea is simply that we apply Population-based Incremental Learning (PBIL) [7] to learn Bayesian networks. So, we first describe PBIL in the following.

PBIL [7] is proposed by Baluja et al. in 1994, which is based on genetic algorithms, but is designed to evolve a probability vector. In PBIL, an individual creature s is defined as a vector $s = (v_1, v_2, \ldots, v_L)$, where $v_i (1 \le i \le L)$ is the i-th element that takes a value 0 or 1, and L is the number of elements that consist of an individual. Let $P = (p_1, p_2, \ldots, p_L)$ be a probability vector where $p_i (1 \le i \le L)$ represents the probability to be $v_i = 1$. Then, the algorithm of PBIL is described as follows:

(1) As initialization, we let $p_i = 0.5$ for all $i = 1, 2, \ldots, L$.
(2) Generate a set S that consists of C individuals, according to P. Namely, element v_i of each individual is determined according to the corresponding probability p_i.
(3) Compute the evaluation value for each individual $s \in S$.
(4) Select a set of individuals S' whose members have evaluation values within top C' in S, and update the probability vector according to the following formula:
$$p_i^{new} = dec(i) \times \alpha + p_i \times (1.0 - \alpha) \tag{1}$$
where p_i^{new} is the updated value of the new probability vector P^{new} (P is soon replaced with P^{new}), $dec(i)$ is a function that takes 1 if more than $C'/2$ individuals have their evaluation values more than 0.5, and takes 0 otherwise, and α is the parameter called learning ratio.
(5) Apply mutations on the new probability vector P: For all pairs of events $(X_i, X_j), i \neq j$, we apply the following formula with probability p_{mut}, where the function $rand()$ generates a random value from [0,1].
$$p_i^{new} = rand() \times \beta + p_i \times (1 - \beta) \tag{2}$$
(6) Repeat steps (2)-(5).

By merging top-C' individuals, PBIL evolves the probability vector such that the good individuals are more likely to be generated. Different from other genetic algorithms, PBIL does not include "crossover" between individuals. Instead, it evolves the probability vector as a "parent" of the generated individuals. Whereas, mutations play the same role as the general genetic algorithms, i.e., which is used to escape from local minimum areas in the search space.

3.2 Proposition: A PBIL-Based Algorithm to Construct Bayesian Network Models

We propose an probability-based GA algorithm by applying BPIL to learn Bayesian network models. Since our problem (i.e., Problem 1) is a little different from the general describtion of PBIL shown in the previous section, a little ajdustment is required.

P		Parent Node					
		X_1	X_2	...	X_i	...	X_N
Child node	X_1	0.0	0.5	...	p_{i1}	...	0.5
	X_2	0.5	0.0	...	p_{i2}	...	0.5
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	...	$\vdots$
	X_j	p_{1j}	p_{2j}	...	p_{ij}	...	p_{Nj}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	X_N	0.5	0.5	...	p_{iN}	...	0.0

Fig. 2. A Probability Vector

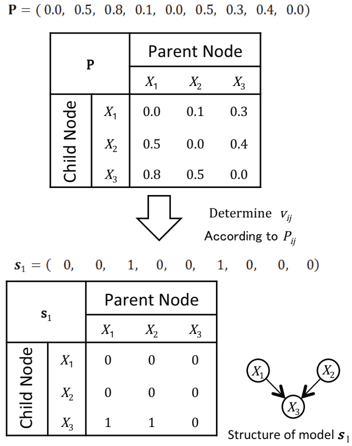


Fig. 3. Step (2):Generating Individuals

In our algorithm, individual creatures correspond to each Bayesian network model. Namely, with the number of events N, an individual model is represented as $s = (v_{11}, v_{12}, \ldots, v_{1N}, v_{21}, v_{22}, \ldots, v_{N1}, v_{N2}, \ldots, v_{NN})$, where v_{ij} corresponds to the edge from events X_i to X_j, i.e., if $v_{ij} = 1$ the edge from X_i to X_j exists in s, and if $v_{ij} = 0$ it does not exist. Similarly, we have the probability vector P to generate individual models as $P = (p_{11}, p_{12}, \ldots, p_{1N}, p_{21}, p_{22}, \ldots, p_{N1}, p_{N2}, \ldots, p_{NN})$, where p_{ij} is the probability that the edge from X_i to X_j exists. A probability vector can be regarded as a table as illustrated in Fig. 2. Note that, because Bayesian networks do not allow self-edges, p_{ij} is always 0 if $i = j$.

The process of the proposed algorithm is basically obtained from the steps of PBIL. Namely, the basic steps are described as follows:

(1) Initialize the probability vector P as $p_{ij} = 0$ if $i = j$ and $p_{ij} = 0.5$ otherwise.
(2) Generate S as a set of C individual models according to P. (This step is illustrated in Fig. 3.)
(3) Compute values of the evaluation criterion for all individual models $s \in S$.
(4) Select a set of individuals S' whose members have top-C' evaluation values in S, and update the probability vector according to the formula (1). (These steps (3) and (4) are illustrated in Fig. 4.)
(5) Apply mutations on P according to the formula (2). (This step is illustrated in Fig. 5.)
(6) Repeat steps (2)-(5).

Same as PBIL, the proposed algorithm evolves the probability vector to be likely to generate better individual models. However, there is a point specific to Bayesian networks, that is, a Bayesian network model is not allowed to have cycles in it. To consider this point in our algorithm, step 2 is detailed as follows:

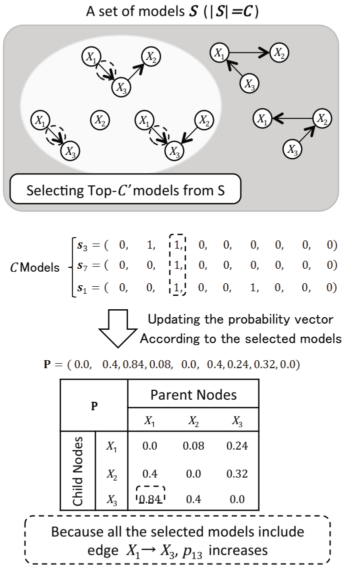


Fig. 4. Step (3)(4): Updating the Probability Vector

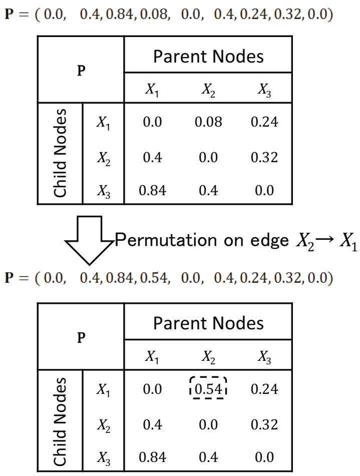


Fig. 5. Step (5): Mutations

(2a) Create a random order of pairs (i, j), where $1 \le i, j \le N$ and $i \ne j$.
(2b) Determine the values of v_{ij} according to P, with the order created in step (2a); every time v_{ij} is determined, if v_{ij} is determined as 1, we check whether this edge from X_i to X_j creates a cycle with all the edges determined to exist so far. If it creates a cycle, let v_{ij} be 0.
(2c) Repeat steps (2a) and (2b) until all pairs (i, j) in the order are processed.

These steps enables us to treat the problem of learning good Bayesian network models within the framework of PBIL. Note that checking the cycle creation in step (2b) can be done efficiently using a simple table that manages the taboo edges that create cycles when added.

4 Evaluation

4.1 Methods

We evaluate the proposed algorithm in comparison with a representative method K2 with its constraint (i.e., the order of events) evolved with genetic algorithms. Namely, in this conventional algorithm, we repeat creating Bayesian network models, in which its constraints are continuously evolved with a typical genetic algorithm over generations. We use Alarm Network [8] shown in Fig. 6, which is a Bayesian network model frequently used as a benchmark problem in this area of study. We create a set of 1000 observations according to the structure of Alarm Network, and then learn Bayesian network models from the observations

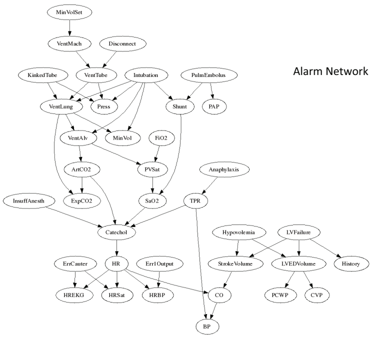


Fig. 6. The Alarm Network

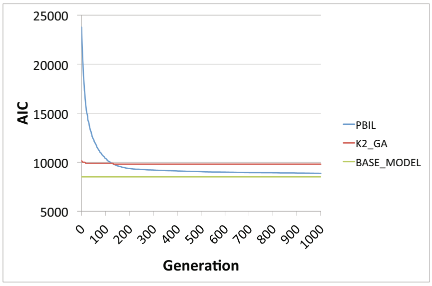


Fig. 7. A Bayesian Network Model

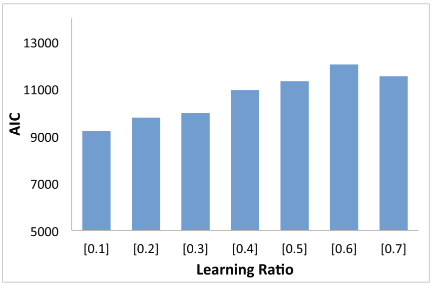


Fig. 8. A Bayesian Network Model

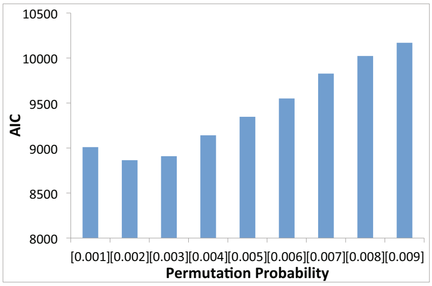


Fig. 9. A Bayesian Network Model

using those two algorithms. As the evaluation criterion, we use AIC, one of the representative criterion in this area. Namely, we compare the AIC values in order to evaluate how good is the Bayesian network models obtained by these two algorithms. As for parameters, we use $C = 100, C' = 1, \alpha = 0.1, \beta = 0.5$, and $p_{mut} = 0.0002$. All results shown are the averages of 10 repetitions with different random seeds.

4.2 Results

The main result is shown in Fig. 7, which indicates the AIC score of the best Bayesian network model found with the growth of generations. In this figure, the AIC score of the original Alarm Network, which is the optimal score, is denoted by "BASE_MODEL." The proposed algorithm (PBIL in the figure) converges to the optimal score as time passes, whereas K2-GA stops improving in the early stage. We can conclude that the performance of the proposed algorithm is better than the conventional algorithm. Note that the running time per generation in the proposed method is far shorter than the conventional method. (The difference is more than 250 times in our implementation.)

Fig. 8 and 9 shows the performance of the proposed algorithm with variation of learning ratio α and mutation probability p_{mut}. These results show the performance of the proposed method depends on α and p_{mut}, which indicates that we should care for the value of α to improve the performance of the proposed algorithm.

4.3 Conclusion

In this paper, we proposed a probability-based genetic algorithm to construct good Bayesian network models by applying the framework of PBIL method. We evaluated the proposed algorithm through comparison of the performances with well-known K2-based algorithms. As a result, we have clarified that the proposed method finds better Bayesian network models than the conventional ones, and also clarified the property that the proposed algorithm converges to the optimal solution whereas the conventional algorithm stops converging. The results indicate the possible efficiency of the probabilictic GA approach for the problem of finding good Bayesian network models.

One of the important future tasks is to explore the possibility of this approach on this problem, through improving the steps of the algorithm.

Acknowledgement. This work was partly supported by "the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry" of NARO (National Agriculture and Food Research Organization), and "the Program for Promotion of Stockbreeding" of JRA (Japan Racing Association).

References

1. Chickering, D.M., Heckerman, D., Meek, C.: Large-Sample Learning of Bayesian Networks is NP-Hard. Journal of Machine Learning Research 5, 1287–1330 (2004)
2. Cooper, G.F., Herskovits, E.: A Bayesian methods for the induction of probabilistic networks from data. Machine Learning 9, 309–347 (1992)
3. Hsu, W.H., Guo, H., Perry, B.B., Stilson, J.A.: A permutation genetic algorithm for variable ordering in learning Bayesian networks from data. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO (2002)
4. Barrière, O., Lutton, E., Wuillemin, P.H.: Bayesian Network Structure Learning using Cooperative Coevolution. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), pp. 755–762 (2009)
5. Tonda, A.P., Lutton, E., Reuillon, R., Squillero, G., Wuillemin, P.-H.: Bayesian network structure learning from limited datasets through graph evolution. In: Moraglio, A., Silva, S., Krawiec, K., Machado, P., Cotta, C. (eds.) EuroGP 2012. LNCS, vol. 7244, pp. 254–265. Springer, Heidelberg (2012)
6. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Proceedings of the 2nd International Symposium on Information Theory, pp. 267–281 (1973)
7. Baluja, S.: Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. Technical Report No. CMU-CS-94-163, Carf Michigan, Ann Arbor (1994)
8. Beinlich, I.A., Suermondt, H.J., Chavez, R.M., Cooper, G.F.: The ALARM monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks. In: Second European Conference on Artificial Intelligence in Medicine, London, Great Britain, vol. 38, pp. 247–256. Springer, Berlin (1989)

A Constraint Programming Approach to the Zahn's Decision Problem

Mhamdi Amel[1] and Naanaa Wady[2]

[1] MIRACL, Sfax, Tunisia
[2] Faculty of Sciences, University of Monastir, Monastir, Tunisia
amel.mhamdi2@gmail.com, wady.naanaa@fsm.rnu.tn

Abstract. Given a graph $G = (S, E)$, the problem dealt with in this paper consists in partitioning S into a disjoint union of cliques by adding or removing a minimum number $z(G)$ of edges. The problem, which is refered to by the Zahn Problem (ZP), is NP-hard in general.

This paper presents a constraint programming approach to ZP. The problem is formulated in terms of a Weighted Constraint Satisfaction Problem (WCSP), a widely used framework for solving hard combinatorial problems. As a seach strategy, we applied a Limited Discrepancy Search coupled with a branch-and-bound algorithm, a combination which has proved to be very advantageous.

We compared our approach to a fixed-parameter tractability algorithm, one of the most used algorithms for solving ZP. The comparison clearly show that our approach is very competitive, especially on large ZP instances.

Keywords: Zahn's problem, Constraint programming, Cliques, Cluster editing, Fixed parameter tractable problems.

1 Introduction

The problem of approximating a symmetric relation by an equivalence relation is specified as follow: let R be a symmetric binary relation over a finite set S. The task is to determine an equivalence relation E, also defined on S, which best fits R. The fitness of E with regard to R is evaluated by means of a distance between E and R. This diatance is known as the *Zahn index* and is defined as follows:

$$\delta(E, R) = |R - E| + |E - R| \tag{1}$$

A solution to the so-called ZP consists, therefore, in an equivalence relation E^* over S such that $\delta(E^*, R)$ is minimal over all possible equivalence relations defined over S [1]. In [7], it has been proven that ZP is NP-hard.

The problem, which has been first defined by C.T. Zahn in 1964, is still of great interest because it arises in the analysis of gene expression data [6]. The input of the latter problem is a similarity graph, where vertices represent genes and weighted edges reflect the extent of similarity between genes. Similar expression

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference,* Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_59,

profiles of two genes can be explained by a similar function of the corresponding protein. Proteins are often part of protein families. This kinship can be used to find related family members which are related by intermediate sequences. This concept is called transitivity and refers to the following property of mathematical relations: If A and B are related as well as B and C, then A and C are also related. So A, B and C are in the same protein families. Similar expression profiles of organisms can be explained by a similar metabolism state which may be caused by similar environmental reasons or diseases. This explains why clustering of similar genes, or organisms based on expression profiles is vital for modern biological data analysis [4].

This work has been motivated by the one presented in [1], where it has been pointed out that the problem can be more easily visualized when translated to its graph theoretic counterpart. This is achieved by associating to each symmetric binary relation, an undirected graph. It has been noticed that the graph of an equivalence relation E consists of a certain number of complete subgraphs, i.e., cliques, that are not connected to each other by any edge. The individual complete subgraphs correspond to the equivalence classes of E. The problem is, therefore, referred to as *graph partitioning* [8] or cluster editing [2]. It has been shown that all these connected problems belong to the class of *Fixed Parameter Tractable* (FPT) problems [14]. Algorithms in the FPT class use the desired modification cost as a controle parameter. This parameter is closely related to the time complexity of the FPT algorithms since the time complexity of these algorithms is, in general, $O(c^k)$, where k is the FPT parameter and c is a constant.

