Computer Science and Information Systems

Com SIS

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 11, Number 2
June 2014

Vol 11, No 2, June 2014

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Consortium:
University of Belgrade:
Faculty of Organizational Science, Belgrade, Serbia
Faculty of Mathematics, Belgrade, Serbia
School of Electrical Engineering, Belgrade, Serbia
Serbian Academy of Science and Art:
Mathematical Institute, Belgrade, Serbia
Union University:
School of Computing, Belgrade, Serbia
University of Novi Sad:
Faculty of Sciences, Novi Sad, Serbia
Faculty of Technical Sciences, Novi Sad, Serbia
Faculty of Economics, Subotica, Serbia
Technical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia
University of Montenegro:
Faculty of Economics, Podgorica, Montenegro

ComSIS Editorial Office:
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Informatics
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
Phone: +381 21 458 888; **Fax:** +381 21 6350 458
www.comsis.org; Email: comsis@uns.ac.rs

Volume 11, Number 2, 2014
Novi Sad

Computer Science and Information Systems

ISSN: 1820-0214

The ComSIS journal is sponsored by:

Ministry of Education, Science and Technological Development of the Republic of Serbia
http://www.mpn.gov.rs/

Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. ComSIS also welcomes survey papers that contribute to the understanding of emerging and important fields of computer science. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2012 two-year impact factor 0.549,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Summon (Serials Solutions),
- EBSCO bibliographic databases,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (MS Word or LaTeX), or three hard-copies of the manuscript written in English, intended for publication and prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be sent to official journal e-mail.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 20 (using the appropriate Word or LaTeX template).

Manuscripts will be refereed in the manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

All authors are requested to sign the "Transfer of Copyright" agreement before the paper may be published. The copyright transfer covers the exclusive rights to reproduce and distribute the paper, including reprints, photographic reproductions, microform, electronic form, or any other reproductions of similar nature and translations. Authors are responsible for obtaining from the copyright holder permission to reproduce the paper or any part of it, for which copyright exists.

Computer Science and Information Systems

Volume 11, Number 2, June 2014

CONTENTS

Regular Papers

Special Section: Contemporary ICT Research Trends Emerging from South-East Europe

Special Section: Computational Intelligence in Business Administration

ComSIS Volume 11, Issue 2: Current Trends in Computer Science and Information Systems

Mirjana Ivanović and Miloš Radovanović

Department of Mathematics and Informatics
Faculty of Sciences
University of Novi Sad, Serbia
E-mail:{mira, radacha}@dmi.uns.ac.rs

1. Introduction

The current issue of the Computer Science and Information Systems journal continues to publish research papers from a wide range of contemporary topics in the fields of computer science and information systems. This issue contains nine regular articles, and thirteen articles in two special sections. The first special section "**Contemporary ICT Research Trends Emerging from South-East Europe**" contains five articles selected from the 6th Balkan Conference in Informatics (BCI 2013), held in Thessaloniki, Greece, September 19–21, 2013. The second special section "**Computational Intelligence in Business Administration**" contains eight papers selected from the International Conference on Modeling and Simulation in Engineering, Economics and Management (MS'10 Barcelona), held at the University of Barcelona, Spain, July 15–17, 2010. We are would like to express our gratitude to the guest editors, article authors and reviewers for their hard work and effort in making this issue of our journal possible.

Regular articles present interesting research achievements and considerable novelties in the following areas: ubiquitous health monitoring data, domain ontologies for natural languages, distributed services for messaging middleware models, selection of appropriate question-based design patterns, collaborative methods for business process model maintenance, visual Web-based approach to the interaction of data warehouse schemas, application and database co-refactoring, agile approach to the design of enterprise data warehouses, and neural network approaches to anomaly intrusion detection.

2. Special Sections

Special section "**Contemporary ICT Research Trends Emerging from South-East Europe**" contains papers that illustrate current research directions and trends in the South-East European region presented at the 6th Balkan Conference in Informatics (BCI 2013). **M**ore than a hundred papers were submitted to the conference; only 30% were accepted and published in the ACM International Conference Proceedings Series.

Five extended and improved versions of these papers went through a new rigorous, anonymous, peer review process and are published in this special section of ComSIS. Papers cover the following interesting research results: an effective prototype selection data reduction technique which improves the efficiency of k-nearest neighbor classification on large datasets, a generic meta-model and standard physical database schema meta-model for representing existing (legacy) information systems, an agent based framework for reasoning services over the Web, an educational computer game that helps young children with autism to learn how to identify facial expressions, and a methodology for addressing the privacy-preserving record linkage problem which is based on a distributed framework.

Special section **"Computational Intelligence in Business Administration"** contains papers selected from the International Conference on Modeling and Simulation in Engineering, Economics and Management (MS'10 Barcelona). Papers present research results based on application of intelligent techniques, modeling and simulation in a wide range of business areas: economics, management, tourism, finance, and engineering. Papers cover the following interesting research results: using fuzzy sets theory for reducing uncertainty in the resource selection problem, a framework for dealing with the theory of affinities in order to assess grouping problems, analysis of impact of Web 2.0 technologies in business performance, application of innovative business intelligence software in a range of business companies, improvements of claim predicting models by using fuzzy numbers and other related techniques, a new aggregation operator that uses intuitionistic fuzzy numbers and its application in business decision making, application of innovative fuzzy systems in the finance domain, and a new approach for propagating ontological and epistemic uncertainty by using risk assessment models and fuzzy time-series techniques.

3. Regular Papers

The regular paper section contains nine articles that cover a wide range of topics and areas in computer science and information systems.

The first regular paper, "Collaborative Method to Maintain Business Process Models Updated" by Nuno Castela et al. discusses collaborative aspects of business process model maintenance through the creation of interaction contexts which enable business actors to discuss business processes and share business knowledge. In fact, the authors defined an organizational support method to explicitly describe the steps and participants involved in the dynamic updating of business processes models. Organizational actors monitor the activities they perform in order to propose updates to the models using annotations, establishing a conversation between the involved actors. The proposed updates are reviewed and evaluated by the actors involved in the business process context and may lead to the creation of new versions of business process models. By examining real case studies (the first case study was conducted at the Social Security District Center, the second case study at Huf Portuguesa, an automotive parts manufacturing company, and the third case study was developed at

Technology School) the authors conclude that the defined method and the developed tool can help organizations effectively maintain a business process model in an updated state.

Clemente R. Borges and José A. Macías, in "Facilitating the Interaction with Data Warehouse Schemas through a Visual Web-Based Approach," tried to resolve several drawbacks existing in the area of data warehousing: implicit drawbacks concerning representation and interaction with data in relational database applications, lack of expressiveness and ease of use in user interfaces to handle such data, and the implicit need for interactive end-user visual tools to query data and avoid dependency on programming languages. They study the problem of database interaction and usability, proposing a web-based tool that manipulates data warehouse schemas by using a visual language to represent the database structure and providing several visualization techniques that facilitate the interaction and creation of queries involving different levels of complexity. They also open new challenging questions for future work, involving investigations of data-model usability in order to obtain new findings on how data warehouse structures can affect overall usability in database applications.

"Building ontologies for different natural languages" by Emhimed Salem Alatrish et al. is concentrated on ontologies as a very important concept in different scientific fields such as knowledge engineering and representation, information retrieval and extraction, knowledge management, agent systems, etc. The authors propose a semi-automatic procedure to create ontologies for different natural languages. The approach combines different resources and tools such as DODDLE-OWL, WordNet, Protégé and XSLT transformations in order to formulate a general procedure to construct a domain ontology for any natural language. The authors also plan to conduct further analysis of the results and compare the obtained ontology trees using different natural languages with the same input text. As crucial future work they will attempt to improve the proposed approach by integrating additional software tools and making certain steps simpler.

The tendency of contemporary society towards increasing numbers of elderly people and people who need continuous health monitoring creates the need for adequate research in the area of ubiquitous health monitoring systems. In "User-Centric Privacy-Preserving Statistical Analysis of Ubiquitous Health Monitoring Data," George Drosatos and Pavlos S. Efraimidis propose a user-centric software architecture for managing ubiquitous health monitoring data (UHMD) generated from wearable sensors in a ubiquitous health monitoring system (UHMS), and examine how these data can be used within privacy-preserving distributed statistical analysis, with the main goals of enhancing patient privacy and managing the large volumes of biomedical data obtained through wearable sensors. They implemented a prototype that performs privacy-preserving statistical analysis on a community of independent personal agents and presented experimental results with a significant number of agents (several hundred) that confirm the viability and the effectiveness of their approach.

Peter Szilagyi in "Iris: A decentralized approach to backend messaging middlewares," introduces the Iris decentralized messaging framework, which takes a midway approach between the two prevalent messaging middleware models: the centralized one represented by the AMQP family and the socket queuing one represented by ZeroMQ. Iris achieves this goal by turning towards peer-to-peer overlays as the internal transport for message distribution and delivery. The main advantage of the proposed system is its significant switching speed, which can be scaled better than that of existing messaging frameworks, whilst incurring effectively zero configuration costs. Nevertheless, the author plans to continue with experiments to further validate and improve the proposed system.

The article "A Direct Approach to Physical Data Vault Design" by Dragoljub Krneta et al. presents a comparative review of contemporary approaches to data warehouse design automation and defines a novel direct approach to physical data vault design based on source data meta-model and rules. Apart from that, the authors are concentrated on a novel agile approach to large scale design of enterprise data warehouses based on a data vault model. The proposed algorithm for incremental design of physical enterprise data warehouses employs source data meta-model and rules, and is used for developing a prototype case tool for data vault design. Finally, an important contribution of the paper is that it provides a concrete basis for the development of tools for designing data warehouses.

Today's computer systems and networks suffer due to rapid increase of attacks. In order to keep them safe from malicious activities or policy violations, emergent research activities go in the direction of developing effective security monitoring systems, such as intrusion detection systems (IDS). Many researchers concentrate their efforts in this area using different approaches to build reliable intrusion detection systems. "Flow-Based Anomaly Intrusion Detection System Using Two Neural Network Stages" by Yousef Abuadlla et al. proposes an intrusion detection system that uses two neural network stages based on flow-data for detecting and classifying attacks in network traffic, with the first stage detecting significant changes in the traffic that could be a potential attack, and the second stage defining whether there is a known attack and classifying the type of attack. The experimental results are promising and demonstrate that the designed models exhibit good accuracy and computational time, with a low probability of false alarms. Therefore, the authors plan to continue their work in order to develop a more accurate model that can be used in real-time environments.

Evolution of software is an unavoidable process during software development and influences all phases of the software lifecycle. On the other hand, refactoring is popular practice for evolving the source code and software architecture highly connected to object-oriented environments. It can be successfully applied in different software systems and environments. Ondrej Macek and Karel Richta, in "Application and Relational Database Co-Refactoring," go a little bit further and analyze the impact of the application of refactoring on relational database schemas and appropriate stored data. The main contribution of their research is a new point of view on the application of refactoring. The authors are particularly concentrated on proposing a model for refactoring of application persistent objects which is capable of migrating the database according to refactoring of the application code. Furthermore, they showed that the automatic co-refactoring is possible not only in the case of basic changes to an application, but also complex refactoring can be successfully applied automatically on the applications' code and database.

Finally, Luka Pavlič et al. in their article "A Question-Based Design Pattern Advisement Approach" propose a question-based design pattern advisement approach, whose goal is to assist developers in identifying and selecting the most suitable design pattern for a given problem. Controlled experiment and two surveys were conducted and results have shown that the proposed approach is beneficial especially to software developers, who have less experience with design patterns. The article also presents extensions to the existing object-oriented design ontology (ODOL), defines a new design pattern advisement ontology and introduces a tool that supports the proposed ontology and the question-based advisement (OQBA) approach. The authors will continue their research activities in the area by developing a holistic methodology for design pattern selection including automatic question forming based on the analysis of paths taken by developers interacting with the system during the question-answer session. Using ontology-based reasoning and techniques they plan to automatically create relevant questions and answers based on information, concepts and relationships stored in the ontology.

4. Conclusions

This issue of Computer Science and Information Systems consisting of regular articles and papers in two special sessions offers variety of interesting research results in a wide range of computer science and information systems topics.

We hope that our readers will enjoy this issue and find interesting and challenging papers for their current and future professional work and research.

Editor-in-Chief
Mirjana Ivanović

Managing Editor
Miloš Radovanović

GUEST EDITORIAL

Special Section:
Contemporary ICT Research Trends Emerging from South-East Europe

This special section of Computer Science and Information Systems Journal hosts extended versions of selected papers presented at the 6th Balkan Conference in Informatics (BCI 2013), held in Thessaloniki, Greece, on 19th to 21st September 2013. The main objective of the BCI conference series is to be the leading conference of South-East Europe in Informatics. BCI provides a forum for dissemination of research accomplishments and promotes interaction, discussions and collaboration among scientists from the Balkan and the rest of the world. Traditionally, BCI conferences call for papers dealing with theory and/or applications in wide areas of ICT.

Initially, more than a hundred papers were considered for presentation in the BCI'13 Conference, out of which 30% were accepted and published in the ACM International Conference Proceedings Series. Five extended and improved versions of these papers went through a new rigorous, anonymous, peer review process and are published in this special issue of ComSIS.

The first paper "Efficient data abstraction using weighted IB2 prototypes" by Stefanos Ougiaroglou and Georgios Evangelidis presents an effective prototype selection data reduction technique which improves the efficiency of k-Nearest Neighbor classification on large datasets. The paper proposes an improved variation of IB2, namely AIB2, that generates new prototypes instead of selecting them.

Sonja Ristić, Slavica Aleksić, Milan Čeliković and Ivan Luković in their paper "Generic and Standard Database Constraint Meta-Models" present one generic meta-model, and one standard physical database schema meta-model which are used for representing existing (legacy) information systems into a higher level of abstraction in order to facilitate their migration to new, updated environments.

Costin Badica, Nick Bassiliades, Sorin Ilie, and Kalliopi Kravari in "Agent Reasoning on the Web using Web Services" introduce a system that provides reasoning services over the Web, which is based on an extension of the EMERALD framework for agent based reasoning services with a Web service interface.

Eirini Christinaki, Nikolas Vidakis, and Georgios Triantafyllidis present a research study in their paper "A Novel Educational Game for teaching Emotion Identification Skills to Preschoolers with Autism Diagnosis" that deals with the use of an educational computer game developed for helping young children with autism to learn how to identify facial expressions.

The last paper "A Distributed Near-Optimal LSH-based Framework for Privacy-Preserving Record Linkage" by Dimitrios Karapiperis and Vassilios S. Verykios, presents a methodology for addressing the Privacy-Preserving Record Linkage problem which is based on a distributed framework. The methodology presented for linking disparate data sets in order to identify accurately common real world entities is both very efficient and preserves privacy of the underlying data.

As editors of this special section we would like to express our gratitude to the Editor-in-Chief of ComSIS, Prof. Mirjana Ivanović, for giving us the opportunity to organize this special section, for her guidance and continuous support. We would also like to thank all the authors for their cooperation and their efforts to improve the extended versions of their papers. Finally we express our special thanks to the reviewers for their valuable feedback, which significantly helped authors to fine-tune their papers.

Guest Editors

Petros Kefalas,
International Faculty, City College,
Thessaloniki, Greece,
kefalas@city.academic.gr.

Demosthenes Stamatis,
Department of Information Technology,
Alexander Technological Educational Institute (ATEI) of Thessaloniki, Greece,
demos@it.teithe.gr.

Konstantinos I. Diamantaras,
Department of Information Technology,
TEI of Thessaloniki, Greece,
kdiamant@it.teithe.gr.

GUEST EDITORIAL

Special Section:
Computational Intelligence in Business Administration

The International Conference on Modelling and Simulation in Engineering, Economics and Management (MS'10 Barcelona) was held at the University of Barcelona, Spain, 15–17 July, 2010. This special section of the Computer Science and Information Systems Journal entitled "Computational Intelligence in Business Administration" comprises extended versions of selected papers presented at the conference. In this edition of the conference more than 100 participants from more than 30 countries attended the conference with presentations concerning different topics of modelling and simulation in a wide range of areas including, economics, management, tourism, finance, engineering and mathematics.

In this section, eight papers have been selected for publication. All of them have gone a careful and rigorous review process. They have been selected based on their quality and their relation to the scope of the special section. The need for computational intelligence tools in business administration is becoming more relevant in the literature. Especially, because the real world demands more efficient models that permit to assess the information in a more complete way in order to maximize the benefits as much as possible. Business administration is a very broad area that encompasses many fundamental topics including management, finance, marketing and information systems. The selection of papers has also been conditioned by this aspect where we have tried to include papers from all the key subareas of business administration.

The first paper, by J. Rojas-Mora and J. Gil-Lafuente, introduces a new approach for reducing uncertainty in the resource selection problem by using fuzzy sets theory. They use triangular fuzzy numbers and weighted averages to assess the information. An application in marketing management is presented focused on market segmentation.

In the second paper, A.M. Gil-Lafuente and A. Klimova introduce a new framework for dealing with the theory of affinities in order to assess grouping problems. Several methodologies are used including fuzzy pretopology and Galois lattices. An application of this approach is developed in regional economic modelling concerning grouping methods in the Russian Federation and Ukraine.

Next paper, by Yejun Xu, Carlos Llopis-Albert and J. González, analyses the impact of Web 2.0 technologies in business performance. They find a connection through innovativeness. A case study in hospitality management is developed by using structural equation modelling in order to prove these propositions.

The fourth paper, by E. Edelhauser and A. Ionica, presents business intelligence software. It is developed with the aim of being very useful for Romanian companies during the economic crisis. It uses both qualitative and quantitative techniques and questionnaires. Several examples regarding the use of the software are also developed.

The fifth paper, by Jorge De Andrés, contributes to the research literature of fuzzy insurance analysis. He develops several methods for quantifying claim provisions of a non-life insurance company under fuzzy environments. He improves the ANOVA claim predicting model by using the expected value of a fuzzy number and other related techniques.

The next paper, by S.Z. Zeng, Leina Zheng, J.M. Merigó and Pan Tiejun, presents new aggregation operators based in uncertain environments that can be assessed with intuitionistic fuzzy information. They suggest a new aggregation operator that unifies the weighted average with the ordered weighted average by using intuitionistic fuzzy numbers. An application in a business decision making problem regarding the selection of optimal strategies is also presented.

Next, A. Terceño, M. Glòria Barberà-Mariné, H. Vigier and Y. Laumann study an application of fuzzy systems in finance. They analyse the stability of beta coefficients in portfolio management under uncertain environments that can be assessed with fuzzy regression models.

In the last paper, V. Georgescu suggests a new approach for propagating ontological and epistemic uncertainty by using risk assessment models and fuzzy time series techniques. Several computational intelligence tools are used in the analysis including probabilistic reasoning methods, fuzzy systems and neural networks.

As guest editors, we would like to express our gratitude to all the authors of accepted and rejected papers of this special section for their contributions and to the reviewers for their valuable effort in taking the time to review these papers and improve their quality by giving additional suggestions. Thanks goes also to the whole organizing team of the conference for a very pleasant event that has received considerable attention by many people. We would also like to thank Prof. Mirjana Ivanović, the editor-in chief of ComSIS, for her support during the preparation of this special section in the journal.

Guest Editors

José M. Merigó
Manchester Business School,
University of Manchester,
Booth Street West
M15 6PB Manchester,
United Kingdom,
Email: jose.merigolindahl@mbs.ac.uk

Anna M. Gil-Lafuente
Department of Business Administration
University of Barcelona
Av. Diagonal 690
08034 Barcelona, Spain
Email: amgil@ub.edu

Computer Science and Information Systems 11(2):461–480 DOI: 10.2298/CSIS130117031C

Collaborative Method to Maintain Business Process Models Updated

Nuno Castela[1], Paulo Dias[2], Marielba Zacarias[3], and
José M. Tribolet[4]

[1] Center for Organizational Design and Engineering, INESC INOVAÇÃO,
Rua Alves Redol, nº 9, 1000-029 Lisbon, Portugal
Polytechnic Institute of Castelo Branco, Technology School, Av. Empresário, 6000-767
Castelo Branco, Portugal
ncastela@ipcb.pt

[2] Polytechnic Institute of Castelo Branco, Av. Pedro Álvares Cabral n.º 12
6000-084 Castelo Branco, Portugal
paulo.dias@ipcb.pt

[3] Center for Organizational Design and Engineering, INESC INOVAÇÃO,
Rua Alves Redol, nº 9, 1000-029 Lisbon, Portugal
Research Center for Spatial and Organizational Dynamics, University of Algarve, Campus de
Gambelas, 8005-139 Faro, Portugal
mzacaria@ualg.pt

[4] Center for Organizational Design and Engineering, INESC INOVAÇÃO,
Rua Alves Redol, nº 9, 1000-029 Lisbon, Portugal
Dep. of Information Systems and Computer Science, IST, University of Lisbon, Campus
Alameda, Av. Rovisco Pais, 1, 1049-001 Lisbon, Portugal
jose.tribolet@inesc.pt

Abstract. Business process models are often forgotten after their creation and its representation is not usually updated. This appears to be negative as processes evolve over time. This paper discusses the issue of business process models maintenance through the definition of a collaborative method that creates interaction contexts enabling business actors to discuss about business processes, sharing business knowledge. The collaboration method extends the discussion about existing process representations to all stakeholders promoting their update. This collaborative method contributes to improve business process models, allowing updates based in change proposals and discussions, using a groupware tool that was developed. Four case studies were developed in real organizational environment. We came to the conclusion that the defined method and the developed tool can help organizations to maintain a business process model updated based on the inputs and consequent discussions taken by the organizational actors who participate in the processes.

Keywords: collaborative business process updating method, groupware modeling tool, shared business process models.

1. Introduction

This work defines an organizational support method to explicitly describe the steps and participants involved in the dynamic updating of business processes models, in which organizational actors monitor the activities they perform in order to propose updates to the models. These updates are made to models, using annotations, in order to establish a conversation between the involved actors. The proposed updates are reviewed and evaluated by the actors involved in the business process context (each one with his own role) and may lead to the creation of new versions of business process models (updated models).

Considering that enterprise models represent several aspects, views or perspectives of organizations, and that the business processes view is one of the most important of enterprise models because it shows the activities flow as well as its informational inputs and outputs, we believe that the business process model could be used permanently as a support for a variety of operational and management tasks if it could provide an accurate representation of the current business processes. This support could be accomplished because business process models could take advantage of its ability to capture, represent and distribute organizational knowledge. However, business process models have been used primarily to support sporadic organizational tasks, rather than serve as a repository of organizational knowledge to withstand the daily organizational tasks [1].

The model as an updated repository may facilitate the collection and sharing of organizational knowledge making it in an essential tool for implementing learning organizations and materializing organizational self awareness. In order to do this, we developed a method to continuously update the as-is process model and to ensure its connection with the "organizational reality" shared by its members in business process contexts.

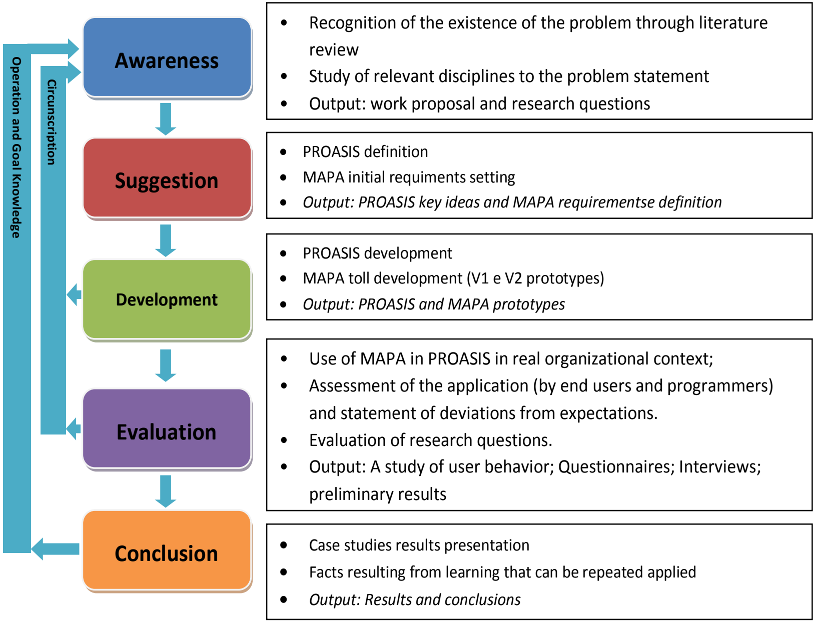
This work led to questions related with the conversation/negotiation among organizational actors and between the actors and the “organizational representation”, in order to maintain the as-is model updated. It was empirically observed that the as-is model is not usually updated in organizations despite being an important organizational asset, causing a permanent misalignment between the processes represented and implemented in real life. This observation was later supported by bibliographic revision carried in the course of this work.