In [14], the authors presented a fixed parameter tractable (FTP) algorithm for cluster editing. The latter algorithm stops when the modified graph converges to an union of cliques The time complexity of this algorithm, which is one of the more sophisticated in the FPT class, is $O(2.27^k)$. It is enhanced by two graph reduction rules. The first rule is based on the neighborhood of every pair of vertices $\{u, v\} \in S$. (1) If u and v have more than k common neighbors, then $\{u, v\}$ has to belong to the target graph. (2) If u and v have more than k non-common neighbors, then $\{u, v\}$ must not belong to the target graph. If u and v have both more than k common and more than k non-common neighbors, then the given instance has no cost k solution.

In this paper, we propose an approach for solving ZP which resorts to the constraint programming (CP) paradigm. This paradigm has been widely employed in solving NP-hard problems such as combinatorial optimization problems. The originality of the CP approach is that it provides a clear separation between problem formulation and problem solution. The approximation of symmetric relations is, therefore, encoded as a weighted constraint satisfaction problem (WCSP) in an easy and declarative way. The problem is, thereafter, solved by means of one of the more efficient constraint solvers, namely, the TOULBAR2 solver [9]. Our approach is flexible and extensible, in that it can easily encode and handle new constraints in order to cope with other variations of ZP. Moreover, to our knowledge, this paper presents the first constraint programming approach to ZP.

The remainder of the paper is organised as follow. Section 2, specifies some constraint programming background. In Section 3, we detail the new constraint programming model of ZP. Section 4 evaluates the performance of our approach then compares its results to the results obtained by FPT algorithms. Finally, Section 5 gives a brief conclusion and directions for future works.

2 Constraint Programming Background

Constraint Satisfaction Problems (CSPs) provide a general framework for modeling and solving numerous combinatorial problems. Basically, a CSP consists of a set of *variables*, each of which can take a value chosen among a set of potential values called *domain*. The *constraints* express restrictions on which combinations of values are allowed. The problem is to find an assignment of values to variables, from their respective domains, such that all the constraints are satisfied. In the valued CSP framework (VCSP) a valuation is associated to every value combination [10]. The set of possible valuations E is assumed to be a totally ordered set with a minimum ($\perp$) and a maximum ($\top$) element, equipped with a single monotonic binary operation $\oplus$ known as aggregation. The VCSP components can be gathered into a valuation structure that can be specified as follows:

Definition 1. *A valuation structure is a tuple* $S = (E, \oplus, \preceq)$*, where* E *is a set of valuations,* $\preceq$ *is a total order on* E *and* $\oplus$ *is a binary commutative, associative and monotonic operator.*

The studied valued constraint satisfaction problem (VCSP) is, in turn, defined as follows:

Definition 2. *A VCSP instance is defined by a tuple* (X, D, C, S) *, where*

- X *is a finite set of variables.*
- D *is a set of finite domains, where* $D_x \in D$ *denotes the domain of* $x \in X$*.*
- C *is a set of valued constraints. Each valued constraint* c *is an order pair* (σ, ϕ) *where* $\sigma \subseteq X$ *is the scope of* c *and* ϕ *is a function from* $\prod_{x\in\sigma} D_x$ *to* E*.*
- $S = (E, \oplus, \preceq)$ *is a valuation structure, where* $\oplus$ *is strictly monotonic.*

A variable x should be assigned a value only from its domain, D_x. If a valued constaint is defined by a function which uses the minimum ($\perp$) and the maximum valuation ($\top$) only then it is said to be a *hard constaint*, otherwise, it is a *soft constaint.* The arity of a valued constraint is the size of its scope. The arity of a problem is the maximum arity over all its constraints. In this work, we are mainly concerned with ternary VCSPs, that is, VCSPs with unary , binary and ternary constraints only.

3 The Constraint Programming Model of ZP

In this section, we propose an encoding of ZP in terms of weighted constraint satisfaction problem (WCSP). We have, therefore, to identify the set of variables, X, the value domain of every variable, D, and finally the set of weighted constraints, C. Remind that the goal is to transform a given graph G into a set of disjoint union of cliques.

We start from the following observation: a graph is an union of disjoint cliques if and only if it does not contain a subgraph which is isomorphic to P_3, the graph with three vertices and two edges (see Figure 1).

The proposed encoding of ZP associates a boolean variable to every pair of vertices in G. Hard ternary constraints will be used in order to avoid the P_3 subgraphs. The objective, which consists in limiting the modifications carried on the original graph, is encoded using unary soft constraints.

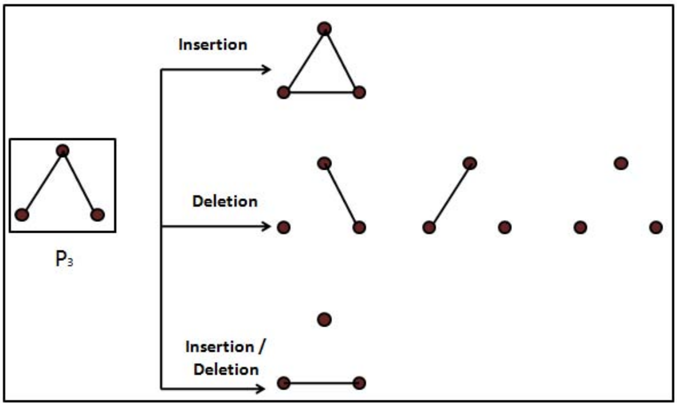


Fig. 1. P_3 Graph and insertion/deletion possibilities

3.1 Valuation Structure

The valuation structure of weighted CSP (WCSP) is, $S(k) = ([0..k], \oplus, \preceq)$ where $K > 0$ is a natural number; $\oplus$ is defined as $a \oplus b = min\{k, a+b\}$; $\preceq$ is the standard orfer among naturals. Observe that in $S(k)$, we have $0 = \bot$ and $k = \top$ [5] .

3.2 The Variables and Their Domain

Let $G = (S, E)$ be the graph of the symmetric relation to be approximated. Every pair of vertices $\{u, v\} \subseteq S$ is associated with a boolean variable x of the WCSP. When x is set to true by the WCSP solver, this means that the decision is to keep $\{u, v\}$ in the target graph or to add it to the target graph. Otherwise, the edge $\{u, v\}$ is either removed or not added to the target graph. The variable set X is, therefore, composed of $|S|(|S| - 1)/2$ boolean variables.

3.3 The Constraints

Two types of constraints are used in our encoding of ZP: soft unary constaints and hard ternary constraints.

- The role of the unary soft constraints is to minimize the number of changes introduced on the original graph. A change is an addition of a new edge or the removal of an existing edge. Each change in the original graph entails an additive unitary cost. We distinguish two situations that entail costs: (1) the case where $\{u, v\} \in E$ and the decision is to remove $\{u, v\}$ by setting the associated boolean variable to 0. (2) the case where $\{u, v\} \notin E$ and the decision is to add $\{u, v\}$ by setting the associated boolean variable to 1. Observe that there are two other situations which do not entail any cost. The first one is when we decide to keep an existing edge in the graph and the second one occurs when we decide to not to add a new edge.
- The role of the ternary hard constraints is to forbid P_3 subgraphs to appear in the target graph. Forbiding P_3 subgraphs can be achieved by enforcing the following rule to apply: If $\{u, v\} \in E$ and $\{v, w\} \in E$ then $\{u, w\} \in E$. The latter rule can be enforced by creating a ternary hard constraint for every triple of variable involving a triple of vertices. More precisely, let x, y and z be the boolean variables associated respectively to the pairs $\{u, v\}, \{v, w\}$ and $\{u, w\}$ of S. Then, we must have, in the encoding WCSP, a ternary constraint whose scope is the triple $\{x, y, z\}$ and whose valuation function is depicted in Table 1. As it can be seen in the latter table, there are three assignements that cannot be taken simultaneously by x, y and z, these are 011, 101 and 110. For this reason, these assignements are associated to an infinite cost. Observe that the three forbiden assignements correspond to the P_3 subgraphs that can be formed by vetices u, v and w.

3.4 The Solution Algorithm

Weighted constraint satisfaction problems are typically solved by a branch-and-bound algorithm which interleaves branching with pruning and lower bound computation. The lower bound is used to prune the search subtrees that do not contain relevent solutions. Both bound computation and the pruning are performed by a soft arc-consistency algorithme called existential directional arc consistency (EDAC*) [11]. To get a more efficient solution algorithm, we used the branch-and-bound algorithm within a Limited Discrepancy Search (LDS) [12]. This search strategy is defined with regard to a value ordering heuristic. The seach algorithm performs a discrepancy when it does not branch through the value indicated by the value ordering heuristic. Limited discrepancy search proceeds by allowing discrepancy in an increasing manner. Thus, the seach path that does not involve any discrepancy is explored first. Then those paths that involve a single discrepancy, etc... The maximal number of descrepancy that can be allowed can be fixed to a given constant before search or can be guessed by the early stopping technique described in [13].

The value ordering heuristic used in this work is the one favoring the value that has the minimal unary cost. This cost is computed, during search, for the values of all the variables by EDAC*.

Table 1. The valuation function of the ternary hard constraints

xyz	cost
000	0
001	0
010	0
011	$+\infty$
100	0
101	$+\infty$
110	$+\infty$
111	0

4 Experimental Results

The graphs investigated in our exprement are random graphs that have the distinction of being similar to real graphs. We used the LFR graph generator developed by santo fortunato which is available at [3]. This generator uses four parameters: the number of vertices n, the average degree $\bar{\delta}$, the maximum degree Δ and the mixing parameter μ. The latter parameter is of great relevance because it controls the fraction of edges that relate dense subgraphs. If $\mu = 0$, all edges are within dense subgraphs that are not related to each other. This results in easy to solve ZP instances. If $\mu = 1$ then the graph does not display any structure and the associated instances are rather hard to solve. The resulting ZP instances are transformed into WCSPs by applying the encoding described in Section 3.

We used the Toulbar2 WCSP solver which is available at [9]. Both the filtering algorithm (EDAC*) and the branching strategy (LSD) employed in this work are implemented within toulbar2. The evaluation criteria are the number of edge change (delition or insertion) and the CPU time in seconds. The FPT algorithm, with which we compared our algorithm, is implemented in C++. All programs run under Ubuntu 13.04 on a 2.0 GHz PC having 6 Go of RAM.

The results of the experiment clearly indicated that the WCSP algorithm outperforms the FPT algorithm on both sparse and dense graphs (see Table 3). For sparse graphs, we noticed that in most cases, the WCSP algorithm gives a better cost that the FPT algorithm (see Figure 2). Moreover, whenever the two algorithms gave the same cost, the WCSP algorithm is always faster as it can be seen from Table 2 (lines 3, 4, 5 and 6).

Table 2. Results on random instances with 50 dordre of the FPT approach and the WCSP approach

n	$\|E\|$	δ	Δ	μ	FPT cost	FPT CPU	WCSP cost	WCSP CPU
50	119	5	10	0.1	89	1	50	4.1
50	623	25	35	0.1	285	40	186	15.08
50	928	40	40	0.1	297	128	297	17.4
50	946	40	47	0.1	279	218	279	7.98
50	1028	45	47	0.1	202	87	202	2.47
50	1101	48	49	0.1	124	216	124	2.4

Table 3. Results on random instances with 100 dordre of the FPT approach and the WCSP approach

n	$\|E\|$	δ	Δ	μ	FPT cost	WCSP cost
100	249	5	10	0.1	178	121
100	303	5	15	0.1	211	125
100	488	10	20	0.1	348	297
100	744	15	20	0.1	535	276
100	971	20	25	0.1	680	503

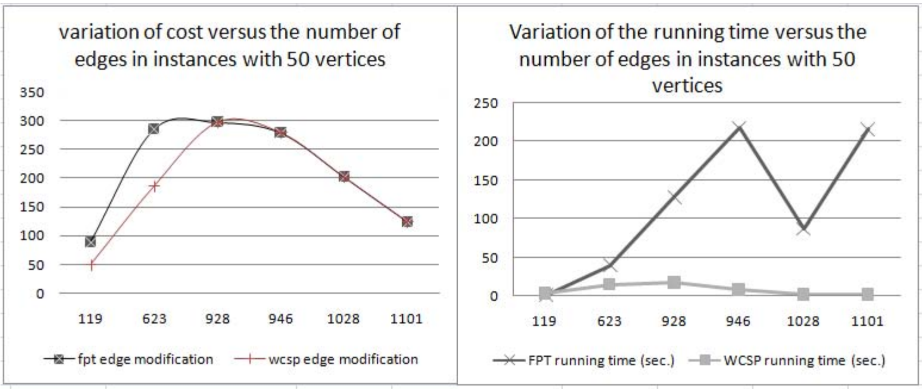


Fig. 2. A comparison of cost and running time with variying of the number of edges in randon instance with 50 order computed by the FPT approach versus WCSP approch

5 Conclusion

We presented a constraint programming approach for solving the Zahn problem. The latter porblem is encoded as a weighted constraint satisfaction problem (WCSP) and solved by an efficient existing WCSP solver named TOULBAR2. We compared our approach to the fixed-parameter tractability algorithm. The results obtained by our approach are clearly competitive, especially on large ZP instances. This work can be extended by designing heuristics espacialy tailored for ZP in order to enable solving larger instances.

References

1. Zahn, C.T.: Approximating symmetric relations. SIAM J. Appel. Math. 12, 840–847 (1964)
2. Sebastian, B.: Weighted cluster editing for clustering biological data, Jena (2007)
3. https://sites.google.com/site/santofortunato/inthepress2
4. Pipenbacher, P., Schliep, A., Schneckener, S., Schonhuth, A., Schomburg, D., Schrader, R.: ProClust: Improved clustering of protein sequences with an extended graph-based approach. Bioinformatics (2002)
5. Larrosa, J., Schiex, T.: Solving Weighted CSP by Maintaining Arc-consistency. Artificial Intelligence (2004)
6. Hansen, P., Jaumard, B.: Cluster analysis and mathematical programming. Math Programming (1998)
7. Krivarek, M., Moravek, J.: NP-hard Problem in hierarchical-tree Clustering. Acta Inform. (1986)
8. Charon, I., Hudry, O.: Noising methods for a clique partitioning problem. Discrete Applied Mathematics 154 (2006)
9. http://mulcyber.toulouse.inra.fr/projects/toulbar2
10. Schiex, T., Fargier, H., Verfaillie, G.: Valued constraint satisfaction problems: Hard and easy problems. In: IJCAI (1995)
11. Sanchez, M., De Givry, S., Schiex, T.: Mendelian error detection in complex pedigrees using weighted constraint satisfaction techniques. Constraints (2008)
12. Harvey William, D., Ginsberg Matthew, L.: Limited discrepancy search. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (1995)
13. Karoui, W., Huguet, M.-J., Lopez, P., Naanaa, W.: YIELDS: A Yet Improved Limited Discrepancy Search for CSPs. In: Van Hentenryck, P., Wolsey, L. (eds.) CPAIOR 2007. LNCS, vol. 4510, pp. 99–111. Springer, Heidelberg (2007)
14. Dehne, F., Langston, M.A., Luo, X., Pitre, S., Shaw, P., Zhang, Y.: The Cluster Editing Problem: Implementations and Experiments. In: Bodlaender, H.L., Langston, M.A. (eds.) IWPEC 2006. LNCS, vol. 4169, pp. 13–24. Springer, Heidelberg (2006)

Multi-agent Model Based on Tabu Search for the Permutation Flow Shop Scheduling Problem

Hafewa Bargaoui[1] and Olfa Belkahla Driss[2]

[1] Stratégies d'Optimisation et Informatique intelligentE (SOIE)
High Institute of Management, University of Tunis
41, Street of Liberty Bouchoucha-City CP-2000-Bardo Tunis, Tunisia
[2] Higher Business School of Tunis, University of Manouba
hafewa.bargaoui@gmail.com, olfa.belkahla@isg.rnu.tn

Abstract. In this work, we suggest a Multi-Agent model based on a tabu search method for solving the permutation flow shop scheduling problem. The problem is strongly NP-hard and its resolution optimally within a reasonable time is impossible. The objective of this work is to minimize the makespan or the total duration of the schedule. The proposed model is composed of two classes of agents: Supervisor agent, responsible for generating the initial solution and containing the tabu search core, and Scheduler agents which are responsible for the satisfaction of the constraints under their jurisdiction and the evaluation of all the neighborhood solutions generated by the Supervisor agent. Computational experiments on different benchmarks data sets demonstrate that the proposed model reaches high-quality solutions.

Keywords: Scheduling, Permutation flow shop, Tabu search, Multi-Agent systems.