Based on the theoretical framework and preliminary empirical evidence, this work seeks to answer the following questions: (1) is the annotation an appropriate mechanism to establish a conversation between the actors and the as-is business process model? If so, its extensions (reviews and evaluations) are appropriated to promote the negotiation between the organizational actors involved in the operational processes where changes are required? (2) Can the organizational actors, using a collaborative process and an appropriate supporting tool, become the modelers of their own organization in a collaborative way?

This work is related with the area of information systems and was guided by the design research methodology [2]. Figure 1 shows the steps of the methodology as well as the expected results in each step for this work.

The first step was designed to recognize problem; the outputs for this step were the work proposal and research questions.

Additionally, it was made an attempt to elaborate the project, which was closely linked with the outcome of the previous stage; the outputs in this work for this step are (1) the key ideas of the method to collaboratively and dynamically update the as-is business process model (that we named PROASIS) and (2) the requirements definition of the tool for monitoring and annotation of processes and activities (that we named MAPA).



- Recognition of the existence of the problem through literature review
- Study of relevant disciplines to the problem statement
- Output: work proposal and research questions

- PROASIS definition
- MAPA initial requiments setting
- *Output: PROASIS key ideas and MAPA requirementse definition*

- PROASIS development
- MAPA toll development (V1 e V2 prototypes)
- *Output: PROASIS and MAPA prototypes*

- Use of MAPA in PROASIS in real organizational context;
- Assessment of the application (by end users and programmers) and statement of deviations from expectations.
- Evaluation of research questions.
- Output: A study of user behavior; Questionnaires; Interviews; preliminary results

- Case studies results presentation
- Facts resulting from learning that can be repeated applied
- *Output: Results and conclusions*

Fig. 1. Design Research Methodology

In addition, we made an attempt to implement the project; the outputs for this step are PROASIS and MAPA.

Once built, these artifacts were evaluated on the evaluation step, through case studies; the outputs for this step are the answers to the interviews, the tool information analysis and some empirical observations.

Finally, in the conclusion step, the overall results and conclusions are presented.

The rest of the paper is structured as follows. Next two sections present the background information and related work to help in the awareness of the problem. Suggestion and development sections define PROASIS and MAPA. The evaluation section presents the case studies carried out in real organizational environment. The results section presents and analyzes the results of case studies. Finally, the conclusion section presents the overall conclusions of this work and points some suggestions for future work..

2. Background

The social, organizational and management sciences have a long tradition in organization modeling [3]. The main purpose of organizational models is to provide analysis forms that enable organizations to produce management theories and principles, and to provide descriptions of the nature, structure and dynamics of organizations.

Morgan [4] defined a set of images or metaphors that simplify the reading and analysis of organizations in order to enhance the dynamics of organizational change. Each image is a part of reality and gives prominence to certain aspects of organizations. The most used metaphors allow organizations to be seen as machines (mechanistic image), organisms, and more recently, as flows and transformation.

The flows and transformation metaphor is based on the principles of complexity. Complexity theory suggests the existence of processes underlying the observable reality, which can explain its state at any instant of time [5]. This metaphor motivates the search for dynamic creation and maintenance of organizations and their environments as concrete social forms.

The complexity is part of our environment, and many scientific fields have dealt with the study of complex systems. There is no single theory of complexity [6], but several theories arising from various natural sciences studying complex systems, such as biology, chemistry, computer simulation and artificial intelligence, evolution, mathematics, physics, economics and social sciences.

Complex Adaptive Systems (CAS) are special cases of complex systems. They are complex because they are diverse and consist of various elements linked by highly nonlinear relationships, and are adaptive because they have the ability to change and eventually learn from experience. The study of systems like CAS leads to the discovery of recurrent patterns of interaction between specific entities. The defining characteristic of CAS is emergence: the notion that at any level of analysis, the order is a property that emerges from the interactions of entities of lower levels. A complex adaptive system consists of a large number of agents, distinguished from each other, whose behavior uses the same set of rules. These rules require that agents must adapt themselves to the behavior of other agents with whom they interact (the principle of co-evolution).

The principles are generic, in that they are common to all complex natural systems. However, the nature of the entities (genes, molecules, numbers, computational agents or systems of human activity) has to be considered, and the application of the principles in each context must to be relevant and appropriate. For example, systems of human activities differ from all other systems because the complex evolutionary human actors can make intentional decisions. Many of the attributes and concepts of complexity theory are well known and belong to the domains of other theoretical frameworks [6]. This theory also describes the distinctive features of complex evolutionary systems, distinguishing them from the complicated systems, like machines. In particular, the concepts of connectivity, interdependence, emergence and feedback are acted on by complexity theory, which adds others, as the co-evolution, space of possibilities exploration, self-organization, multi-dimension, far-from-equilibrium, historicity and time, are used to provide a coherent description.

Kauffman mentions the importance of self-organization in the evolutionary process [7] cit. in [6]. Self organization means that no agent dictates the collective behavior of the system because the system self-organizes itself. "In an organizational context, self-organization can be described as the spontaneous junction of individuals in a group to accomplish a task, for which the group decides what to do, how and when they do, and nobody outside the group manages these activities". The emergent properties are the qualities, structures or patterns that appear at macro level as a result of interactions at micro level. The relationship between micro-events and macro-structure is interactive. This relationship is a co-evolutionary process in which individual entities and macro-structures are created, influencing each other through their interaction in an interactive process. Emergence is the process that creates new order together with self-organization.

A promising approach to the problem of emergence dynamics is provided by models of self-organization [8]. Self-organization can be defined as an organizational spontaneous process (i.e., not driven by an external system), allowing the development of an organized structure. The spontaneous creation of an "organized whole" from various disordered parts, as witnessed in the self-organizing systems in physics, chemistry, biology and sociology, is one of the important parts of emergency dynamics. However, another essential characteristic of emergence as understood in systems theory is its hierarchical and multi-level nature: an emergent whole to a certain level is only one component of an emerging system at the level immediately above. Realistic complex systems (organisms, societies, companies) are characterized by a multi-level structure.

The emergence of a human system tends to create structures, ideas, relationships and irreversible organizational forms, which become part of the history of individuals and organizations that affect the future evolution of these entities [6]. The generation of knowledge and innovative ideas when a team works together can be described as an emergent property in that it arises from the interaction of individuals and not just as a sum of existing ideas, but may well be something new and possibly unexpected. When ideas are articulated become part of the history of each individual and part of the shared history of the team. This process is not reversible, and these ideas and new knowledge can be built to generate new ideas and knowledge.

The theory of communicative action of Habermas [9], acknowledges the existence of rules that support and constrain the interactions between individuals, and that these rules are imposed both internally by interacting individuals and externally by social, political and economic structures of systems. Habermas also explains the set of requirements for a successful communication, and sets the validity claims as requirements to be applied in restoring communication when the communication action is interrupted. The concept of power structures, the validity claims and correctness emphasize the nature of interactions between individuals.

The underlying objective of language, according to Habermas, is to get to a common understanding among the individuals involved [10]. This implies that there is coordination of actions in the form of interaction to allow an orderly revelation, so the speech fulfills this function, since the meanings of the expressions are based on reasons. Habermas refers to this vision as "the basis of the validity of meaning". The speech is considered the basic unit of linguistic communication, in which each act is composed of one illocutionary or rhetorical force and a propositional content.

3. Related Work

Enterprise model allows communicating, documenting and understanding the organizations activity. Enterprise model have to accommodate different points of view: the individual, the organizational, and the views of various groups of actors grouped in organizational contexts, but at the same time the consistency of the whole model has be guaranteed, allowing to represent the organizational self awareness [11]. The primitive, the syntax and semantics of the business model should allow simple and immediate verification, from each of the organizational actors, from the reality of their continuing action, once the basis of representation must be developed in concrete activities, because this is the only truly verifiable and comprehensive basis which organizational stakeholders can use [12].

According to Adamides [13], modeling has always been in the core of processes management activities and methodologies, because process models have been used in improvement, re-engineering, certification and IT implementation initiatives. Therefore there has been considerable discussion about the characteristics and suitability of the different modeling formalisms. However, has been given little attention to the modeling process itself as a socio-cognitive process.

Rittgen [14] refers that modeling is a kind of conversational negotiation. The four above steps of organizational semiotics ladder were considered in his experience: syntactic, semantic, pragmatic, and social. Most activities at pragmatic level are associated with negotiation. An analysis of workflows at pragmatic level revealed a structure that goes beyond the simple identification of generic activities, so the negotiation process follows a certain pattern, consisting of an initial and rejection state, in a state where acceptance is favored, a state where rejection is favored, a sub-recursive state to negotiate a counter-proposal and a state of acceptance. Each of the states allows a certain set of activities that drives the pragmatic negotiations to different states [14].

According to Borghoff and Schlichter [15], the widespread use of personal computers and associated networks meant that these resources began to be used to work collaboratively, so the designations computer supported cooperative work (CSCW) and groupware were introduced. The groupware designation refers to the solutions and tools designed to support collaborative work in practice, where the role of individual members of groups is an important aspect in the development goupwere tools, because the roles help to structure the interactions between team members and to define the functionalities and access rights of the group. The roles define the social function of individuals in relation to group process, to group organization and relatively to other group members. The roles define rights and obligations in relation to group process. Groupware systems send notifications when something changes. This feature eases the self-awareness of the individual and the group itself and, moreover, the actions that occur in competition. The notifications inform all users involved in a group of changes made to a shared environment.

Collaborative modeling can enhance productivity and quality of modeling by helping to construct agreement and a sense of model ownership among stakeholders. In order to reach consensus and agreement, modelers need to commit themselves to work as a team and abide by their collective knowledge, conventions and decisions. Their communication strategy sets the goals and rules (explicitly or implicitly) for a conversational dialog in which the modelers propose and argue about (negotiate) the different positions raised. This communication may result in (dis)agreement with, and acceptance/rejection of, the ideas proposed [16].

Despite process models are considered to have great potential for the sharing of knowledge in organizations its potential usage is generally neglected in organizations, mainly because are used by few people such as analysts, developers and managerial staff, who use BPM systems to manage business processes [17]. As a result, process participants or new employees, who need to know about organizational processes or want to give feedback on business processes, are unaware of models as sources for information on business processes. Some reasons that lead to this situation, derive from several factors, such: in the repositories accessible for all stakeholders organizations, such as knowledge management systems, models are usually neglected; the majority of staff in organizations does not accept the use of models, as they often perceive models as technical artifacts used by modelers and analysts; models are closely bound to the modelers, who are in control of changes to them, thus making other staff bystanders in process documentation and change, consequently, getting feedback for business process models is costly and has to be done by modelers personally asking people for their feedback.

These issues can be overcome with the use of groupware solutions, namely with the use of groupware systems for collective sense-making, which implements collaborative modeling with a component of negotiation that facilitates the structuring of arguments and decisions regarding modeling choices [14].

In addition to the collaborative and negotiation features, these tools have to take into account the main characteristics and features that are usually found in the tools for business processes modeling, which can be defined as an automated system that provides capabilities to build business process models [18].

Gonzalez [18] identifies a set of features necessary to collaborative tools for modeling business processes and analyzes a set of commercial tools to check whether these features identified are implemented or not. The requirements of collaborative tools for modeling business processes can be classified according to the spatial-temporal matrix of groupware [19], because the interaction can happen at the same time (synchronous) or at different times (asynchronous) and the participants in interaction can be in the same location or different locations. After reviewing the available tools, three categories of business process modeling collaborative tools were identified [18]: Web tools with modeling support, client/server local tools, and export/import documents tools.

Gonzalez [18] analyzed 35 tools from the market for evaluating the collaborative features considered most important, which include: Web Publishing: the tool contains a module for Web publishing or for exporting to HTML, allowing the model presentation without having to edit them; Model Viewer: the tool contains a module that allows the model displaying without editing possibility; Information Reports: The tool generates reports, including graphics or information about models; Version Control: this feature allows the comparison of different versions of the same model; User Profiles: the user control allows participants to do just the actions that allowed by their profile; comments and notes: these artifacts allow participants to include information about models asynchronously; Ability to disaggregate and aggregate: this feature allows the splitting of the model into parts, so that participants only work on their part accordingly; Other collaborative features not included in the analysis: notification to alert the participants that changes were made to the model during the asynchronous work and chat for discussion among participants. This study revealed that, in general, most of the tools analyzed provide collaborative features, but many of the features are not present unless, in some cases, all modules have to be purchased, or in other cases, some modules have to be purchased from other vendors.

4. Suggestion and Development

The background section consolidates the motivation behind this work: the need to develop a method and a supporting tool to update the business processes of organizations in a distributed and collaborative way. These method and tool could enable the emergence of new representations of business processes more closely aligned with the operational reality of organizations, based on a negotiation process involving discussion to lead to agreements.

The section of related work provided the guidelines for designing the collaborative updating method and for defining the supporting tool requirements.

To solve the problem of obsolescence of as-is business process models, we decided to build a collaborative method that provides the distribution of processes to stakeholders, allowing the individual proposal for changes, thus this also provides the basis for a participatory discussion extended to all stakeholders, which leads to a continuously update of business processes reducing the gap between the representation and the execution of business processes.

Figure 2 shows the generic activities in PROASIS and shows the negotiation pattern involved in its review and evaluation steps. The characteristics of PROASIS are:

(a) It is executed by people that exists in the operational dimension and share a common representation of business processes.

(b) The mechanisms that they can use in this process to declare misalignments are the annotations. The annotations are used to build updating proposals to the model in order to align it with the reality perceived by each organizational actor. These proposals aim to make the corrective maintenance of the business process model and can have two goals: to correct the model or to increase its detail.

(c) After making an annotation on a modeling element, a negotiation with the actors who eventually share the same context of action may exist. This negotiation/discussion will be made by all stakeholders of the annotated element in order to clarify the original purpose of the annotation. All actors involved in this review should declare the agreement or disagreement with the annotation made to the model element, justifying it.

(d) After the review of the annotation, the annotation should be evaluated by the actors enabled to do so, having some degree of responsibility on the executed activities or on the organizational actors involved. If the evaluation of the annotation (and any reviews made to it) results in an approval, the changes requested in the annotation could be incorporated in the new version of the process model by the modeler.

(e) These reviews and evaluations made to annotations can improve organizational self awareness since all actors involved in the update context of a particular model or modeling element could collaboratively participate in the update process, trough the analysis and discussion of the model or modeling element.

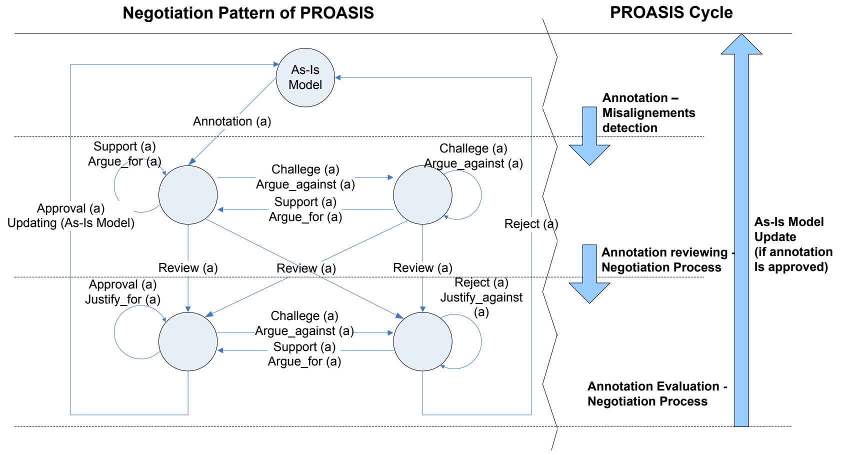


Fig. 2. Activities and negotiation pattern of PROASIS [1]

The set of modeling elements that can be annotated and the organizational actor roles that can be considered as standard annotators in PROASIS, are represented in figure 3.

The organizational roles presented can take on different roles on PROASIS depending on the particular modeling element annotated. Notice that an annotation is always created by individual initiative in a particular context, involving diverse actors in the later stages of reviewing and evaluation. This means that the updating context (PROASIS) captures the actors involved in the action context (operational level), consisting in a subset of actors of the operational model - people who participate in the reviewing and evaluation of the annotations.

To define a dynamic update process whose use is as comprehensive as possible it was necessary to consider the various levels of granularity that a business processes model can provide. These levels of detail derive from the contexts of the operational model (Process, Activity and Organizational Unit), and are considered to support the as-is business process model updating.

One of the goals in defining PROASIS was to approach as much as possible its collaborative updating process to the problem domain of business processes modeling. To achieve this, some options were taken: the annotation remain attached to the annotated modeling elements; review and evaluations remain attached to the annotation made; two categories of annotations were defined (correction, detail augmentation and adaptation); two types of annotations were defined (textual or graphical, allowing actors to annotate the model through a draft model diagram containing the proposed corrections, thus restricting the universe of discourse because this diagram must comply with the notational language used for process modeling); two types of reviews: agreement or disagreement, both complemented with text; two types of evaluations: approval or disapproval, both complemented with text; creation and maintenance of model versioning linked to the annotations that lead to updates.

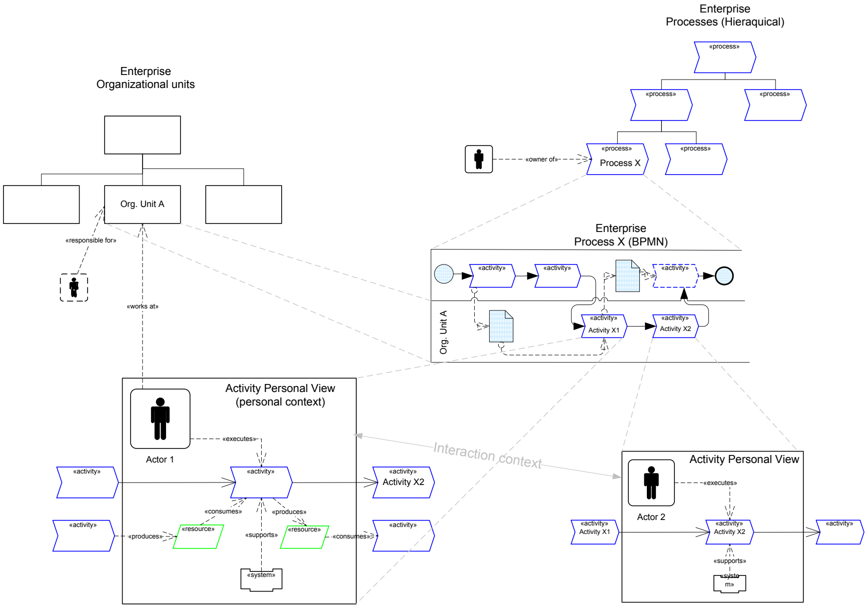


Fig. 3. Business process model [1]

To support PROASIS, the web-based groupware prototype tool MAPA was developed and two versions were created (first we created the v1version, which later has evolved into v2 version).

MAPA v1 works at the activity level of detail. The main view of the version 1 of the tool used when operating at the activity level of detail (figure 4), provide to the activities executors, a view to APV (Activity Personal View) diagrams that aggregates information from the activities and its executing context (the documents used and produced, information systems used to support the activity, previous and subsequent activities, annotations, etc.).

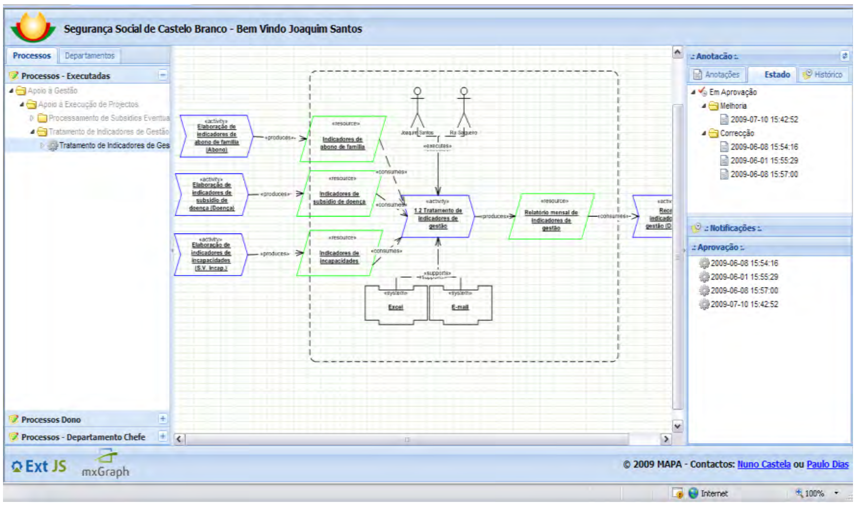

Segurança Social de Castelo Branco - Bem Vindo Joaquim Santos
Processos
Departamentos
Processos - Executadas
Apoio à Gestão
Apoio à Execução de Projectos
Tratamento de Indicadores de Gestão
Tratamento de Indicadores de Ges
Processos Dono
Processos - Departamento Chefe
Anotação
Anotações
Estado
Histórico
Em Aprovação
Melhoria
2009-07-10 15:42:52
Correcção
2009-06-08 15:54:16
2009-06-01 15:55:29
2009-06-08 15:57:00
Notificações
Aprovação
«activity»
«resource»
«produces»
«consumes»
«executes»
«supports»
«system»
Indicadores de abono de família
Indicadores de subsídio de doença
Indicadores de incapacidades
1.2 Tratamento de indicadores de gestão
Relatório mensal de indicadores de gestão
Excel
E-mail
Ext JS
mxGraph
© 2009 MAPA - Contactos: Nuno Castela ou Paulo Dias
Internet
100%

Fig. 4. MAPA v1 screenshot

Notice that although MAPA V2 tool allow modeling with BPMN 2.0 specification, the diagrams shown in the figure 5 and in the following figures were modeled by organizational actors involved in the processes without the care to fully meet this notation.

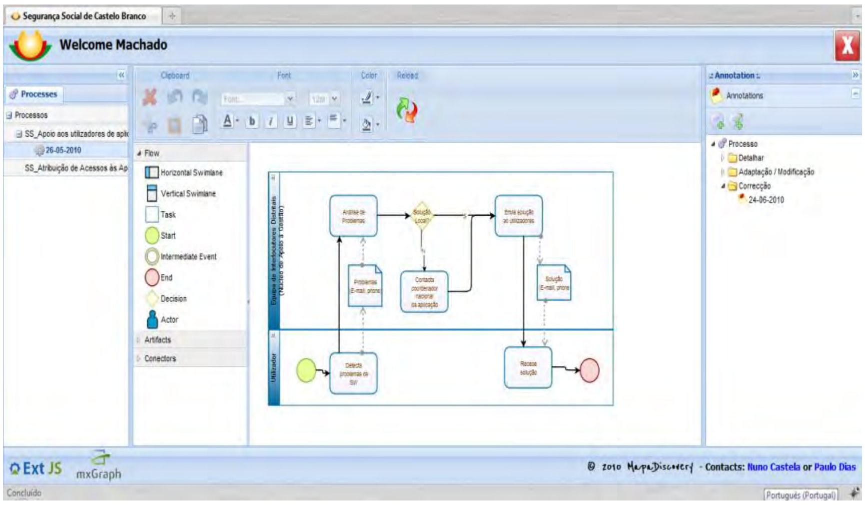


Fig. 5. MAPA v2 screenshot

This version was designed to operate at the process level of detail. Although the MAPA v2 tool does not fully implement the activity detail level, it allows annotating activities and other modeling elements that exists in the process model (in BPMN), either individually or grouped.