1 Introduction

The scheduling problem is one of the most encountered problems in the management of production system. It consists in allocation number of jobs to machines taking into consideration a set of constraints. One of the most thoroughly difficult scheduling problem we cite the Permutation Flow Shop scheduling Problem (PFSP) which has been the subject of a significant amount of literature in the operations research field. It concerns a set N = {1… n} of n independent jobs that has to be processed on a set of M = {1… m} machines in the same order. The processing time for job i on machine j is denoted $t_{i,j}$. In the PFSP, the processing sequence of jobs on the first machine is maintained throughout the remaining machines. The objective is to find a permutation that minimizes one or more criterion. The studied problem is expressed as $Fm|prmu|C_{max}$ following the notation introduced by [1] where the objective is to find a permutation that minimizes the makespan (the completion time of the last job on the last machine). This problem is known to be NP-hard for $m \geq 3$ [2].

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_60,

2 Literature Review

Since the pioneering algorithm of [3] to solve the two machines problem optimally, many exact methods have been developed such as [4]. However, due to the difficulty and the huge computation time, of exactly resolving the problem has led researchers to focus their energies in the development of heuristic procedures which provide high quality solution to large size problems in a short time. The heuristic methods can be classified either as constructive or improvement approaches. Constructive heuristic builds a feasible solution starting from scratch, step by step. Among them the Campbell, Dudek and Smith (CDS) heuristic [5] as an extension of the Johnson algorithm for m > 2 and the algorithm developed by [6] known as Rapid Access (RA) which can be presented as variant of CDS procedure. Also, Nawaz, Enscore and Ham presented the NEH heuristic [7] which is considered as the best constructive heuristic for the PFSP. On the other hand, the improvement heuristic starts with a feasible solution which they aim to improve by exploring its neighborhood. Methods of this type include [8] which is a variant of NEH heuristic. Meta-heuristics for the PFSP appeared much later to generate good results. Some of the most recent are the iterated greedy methods by [9], the tabu search proposed by [10] and the bee colony developed by [11].

In this paper, we propose a tabu search for the PFSP with the makespan minimization criterion. The tabu search (TS) originally proposed by [12] is an iterative meta-heuristic approach based on the principle of the local search and that proved to be successful in solving hard optimization problem. The remainder of this paper is organized as follow: the next section includes the proposed Multi-Agent model. In section 4, we present the different elements of the tabu search. Section 5 discusses experimental results on the performance of the proposed model. Finally, section 6 gives our conclusions and remarks with some future directions.

3 The Multi-agent Model

The proposed model named Multi-Agent model for Flow Shop Problem (M.A.F.S.P) as illustrated in Fig.1 involves two types of reactive agents: One Supervisor agent and a set of (n-1) Scheduler agents with n refers the number of jobs. We used (n-1) Scheduler agents because for each current solution of the TS algorithm the Supervisor agent generates (n-1) neighboring solutions.

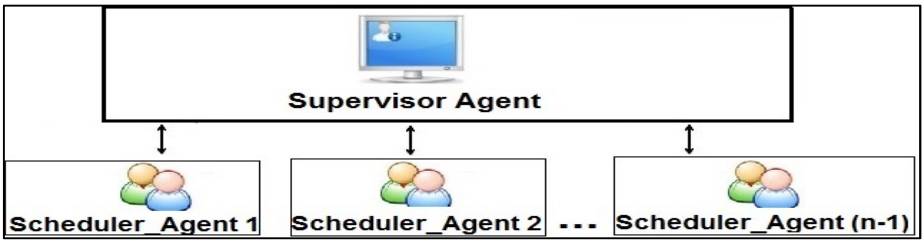


Fig. 1. M.A.F.S.P architecture

Each agent has acquaintances (the agents that it knows and with which it can communicate) and its own behavior. Before starting the distributed solving process the Supervisor agent generate an initial solution and then try to ameliorate it iteratively by applying the global optimization process based on TS approach in order to minimize the makespan. The different types of agents are explained in detail below.

3.1 Supervisor Agent

Supervisor agent contains the core of tabu search; it's unsatisfied as long as the maximal number of iterations is not reached. It aims to launch the program and create as many Scheduler agents as neighboring solutions (i.e. number of jobs -1). Furthermore, it provides all the necessary information needed for each Scheduler agent namely the execution time of each job on all machines, the current solution, the job to move and its new position. It also aims to detect that the problem has been resolved and output the corresponding scheduling solution and its makespan value.

As the core of the global optimization, i.e. the tabu search, of M.A.F.S.P is located at the Supervisor agent, the knowledge of the latter are composed mainly of the tabu search parameters. Static knowledge of the Supervisor agent is composed of: the execution time of each job on all machines, the initial solution from which begins the optimization process, the used stopping criterion, i.e. the maximum number of iterations allowed by the tabu search algorithm, the size of the tabu list and the used diversification criterion, i.e. the number of iterations between two successive improvements. Its dynamic knowledge consists of: neighboring solutions and their makespans, the current solution and its makespan, the best found schedule and its makespan, the performed number of iterations, the tabu list elements, and the number of iterations after the last improvement.

The Supervisor agent is satisfied when the stopping criterion is reached; in this case it provides the best found-solution to the user. Otherwise, at each iteration of the tabu search, the Supervisor agent generates a set of (n-1) neighboring solutions and orders the scheduler agents to evaluate the cost of each neighbor in a parallel way.

3.2 Scheduler Agents

The Scheduler agent plays an important role in our model. Indeed, it has been shown that the time needed to evaluate the value of the makespans of the neighborhood of the current solution is almost the entire calculation time [14]. For this reason, we assign for each Scheduler agent the task of evaluating and calculating the cost of one neighbor; according to this, we estimate, reduce the total time required to search for the best neighbor. The scheduler agent behaviors are divided into two types:

- Local plan which describes the behavior of an isolated scheduler agent, i.e. constructing a neighbor solution, calculating it's cost and the satisfaction of constraints under their responsibility, and it does not require interaction with other agents,
- Communication protocol that describes the behavior of the Scheduler agent interacting only with the Supervisor agent and requires sending and receiving messages.

In M.A.F.S.P, the evaluation of the whole neighborhood solutions is achieved by cooperation among the Scheduler agents. In fact, each Scheduler agent receives a message from the Supervisor agent that contains the current solution, the job j_i to be moved as well as its new position. The agent Scheduler, responsible for scheduling calculations, put the job j_i in its new position and calculates the cost of the new solution. Then, it sends a message to the Supervisor agent containing the new scheduling and its cost. The latter, after receiving all the neighboring solutions, calculated by the Scheduler agents, chooses the best non tabu neighbor from the neighborhood of the current solution to start the next iteration.

The knowledge of the Scheduler agent are mainly composed of the current solution provided by the Supervisor agent, the execution time of each job on all machines, the job to move and its new position.

4 Elements of the Tabu Search Approach

In what follows, we will give the adaptation of the different parameters of the tabu search for the permutation flow shop scheduling.

4.1 Initial Solution

It has been shown that the effectiveness of the approach based on the principle of local search depends heavily on the quality of the initial solution. Based on the review of flow shop heuristics by [13], the NEH heuristic [7] is regarded as one of the best constructive heuristic for the permutation flow shop problem. The NEH heuristic can be described by the following four steps:

1. Sort the n jobs by decreasing total processing time on the m machines
2. Take the first two jobs and schedule them in order to minimize the partial makespan as if there were only these two jobs
3. For the k-th job, $k = 3$ to n do
4. Insert the k-th job into the place, among k possible ones, which minimizes the partial makespan

The complexity of NEH heuristic is $O(n^3m)$ which can lead to considerable computation times for large instances. However, it is possible to reduce the complexity of NEH to $O(n^2m)$ by using the implementation of [14] called NEHT, by means of three matrices, named e (heads), q (tails) and f (heads + processing times).

In M.A.F.S.P, we use NEHT [14] algorithm since it is the best polynomial heuristic in practice and it is rather implemented straightforwardly. At the initial stage, the Supervisor agent generates an initial scheduling then searches through its neighborhood, a permutation with the lowest makespan by applying TS mechanism.

4.2 Neighborhood Definition

The function adopted to generate the neighborhood of a current solution is the key factor for the success of an approach based on TS. In the literature, we can meet many types of moves. Among the most used ones we can cite: Swapping move, Exchange move and Insertion move. Experiments show that the insertion move, which consists in removing the job, placed at position j and insert it to position k, is regarded as the best neighborhood structure for the PFSP so we adopted this strategy. In order to remove a job from a sequence to reinsert it elsewhere efficiently we use the extension of Taillard's implementation presented in [15] which can be described as follows:

1. Compute the earliest completion time $e_{i,j}$ of the i-th job on the j-th machine; the starting time of the first job on the first machine is 0
   $e_{0,j} = 0$; $e_{i,0} = 0$; $e_{i,j} = \max(e_{i,j-1}, e_{i-1,j}) + t_{i,j}$; $i = 1 \ldots k$; $j = 1 \ldots m$
2. Compute the tail $q_{i,j}$, i.e. the duration between the starting time of the i-th job on the j-th machine and the end of the operations
   $q_{k+1,j} = 0$; $q_{i,m+1} = 0$; $q_{i,j} = \max(q_{i,j+1}, q_{i+1,j}) + t_{i,j}$; $i = k \ldots 1$; $j = m \ldots 1$
   The value of the partial makespan M'_i when removing the job at the i-th position is: $M'_i = \max_j(e_{i-1,j} + q_{i+1,j})$; $i = 1 \ldots k$; $j = 1 \ldots m$.

In case of job remove [15] proposed a relative criterion which consists in maximizing the gain of makespan relatively to the sum of processing times of the removed job. The relative criterion is described as follows: remove the job at the p-th position, such that: $M_p = \max_{i=1..k}\left(\frac{M-M'_i}{\sum_{j=1..m} t_{i,j}}\right)$ where M is the makespan of the permutation before removing the job. In M.A.F.S.P once the initial permutation is determined, it is considered as a current solution. The Supervisor agent selects the less well-inserted job and sends a message to the (n-1) Scheduler agents each of which contains the current solution, the job to move and its new position (remove the job at position p1 and insert it at position p2 such as p2∉p1), see Fig. 2. Then all the Scheduler agents have the responsibility to evaluate the cost of each neighbor in a parallel way.

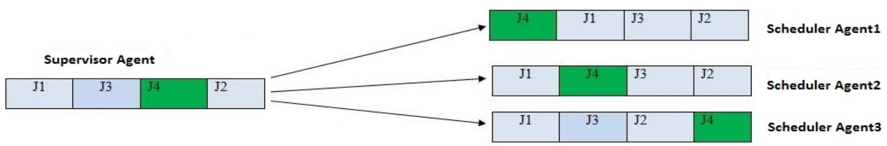


Fig. 2. The proposed neighborhood structure

4.3 Move Strategy

Move strategy determines the manner in which the neighbors are examined. In M.A.F.S.P, first, the makespan for each neighborhood solution is calculated then the non-tabu permutation that has the smallest makespan among all possible neighbors is selected as a current solution for the next iteration. Once the calculation of makespan of all possible neighbor solutions is done, they will be sent to the Supervisor agent. The latter chooses the non-tabu permutation with the smallest cost to start a new iteration and inserts its makespan in the tabu list.

When a new best solution cannot be found after a predetermined threshold, the search switches to diversification phase which will be explained in Section 4.6.

4.4 Tabu List

The tabu list is one of the mechanisms that attempted to avoid local optimality. In M.A.F.S.P, we propose to apply a tabu list with a fixed size which stores makespans of permutations already visited. The FIFO rule is adopted; once the list is full, the oldest element is removed and a new one is added. In our implementation we used a tabu list of size7.

4.5 Stopping Rule

In M.A.F.S.P, when the maximum number of iterations, called maxiter, is reached the Supervisor stops the algorithm and displays the results.

4.6 Diversification Scheme

Despite the effectiveness of the TS method in PFSP, some limitations have been detected. Indeed, the main inconvenience is summed up in the absence of an effective diversification technique that encourages the search process to examine unvisited regions. In order to provide a wide exploration of the search space and to get out of a local optimum, NEHT heuristic is hybridized into TS to find a new solution. In M.A.F.S.P, when no improvement in makespan is obtained for a number of iterations (which is fixed to twenty), it means that best-so-far solution is not replaced by neighborhood solutions and that TS is entrapped in local optimum. In this case, the algorithm jumps back to the best permutation obtained so far and selects randomly two integers L_1 and L_2 ($0< L_1, L_2 \leq n$ and $L_1 < L_2$) representing the start and end point of the sub-sequence of jobs to reinsert by applying the insertion phase of the NEHT heuristic until a complete permutation of n jobs is obtained as described in Fig. 3.

Fig. 3. Example for the application of diversification scheme

In M.A.F.S.P, the Supervisor agent applies the diversification phase on the current solution. The new obtained scheduling serves as a new current solution of TS for the next iteration in order to get rid of the local optimum.

5 Experimental Evaluation

The M.A.F.S.P was implemented in the JADE platform and all tests were conducted on a Core2Duo 2.20 GHz with 3.0 GB. We fixed M.A.F.S.P maxiter to 2000. The proposed M.A.F.S.F is compared with 3 other powerful methods: NEHT heuristic [7] and 3XTS tabu search [10] and the HDABC bee colony algorithm developed by [11]. In the first experiments we compared M.AF.S.P with NEHT and 3XTS algorithms and tests were conducted on Taillard's benchmarks [16]. For the second tests, we compare M.A.F.S.P with the HDABC bee colony algorithm and computational simulations were carried out with instances designed by Reeves and Yamada [17].

Table 1. Results for NEHT, 3XTS and M.A.F.S.P

Problem	size($n \times m$)	NEHT	3XTS	M.A.F.S.P
Ta001		1286	1278	**1278***
Ta002		1365	1359	**1359***
Ta003	20*05	1159	1081	**1081***
Ta004		1325	1293	**1293***
Ta005		1305	1235	**1235***
Ta011		1680	1582	**1582***
Ta012		1729	1659	**1659***
Ta013	20*10	1557	1496	**1496***
Ta014		1439	1377	**1377***
Ta015		1502	1419	**1419***
Ta021		2410	2297	**2297***
Ta022		2150	2103	**2099***
Ta023	20*20	2411	2330	2328
Ta024		2262	2229	**2223***
Ta025		2397	2291	**2291***
Ta031	50*5	2733	2724	**2724***
Ta041	50*10	3135	3025	**3025**
Ta051	50*20	4082	3893	3895
Ta061	100*5	5519	5493	**5493***
Ta071	100*10	5846	5770	**5770***
Ta081	100*20	6541	6300	**6283**
Ta091	200*10	10 942	10869	10872
Ta101	200*20	11 594	11251	11 331

Table 2. Results for HDABC and M.A.F.S.P

Instances	size ($n \times m$)	C*	HDABC	M.A.F.S.P
Rec01		1247	1247	**1247***
Rec03	20*05	1109	1109	**1109***
Rec05		1242	1242	**1242***
Rec07		1566	1566	**1566***
Rec09	20*10	1537	1537	**1537***
Rec11		1431	1431	**1431***
Rec13		1930	1932	**1930***
Rec15	20*15	1950	1963	**1950***
Rec17		1902	1917	**1902***
Rec19		2093	2101	**2099**
Rec21	30*10	2017	2046	**2019**
Rec23		2011	2020	**2018**
Rec25		2513	2542	**2522**
Rec27	30*15	2373	2392	**2379**
Rec29		2287	2310	**2289**
Rec31		3045	3101	**3053**
Rec33	50*10	3114	3126	**3114***
Rec35		3277	3277	**3277***
Rec37		4951	5104	**5014**
Rec39	75*20	5087	5180	**5134**
Rec41		4960	5106	**5031**

The results of comparison between NEHT, 3XTS and M.A.F.S.P are given in table1. The performance comparison between HDABC and M.A.F.S.P is shown in Table 2. The asterisk (*) indicates that optimal solution was found. Table1 shows the solutions for the makespan obtained by M.A.F.S.P as well as the optimal solution for 17 instances out of 23. Table1 shows also that results obtained by M.A.F.S.P are better than those obtained by 3XTS approach in 17% of instances while 3XTS approach outperforms our approach in 13% of instances.

From Table2, it can be concluded that the performance of M.A.F.S.P is better than HDABC. Tests demonstrate that M.A.F.S.P surpassed HDABC in 67% of instances in terms of makespan. Also, we can note that the proposed model provided the optimal solution for 11 instances out of 21.

6 Conclusions and Future Research

In this paper we have proposed a Multi-Agent model based on tabu search metaheuristic to solve the n*m permutation flow shop scheduling problem with makespan criterion. M.A.F.S.P consists of Supervisor agent and Scheduler agents in interaction, trying to find the best possible permutation.

Our model was tested against another tabu search and a bee colony algorithm. The results are very promising and show that the M.A.F.S.P is competitive with other successful methods. In the future, we plan to investigate the applicability of the proposed model to the multi-objective permutation flow shop scheduling problem.