Some of the main functionalities implemented in MAPA tool are [20]:

(1) Annotation editing functions: actors need support to make immediate annotations in the context where experiences occur. Therefore, an annotation creation, modification and deleting system was created to be used by organizational users; mapping annotations to modeling elements: is essential to know to what modeling elements corresponds each annotation;

(2) Different levels of granularity: it should be possible to annotate any modeling element (process, activity, role, resource, relation) in the process model as well as any attribute of each object;

(3) Selective distribution of diagrams and modeling elements: users only access information that concerns to them, depending on the role played (executor, process owner and organizational unit responsible);

(4) Access rights: to protect the annotations authors, different levels of access rights should be addressed. Only the author of an annotation must be able to delete or modify it;

(5) Ability to save the entire history of models and their annotations, and the corresponding reviews and evaluations;

(6) Notifications: all actors related to process are notified by e-mail when annotations, reviews, evaluations or updates are made.

(7) Diagramming capabilities: to allow annotator actors to make graphical annotations with proposed changes to models and to allow modeler actors to directly change the diagrams if the proposals for changing the model are approved.

5. Evaluation

Four case studies were set up to evaluate the present work (figure 6).

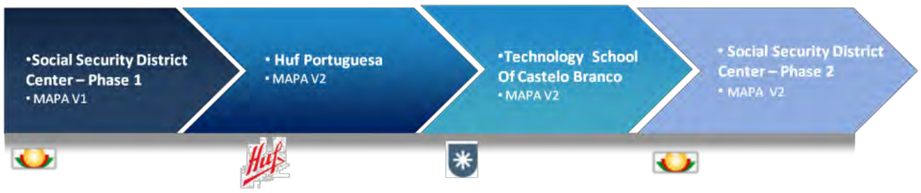

•Social Security District Center – Phase 1
• MAPA V1
• Huf Portuguesa
• MAPA V2
•Technology School Of Castelo Branco
• MAPA V2
• Social Security District Center – Phase 2
• MAPA V2
Huf

Fig. 6. Sequence of implementation of case studies

The first case study was conducted at Social Security District Center (SSDC), with MAPA v1 (phase 1) where processes and activities were modeled from scratch with top down and bottom up modeling approaches at the activity level of detail.

The second case study was conducted at Huf Portuguesa, an automotive parts manufacturing company. In this case study the main idea was to model business processes starting from an existing outdated model. MAPA tool helped in the translation of the existing model to build a BPMN end to end business process model. When this model is ended, we can apply PROASIS to maintain processes updated.

The third case study was developed at Technology School of Castelo Branco. It intend to apply PROASIS with version 2 of MAPA tool to the Business Processes of

Academic Services modeled from scratch, using a top down modeling approach at the process level of detail. In this case study we modeled two business processes, involving six users.

The social security case study (phase 2), used MAPA v2 in two of the processes already modeled in phase 1 and in 17 new business processes. The modeling approach used was the same that were used in phase 1, but the level of detail used was the process level of detail. We finish modeling 2 processes from District interlocutors Team, 12 processes from Financial Management Team and 5 processes from Administration and Assets Team.

To illustrate the dynamics of interaction provided by the tool, we show some examples of annotations, and subsequent reviews and approvals made in the context of a business process of social security. Figure 7 show the initial diagram of "Procurement and Materials Management" business process, modeled in MAPA tool. This process shows the interaction between the Social Security District Center, the procurement and patrimony team (equipa de aprovivionamento e património, in Portuguese), the financial management team (equipa de gestão financeira, in Portuguese), the social security central services (serviços centrais, in Portuguese) and the suppliers (fornecedor, in Portuguese) necessary to manage the acquisition of services and equipments for the Social Security District Center.

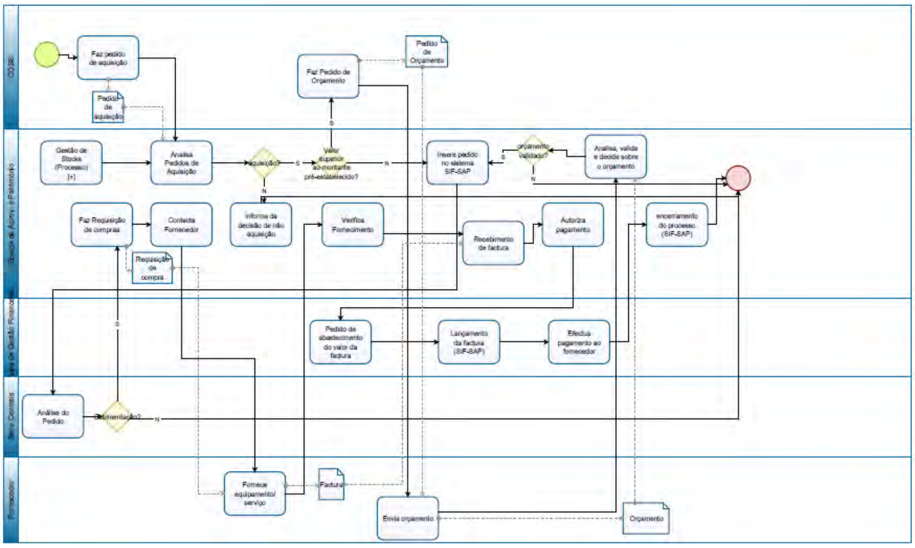
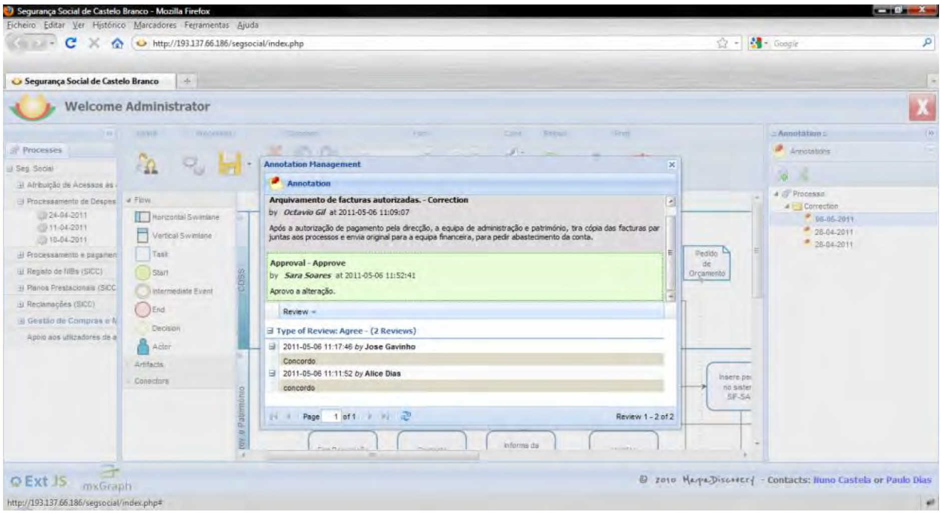

Faz pedido de aquisição
Pedido de aquisição
Faz Pedido de Orçamento
Pedido de Orçamento
Gestão de Stocks (Processo) [+]
Analisa Pedidos de Aquisição
aquisição?
S
Valor superior ao montante pré-estabelecido?
N
Insere pedido no sistema SIF-SAP
orçamento validado?
Analisa, valida e decide sobre o orçamento
Faz Requisição de compras
Contacta Fornecedor
Informa da decisão de não aquisição
Verifica Fornecimento
Recebimento de factura
Autoriza pagamento
encerramento do processo (SIF-SAP)
Requisição de compra
Pedido de abastecimento do valor da factura
Lançamento da factura (SIF-SAP)
Efectua pagamento ao fornecedor
Análise do Pedido
Cabimentação?
Fornece equipamento/ serviço
Factura
Envia orçamento
Orçamento
CDSS
Equipa de Aprov. e Património
Equipa de Gestão Financeira
Serv. Centrais
Fornecedor

Fig. 7. First version of "Procurement and Materials Management" business process

Figure 8 show the (textual) annotation made by Octávio Gil "After payment authorization by the management, the procurement team and heritage takes to gather copies of invoices to process and send the original to the team financial to ask for account supply", and the corresponding reviews made by 2 colleages that participate in the same business process that agreeded with the annotation made. In the same figure we can also see the evaluation (approval) made to the annotation by the responsible of the team Sara Soares.


Segurança Social de Castelo Branco - Mozilla Firefox
http://193.137.66.186/segsocial/index.php
Welcome Administrator
Processes
Annotation Management

Annotation

Arquivamento de facturas autorizadas. - Correction
by *Octavio Gil* at 2011-05-06 11:09:07

Após a autorização de pagamento pela direcção, a equipa de administração e património, tira cópia das facturas par juntas aos processos e envia original para a equipa financeira, para pedir abastecimento da conta.

Approval - Approve
by *Sara Soares* at 2011-05-06 11:52:41

Aprovo a alteração.

Review

Type of Review: Agree - (2 Reviews)

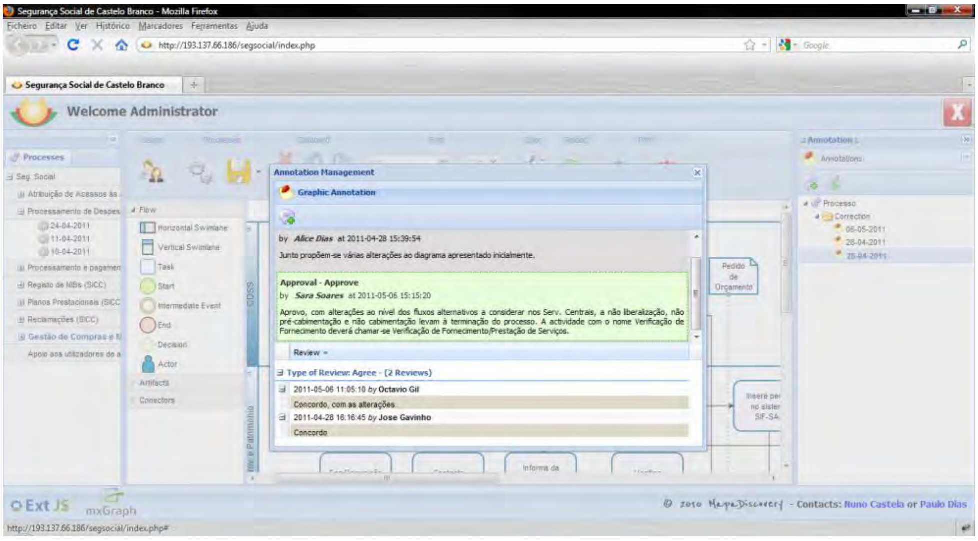
2011-05-06 11:17:46 *by* **Jose Gavinho**
Concordo

2011-05-06 11:11:52 *by* **Alice Dias**
concordo

Page 1 of 1 — Review 1 - 2 of 2

Fig. 8. Textual annotation, reviews and evaluation made to the business process

Figure 9 shows a graphical annotation made in the scope of the same version of the "Procurement and Materials Management" business process by Alice Dias. We can also see the reviews and the evaluation made to this annotation (the team responsible approved the annotation with some comments to improve the process representation) that in addition to the textual description of figure 9, contains a graphic description shown in figure 10.


Segurança Social de Castelo Branco - Mozilla Firefox
http://193.137.66.186/segsocial/index.php
Welcome Administrator
Processes
Annotation Management

Graphic Annotation

by *Alice Dias* at 2011-04-28 15:39:54

Junto propõem-se várias alterações ao diagrama apresentado inicialmente.

Approval - Approve
by *Sara Soares* at 2011-05-06 15:15:20

Aprovo, com alterações ao nível dos fluxos alternativos a considerar nos Serv. Centrais, a não liberalização, não pré-cabimentação e não cabimentação levam à terminação do processo. A actividade com o nome Verificação de Fornecimento deverá chamar-se Verificação de Fornecimento/Prestação de Serviços.