References

1. Graham, R.L., Lawler, E., Lenstra, J.K., Rinnooy Kan, A.: Optimization and Approximation in Deterministic Sequencing and Scheduling: A Survey. Annals of Discrete Mathematics 5, 236–287 (1979)
2. Garey, M.R., Johnson, D.S.: Computers and intractability: A guide to the theory of NP-completeness. Freeman, San Francisco (1979)
3. Johnson, S.M.: Optimal two- and three-stage productionschedules with setup times included. Naval Research Logistics Quarterly 1, 61–68 (1954)
4. Ladhari, T., Haouari, M.: A computational study of the permutationflow shop problem based on a tight lower bound. Computers & Operations Research 32, 1831–1847 (2005)
5. Campbell, H.G., Dudek, R.A., Smith, M.L.: A heuristicalgorithm for the n job, m machine sequencing problem. Management Science 16(10), B630–B637 (1970)
6. Dannenbring, D.G.: An evaluation of flow shop sequencing heuristics. Management Scienc 23(1), 1174–1182 (1977)
7. Nawaz, M., Enscore Jr., E.E., Ham, I.: A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem. OMEGA: The International Journal of Management Science 11(1), 91–95 (1983)
8. Dong, X., Huang, H., Chen, P.: An improved NEH-based heuristic for the permutation flow shop problem. Computers & Operations Research 35(12), 3962–3968 (2008)
9. Ruiz, R., Stützle, T.: A simple and effective iterated greedy algorithm for the permutation flow shop scheduling problem. European Journal of Operational Research 177, 2033–2049 (2007)
10. Ekşioğlu, B., Ekşioğlu, S.D., Jain, P.: A tabu search algorithm for the flow shop scheduling problem with changing neighborhoods. Computers & Industrial Engineering 54(1), 1–11 (2008)
11. Liu, Y.F., Liu, S.Y.: A hybrid discrete artificial bee colony algorithm for permutation flowshop scheduling problem. Applied Soft Computing 13(3), 1459–1463 (2013)
12. Glover, F.: Future paths for integer programming and links to artificial intelligence. Computers and Operations Research 13, 533–549 (1986)
13. Ruiz, R., Maroto, C.: A comprehensive review and evaluation of permutation flowshop heuristics. European Journal of Operational Research: Project Management and Scheduling 165(2), 479–494 (2005)

14. Taillard, E.: Some efficient heuristic methods for the flowshop sequencing problem. European Journal of Operational Research 47, 65–74 (1990)
15. Deroussi, L., Gourgand, M., Norre, S.: Une adaptation efficace des mouvements de Lin et Kernighan pour le flow-shop de permutation. In: 8éme Conférence Internationale de MOdélisation et SIMulation: MOSIM, Hammamet, Tunisie (2010)
16. Taillard, E.: Benchmarks for basic scheduling problems. European Journal of Operational Research 64, 278–285 (1993)
17. Reeves, C., Yamada, T.: Genetic algorithms, path relinking and the flow shop sequencing problem. Evolutionary Computation 6, 45–60 (1998)

Neural-Based Method of Measuring Exchange-Rate Impact on International Companies' Revenue

Svitlana Galeshchuk[1,2]

[1] Faculty of Accounting and Audit, Ternopil National Economic University, Ternopil, Ukraine
[2] Université Lille 1 - Sciences et Technologies, Lille, France
svitlana.galeshchuk@gmail.com

Abstract. Measuring the impact of exchange-rate changes on the total revenue of an international company by the development of a neural-based method which predicts the company's revenue taking into account the currency exchange-rate changes is considered in this paper. The analysis of financial indices which influence on international company's revenue including the currency exchange-rate changes is fulfilled in order to provide input data for neural network training. The structures of multilayer perceptron and recurrent neural network are presented. The simulation modeling results show good quality prediction of the revenue of the international company Ryan Air taking into account the currency exchange-rate changes.

Keywords: Currency exchange-rate changes, revenue prediction, multilayer perceptron, recurrent neural network.

1 Introduction

Problems of financial markets, currency wars, and volatility in the world financial markets caused necessity to measure an impact of exchange-rate changes on the total revenue of international companies. Moreover, currency risk gets more significance with the number of currencies in which the company operates and subsidiaries in other countries. One of the methods to measure this impact is to provide good quality prediction of future revenue of an international company based on the past currency exchange-rate changes. Therefore the prediction of the companies' revenue taking into account the exchange-rate changes became an important tool to hedge the currency risks.

From macroeconomic point of view such prediction allows measuring the stake of monetary integration in the profitability of the international companies as well as the influence of the currencies volatility and central bank's decisions on the economic activities both in the external and internal markets. Also national regulation authorities can form appropriate national monetary policy knowing how the changes of exchange rates will be reflecting in the profits of the national-based companies which operates in the world markets. With the foreseen exchange rate adjustments international companies and investors can diversify their risks according to the assets

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_61, © Springer International Publishing Switzerland 2014

portfolio purposes. National economic bodies can take the prediction results into consideration as well in order to prevent a shift of the international money flows to the other markets and to develop attractive investment climate.

Prediction of different financial indices based on time series is one of the most urgent research directions in economic sciences today [1]. Last researches show that the application of computational intelligence methods as a base for predicting models, in particular, neural networks, evolutionary computations, fuzzy logic and others is very crucial due to their principal differences from existing mathematical approaches [2]. For example, neural networks have inherent learning abilities that allow effectively capturing the dynamic, nonlinear and complicated features of financial data; their self-training and self-adaptation properties provide a universality of the prediction model within the class of forecasting problems [3]. While the prediction model based on mathematical expression should be re-calculated for every prediction case. Finally, some prediction tasks in economics cannot be predicted by mathematical methods due to a number and diversity of input parameters.

In the last years, more than thousand scientific contributions have been proposed on the application of neural networks (NNs) in finance. The paper [4] provides a detailed survey of the 25 contributions available in scientific literature on the NNs application in the financial field for the last few years. Their conclusions have confirmed that NN-based solutions outperform other statistical and mathematic approaches in the most cases. The biggest number of NN applications in finance is for exchange rate [5-6], stock price [4, 7] and sales [8-9] predictions. The main purpose of sales forecasting is to estimate company's sales which are going to be in the future. Knowing the future company's sales allows making intelligent business decisions, planning appropriate financial activities and estimating expected profit/revenue indices. Several papers [10-11] also refer to revenue prediction and show that the revenue depends on different input financial indices.

The goal of this paper is to investigate the impact of exchange-rate changes on the total revenue of an international company by the development of a neural-based method which predicts the company's revenue taking into account the currency exchange-rate changes. The rest of the paper is organized as follows. The input data analysis and preparation for NN training is illustrated in Section 2. The NN models used for prediction are described in section 3. The results of prediction are presented in Section 4. Section 5 concludes this paper.

2 Input Data Analysis and Preparation

As a case study for our revenue prediction method development taking into account the currency exchange-rate changes we have chosen data about the revenue of international company Ryan Air for 2000-2013 years. All financial indices were downloaded from the Ryan Air financial reports [12]. The reports include the data on a quarter basis. Therefore we have downloaded and collected all available data about company's revenue for 52 quarters from Dec 31, 2000 until Sep 30, 2013.

As we research the revenue taking into account the currency exchange-rate changes, it is expedient to choose the GBP/EUR exchange rate in a case of Ryan Air Company because many operations of the company are fulfilled in GBP in England while the main sales mostly are in EUR. So we have collected the monthly average exchange rate GBP/EUR for all months from Oct 2000 till Sep 2013 using the web-sites [13, 14]. We have analyzed the monthly average exchange rates for three months in each quarter and noticed that the data are very similar in each month inside one quarter in the most cases. Therefore we should see experimentally, probably it is not necessary to use data about monthly average exchange rate for all months since they will bring the same information to NN. Instead it is necessary to use the monthly average exchange rate only for one (last) month within each quarter.

The analysis of Ryan Air financial reports (which includes about 100 different financial indices) has showed that the following indices are related to the international transactions which may be affected by the currencies exchange-rate changes. Thus these additional financial indices should be accounted in our model because they could better describe the impact of the currency exchange-rate changes to the revenue of an international company:

- Foreign exchange gain/(loss) – it is a parameter which shows whether the company got an income or lose some amount with the operations with foreign currencies in a current quarter;
- Trade receivables – it is an amount received by all operations from debtors in a current quarter.

All three parameters, the revenue, the foreign exchange gain/(loss) and the trade receivables are measured in Million Euro (mEuro).

The visualization of the mentioned financial indices and the monthly average exchange rate GBP/EUR in the last month of each quarter is depicted in Fig. 1. As we can see, the revenue has an obvious periodic character which is connected maybe with seasonal sales of Ryan Air Company. Thus, to predict this parameter we should provide its historical retrospective to the input of NN because we have to account its seasonal nature somehow. The foreign exchange gain (loss), the gain with positive data within Y axes and the loss with negative data within Y axes, has chaotic character, but it also should be used for NN training because it correlates with exchange rate change and shows whether the company gained or loosed some amounts in the analyzed quarter. The trade receivable and the monthly average exchange rate have increasing character and show why the increasing number of operations from debtors within permanently increasing average exchange rate lead to the increasing the trend of the revenue.

Thus our analysis shows that the considered four parameters could provide enough information for NN training in order to predict the revenue of international company taking into account the currency exchange-rate changes. The experimental researches using these data are considered in Section 4 below.

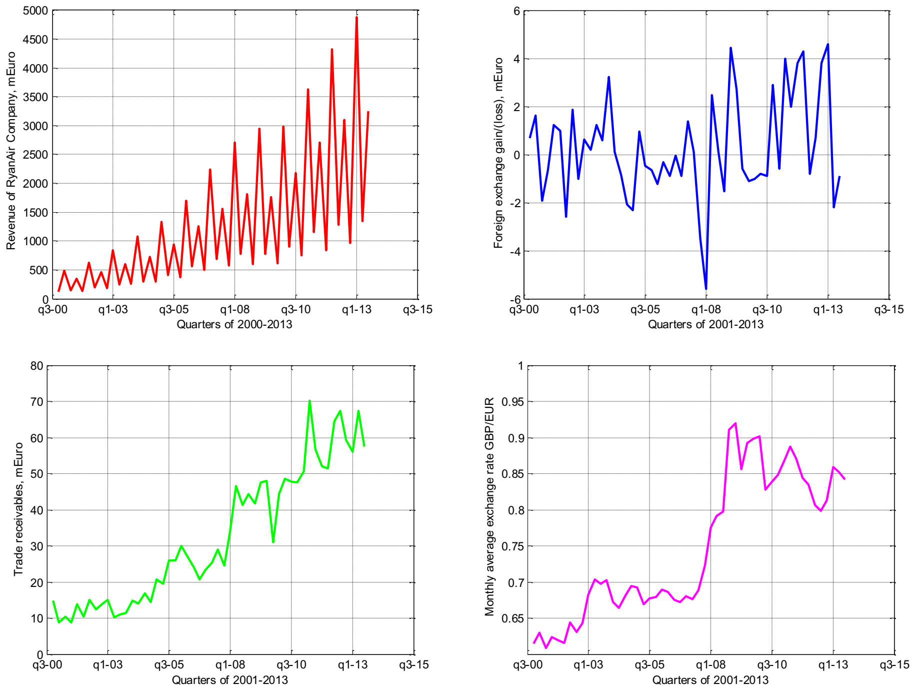


Fig. 1. Plots of financial indices of Ryan Air Company and monthly average exchange rate GBP/EUR

3 Neural Network Models

We have used two standard models of NNs: a multilayer perceptron, MLP (Fig. 2) and a recurrent neural network, RNN (Fig. 3) [3]. These models are well-researched and they are capable to fulfill approximation tasks with any required level of accuracy.

The output value of the three-layer perceptron can be formulated as:

$$y = F_3\left(\sum_{j=1}^{n} v_j h_j - b_3\right), \quad h_j = F_2\left(\sum_{i=1}^{m} w_{ij} x_i - b_{2j}\right),$$

where F_3, F_2 is the logistic activation function $F(x) = 1/(1+e^{-x})$ of the neurons of the output and hidden layer respectively, n is the number of neurons in the hidden layer, v_j is the weight coefficient from j-neuron of the hidden layer to the output neuron, h_j is the output value of j-neuron of the hidden layer, b_3 is the threshold of the output neuron, m is the number of neurons in the input layer, w_{ij} are the weight coefficients from the i-input neuron to j-neuron of the hidden layer, x_i are the input values, b_{2j} are the thresholds of the neurons of the hidden layer [3, 15].

The output value of RNN can be formulated as:

$$y = F_3\left(\sum_{j=1}^{n} v_j h_j - b_3\right),\; h_j = F_2\left(\sum_{i=1}^{m} w_{ij} x_i + \sum_{k=1}^{n} v'_{kj} h_k(t-1) + w3_j\, y(t-1) - b_{2j}\right),$$

where F_3, F_2, n, v_j, h_j, b_3, m, w_{ij}, x_i, b_{2j} have the same meaning as in the MLP above (the output value of j-neuron of the hidden layer h_j is computed in the different way), v'_{kj} is the synapse from k context neuron of the hidden layer to the j- neuron of the same (hidden) layer, $h_k(t-1)$ is the output value of k context neuron of hidden layer in the previous moment of time $t-1$, $w3_j$ is the synapse from context output neuron to the j-neuron of the hidden layer, $y(t-1)$ is the value of context output neuron in the previous moment of time $t-1$ [3, 16].

The standard batch pattern back propagation training algorithm is used for the training. We have used the software [17] implementing these models.

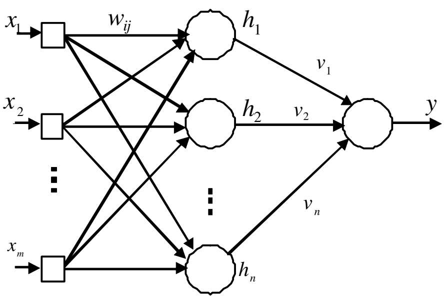


Fig. 2. Structure of three-layer perceptron

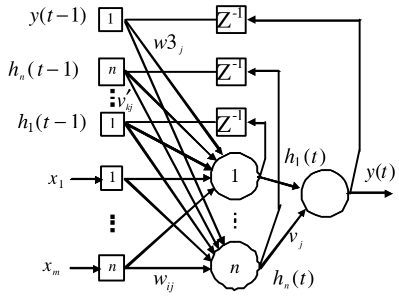


Fig. 3. Structure of recurrent neural network

4 Experimental Results

Experimental researches were fulfilled using both models of NNs with different numbers of hidden layer neurons and the data of Ryan Air Company described in Section II above. The data from 44 quarters within 2000-2011 years are used for the training, the predicting interval is chosen for 8 quarters: 4th quarter of 2011, all (four) quarters of 2012 and three (1st -3rd) quarters of 2013. A one-step prediction mode was used. The NNs were re-trained after each prediction step using the real values of current quarter. However in this experiment we have used monthly average exchange rate GBP/EUR not only for the last 3rd month within each quarter as it was specified in Section II above, but all three values of exchange rates for each month within each quarter. Therefore the number of input neurons in NN models is equal to 6. The best result was provided by RNN, 6 input neurons, 10 hidden neurons and 1 output neuron. The neurons of the hidden layer and the output neuron have logistic activation functions. The RNN was trained up to 10^{-4} sum-squared training error, the learning rate value was equal to 0,1. The average relative prediction error on 8 predicting quarters was 20,8%, the maximum prediction error was 70,5%.

This prediction result is not very good because the prediction errors are quite high, however the NN follows the trend of revenue quite well. This prediction mode does not account the seasonal change of revenue as we described in Section II above because the data about only one last quarter is given to the input of NN. Therefore it is expedient to put the data at least for one season, 4 quarters, to the input of NN simultaneously. For this purpose we have increased the number of inputs of NN to 16. Thus we have first four inputs the (i) revenue, (ii) foreign exchange gain (loss), (iii) trade receivable and (iv) monthly average exchange rate (3^{rd} month only) for the current quarter Q, second four inputs – the four parameters above for the previous quarter in the past Q-1, third four inputs – the four parameters above for the past quarter Q-2, and the last fourth four inputs – the four parameters above for the past quarter Q-3.

The experimental research for this new NN architecture is fulfilled by the same way as previous one. The data from 40 quarters within 2000-2011 years are used for the training due to extended numbers of input neurons. We have made several experiments with different number of hidden layer neurons to research the generalized properties of the NN. The MLP model 16-5-1 showed 17,2% of the average relative prediction error, RNN model 16-5-1 showed 27,2%, RNN model 16-10-1 showed 15,5%, MLP model 16-15-1 showed 14,3%, RNN model 16-15-1 showed 20,5% and MLP model 16-20-1 showed 16,3%. The best result was provided by MLP, 16 input neurons, 10 hidden neurons and 1 output neuron. The hidden layer neuron and the output neuron have logistic activation functions. The MLP was trained up to $2x10^{-4}$ sum-squared training error, the learning rate value was equal to 0,5. The average relative prediction error on 8 predicting quarters was 13,3%, the maximum prediction error was 48,8% (Fig 4.). The explanation q4-11 on the Fig. 4 means 4^{th} quarter of 2011 year. The relative prediction error of revenue of Ryan Air Company was less than 20% for 7 quarters from 8, only one quarter was predicted with relative prediction error 48,8%.

The analysis of the similar solution in paper [10] shows, that the prediction of the revenue of the MediaTek Corporation in Taiwan using a least-squares support vector regression model in a fuzzy environment has showed 14,74%, 18,34% and 19,68% (for different kernels of support vector machine) of the average relative prediction error on 12 predicting months. The work [10] is similar to our approach by the small available data used in the predicting model since the data can be gathered only from the monthly or quarterly financial reports of appropriate companies. We are using 40 data for the NN training and 48 data for the training is used in paper [10]. Our difference with the paper [10] is in a bit different goal of our research: we show the impact of exchange rate change on the revenue of an international company. That's why we are using additional parameters, foreign exchange gain (loss), trade receivable and monthly average exchange rate in our predicting model. Instead, the purely information about the revenue is used in paper [10].

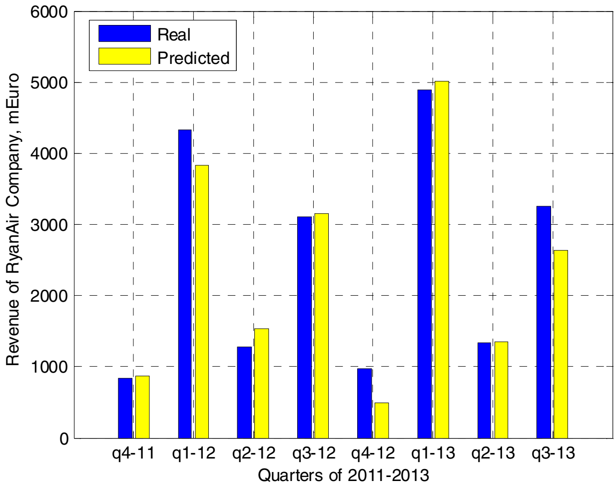


Fig. 4. Prediction results for revenue of Ryan Air Company for 2011-2013

5 Conclusions

The method of measuring the exchange-rate impact on the revenue of an international company is presented in this paper. The method is investigated by a neural-based prediction of Ryan Air Company's revenue taking into account the currency exchange-rate changes. The analysis of financial indices which influence on international company's revenue provides the input data for neural network training. The models of multilayer perceptron and recurrent neural network are used for experimental research. We have formed the input data for the neural network training by the special way trying to account the season nature of revenue of Ryan Air Company. The average relative prediction error of the company's revenue is 13,3% for 8 predicting quarters. The obtained prediction results allow us (i) to measure the impact of the exchange-rate changes on the total revenue of Ryan Air international company and (ii) to apply the proposed approach to investigate the impact of exchange-rate changes on the total revenue of other international companies. We consider the latter as our future research activity.

References

1. Štěpnička, M., Cortez, P., Peralta Donate, J., Štěpničková, L.: Forecasting seasonal time series with computational intelligence: On recent methods and the potential of their combinations. Expert Systems with Applications 40(6), 1981–1992 (2013)
2. `http://www.forecasts.org/help/faq.htm`

3. Haykin, S.: Neural Networks and Learning Machines, 3rd edn., p. 936. Prentice Hall (2008)
4. Gurusen, E., Kayakulu, G., Daim, T.U.: Using artificial neural network model in stock market index prediction. Expert Systems with Applications 38(8), 10389–10397 (2011)
5. Majhi, B., Rout, M., Majhi, R., Panda, G., Fleming, P.J.: New robust forecasting models for exchange rates prediction. Expert Systems with Applications 39(16), 12658–12670 (2012)
6. Zhang, Y.-Q., Wan, X.: Statistical fuzzy interval neural networks for currency exchange rate time series prediction. Applied Soft Computing 7(4), 1149–1156 (2007)
7. Turchenko, V., Beraldi, P., De Simone, F., Grandinetti, L.: Short-term Stock Price Prediction Using MLP in Moving Simulation Mode. In: Proc. 6th IEEE Intern. Conf. on Intelligent Data Acquisition and Advanced Computing Systems, Czech Republic, pp. 666–671 (2011)
8. Lu, C.J., Lee, T.S., Lian, C.M.: Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks. Decision Support Systems 54(1), 584–596 (2012)
9. Chang, P.C., Liu, C.H., Fan, C.Y.: Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry. Knowledge-Based Systems 22(5), 344–355 (2009)
10. Lin, K.P., Pai, P.F., Lu, Y.M., Chang, P.T.: Revenue forecasting using a least-squares support vector regression model in a fuzzy environment. Information Sciences 220, 196–209 (2013)
11. Zhang, L., Luo, J., Yang, S.: Forecasting box office revenue of movies with BP neural network. Expert Systems with Applications 36(3), 6580–6587 (2009)
12. `http://www.ryanair.com/index.php/en/investor/download/2013`
13. `http://www.x-rates.com/average/?from=EUR&to=GBP&amount=1.00&year=2009`
14. `http://www.oanda.com/currency/average`
15. Turchenko, V., Grandinetti, L.: Scalability of enhanced parallel batch pattern BP training algorithm on general-purpose supercomputers. In: de Leon F. de Carvalho, A.P., Rodríguez-González, S., De Paz Santana, J.F., Rodríguez, J.M.C. (eds.) Distributed Computing and Artificial Intelligence. AISC, vol. 79, pp. 518–526. Springer, Heidelberg (2010)
16. Turchenko, V., Grandinetti, L.: Parallel batch pattern BP training algorithm of recurrent neural network. In: Proc. 14th IEEE Intern. Conf. on Intelligent Engineering Systems, Spain, pp. 25–30 (2010)
17. `http://uweb.deis.unical.it/turchenko/research-projects/pagalinnet/`

Parallel Batch Pattern Training Algorithm for MLP with Two Hidden Layers on Many-Core System

Volodymyr Turchenko

Research Institute for Intelligent Computing Systems,
Ternopil National Economic University,
3 Peremogy Square, Ternopil, Ukraine
vtu@tneu.edu.ua

Abstract. The development of parallel batch pattern back propagation training algorithm of multilayer perceptron with two hidden layers and the research of its parallelization efficiency on many-core system are presented in this paper. The model of multilayer perceptron and batch pattern training algorithm are theoretically described. The algorithmic description of the parallel batch pattern training method is presented. Our results show high parallelization efficiency of the developed algorithm on many-core parallel system with 48 CPUs using MPI technology.

Keywords: Parallel batch pattern training, neural networks, parallelization efficiency.

1 Introduction

Artificial neural networks (NNs) represent a very good alternative to traditional methods for solving complex problems in many fields, including image processing, predictions, pattern recognition, robotics, optimization, etc [1]. However, NN models require high computational load in the training phase (on a range from several hours to several days) depending on the solving problem. This is, indeed, the main obstacle to face for an efficient use of NNs in real-world applications. The development of parallel algorithms to speed up the training phase of NNs is one of the ways to outperform this obstacle. Due to the last technological achievements, many-core high performance computing systems have a widespread use now in research and industry communities. Therefore the estimation of the parallelization efficiency of such parallel algorithms on many-core high performance systems is an actual research task.

Taking into account the parallel nature of NNs, many researchers have already focused their attention on NNs parallelization. The authors of [2] investigate parallel training of multi-layer perceptron (MLP) on SMP computer, cluster and computational grid using MPI (Message Passing Interface) parallelization. The development of parallel training algorithm of Elman's simple recurrent neural network based on Extended Kalman Filter on multicore processor and Graphic Processing Unit is presented in [3]. The authors of [4] have presented the development of parallel training algorithm of fully

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_62, © Springer International Publishing Switzerland 2014

connected RNN based on linear reward penalty correction scheme. In our previous works, within the development of the parallel grid-aware library for neural networks training [5], we have developed the batch pattern back propagation (BP) training algorithm for the multi-layer perceptron with one hidden layer [6], recurrent neural network [7], recirculation neural network [8] and neural network with radial-basis activation function [9] and showed their good parallelization efficiency on different high performance computing systems.

However the analysis of the state-of-the-art has showed that the parallelization of the model of MLP with two hidden layers of neurons was not properly investigated yet. For example, the authors of [2] parallelized the MLP architecture 16-10-10-1 (16 neurons in the input layer, two hidden layers with 10 neurons in each layer and one output neuron) on the huge number of the training patterns (around 20000) coming from Large Hadron Collider. Their implementation of this relatively small NN with 270 internal connections (number of weights of neurons and their thresholds) does not provide positive parallelization speedup due to large communication overhead, i.e. the speedup is less than 1. In our opinion this overhead is caused by the fact, that the "communication part" of their algorithm is not optimized and contains three separate messages.

According to the theorem of universal approximation [10-11], the MLP with two hidden layers provides better control on approximation process instead the MLP with one hidden layer [1]. Also the MLP with two hidden layers allows processing the "global features", i.e. the neurons of the first hidden layer gather the "local features" from input data and the neurons of the second hidden layer generalized their outputs providing higher level representation. This is especially urgent for solving object recognition and computer vision tasks. Therefore the research of efficiency of parallel training algorithm for MLP with two hidden layers of neurons is an urgent research problem.

Taking into account that the batch pattern BP training scheme showed good parallelization efficiency on the number of NN architectures, the goal of this paper is to apply this scheme for the MLP with two hidden layers and assess whether its parallelization efficiency is better/worse than the efficiency of the MLP with one hidden layer. The rest is organized as follows: Section 2 details the mathematical description of batch pattern BP algorithm, Sections 3 describes its parallel implementation, Section 4 presents the obtained experimental results and Section 5 concludes this paper.

2 Batch Pattern BP Training Algorithm for Multilayer Perceptron with Two Hidden Layers

The batch pattern training algorithm updates neurons' weights and thresholds at the end of each training epoch, i.e. after all training patterns processing, instead of updating weights and thresholds after processing of each pattern in the sequential training mode [6].

The output value of the MLP with two hidden layers (Fig. 1) is described by [1]:

$$y_3 = F_3(S_3),\ S_3 = \sum_{k=1}^{K} y_{2k} \cdot w_{3k} - T_3,\ y_{2k} = F_2(S_{2k}),\ S_{2k} = \sum_{j=1}^{N} y_{1j} \cdot w_{2jk} - T_{2k}\ , \quad (1)$$

$$y_{1j} = F_1(S_{1j}),\ S_{1j} = \sum_{i=1}^{M} x_i \cdot w_{1ij} - T_{1j}\ , \quad (2)$$

where F_3, F_2 and F_1 are the logistic activation functions $F(x) = 1/(1+e^{-x})$ which used for the neurons of the output and two hidden layers respectively, S_3, S_{2k} and S_{1j} are the weighed sums of the neurons of the output and two hidden layers respectively, K, N and M are the number of neurons of two hidden and input layers respectively, y_{2k} are the outputs of the neurons of the second hidden layer, w_{3k} are the weight coefficients from the k-neuron of the second hidden layer to the output neuron, T_3 is the threshold of the output neuron, y_{1j} are the outputs of the neurons of the first hidden layer, w_{2jk} are the weights from the neurons of the first hidden layer to the neurons of the second hidden layer, T_{2k} are the thresholds of the neurons of the second hidden layer, x_i are the input values, w_{1ij} are the weights from the input neurons to the neurons of the first hidden layer and T_{1j} are the thresholds of the neurons of the first hidden layer.

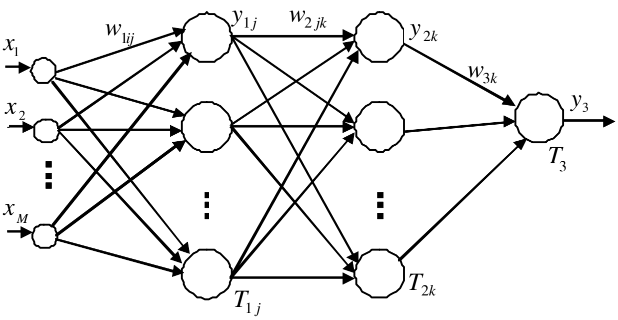


Fig. 1. The structure of a MLP with two hidden layers

The batch pattern BP training algorithm for this MLP consists of the following steps:

1. Set the desired Sum Squared Error SSE= $E_{\min}$ and the number of training epochs t;
2. Initialize the weights and thresholds of the neurons with values in range (0…0.5) [13];
3. For the training pattern pt:

3.1. Calculate the output value $y_3^{pt}(t)$ by expressions (1), (2);

3.2. Calculate the error of the output neuron $\gamma_3^{pt}(t) = y_3^{pt}(t) - d^{pt}(t)$, where $y_3^{pt}(t)$ is the output value of the MLP and $d^{pt}(t)$ is the target output value;

3.3. Calculate the errors of the second hidden layer neurons' $\gamma_{2k}^{pt}(t) = \gamma_3^{pt}(t) \cdot w_{3k}(t) \cdot F_3'(S_3^{pt}(t))$, where $S_3^{pt}(t)$ is the weighted sum of the output neuron and the errors of the first hidden layer neurons' $\gamma_{1j}^{pt}(t) = \sum_{k=1}^{K} \gamma_{2k}^{pt}(t) \cdot w_{2jk}(t) \cdot F_2'(S_{2k}^{pt}(t))$, where $S_{2k}^{pt}(t)$ are the weighted sums of the k -neuron of the second hidden layer;

3.4. Calculate the delta weights and delta thresholds of all neurons and add the result to the value of the previous pattern

$$s\Delta w_{3k} = s\Delta w_{3k} + \gamma_3^{pt}(t) \cdot F_3'(S_3^{pt}(t)) \cdot y_{2k}^{pt}(t),\ s\Delta T_3 = s\Delta T_3 + \gamma_3^{pt}(t) \cdot F_3'(S_3^{pt}(t)), \quad (3)$$

$$s\Delta w_{2jk} = s\Delta w_{2jk} + \gamma_{2k}^{pt}(t) \cdot F_2'(S_{2k}^{pt}(t)) \cdot y_{1j}^{pt}(t),\ s\Delta T_{2j} = s\Delta T_{2j} + \gamma_{2k}^{pt}(t) \cdot F_2'(S_{2k}^{pt}(t)), \quad (4)$$

$$s\Delta w_{1ij} = s\Delta w_{1ij} + \gamma_{1j}^{pt}(t) \cdot F_1'(S_{1j}^{pt}(t)) \cdot x_i^{pt}(t),\ s\Delta T_{1j} = s\Delta T_{1j} + \gamma_{1j}^{pt}(t) \cdot F_1'(S_{1j}^{pt}(t)); \quad (5)$$

3.5. Calculate the SSE using $E^{pt}(t) = \frac{1}{2}\left(y_3^{pt}(t) - d^{pt}(t)\right)^2$; (6)

4. Repeat the step 3 above for each training pattern pt, where $pt \in \{1,...,PT\}$, PT is the size of the training set;
5. Update the weights and thresholds of neurons using the expressions $w_{3k}(PT) = w_{3k}(0) - \alpha_3(t) \cdot s\Delta w_{3k}$, $T_3(PT) = T_3(0) + \alpha_3(t) \cdot s\Delta T_3$, $w_{2jk}(PT) = w_{2jk}(0) - \alpha_2(t) \cdot s\Delta w_{2jk}$, $T_{2k}(PT) = T_{2k}(0) + \alpha_2(t) \cdot s\Delta T_{2k}$, $w_{1ij}(PT) = w_{1ij}(0) - \alpha_1(t) \cdot s\Delta w_{1ij}$, $T_{1j}(PT) = T_{1j}(0) + \alpha_1(t) \cdot s\Delta T_{1j}$ where $\alpha_3(t)$, $\alpha_2(t)$ and $\alpha_1(t)$ are the learning rates;
6. Calculate the total SSE $E(t)$ on the training epoch t using $E(t) = \sum_{pt=1}^{PT} E^{pt}(t)$;
7. If $E(t)$ is greater than the desired error $E_{\min}$ then increase the number of training epochs to $t+1$ and go to step 3, otherwise stop the training process.

3 Parallel Batch Pattern BP Training Algorithm of MLP with Two Hidden Layers

Similarly to the MLP with one hidden layer [6], the sequential execution of steps 3.1-3.5 for all training patterns is parallelized, because the sum operations $s\Delta w_{3k}$, $s\Delta T_3$, $s\Delta w_{2jk}$, $s\Delta T_{2k}$, $s\Delta w_{1ij}$ and $s\Delta T_{1j}$ are independent of each other. The computational work is divided among the *Master* (executing assigning functions and calculations) and the *Workers* (executing only calculations) processors. The flowchart of the parallel batch pattern BP training algorithm of MLP with two hidden layers is similar to the flowchart of the parallel algorithm of MLP with one hidden layer presented in

[6, 14, 15, 17] except of additional deltas that should be calculated within the step 3.4. Therefore we provide only textual description of the parallel algorithm in this section.