Review

Type of Review: Agree - (2 Reviews)

2011-05-06 11:05:10 *by* **Octavio Gil**
Concordo, com as alterações

2011-04-28 16:16:45 *by* **Jose Gavinho**
Concordo

Fig. 9. Graphical annotations with reviews and evaluation

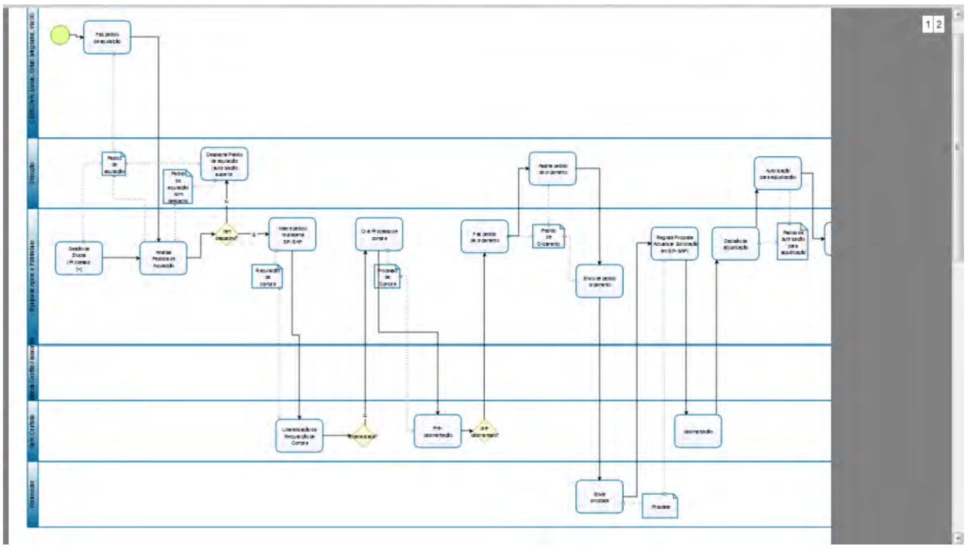


Fig. 10. Graphical change proposals of the annotation made

Figure 11 shows the new version of the “Procurement and Materials Management” business process, modeled in MAPA tool by the team responsible taking in account all the annotations and reviews approved.

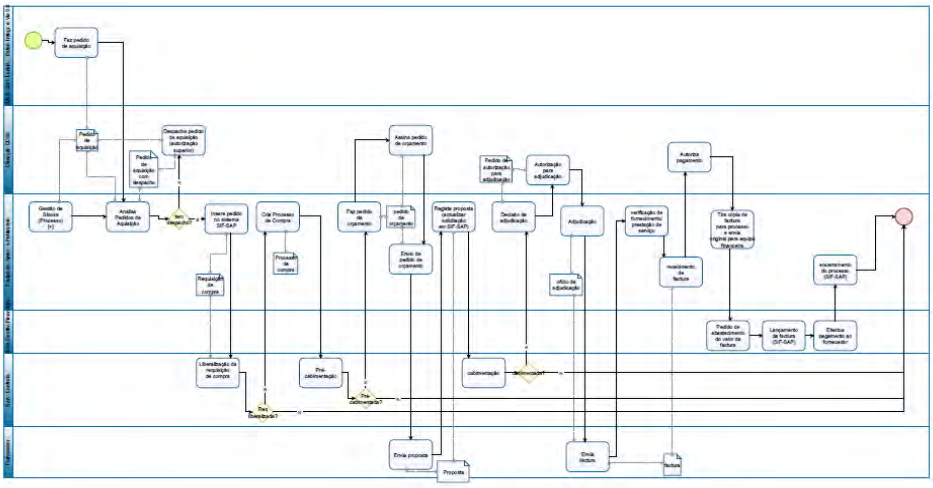

Faz pedido de aquisição
Pedido de aquisição
Despacha pedido de aquisição (autorização superior)
Pedido de aquisição com despacho
Gestão de Stocks (Processo) [+]
Analisa Pedidos de Aquisição
tem despacho?
Insere pedido no sistema SIF-SAP
Cria Processo de Compra
Requisição de compra
Processo de compra
Faz pedido de orçamento
pedido de orçamento
Assina pedido de orçamento
Envio de pedido de orçamento
Regista proposta (actualizar solicitação em SIF-SAP)
Pedido de autorização para adjudicação
Autorização para adjudicação
Decisão de adjudicação
Adjudicação
ofício de adjudicação
verificação de fornecimento/ prestação de serviço
recebimento de factura
Autoriza pagamento
Tira cópia de factura para processo e envia original para equipa financeira
encerramento do processo (SIF-SAP)
Pedido de abastecimento do valor da factura
Lançamento da factura (SIF-SAP)
Efectua pagamento ao fornecedor
Liberalização da requisição de compra
Req liberalizada?
Pré-cabimentação
Pré-cabimentada?
cabimentação
Envia proposta
Proposta
Envia factura
factura

Fig. 11. New version of Procurement and Materials Management” business process

Table 1 shows the summary of the results obtained globally in the two case studies developed where MAPA v2 was used to maintain the as-is business process model updated: Social Security and Technology School of Castelo Branco.

This table discards the results of Social Security phase 1 case study because this case used MAPA v1 and Huf Portuguesa because in this latter case study MAPA v2 was still only used to translate the former business process model to a new business process model modeled with BPMN.

Table 1. Summary of Case Studies Results

PROASIS Cycle (2)	Summary of results (CDSSCB - phase 2 + ESTCB)	
Modeling	Number of Processes	21
Annotation	Number of annotated processes	16
	Number of annotations made	36
	Percentage of annotated processes	76,2%
	Number of textual annotations	22
	Percentage of textual annotations	61,1%
	Graphical annotations	14
	Percentage of graphical annotations	38,9%
Review	Number of reviews	45
	Percentage of reviewed annotations	69,4%
	Reviews of type “I agree”	44
	Reviews of type “I do not agree”	1
Evaluation	Number of evaluations	29
	Percentage of evaluated annotations	80,6%
	Avaliações of type “approval”	30
Modeling	New process versions (by cycle)	16+2
	Percentage of updated processes (1st cycle)	76,2

From the 36 annotations made, 52.8% were of type “correction”, 22.2% of type “adaptation” and 25.0% were of type “detail”. These results are directly connected with the validation of first business process modeling versions that MAPA allowed.

Also relevant is the relatively high percentage of graphical annotations: 38.4%.

Interviews with organizational actors involved in the case studies were conducted in order to systematize the qualitative results.

From the responses to the interviews conducted to validate the results, 93% of organizational actors have made textual annotations and 40% used graphical annotations, we can see that all actors considered important to discuss the process updating and the widespread use of the MAPA tool to all processes of the organization where they work. 94% of organizational actors considered important the discussion possibility that the reviews allow. The organizational actors also expressed their opinion about the benefits that regular use of the MAPA tool could bring to organizations (figure 12).

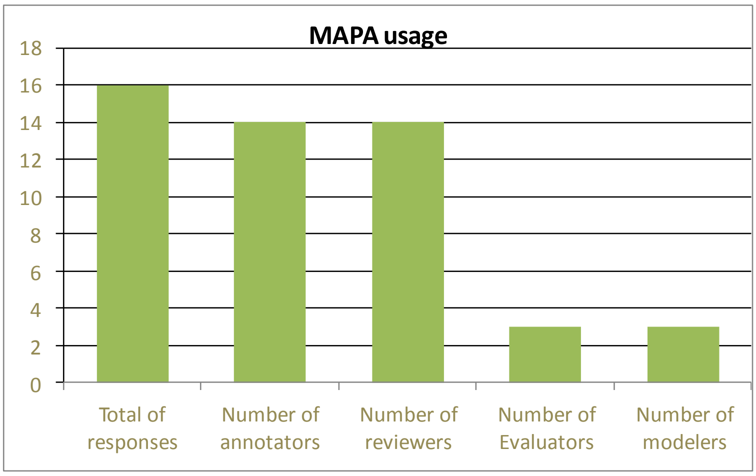


Fig. 12. Number of Answers from users interviews

Some empirical observations were extracted from each case study: in Social Security Phase 1 there was a great ease in the recognition of the process models within the MAPA tool mainly because the executers and leaders were involved in the initial modeling process; in Huf Portuguesa, the operation of the MAPA tool was initially tried with business processes modeled within the Department of Information Systems, but that has revealed to prove fruitless due to the non recognition of processes by their executors and leaders, but we found that MAPA tool can be used as a modeling tool; in Social Security Phase 2, organizational actors showed a greater easiness in interacting with the tool and a great acceptance in using graphical annotations at the expense of textual annotations because they were involved in the initial modeling stages, and we found that the assignment of the modeler role to the process owner provided a considerable improvement in updating processes.

6. Conclusions and Future Work

The results of case studies have demonstrated that the executors of activities, through their participation in PROASIS, established a basis for understanding the business process models and found a channel to express their own vision of the models. From the interviews conducted we found that 88% of the organizational actors involved have proposed updates to the models through annotations, and said they found it a proper way to formalize their proposals. So we can consider that the annotation an appropriate mechanism to express model updates and that the reviews and evaluations are appropriate to promote the negotiation necessary to make the annotations effective.

The annotators have demonstrated availability to make graphical annotations and the modelers have demonstrated availability to create new versions of model (25% of the organizational actors involved in the case studies made updates to the models and 25% of organizational actors who proposed updates to the models have made it through graphical annotations). These actors see the possibility to actively model collaboratively the organization as important. MAPA has an important role in gathering the information needed to update the model, opening of communication channels that encourages the collection and sharing of knowledge about organizational activities, allowing actors to play the role of active modelers in a collaborative and distributed way, making them organizational modelers.

PROASIS is important in the growth of self awareness, providing explicit representations to the organizational actors that are left with a better sense of what they do and the surrounding context. PROASIS also increase group awareness around processes and activities, creating the history of annotations (and its negotiation/discussion) that culminate with the proper evolution of the modeled processes, aligning them with their implementation in practice.

Some of the concluding remarks that we can make about this work could lead to future work in this area. For instance, PROASIS could have an important role in helping the redesign of business processes in the BPM lifecycle, which is usually based on initial modeling techniques including interviews and meetings, and is triggered by management needs. Instead, PROASIS can run continuously in the background, being triggered asynchronously by the actors, and can significantly reduce the effort in redesigning and updating the as-is process model.

We also recognize the action level of detail as important to try to understand the mental mechanisms used by organizational actors to act in the context of PROASIS, so it is important to consider for future development, personal areas in which organizational actors can declare their actions as they see it in the context of their work.

Apart from the case studies presented, PROASIS and MAPA are being used autonomously in other organizations for the collaborative construction and updating of business processes; MAPA tool is being used at Viriato Theatre Company of Viseu, where nine business processes were modeled, some with multiple versions due to PROASIS; It is also being used at the University of Algarve, where a total of 44 business processes have been modeled, which are being annotated and reviewed by a group.