The *Master* starts with definition (i) the number of patterns *PT* in the training data set and (ii) the number of processors *p* used for the parallel execution of the training algorithm. The *Master* divides all patterns in equal parts corresponding to the number of the *Workers* and assigns one part of patterns to itself. Then the *Master* sends to the *Workers* the numbers of the appropriate patterns to train.

Each *Worker* executes the following operations for each pattern *pt*:

1. calculate the steps 3.1-3.5 and 4, only for its assigned number of training patterns. The values of the partial sums of delta weights $s\Delta w_{3k}$, $s\Delta w_{2jk}$, $s\Delta w_{1ij}$ and delta thresholds $s\Delta T_3$, $s\Delta T_{2k}$, $s\Delta T_{1j}$ are calculated there;
2. calculate the partial SSE for its assigned number of training patterns.

After processing all assigned patterns, only one all-reduce collective communication operation with summation is executed automatically providing the synchronization with other processors by its internal implementation [14]. Then the summarized values $s\Delta w_{3k}$, $s\Delta T_3$, $s\Delta w_{2jk}$, $s\Delta T_{2k}$, $s\Delta w_{1ij}$, $s\Delta T_{1j}$ are sent to all processors working in parallel and placed into the local memory of each processor. Each processor uses these values for updating the weights and thresholds according to the step 5 of the algorithm to use them in the next epoch of the training. As the summarized value of $E(t)$ is also received as a result of this operation, the *Master* decides whether to continue the training or not.

The software is developed using C programming language with the standard MPI functions. The parallel part of the algorithm starts with the call of *MPI_Init()* function. The *MPI_Allreduce()* function reduces the deltas of weights $s\Delta w_{3k}$, $s\Delta w_{2jk}$, $s\Delta w_{1ij}$ and thresholds $s\Delta T_3$, $s\Delta T_{2k}$, $s\Delta T_{1j}$, summarizes them and sends them back to all processors in the group. Since the weights and thresholds are physically located in the different matrixes of the routine, we have done pre-encoding of all data into one communication message before sending and reverse post-decoding the data to the appropriate matrixes after receiving the message in order to provide only one physical call of the function *MPI_Allreduce()* in the communication section of the algorithm. Function *MPI_Finalize()* finishes the parallel part of the algorithm.

4 Experimental Design

The many-core parallel supercomputer *Remus*, located in the Innovative Computing Lab, the University of Tennessee, USA is used for the research. *Remus* consists of two socket G34 motherboards RD890 (AMD 890FX chipset) connected each other by AMD Hyper Transport. Each motherboard contains two twelve-core AMD Opteron 6180 SE processors with 2500 MHz clock rate and 132 GB of local RAM. Thus the total number of computational cores is 48. Each processor has the L2 cache of 12x512 Kb and the L3 cache of 2x6 Mb. We run the experiments using message passing library Open MPI 1.6.3 [15].

In our previous research [16] we have used many parallelization scenarios starting from very small MLP 5-5-1 (36 connections) and 100 training patterns to huge MLP 60-60-1 (3721 connections) and 10000 patterns in order to investigate small and large parallelization problems and to assess the scalability of the parallel algorithm [6]. These experiments were fulfilled on SMP machines, ccNuma, Cluster Gigabit Ethernet and Cluster Infiniband architectures. There were no experiments for MLP on many-core architecture, except of RCNN only [8]. Thus it is not possible to use the results obtained on other architectures in the past in order to compare them accurately with the efficiency of the MLP with two hidden layers of neurons. Therefore the scenarios in Table 1 are chosen for the experimental research. These scenarios are similar with previous researches in order to have the possibility to compare the results between different parallel architectures for the case of MLP with one hidden layer. The architectures of MLPs were chosen in order to provide similar size of the communication message and to assess the parallelization efficiency taking into account only the computational part of both algorithms.

Table 1. The scenarios for the experimental research

Scenario	MLP with one hidden layer	MLP with two hidden layers
Small	16-7-1, 127 conn., 200 patterns	16-5-5-1, 121 conn., 200 patterns
Middle	16-22-1, 397 conn., 1000 patterns	16-13-13-1, 417 conn., 1000 patterns
Large	16-145-1, 2611 conn., 5000 patterns	16-50-34-1, 2619 conn., 5000 patterns

The real data about prediction of revenue of an international company taking into account the exchange-rate changes were used as training patterns. The number of input neurons in all scenario is equal to 16 because of 16 input parameters are analyzed. The neurons of the hidden and output layers have logistic activation function. The number of training epochs is fixed to $3{\cdot}10^4$. The learning rates are constant and equal $\alpha_1(t) = 0.05$, $\alpha_2(t) = 0.01$ for the MLP with one hidden layer and $\alpha_1(t) = 0.1$, $\alpha_2(t) = 0.1$, $\alpha_3(t) = 0.01$ for the MLP with two hidden layers.

The expressions $S=Ts/Tp$ and $E=S/p\times100\%$ are used to calculate the speedup and efficiency of parallelization, where Ts is the time of sequential executing the routine, Tp is the time of executing the parallel version of the same routine on p processors of parallel system. We have calculated the values of speedup and parallelization efficiency for each parallelization scenario from Table 1 on 2, 4, 8, 16, 32 and 48 cores for both MLPs and depicted the dependence of parallelization efficiency on the number of cores of multi-core system. The results (Fig. 2) showed (i) high parallelization efficiency of the MLP with two hidden layers of neurons: more than 93% on 48 cores for the large parallelization scenario and more than 76% on 48 cores for the middle parallelization scenario and (ii) that the parallelization efficiency is decreasing with increasing the number of cores. The parallelization efficiency of the small parallelization scenario remains more than 50% up to execution on 8 cores, the efficiency has lower values up to 10% while using up to 48 cores for the small scenario. The parallelization efficiency of the MLP with two hidden layers is bigger on several percents in average (1.2% for the small scenario, 2.6% for the middle scenario and 3.6% for the large scenario) than the efficiency of the MLP with one hidden layer at the same size of the communication message.

The theoretical analysis of the computational complexity of the computational parts of both algorithms using the BSP approach [17] showed that the MLP with two hidden layers has the bigger number of operations (for example, on 936K operations for the middle parallelization scenario). Therefore its training algorithm theoretically should have better parallelization efficiency than the algorithm for MLP with one hidden layer. The results of the experimental research correspond to this theoretical investigation.

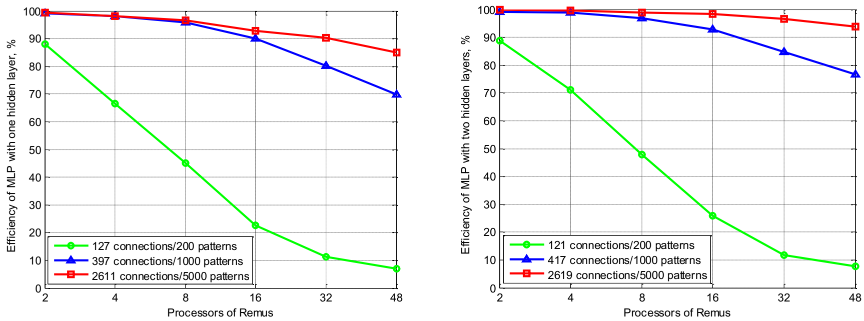


Fig. 2. Parallelization efficiencies for MLPs on many-core architecture

5 Conclusions

The development of parallel batch pattern back propagation training algorithm of multilayer perceptron with two hidden layers and its parallelization efficiency research on many-core system are presented in this paper. The model of multilayer perceptron and the batch pattern training algorithm are theoretically described. The algorithmic description of the parallel batch pattern training method is presented. Our results show high parallelization efficiency of the developed algorithm on many-core parallel system with 48 CPUs using MPI technology: the parallelization efficiency is more than 93% for the large parallelization scenario with 2619 connections and 5000 patterns and more than 76% for the middle parallelization scenario with 417 connections and 1000 patterns. The parallelization efficiency of MLP with two hidden layers is bigger on several percents in average (1.2% for the small scenario, 2.6% for the middle scenario and 3.6% for the large scenario) than the efficiency of MLP with one hidden layer for the scenarios with the same size of the communication message.

The future direction of research can be considered as an investigation of parallelization efficiency of parallel training algorithms for other multilayer architectures, in particular convolution neural network within data classification and recognition tasks.

References

1. Haykin, S.: Neural Networks and Learning Machines, 3rd edn., p. 936. Prentice Hall (2008)
2. De Llano, R.M., Bosque, J.L.: Study of neural net training methods in parallel and distributed architectures. Future Generation Computer Systems 26(2), 183–190 (2010)
3. Čerňanský, M.: Training recurrent neural network using multistream extended Kalman filter on multicore processor and CUDA enabled GPU. In: Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G. (eds.) ICANN 2009, Part I. LNCS, vol. 5768, pp. 381–390. Springer, Heidelberg (2009)
4. Lotrič, U., Dobnikar, A.: Parallel implementations of recurrent neural network learning. In: Kolehmainen, M., Toivanen, P., Beliczynski, B. (eds.) ICANNGA 2009. LNCS, vol. 5495, pp. 99–108. Springer, Heidelberg (2009)
5. `http://uweb.deis.unical.it/turchenko/research-projects/pagalinnet/`
6. Turchenko, V., Grandinetti, L.: Scalability of enhanced parallel batch pattern BP training algorithm on general-purpose supercomputers. In: de Leon F. de Carvalho, A.P., Rodríguez-González, S., De Paz Santana, J.F., Rodríguez, J.M.C. (eds.) Distributed Computing and Artificial Intelligence. AISC, vol. 79, pp. 525–532. Springer, Heidelberg (2010)
7. Turchenko, V., Grandinetti, L.: Parallel batch pattern BP training algorithm of recurrent neural network. In: Proc. 14th IEEE Intern. Conf. on Intelligen. Engin. Syst., Spain, pp. 25–30 (2010)
8. Turchenko, V., Bosilca, G., Bouteiller, A., Dongarra, J.: Efficient Parallelization of Batch Pattern Training Algorithm on Many-core and Cluster Architectures. In: Proceedings of the 7th IEEE International Conference IDAACS 2013, Germany, pp. 692–698 (2013)
9. Turchenko, V., Golovko, V., Sachenko, A.: Parallel training algorithm for radial basis function neural network. In: Proceedings of the 7th ICNNAI, Belarus, pp. 47–51 (2012)
10. Funahashi, K.: On the approximate realization of continuous mappings by neural network. Neural Networks 2, 183–192 (1989)
11. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366 (1989)
12. Sontag, E.: Feedback stabilization using two-hidden-layer nets. IEEE Transactions on Neural Networks 3, 981–990 (1992)
13. Golovko, V., Galushkin, A.: Neural Networks: Training, Models and Applications, Moscow, Radiotechnika (2001) (in Russian)
14. Turchenko, V., Grandinetti, L., Bosilca, G., Dongarra, J.: Improvement of parallelization efficiency of batch pattern BP training algorithm using Open MPI. Procedia Computer Science 1(1), 525–533 (2010)
15. Turchenko, V., Grandinetti, L., Sachenko, A.: Parallel batch pattern training of neural networks on computational clusters. In: Proc. of the 2012 Intern. Conf. HPCS 2012, Spain, pp. 202–208 (2012)
16. `http://www.open-mpi.org/`
17. Turchenko, V., Grandinetti, L.: Application of BSP-based computational cost model to predict parallelization efficiency of MLP training algorithm. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010, Part III. LNCS, vol. 6354, pp. 327–332. Springer, Heidelberg (2010)

A Bee-Inspired Data Clustering Approach to Design RBF Neural Network Classifiers

Dávila Patrícia Ferreira Cruz[1], Renato Dourado Maia[2], Leandro Augusto da Silva[1], and Leandro Nunes de Castro[1]

[1] Natural Computing Laboratory (LCoN), Graduate Program in Electrical Engineering, Mackenzie University, São Paulo, Brazil
davilapatricia11@gmail.com,
{prof.leandro.augusto,lnunes}@mackenzie.br
[2] Computer Science Department, State University of Montes Claros, MG, Brazil
renato.dourado@unimontes.br

Abstract. Different methods have been used to train radial basis function neural networks. This paper proposes a bee-inspired algorithm to automatically select the number and location of basis functions to be used in such RBF network. The algorithm was designed to solve data clustering problems, where the centroids of clusters are used as centers for the RBF network. The approach presented in this paper is preliminary evaluated in three synthetic datasets, two classification datasets and one function approximation problem, and its results suggest a potential for real-world application.

Keywords: RBF, optimal data clustering, neural network, bee-inspired algorithms.

1 Introduction

Radial basis function (RBF) neural networks have been used in a variety of domains, from function approximation [1] and pattern classification [2], to time series prediction [3]. RBF present two main parameters to be adjusted [4]: 1) the number, location, shape, and width of basis functions; 2) the output layer weights. This paper proposes the use of cOptBees, a bee-inspired clustering algorithm [5, 6], to select the number and locations of the centers of radial basis functions in RBF networks.

cOptbees, an adaptation of the OptBees optimization algorithm [7], was designed to solve data clustering problems and is inspired by the collective decision-making process in bee colonies. OptBees is able to generate and maintain the diversity of solutions by finding multiple suboptimal solutions in a single run, a feature useful for solving multimodal optimization problems.

This paper is organized as follows: Section 2 briefly reviews RBF networks; Section 3 introduces cOptBees and applies it to design RBF networks; Section 4 presents the experimental results; and Section V brings the conclusions.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_63,

2 Radial Basis Function Networks

In a RBF network there is a single set of adjustable weights between the units of the hidden layer and the units of the output layer (Fig. 1). For an input vector $\mathbf{x} = [x_1, x_2, ... ,x_d]$, where $\mathbf{x} \in \mathbf{X}$, $\mathbf{X} \in \mathbb{R}^{N \times d}$, and X is the training set, presented to the input layer, the RBF output, **y**, is the inner product between the adjustable weight vector, **w**, and the vector of responses of the basis functions, g, $\langle \mathbf{w}, \mathbf{g} \rangle$:

$$\mathbf{y} = \mathbf{w}_i^T \mathbf{g}, \tag{1}$$

where $\mathbf{w}_i = [w_{i1}, ..., w_{im}]^T$, $i = 1,...,o$, are the network output weight vectors, $\mathbf{g} = [g_1, g_2, ... ,g_m]^T$ is the vector of basis functions, and o is the number of RBF outputs.

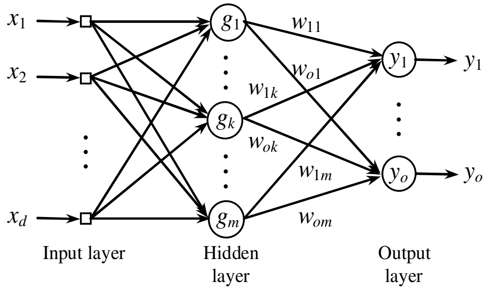


Fig. 1. RBF network with multiple outputs

Given a set of center vectors $\mathbf{c}_j \in \mathbb{R}^d$ and their corresponding standard deviations σ_j, $j = 1,..., m$, the output of each radial basis function is

$$g_j = h(\|\mathbf{x} - \mathbf{c}_j\|, \sigma_j), j = 1, \ldots, m, \tag{2}$$

where $h_j(\cdot)$ is the basis function and $\|\cdot\|$ is a norm defined on the input space.

One form of finding a regularized solution to the problem of defining the output layer weight is by using the following pseudo-inverse [8, 9, 10, 11, 12]:

$$\mathbf{w}^* = \mathbf{H}^+ \mathbf{d} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{d}, \tag{3}$$

where $\mathbf{w}^*$ is the weight matrix that solves a least-squares data-fitting problem, $\mathbf{H}$ is an N-by-m matrix, called interpolation matrix, and $\mathbf{H}^+$ its pseudo-inverse [10].

It has been argued that experience with this method showed it is relatively insensitive to the choice of regularization parameters, as far as an appropriate choice of the RBF centers is performed [13]. Therefore, the approach used here to train a RBF neural network is to assume fixed radial basis functions as the activation of the hidden units. The number and locations of the centers will be chosen using a bee-inspired clustering algorithm to be described in the next section. A Gaussian function will be used for the radial basis with fixed standard deviation according to the spread of the centers: $\sigma_j = d_{\max}.(2.m_1)^{-1/2}$, where $j = 1,...,m_1$, is the number of centers, σ_j is the standard deviation (equal for all basis functions), and $d_{\max}$ is the maximum distance between the chosen centers. This choice for σ_j guarantees that the individual radial basis functions are not too peaked or too flat; cases that should be avoided [13].