References

1. Castela, N., Zacarias, M., Tribolet, J.: Proasis, As-is business process model maintenance. In. Frank Harmsen, Knut Grahlmann, Erik Proper (Eds.) Lecture Notes in Business Information Processing (LNBIP), Proceedings of the 3rd Conference on Practice-driven Research on Enterprise Transformation, Luxembourg, Springer, Berlin Heidelberg, Vol. 89, 53-82. (2011)
2. Vaishnavi, V. and Kuechler, W.: Design Science Research in Information Systems, Association for Information Systems (2004) [online]. Available: http://desrist.org/desrist (current October 2011)
3. Zacarias, M, Magalhães, R., Caetano, A., Pinto, H.S., and Tribolet J.: Towards organizational self-awareness: An initial architecture and ontology. In P. Rittgen (Ed) Ontologies for Business Interactions, IDEA Group, 101-121. (2008)
4. Morgan, G.: Images of the Organization. Sage, Thousand Oaks, Canada. (1996)

5. Bohm, D: Wholeness and the Implicate Order. Routledge. London, UK. (1980)
6. Mitleton-Kelly, E. and Land, F. Complexity & Information Systems: Blackwell Encyclopedia of Management on Management Information System. Blackwell (2nd edition), London, UK. (2004)
7. Kauffman, S.: The Origins of Order: Self-Organisation and Selection in Evolution. University Press. Oxford, UK. (1993)
8. Heylighen F.: Self-Organization, Emergence and the Architecture of Complexity. In Proceedings of the 1st European Conference on System Science, (AFCET), , Paris, France, 23-32. (1989)
9. Habermas, J: The Theory of Communicative Action: Reason and the Rationalization of Society.Beacon Press, Boston, USA. (1984)
10. Inlayson, J. G.: Habermas. University Press, Oxford, UK. (2005)
11. Magalhães, R., Sousa, P., Tribolet, J.: The Role of Business Processes and Enterprise Architectures in the Development of Organizational Self-Awareness, Polytechnical Studies Review, Vol. VI, No. 9, 1-22. (2008)
12. Tribolet, J.: Organizações, Pessoas, Processos e Conhecimento: Da Reificação do Ser Humano como Componente do Conhecimento à "Consciência de Si" Organizacional, in Amaral, L (Ed.), Sistemas de Informação Organizacionais, Sílabo, Lisbon, Portugal. (2005)
13. Adamides, E.D., Karacapilidis, N.: A knowledge centred framework for collaborative business process modelling, Business Process Management Journal, Vol. 12, No. 5, 557-575. (2006)
14. Rittgen, P. Negotiating Models. In Krogstie, John; Opdahl, Andreas; Sindre, Guttorm (Eds.): Advanced Information Systems Engineering, 19th International Conference, CAiSE 2007, Trondheim, Norway, June 2007, Proceedings, LNCS No. 4495, Springer, Berlin, Germany, 561-573. (2007)
15. Borghoff, U., Schlichter, J.: Computer-Supported Cooperative Work: Introduction to Distributed Applications, Springer, Berlin, Germany. (2000)
16. Ssebuggwawo, D., Hoppenbrouwers, S., & Proper, E: Interactions, goals and rules in a collaborative modelling session. In The Practice of Enterprise Modeling. Springer, Berlin Heidelberg, Germany, 54-68. (2009)
17. Giorgio Bruno, Frank Dengler, Ben Jennings, Rania Khalaf, Selmin Nurcan, Michael Prilla, Marcello Sarini, Rainer Schmidt and Rito Silva: Key challenges for enabling agile BPM with social software, J. Softw. Maint. Evol.: Res. Pract. Vol. 23, John Wiley & Sons, USA, 297–326. (2011)
18. Gonzalez, P.P., Framinan, J.M.: Business Information Systems: Concepts, Methodologies, Tools and Applications, Vol. 4. Ed. Information Resources Management Association, USA, 636-648. (2010)
19. Ellis, C.A., Gibbs, S.J. and Rein, G.L. Groupware: Some Issues and Experiences. Communications of the ACM. Vol. 34, No.1, 38-58. (1991)
20. Castela, N., Dias, P., Zacarias, M. and Tribolet, J.M.: Collaborative maintenance of business process models, Int. J. Organisational Design and Engineering, Vol. 2, No. 1, 61–84. (2012)

Nuno Castela is an Assistant Professor at Polytechnic Institute of Castelo Branco, Portugal, and an Invited Researcher in Center for Organizational Design & Engineering of INESC Inovação. Currently he is Vice President of Polytechnic Institute of Castelo Branco and he was President of the Scientific Council (2012-14) at Technology School of the Polytechnic Institute of Castelo Branco, where he teaches several courses in the area of information systems. He holds an Electrical Engineer degree (1996), from University of Coimbra, a Computer Engineer MSc degree (2001) and an Information Systems and Computer Engineering PhD degree (2011), both from Instituto Superior Técnico - University of Lisbon. His main academic interests are in the areas of Business Process Modeling, Enterprise Architecture and Enterprise Engineering, and he has published several refereed papers in those areas.

Paulo Dias is an Information Systems Technician and an Invited Teaching Assistant at Polytechnic Institute of Castelo Branco. He holds an Information Technology Engineer Degree and a Information Systems and Technology MSc degree, from the University of Beira Interior, Covilhã, Portugal.

Marielba Zacarias is a Computer Engineer (1982) with a MSc degree in Systems Engineering (1996), and a PhD in Informatics and Computers Engineering (2008). She has a professional experience of over 25 year in the Information Systems field. She is an assistant professor of Electronics and Informatics Engineering Department and member of Research Center for Spatial and Organizational Dynamics (CIEO), at the Algarve University. After her PhD, she has worked mainly in organizational modeling and ontologies, focusing particularly in modeling organizational contexts, and has published 28 refereed scientific papers in those areas.

José M. Tribolet is Full Professor of Information Systems at IST, (Instituto Superior Técnico), the Engineering School of University os Lisbon in Portugal, and holds a joint appointment as Full Professor at the Engineering and Management Department of IST. His main academic interests are in the areas of Enterprise Architecture, Enterprise Engineering and Enterprise Governance, with emphasis on IS Architecture, Business Process Engineering, Enterprise Transformation and Change Management. He holds a Ph.D. in Electrical Engineering and Computer Science from MIT (1977). He spent a full sabbatical year (1997-98) at MIT´s Sloan School of Management and was Visiting Professor at the University of St. Gallen, Switzerland, in the spring term of 2011-2012. – He founded in 1980 the first non-state owned research institute in Portugal, INESC – Institute for Systems and Computer Engineering. INESC is today a holding of six research institutes nationwide, three of them having become formal Associated Laboratories of the Portuguese Science System. Dr. Tribolet is the President of INESC. He is a founding member of the Portuguese Engineering Academy and a founder of the Informatics Engineering College of the Portuguese Engineers Professional Association.

Received: January 17, 2013; Accepted: May 30, 2014

Computer Science and Information Systems 11(2):481–501 DOI: 10.2298/CSIS131130032B

Facilitating the Interaction with Data Warehouse Schemas through a Visual Web-Based Approach

Clemente R. Borges and José A. Macías

Escuela Politécnica Superior, Universidad Autónoma de Madrid
Francisco Tomás y Valiente 11.
28049 Madrid, Spain
{clemente.borges, j.macias}@uam.es

Abstract. In most respects, there are implicit drawbacks concerning representation and interaction with data in relational-database applications. On the one hand, there is a lack of expressiveness and ease of use in the user interfaces that handle such data. On the other hand, there is an implicit need for interactive end-user visual tools to query data and avoid dependency on programming languages. The main aim of this work is to study the problem of database interaction and usability, comparing existing solutions and providing a new approach that overcomes existing problems. We propose a web-based tool that manipulates Data Warehouse schemas by using a visual language to represent the database structure and providing several visualization techniques that facilitate the interaction and creation of queries involving different levels of complexity. We based our research on an End-User Development approach that has been evaluated to obtain some initial usability indicators.

Keywords: Human-Computer Interaction, User-Centered Design, Usability, User Interface for Data Warehouse.

1. Introduction

Over the last decade, database research has produced notable results concerning architectural and storage facilities. This has enabled database systems to deal with large volumes of information for different purposes. Nonetheless, this advance does not necessarily mean that the data can be quickly and easily accessed by end-users. In this sense, it is essential to take into consideration the logical data models, which are the basis for a successful modeling and management of information. The Relational Model [1] can be considered as one of the most commonly used. This model has been worldwide used and extended for the management of large volumes of tables and complex relationships. In general, to outweigh the use of the Relational Model, it is necessary to consider two different problems that claim a common solution: the management of data models and the effective construction of query-based applications that make easier the way end-users can obtain information from the data models.

The Data Warehouse (DW) [2], [3] was actually conceived as a solution for the management of large relational models, as it comprises a specialized data structure to easily access relevant business information and serve as a solution to improve decision-making processes. DW models are created using a different database modeling technique that allows having a reduced number of tables, also producing data that are easier to understand and represent.

The collection of tables in a DW is called a schema. A typical schema consists of a fact table and several dimension tables that are related to the fact table. The fact table represents the business measures (numeric values), whereas the dimension tables contain dimensional attributes that describe the facts (textual descriptions). This type of structure is commonly known as a star schema, and it is based on the dimensional modeling. One of the most important characteristics of this type of model is its simplicity [4]. Users who utilize a star schema are benefiting from the simplicity of data that are easier to understand, navigate and query [5]. In addition, queries based on a star schema become simpler and facilitate the visual authoring in the query-building process.

All in all, there are also some implicit problems concerning the understanding of data models by end-users. Such difficulties are somehow related to the applications and tools being used to manage the data. Everyday more non-expert-in-query-language users deal with data tools to face daily problem-solving activities. In general, a trade-off between expressiveness and ease of use is mandatory in order for end-users to manage data and thus reduce the user's cognitive load regarding the interaction with database systems. In other words, it is necessary to get the balance right between the richness of the visual languages and the implicit ease of use provided by the application functionality.

The aim of our work is to facilitate the process of query building to non-expert end-users in DW environments. To validate our approach, we have developed a tool called VISQUE (VIsual Star schema QUery by Example) [6], [7], which aims at creating queries by example focused on Data Warehouses and more specifically on the star schema, allowing database users, not necessarily experts in SQL, to build queries in a visual and easy way, but featuring fewer expressive limitations than other similar tools.

We think that star-schema models are easier to represent in a visual manner, and also they seem to be more natural for end-user understanding. To this end, we have based on the End-User Development (EUD) paradigm [8] as a solution that provides a set of methods, techniques, and tools that allow users of software systems, which act as non-professional software developers, to create, modify, or extend in some way a software artifact.

Our research is focused on specific instances of EUD approach such as the Programming by Example [9], [10]. In Programming by Example (PBE), the application observes and gathers information while the user proves how to perform a task that is used by the system to automatically complete similar tasks later. A PBE system can make the programming process easier by providing a visual representation of the objects and avoiding the need to understand programmatic specifications, therefore inferring the user's intentions based on (mostly graphical) demonstrations. In consequence, our research is more specifically focused on the Query by Example paradigm (QBE) [11], which is an extension of PBE dealing with the creation of queries by using graphical demonstrations and visual elements.

More specifically, our research can be summarized in the following objectives:

1) Provide end-users, who do not have to be necessarily experts in query languages, with facilities for the construction of queries based on DW star-schema models.
2) Offer appropriate mechanisms for the visualization and interaction with star schemas, including a visual language that facilitates the interaction between end-users and the database.

3) Improve the expressiveness in the creation of visual queries by considering advanced operators (such as set-based and nested queries) that are commonly unsupported or unavailable in other existing approaches.
4) Reduce the number of steps required to build queries and provide adequate feedback to end-users based on the visualization of immediate real-time results and the multi-visualization of queries, as an effort to improve the interaction process and reduce the learning curve.
5) Design a query-based visual tool that has been evaluated with real users in order to have early usability clues for further improvement, carrying through a user-centered development approach.

These objectives imply a compelling challenge, as the proposed tool should be aimed at end-users with general knowledge on databases and related tools such as Microsoft Access or Excel, for instance, but not on Data Warehousing. Moreover, the users may have knowledge on SQL or just a basic level of it.

However, development issues have to be considered as well, taking into account usability features, allowing a high level of user satisfaction and reducing the learning burden at the same time [12].

This paper is organized as follows. In Section 2, we present previous work along with a study of some related tools in the field. Section 3 describes our approach in terms of the elements of the visual language and a worked use case. In Section 4, we present a comprehensive evaluation of the tool, discussing the most important results. Finally, Section 5 reports on the main conclusions and activities to be considered as future work.

2. Related Work

Over the last decades, several visual tools and languages using different interaction methods have been created to assist users in building queries for different domains [13]. In this sense, our aim was to study the most important works by means of a comparative evaluation to find common drawbacks and carry through an innovative solution for querying Data Warehouse schemas.