3 A Bee-Inspired Algorithm for Optimal Data Clustering

The cOptBees [6] is an algorithm that solves data clustering problems inspired by the foraging behavior of bee colonies. This algorithm is an adaptation of OptBees, originally designed to solve continuous optimization problems and, additionally, it is able to generate and maintain the diversity of candidate solutions, finding multiple local optima without compromising its global search capability [7]. The main features and behavior of OptBees were successfully adapted to solve clustering problems [5], where the generation and maintenance of diversity are important. The OptBees algorithm is summarized in Table 1.

Table 1. ObtBees Algorithm

OptBees Algorithm
Input Parameters:
• n_{min}: initial number of active bees.
• n_{max}: maximum number of active bees.
• ρ: inhibition radius.
• α: recruitment rate.
• n_{mean}: average foraging effort.
• p_{min}: minimum probability of a bee being a recruiter.
• p_{rec}: percentage of non-recruiter bees that will be actually recruited.
Output Parameters:
• Active bees and the respective values of the objective function.
1. Randomly generate a swarm of N bees.
2. **while** (stopping criterion is not attained) **do**
3. Evaluate the quality of the sites being explored by the active bees.
4. Apply local search.
5. Determine the recruiter bees.
6. Update the number of active bees.
7. Determine the recruited and scout bees.
8. Perform the recruitment process.
9. Perform the exploration process.
end while
10. Evaluate the quality of the sites being explored by the active bees.
11. Apply local search.

In cOptBees, as in OptBees, the active bees can belong to one of three groups, according to their task: 1) recruiters, who attract other bees to explore a promising region of the search space; 2) recruited, who are recruited by recruiters to explore a promising region of the search space; or 3) scout bees, who randomly look for new promising regions of the space [5]. The main features of cOptBees are described in the following sections.

3.1 How cOptBees Work

In cOptBees, each bee is composed of a set of prototypes that encodes a potential clustering. A bee is defined by a matrix $\mathbf{B}_{i \times j}$, where $i = (d + 1)$, d being the number of attributes of the input data, and j is the maximum number of clusters in a clustering (*rMax*). Thus, in a given column *j,* lines 1 to *d* represent the dimension of prototype $\mathbf{C}_j$ and the last line represents a threshold value, $L_j \in [0,1]$, that defines if the centroid $\mathbf{C}_j$ is active or not. The centroid $\mathbf{C}_j$ is active when its threshold is greater than or equal to 0.5. Fig. 1 shows the matrix representation of a bee [6], [14].

$$B = \begin{bmatrix} C_{1,1} & \cdots & C_{1,rMax} \\ C_{2,1} & \cdots & C_{2,rMax} \\ \vdots & \cdots & \vdots \\ C_{d,1} & \ddots & C_{d,rMax} \\ L_1 & \cdots & L_{rMax} \end{bmatrix}$$

Fig. 2. Matrix representation of a Bee in cOptBees

The swarm is composed of *N* bees and, for each bee, the objects in the database are associated with the nearest prototype. The initial swarm is randomly generated, respecting the maximum number of clusters, *rMax* (an input parameter introduced in cOptBees) [6].

The recruiter bees explore promising regions of the search space and recruit the closest bees. The number of recruited bees for each recruiter is proportional to the quality of the food sources found in the search space. Determining the recruiter bees involves three steps. In the first step, a probability p_i of being a recruiter bee is associated with each active bee [7]:

$$p_i = \left(\frac{1-p_{min}}{Q_{max}-Q_{min}}\right) \cdot \ (q_i - Q_{min}) + p_{min}, \qquad (4)$$

where q_i represents the quality of the site being explored by bee i, Q_{min} and Q_{max} represent, respectively, the minimum and maximum qualities among the sites being explored by each active bee in the current iteration (these quality values are determined using the objective-function value) [7].

In the second step the bees are processed and, according to the probabilities calculated in the previous step, are now classified as recruiters or non-recruiters. In the third step, the recruiter bees are processed in accordance with the corresponding site qualities, from best to worst and, for each recruiter bee, the other recruiters who have a high similarity are inhibited, i.e., they are classified as non-recruiters. The inhibition process happens when the similarity between two bees is greater than or equal to the inhibition radius ρ, an input parameter that represents a percentage of the maximum possible value for the similarity [7].

Updating the number of active bees aims to adapt the foraging effort in accordance with the number of recruiters and the maximum number of active bees. After the determination of the recruiter bees n_r in a given iteration, the number $n_d = (n_r + 1) \cdot n_{mean}$ determines the desired number of active bees, where n_{mean}, the average foraging effort, determines the desired number of non-recruiter bees for each recruiter bee. If n_d is greater than the current number of active bees, $n_{adjust} = n_d - n_{active}$ is the necessary number of bees that have to become active in order to achieve n_d active bees; if this number is less than the current number of active bees, $n_{adjust} = n_{active} - n_d$ is the necessary number of bees that have to become inactive in order to achieve n_d active bees. This process is constrained by the maximum (n_{max}) and minimum (n_{min}) numbers of active bees [7].

After the classification of bees as recruiter or non-recruiter, a percentage of non-recruiter bees are classified as recruited and exploit promising regions already found. The other non-recruiters are classified as scout bees and explore the search space to find new promising regions [7].

The number of non-recruiter bees, in Step 6, is determined by $n_{nr} = n_{active} - n_r$. The number of recruited bees is $n_r = [p_{rec} \cdot n_{nr}]$, where p_{rec} is the percentage of non-recruiter bees that will be recruited and [.] denotes the nearest integer function. The number of scout bees is $n_s = n_{nr} - n_r$. The process for determining the recruited bees works as follows. First, the number of recruited bees to be associated with each recruiter is determined. The relative quality of the site operated by each recruiter in relation to the others determines this number: each recruiter recruits a number of bees proportional to the quality of the site that it explores. With these numbers already determined, the non-recruiter bees are processed and associated with the most similar recruiter. After these procedures, the remaining n_s non-recruiter bees are considered scout bees [7].

In the recruitment process, the recruiter bees attract the recruited bees to the sites they explore. This recruitment process is implemented by Eq. (5) or (6), each with 50% probability, where α is the recruitment rate, an input parameter, $\mathbf{B}_i$ is the recruited bee, Y is the recruiter bee, u is a random number with uniform distribution in the interval [0, 1], U is a matrix whose elements are random numbers with uniform distribution in the interval [0, 1] (U has the same dimension as $\mathbf{x}_i$ and $\mathbf{y}$) and the symbol ⊗ denotes the element-wise product [7].

$$\mathbf{B}_i = \mathbf{B}_i + u \cdot \alpha \cdot (\mathbf{Y} - \mathbf{B}_i) \tag{5}$$

$$\mathbf{B}_i = \mathbf{B}_i + \alpha .\cdot \mathbf{U} \otimes (\mathbf{Y} - \mathbf{B}_i) \tag{6}$$

In the exploration process, the scout bees are moved to a random position in the search space. By doing this, the exploration process allows the scout bees to explore new regions in the search space [6].

3.2 Combining cOptBees with an RBF Neural Network

The goal of this paper is to use cOptBees to automatically select the number and location of basis functions to be used in the RBF network. The clusters centroids found by cOptBees are used as centers for the RBF network.

The approach proposed in this paper can be summarized as follows:

1. Apply the cOptBees clustering algorithm to the input training data set. The centroids found by cOptBees provide the number and location of centers for the hidden layer of the RBF network;
2. Determine the output of the hidden units (Eq. (2)); and
3. Determine the weights in the linear part of the RBF network according to Eq. (3).

4 Experimental Results

Some experiments were performed to preliminary evaluate the use of cOptBees to select centers for RBF networks. It was applied to three artificial two-dimensional datasets. cOptBees was run for all datasets with the following input parameters: n_{min} = 50; n_{max}= 100; n_{min} = 10; p_{min} = 0.01; p_{rec} linearly varying between 0.1 and 0.5 with the number of iterations; ρ linearly varying between 0.1 and 0.5 with the number of iterations; number of iterations = 50 and α = 0.5. After training, the hybrid approach was run to classify a dense dataset generated randomly, so that one can observe the decision boundaries formed by the objects.

For the first dataset the hybrid algorithm proposed in this paper was run with *rMax* = 20. The results are presented in Fig. 3.

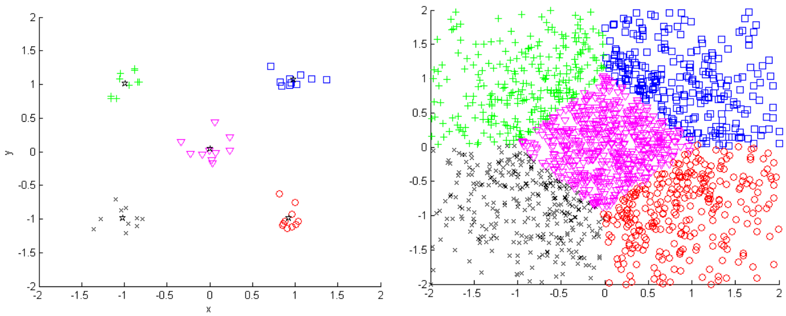


Fig. 3. Training input vectors and RBF centers (black stars) obtained with cOptBees. On the right side is the decision boundary found.

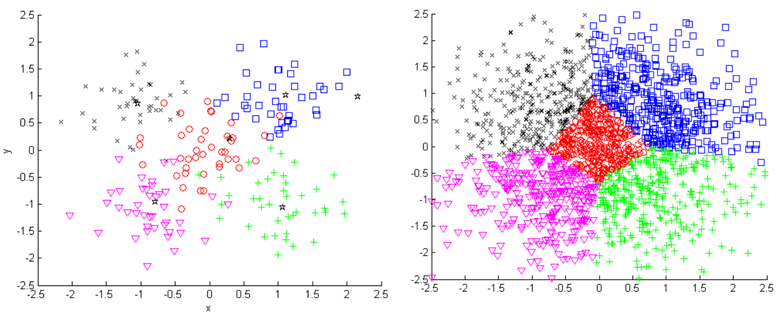


Fig. 4. Training input vectors and RBF centers (black stars) obtained with cOptBees. On the right side is the decision boundary found.

The second dataset is similar to the first one, but present an overlap among the groups. cOptBees was run with *cMax* = 20. Fig. 4 shows the results for this dataset.

The last problem evaluated was a function approximation problem, instead of data classification. cOptBees was run using *cMax* = 20 and the results are shown in Fig. 5. The cOptBees found seven centers and the RBF network provided a smooth approximation to the sine function.

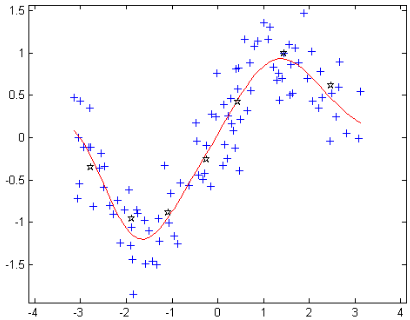


Fig. 5. Training input vectors, RBF centers (black stars) and final function approximated by the hybrid approach

5 Discussion and Future Research

This paper proposed a hybrid approach to determine the number and location of basis functions to be used in RBF neural networks. The proposed approach uses a bee-inspired clustering algorithm, named cOptBees, to optimize the number and location of basis functions for RBF networks. The cOptBees clustering algorithm is inspired by the foraging behavior of bee colonies and is able to find high quality cluster partitions in datasets without the need to previously inform the number of clusters.

The algorithm was preliminary assessed with three artificial bi-dimensional datasets. The graphical results show that the algorithm was capable of suitably determining the number and positioning the RBF centers so that the neural network could solve the pattern classification and function approximation problems.

We do recognize the simplicity of the problems evaluated and the need to perform experiments with real-world problems and comparisons with other approaches from the literature. These are the following main steps of this research.

Acknowledgments. The authors thank Capes (Proc. n. 9315/13-6), CNPq, MackPesquisa, Fapesp #2013/12005-7, Fapesp #2013/05757-2 and Fapemig for the financial support.

References

1. Yingwei, L., Sundararajan, N., Saratchandran, P.: A sequential learning scheme for function approximation using minimum radial basis function neural networks. Neural Computation 9(9), 461–478 (1997)
2. Hwang, Y.-S., Bang, S.-Y.: An Efficient Method to Construct Radial Basis Function Neural Network Classifier. Neural Networks 10(8), 1495–1503 (1997)
3. Whitehead, B.A., Choate, T.D.: Cooperative-Competitive Genetic Evolution of Radial Basis Function Centers and Widths for Time Series Prediction. IEEE Trans. on Neural Networks 7(4), 869–880 (1996)
4. de Castro, L.N., Hruschka, E.R., Campello, R.J.G.B.: An evolutionary clustering technique with local search to design RBF neural network classifiers. In: Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, July 25-29, vol. 3, pp. 2083–2088 (2004)
5. Cruz, D.P.F., Maia, R.D., Castro, L.N.: A New Encoding Scheme for a Bee-Inspired Optimal Data Clustering Algorithm. In: 1st BRICS Countries Congress (BRICS-CCI), Recife (2013)
6. Cruz, D.P.F., Maia, R.D., Castro, L.N., Szabo, A.: A Bee-Inspired Algorithm for Optimal Data Clustering. In: IEEE World Congress on Computational Evolutionary, Cancún, México, pp. 3140–3147 (2013)
7. Maia, R.D., Castro, L.N., Caminhas, W.M.: Collective decision-making by bee colonies as model for optimization - the OptBees algorithm. Applied Mathematical Sciences 7(87), 4327–4351 (2013)
8. Micchelli, C.A.: Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions. Construct. Approx. 2, 11–22 (1986)
9. Orr, M.J.L.: Regularisation in the Selection of Radial Basis Function Centres. Neural Computation 7(3), 606–623 (1995)
10. Broomhead, D.S., Lowe, D.: Multivariable Functional Interpolation and Adaptive Networks. Complex Systems 2, 321–355 (1988)
11. Orr, M.J.L.: Regularisation in the Selection of Radial Basis Function Centres. Neural Computation 7(3), 606–623 (1995)
12. Chen, S., Chng, E.S., Alkadhimi, K.: Regularized Orthogonal Least Squares Algorithm for Constructing Radial Basis Function Networks. Int. Journal of Control 64(5), 829–837 (1996)
13. Haykin, S.: Neural Networks – A Comprehensive Foundation, 2nd edn. Prentice-Hall (1999)
14. Maia, R.D., Barbosa, W.O., De Castro, L.N.: Colônias de Abelhas como Modelo para Agrupamento Ótimo de Dados: Uma proposta e análise paramétrica qualitativa. In: XIX Congresso Brasileiro de Automática, Campina Grande (2012)

An Item Influence-Centric Algorithm for Recommender Systems

Na Chang, Mhd Irvan, and Takao Terano

Dept. of Computational Intelligence and Systems Science,
Tokyo Institute of Technology, Yokohama, Japan
{Changna,irvan}@trn.dis.titech.ac.jp,
terano@dis.titech.ac.jp

Abstract. Collaborative filtering recommendation algorithms are the most popular approaches in the area of recommender systems and have been extensively discussed by researchers. In this paper, we focus on the analysis of items influence received from neighborhood and the corresponding iterative preference prediction based on the influence. Specifically speaking, the proposed approach uses influence coefficient to measure an item's ability to influence neighbors' acceptance by users, and predicts a user's preference for an item based on the user's ratings on these items which have influence on the target item. In the meanwhile, the proposed approach distinguishes influence into persuasive influence and supportive influence, and takes into account the combined effect of the two types of influence. Under this methodology, we verified that the proposed algorithm obviously outperforms standard collaborative filtering methods through 5-fold cross validation.

Keywords: recommender systems, influence coefficient, supportive influence, persuasive influence.

1 Introduction

Recommender systems (RS) are personalized information filtering systems which are used to solve information overload problem of web sites. The aim of recommender systems is to learn users' preference based on various resources, such as users' individual information and items' contents [1]. The basic approaches for recommender systems are content-based approach (CB) and collaborative filtering approach (CF). CB approach has its roots in information retrieval and recommends the right items to users through matching users' profile with items' contents (features) [2]. On the other hand, CF approach is used to predict users' preference on the basis of user-item rating matrix [1], [3], [8].

On the basis of these traditional recommendation algorithms, many extensions have been proposed by incorporating theories or methodologies from other areas. Among them, influence-based recommendation attracts researchers' attention recent years. Most of such research focuses on users' influence on neighborhood. For instance, researchers assume that an influential user will easily exert influence or

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence*, *11th International Conference*, Advances in Intelligent Systems and Computing 290,
DOI: 10.1007/978-3-319-07593-8_64,

impact on the other users and propose various methods to find influential users [4], [5], [6]. However, items influence is rarely mentioned in such area. But the fact is that items also can have influence on the other items and differs in different influence level [7]. In [7], the author defined item rating frequency as its level of influence and take paper citation as an example: a paper may be influential if it is cited often and perhaps cited by many other influential papers.

Based on this situation, we consider item influence-centric recommendation algorithm (item-IR), i.e. a user's preference for an item is approximated based on the user's ratings on these items which have influence on the target item. The main difference of our proposed algorithm and standard item-based algorithm is that each rating is weighted by the corresponding influence of rated item on target unrated item instead of their similarity. And furthermore, as an extension of the other influence-based algorithms, we distinguish these influence into persuasive influence and supportive influence, and take into account the total influence of these two types of influence in the preference prediction process. The experiment shows that our proposed algorithm can make a better performance than the other standard CF methods in terms of *RMSE* (*Root Mean Square Error*).

2 Related Work

In this section, we briefly present some of the research literatures related to collaborative filtering and influence-based recommendation methods.

Collaborative filtering methods can be grouped in the two general classes: memory-based and model-based methods. In memory-based collaborative filtering [1, 3, 8], the user-item ratings stored in the system are directly used to predict ratings for new items. User-based CF and item-based CF recommendation algorithms are the basic CF methods. User-based CF algorithms predict the interest of a user for an item using the ratings for this item by other users who have similar rating patterns. Item-based CF approaches predict the rating of a user for an item based on the ratings of the user for items similar to the target item. In contrast to memory-based systems, model-based approaches use user-item ratings to learn a predictive model. The model can be a data mining or machine learning algorithm. Well-known model-based CF techniques include Bayesian Clustering [9], Support Vector Machines [10], and Singular Value Decomposition (SVD) [11, 12]. Among the variety of CF methods, user-based CF, item-based CF, and SVD are usually considered as baseline methods and researchers compared their proposed algorithms with these algorithms.

On the other hand, influence-based recommendation approach is one important branch of this area. Shang [13] proposed influence-based recommendation models on the basis of social contagion and social influence network theory. In the first model for individuals, they improved the result of collaborative filtering prediction with social contagion outcome. In the recommendation model for groups, they applied social influence network theory to take interpersonal influence into account to form a settled pattern of disagreement, and then aggregate opinions of group members. Rubens [14] proposed a method to select informative items that are not only uncertain but also influential by estimating the influence of an item by changing its rating and observing the changes in the approximated ratings of other unrated items. Rashid [15] presented two ways to measure a user's influence over others. One way is to exclude a user and measure the net changes in predictions caused by the removal. The other way is a predictive model based on several quality factors that may affect users' influence levels, such as number of ratings, degree of agreement with others, rarity of the rated items, and so on, SVM regression. Arora [16] defined the influence rating of a user on the community as the degree to which its item ratings match that of an average user's item ratings and then determined the most influential users who have the maximum influence on the evaluation of other user's recommendations.

3 Proposed Algorithm

3.1 Problem Definition

Consider that there are *m* users and *n* items in a recommender system. We can define the set of users *U* as the set of integers {*1, 2... m*} and the set of items *I* as the set of integers {*1, 2... n*}. The ratings of users for items r_{ui} are stored in a $m\times n$ rating matrix $R=\{r_{ui}|1\le u\le m;1\le i\le n\}$. The objective of a recommendation algorithm is to predict the missing value of *R*. In the following sections, we use letters to represent users as *u*, *v*, items as *i*, *j*, r_{ui}^{o} and r_{ui} to represent user's initial preference and prediction preference respectively.

In addition to rating matrix, we also introduce an acceptance state o_{ui}, which indicates whether an item *i* is liked or disliked by a user *u*. On the other hand, the value o_{ui} can be seen as users' preference state from users' point of view. To obtain an item's acceptance state by a user, we need to define a mean value λ to distinguish whether the user likes or dislikes it as follows:

$$o_{ui}=\begin{cases}1, & if\ r_{ui}\ge\lambda\\ -1, & if\ 0<r_{ui}<\lambda\end{cases} \tag{1}$$

where $0<\lambda\le R$. If *u* has not rated item *i*, then we use the initial preference r_{ui}^{o} instead of the real rating r_{ui} in formula (1). For instance, if r_{ui} = 5 with the rating scale is from 1 to 5, and we set $\lambda=4$, then we can have item *i*'s acceptance state by *u* is 1. In the items' neighbor selection phase, we define *i*'s neighbors which have been rated by user *u* as follows:

$$N(i;u)=\{j\,|\,s_{ij}>0,r_{uj}>0,j\in I\} \tag{2}$$

where s_{ij} is the similarity between *i* and its neighbor *j*. From this definition, we can see that the number of neighbors of an item is not static and adaptively varies with the target user.

As the introduction addressed in section 1, the proposed algorithm can be descripted as that user's preference for an item is the weighted ratings of the user's ratings on these items, which have influence on the target item, and the total influence effect of persuasive influence and supportive influence. The main procedure of the algorithm is as follows:

- Computation of item's influence coefficient which denotes an item's ability to influence another item, and persuade to have the same acceptance rate with the item.
- Computation and analysis of persuasive influence and supportive influence, and the total influence of these two types of influence.
- Items' preference prediction based on neighbors' weighted ratings given by the target user and total influence effect.

3.2 Influence Coefficient

Items' influence coefficient denotes the importance of items in predicting users' preference for other items. In this paper, we use influence coefficient to measure an item's ability to influence another item, and persuade it to have the same acceptance rate with the item. The influence coefficient of an item *j* on item *i* is defined as follows:

$$p_{ij} = |N(j)| \times |Y_{ij}| / \sum_{j \in N(i;u)} (|N(j)| \times |Y_{ij}|) \tag{3}$$

where $|N(j)|$ is the number of *j*'s neighbors and $|Y_{ij}|$ is the number of users who have the same preference state for both *i* and *j*. Apparently, items' influence coefficient is normalized, i.e. $0 \le p_{ij} \le 1$, $\sum_{j \in N(i;u)} p_{ij} = 1$ in our definition.

3.3 Persuasive Influence or Supportive Influence

An item's influence on another item is related to the closeness between them. In addition to similarity, we define the following expression to measure the closeness of two items:

$$c_{ij} = s_{ij} \times |Y_{ij}| / |Y_j| \tag{4}$$

where s_{ij} is the similarity between *i* and *j*, $|Y_{ij}|$ is the number of users who have the same preference state for both *i* and *j*, and $|Y_j|$ is the number of users who have rated *j*.

On the basis of social impact theory [18], if a user *u*' preference state for an item j is the same as *u*'s preference state for the target item *i*, we can say that *j* have supportive influence on *i*. On the other hand, if *u*' preference state for an item *j* is different from *u*'s preference state for *i*, then we can say that *j* have persuasive influence on *i*. Except for influence coefficient and closeness defined in [18], we also consider about the popularity of items, i.e. the more popular an item is, the less influence the other items will have on it, and the vice versa. So if we use the number of ratings item *i* received to measure the popularity of the item, then we can define item *j*'s persuasive influence or supportive influence on *i* with respect to user *u* as follows:

$$w_{ij}^{p} = p_{ij}c_{ij} / log(a+|X_i|) \tag{5}$$

$$w_{ij}^{s} = p_{ij}s_{ij} / log(b+|X_i|) \tag{6}$$

where $|X_i|$ indicates the number of ratings item *i* received, parameter *a* and *b* are used to distinguish the difference between persuasive influence and supportive influence. Intuitively, the value of *a* is bigger than that of *b*, since it is harder to persuade to change than support to keep the status. We set *a=250* and *b=20* through our empirical analysis.

Additionally, we also assume that the combined effect of persuasive influence and supportive influence plays an important role in the prediction of preference. On the basis of social impact theory, the total influence item *i* received with respect to user *u* could be calculated in the following forms:

$$w_i = N_p^{\eta} \sum_{j\in N(i:u)_p} w_{ij}^{p} - N_s^{\eta} \sum_{j\in N(i:u)_s} w_{ij}^{s} \tag{7}$$

where $N(i:u)_p$ is the set of items which have opposite acceptance state with *i* by user *u*, N_p is the number of the set, $N(i:u)_s$ is the set of users who have the same acceptance state with *i* by user *u*, N_s is the number of the set, and η is a parameter controlling the contribution of numbers of items who have opposite or the same acceptance state with *i* by *u*.

There are two possible situations by taking combined effect of persuasive influence and supportive influence into account:

- If $w_i \geq 0$, this indicates that the pressure in favor of the preference state change overcomes the pressure to keep the current preference state. Additionally, the bigger the value of w_i, the higher possibility user *u* changes his/her preference state for item *i*.
- If $w_u < 0$, this shows that the pressure to keep the current preference state overcomes the pressure in favor of the preference state change. In addition, the bigger the value of $|w_i|$, the higher possibility user *u* keeps his/her preference state for item *i*.

Thus, considering the combined effect of total influence, we use a parameter ρ to control the contribution the effect of the total influence and have the following formula for the prediction of a user *u*'s preference for item *i* based on *u*'s preference for i's neighbors:

$$\tilde{r}_{ui} = \frac{\sum_{j \in N(i;u)_p} r_{uj} w_{ij}^p + \sum_{j \in N(i;u)_s} r_{uj} w_{ij}^s}{\sum_{j \in N(i;u)_p} w_{ij}^p + \sum_{j \in N(i;u)_s} w_{ij}^s} - \rho o_{ui} \times \frac{1-e^{-w_i}}{1+e^{-w_i}} \tag{8}$$

where the first term indicates the weighted preference of item i's neighborhood, and the second term addresses the effect of total influence received from neighborhood.

3.4 Preference Prediction

In this paper, we use the following iterative formula to predict a user's preference by repeatedly adjusting estimations:

$$\hat{r}_{ui}^{k} = \begin{cases} \alpha \tilde{r}_{ui}^{k-1} + (1-\alpha)\hat{r}_{ui}^{k-1}, & k \geq 2 \\ r_{ui}^{o} & , k = 1 \end{cases} \tag{9}$$

where $\tilde{r}_{ui}^{k-1}$ is the prediction of user u's preference for item i based on i's neighbors' preference and their influence on i during the k-th iteration. r_{ui}^{o} is user's initial preference, which represents user's preference tendency based on the rating distribution or other recommendation algorithms. Parameter α is used to control the contribution of the neighbors' influence during each iteration and $0 \leq \alpha \leq 1$. In order to compare our algorithms with other standard collaborative filtering algorithms, we use a simple formula to get r_{ui}^{o} as follows:

$$r_{ui}^{o} = \beta \overline{r}_u + (1-\beta)\overline{r}_i \tag{10}$$

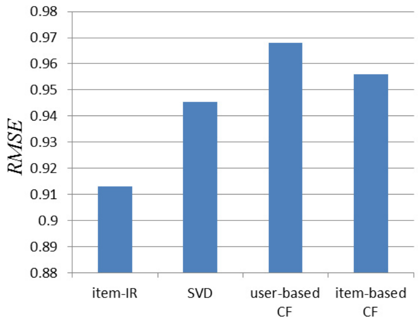
where $0 \leq \beta \leq 1$, $\overline{r}_u$ is the average rating user u has given and $\overline{r}_i$ is the average rating item i received.

4 Experiment Evaluation

In this paper, we used the data set from MovieLens[1] to verify our research. The data set contains 100,000 records (user-item-rating) with rating scale 1-5 and consists of 943 users and 1,682 movies. We firstly used 5-fold cross validation by randomly selecting 80% of the data set as training data and 20% of it as test data to measure the accuracy performance (RMSE) of our proposed algorithms with the other standard recommendation algorithms: user-based CF, item-based CF and SVD. Secondly, to study the proposed algorithms' sensitivity of density on the data set, we carried out an experiment where we varied the value of the training/test ratio from 3/7 to 9/1. The parameters in our proposed algorithm are decided through empirical analysis in which a good recommendation performance is obtained (i.e. $\eta = -0.2$, $\rho = 1.3$, $\alpha = 0.88$,

[1] `http://grouplens.org/datasets/movielens/`

$\beta = 0.25$). In the neighbor selection process, the similarity is computed by adjusted cosine similarity [3]. The experiment results are as Fig. 1 shows.



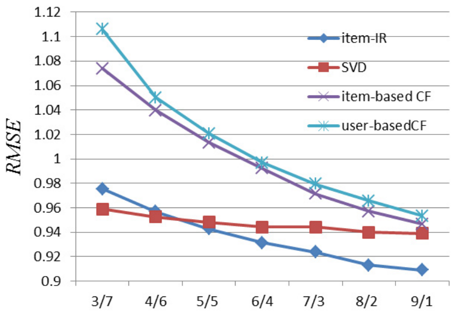


Fig. 1. Accuracy comparison and performance comparison at different training/test ratio

From the first experiment result, we can see that the proposed item-IR has an obvious improvement in terms of RMSE. This result indicates that the proposed approach, which focuses on the influence coefficient of each user pair or item pair and takes into account the combined effect of persuasive influence and supportive influence, is reasonable and effective.

With respect to the performance of sensitivity of density, we have the following results: 1) item-IR shows better performance than standard item-based CF and user-based CF at all values of training/test ratio; 2) when the training/test ratio is bigger than 5/5, the item-IR algorithm gives better performance than the standard item-based CF, user-based CF and SVD, but when the training/test ratio is smaller than 5/5, it can be observed that SVD performs better than the proposed methods. This result indicates that the proposed algorithms still confront sparsity problem like the other neighborhood-based algorithms such as item-based CF and user-based CF.

5 Conclusions

In this paper, we presented an iterative influence-centric recommendation methodology, which uses influence coefficient to measure an item's ability to influence neighbors' acceptance, and takes into account the combined effect of persuasive influence and supportive influence. From the experiments, it is clearly observed that the proposed algorithm outperforms the other standard collaborative filtering algorithms. But, the new algorithms still confront the sparsity problem like user-based CF and item-based CF algorithms.

In the future, we will verify the performance of the proposed algorithm on the other standard datasets. In addition, the sparsity problem of the proposed approach should be our main work. We will seek solutions for this problem.

References

1. Adomavicius, G., Tuzhili, A.: Toward the next generation of recommender systems: Asurvey of the state-of-the-art and possible extensions. IEEE Transaction on Knowledge and Data Engineering 17(6), 734–749 (2000)
2. Lops, P., Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. Recommender Systems Handbook. Springer Science and Business Media (2011)
3. Sarwar, B., Kar, Y.G., Konstan, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference, New York, USA (2001)
4. Chen, D., Lu, L., Shang, M.S., Zhang, Y.C., Zhou, T.: Identifying influential nodes in complex networks. Statistical Mechanicis and its Applications 391(4), 1777–1787 (2011)
5. Faust, K.: Centrality in affiliation networks. Social Networks 19, 167–191 (1997)
6. Morid, M.A., Shajari, M., Golpayegni, A.H.: Who are the Most Influential Users in a Recommender System? In: Proceedings of the 13th International Conference on Electronic Commerce. ACM, New York (2012)
7. Rashid, A.M.: Mining influence in recommender systems. Doctoral Dissertation. University of Minnesota Minneapolis (2007)
8. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing 7(1), 76–80 (2003)
9. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Annual Conf. on Uncertainty in Artificial Intelligence, pp. 43–52. Morgan Kaufmann (1998)
10. Grcar, M., Fortuna, B., Mladenic, D., Grobelnik, M.: k-NN versus SVM in the collaborative filtering framework. Data Science and Classification, 251–260 (2006)
11. Koren, Y.: Factorization meets the neighborhood: A multifaceted collaborative filtering model. In: 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA (2008)
12. Paterek, A.: Improving regularized singular value decomposition for collaborative filtering. In: Proceedings of the KDD Cup and Workshop (2007)
13. Shang, S., Hui, P., Kulkarni, S.R., Cuff, P.W.: Wisdom of the Crowd: Incorporating Social Influence in Recommendation Models. CoRR abs/1208.0782 (2012)
14. Rubens, N., Sugiyama, M.: Influence-based collaborative active learning. In: Proceedings of the 2007 ACM Conference on Recommender Systems, pp. 145–148 (2007)
15. Rashid, A.M., Karypis, G., Riedl, J.: Influence in Ratings-based Recommender Systems: An Algorithm-Independent Approach. In: Proceedings of 5th SIAM International Conference on Data Mining, Newport Beach, California (2005)
16. Arora, I., Panchal, V.K.: An MIU Most Influential Users-Based Model for Recommender Systems. In: IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, pp. 638–643 (2010)
17. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Analysis of recommendation algorithms for E-commerce. In: Proceedings of the ACM E-Commerce, Minneapolis, Minnesota (2000)
18. Latané, B.: The psychology of social impact. American Psychologist 36, 343–356 (1981)

Author Index