For a long time, QBE has been the inspiration for several visual query tools such as XQBE (XQuery By Example) [14], among others. Generally, example-based languages have been traditionally used to encourage users to build queries by specifying a visual example of them on a user interface. There are several example-based database query languages [15], which are principally based on a tabular representation of relations with boxes that are filled in by users, including constants and example elements (such as variables and field names). Also, there have been proposed several dynamic queries interfaces [16] allowing database users to obtain interactive control over the construction of different kinds of queries. Such approaches include many principles of direct manipulation that reduces the difficulties in programming [17], such as visual presentation of the query´s components, visual presentations of results, selection by pointing (versus typing), and immediate and continuous feedback, among others.

Typical example-based tools can be found in Microsoft Access [18] or SQL-Server Query Designer [19], which are specifically based on QBE and thus have limitations in the construction of complex queries visually such as set-based or nested ones.

By contrast, Hyperion Interacting Reporting tool [20] provides a manipulation metaphor based on drag & drop, which represents the database tables as boxes (a common feature in many query-building tools), and includes a selection area with fields and filters that allow to build more complex queries. However, this approach forces the user to select and relate the data model tables, increasing the level of skill required for end-users to create a query. There are other open-source tools like Eclipse Birt Project [21], which provides facilities for creating queries, but it also lacks of visual expressiveness and is not suitable for unskilled users, as it demands to manipulate SQL code manually.

A different approach is presented in Spago BI QbE Enterprise Scope [22], which proposes the creation of queries through a wizard that significantly reduces the expressiveness of query constructions. This tool provides an initial data model as a structure in which there is a clear distinction between measures and dimensions, but by contrast it lacks the visual richness and variety of (relational) query operations as well. Additionally, Tableau Desktop [23] is the commercial version of a research tool called Polaris [24], an application focused on Data Warehouse schemas. However, this tool utilizes a classification method for measures and dimensions based on the kind of the data related to the table field, implying that even relationship fields are displayed as measures being non-relevant for the query building.

Spotfire [25] is another related tool including a database exploration system based on interactive information visualization. It allows working with different types of data sources and displays default visualizations along with several objects that users can manipulate according to the data type of each field in order to adapt the visualization to their needs. However, users have to select and manipulate tables and fields when importing data from a database. On the other hand, the Active Query Builder tool [26] includes an environment to address a variety of complex queries visually. However, this tool utilizes the typical representation of tables as drag & drop boxes in which users have to deal with several complex concepts and relations, and they usually need to understand the model very well in order to extract valuable information.

In order to highlight potential shortcomings, we now analyze some well-known tools and applications considering their most important functionality related to usability issues. All those existing products have already been fully implemented, and therefore they can be tested and analyzed in detail. The competitive analysis is a HCI analysis technique that evaluates existing tools to see how acceptable they are in terms of the functionality and interaction techniques that they provide, in order to improve and optimize requirements for a new system [27].

The objective of the competitive analysis that we report here is to gather different issues provided by some common used tools, in order to identify usability design problems in the user interface. In our case, this analysis helped to provide ideas for the design of VISQUE, obtaining a list of successful guidelines and also those to be avoided. The idea was to assess the cognitive load and the sequence of steps that are typically required to build queries (specifically focused on Data Warehouses). To do this, we selected the following popular tools:

1. **Microsoft Access- Query By Example**
2. **Hyperion Interactive Reporting**
3. **Spago BI QbE Enterprise Scope**
4. **Crystal Reports**
5. **Eclipse Birt Project**

6. **Tableau Desktop**

In order to carry out the competitive analysis, we performed a typical sequence of steps for the DW query building on each of the above-mentioned tools. In short, brief descriptions of the interactive aspects reviewed are outlined:

- **Database table visualization method**: indicates the method used to visualize the data-model tables that are available to create a query. Drag & drop boxes are the most typical representations of database tables used. Moreover, the tree view is another typical mechanism to display database tables.
- **Table selection method**: implies the type of interaction used to select the tables involved in a query (i.e., tables included in the clause FROM in a SQL query).
- **Field selection method**: indicates the type of interaction used to select fields from the tables available for the creation of a query (i.e., fields that follow the clause SELECT in a SQL query).
- **Filter selection method**: represents the type of interaction used to select the filters that apply to a particular query (i.e., the logical conditions that are part of the clause WHERE in a SQL query).
- **Function selection method**: indicates the type of interaction used to apply an aggregation function, which will be applied to a particular field in the query (i.e., functions such as SUM, MAX, MIN, AVG, etc., in a SQL query).
- **Order selection method**: embodies the type of interaction used to apply specific sort criteria over any of the fields included in a query (i.e., fields that are part of the clause ORDER BY in a SQL query).

On the other hand, a couple of specific features were considered, as they are rare or difficult to find in most query-building applications. Such features have been taken into account in the design of VISQUE. Specifically, we have evaluated the following features:

- **Set-based operations**: this functionality refers to the ability to create set-based operations, such as union and intersection, in a visual way.
- **Nested queries**: refers to the ability to visually represent nested queries (SQL clauses IN and NOT IN).

Based on the competitive analysis conducted, we present a summary of the main problems found in the reviewed tools:

- In the selection of aggregation functions in fields, we found several problems. First, in many tools it is not obvious to find this functionality. In the case of Hyperion Interactive Reporting, the action is achieved by following a long sequence of steps, which does not facilitate the memorization by end-users. Finally, in several cases, the user is allowed to apply aggregation functions in fields where the data type is not supported, and therefore this does not help preventing errors. For example, it does not make any sense to apply an Average (AVG) aggregation function to a text field.
- In the selection of query filters, the main shortcoming found in most tools is the inability to suggest possible values existing in the database to create a filter on a field. As a result, the user needs to know or guess these values in advance.
- The Eclipse Birt Project tool does not provide a good visual SQL query abstraction. For example, when the user selects fields, these are automatically

added to an SQL sentence that is displayed. Users must add commas to separate the fields, so that it forces the user to directly manipulate the SQL code.

- The Crystal Reports tool forces the user to perform complicated tasks, such as set up the relationships between tables, so involving the user in the design of the Entity-Relationship database model. This is obviously a difficult task for non-expert users or those who are unfamiliar with the logical data model.
- The Spago BI tool presents some inconsistencies in the visualization of data schemas. It shows the same dimensions at different levels in the tree view representing the DW schema. This is not really necessary, and it might cause confusion when visualizing the data model.
- The Tableau Desktop tool shows all numeric fields as fact table measures, and the rest of non-numeric fields are displayed as dimensional attributes. Consequently, numeric fields belonging to dimension tables (serving as relationships to the fact table) are included as measures, when they are really not.
- Hyperion Interactive Reporting is the only tool providing a visual representation of set-based operations. However, none of the tools analyzed provides a way to visualize nested queries.

In general, all the mentioned approaches offer a great deal of features and capacities to manipulate data by experts, also increasing the final complexity and the level of expertise required for end-users to create queries.

As we will see in the next section, our approach seeks to overcome these common shortcomings, improving the deficiencies found in the analysis, and taking into account several critical interaction points for the creation of queries. To do that, our approach encourages the automatic representation and visualization of the star schema in order to avoid the selection of tables, and the explicit creation of relationships between tables of the DW data schema as a prelude to the query-building process. In addition, we propose a visual interaction mechanism, which clearly and uniquely identifies each of the elements of the DW data schema. Also, our aim is to reduce the steps required to apply aggregate functions (SUM, AVG, etc.) on the measures, noting that this kind of steps are typically carried out in the construction of queries for DW, as well as the need to restrict these functions only to numeric fields. Additionally, the tool proposes the construction of visual queries involving set-based operations and nested queries, which are visual features unavailable in most common tools, and particularly in those previously reviewed.

3. Our Approach

VISQUE is a web-based tool created through a user-centered development process. In addition to competitive analysis and later usability evaluation, by using paper prototyping [28] it was possible to obtain several versions of the tool appearance from early stages of the development (interaction analysis), simulating the behavior of the user interface, as an attempt to ensure usability and reduce some of the commented shortcomings appearing in other common tools. In this section, the main elements of the visual language will be addressed. Finally, a detailed use case will be presented, showing the creation of a complete query by using VISQUE.

3.1. Elements of the Visual Language

VISQUE includes several main graphical elements comprising the Web user interface at the front-end, namely: a) the visual-schema interactive component, b) the operation selection boxes, c) the results table, d) the set-based and nested queries operations, e) the query tabs, and f) the operation visualization.

a) Visual-schema interactive component

A star schema can be represented as a hierarchical structure. This way, the root node represents the fact table, and the subsequent levels correspond to the measures and dimensions. The tree view is a common graphical user interface model presenting a hierarchical view of information, and it has used for many years in different types of applications.

It is essential to make a clear distinction between measures and dimensions as each of these concepts plays a different role. Figure 1 shows an example of the visual-schema interactive component with several descriptive icons and different colors. The fact table is represented as a yellow root node. Items related to the fact table (measures) are represented in green, and items related to dimensions in red color. When the user clicks on any measure, the node expands and displays a series of icons representing the operations that can be performed on this measure. In the case of dimensions, when expanding one of the nodes representing a dimension table, dimensional attributes are shown initially and then, when selecting a specific dimensional attribute, this expands a new level showing the operations that the user can perform on the dimensional attribute.

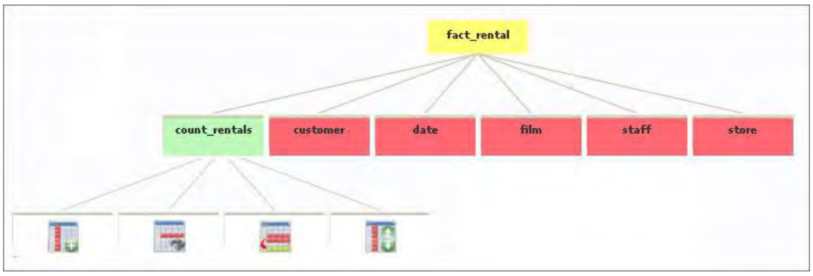


Fig. 1. Visual-schema interactive component

We have created four different icons representing field operations to build queries. Figure 2 depicts the field operators represented by icons that determine specific operations in our tool. The icon numbered as 1 represents a simple selection of fields (SELECT clause), icon 2 concerns the application of filters (WHERE clause). Icon 3 allows the grouping of rows (GROUP BY clause). Finally, icon 4 implies the order to arrange the final results (ORDER BY clause).

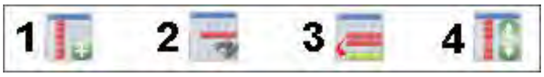
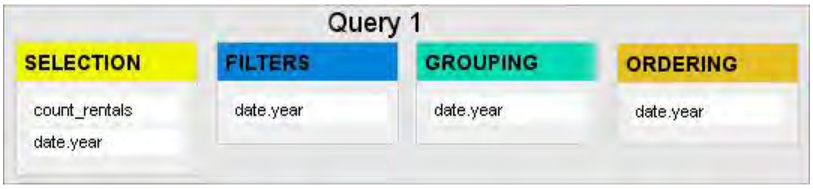


Fig. 2. Icons for field operations

In short, this mechanism provides facilities to end-users in the selection of elements of a data schema, specifying a familiar interactive tree-view based mechanism, clearly differentiating between measures and dimensional attributes, and restricting the kind of operations that can be performed on each of these elements.

b) Operation selection boxes

When selecting a particular action over a field –either a measure or a dimensional attribute, the selection is registered in a specific operation selection box. The selected measures are displayed with a unique name. By contrast, in the case of dimensions, it is displayed the name of the dimension table, followed by a point and the name of the selected dimensional attribute.

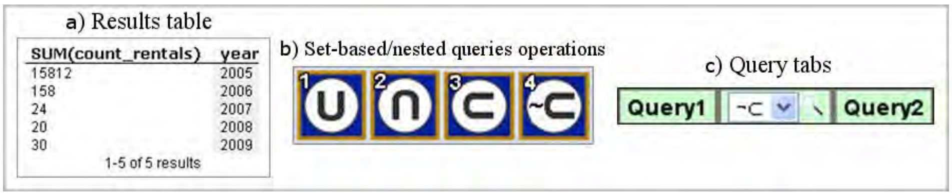


SELECTION	FILTERS	GROUPING	ORDERING
count_rentals	date.year	date.year	date.year
date.year			

Fig. 3. Operation selection boxes

In this case, Figure 3 depicts some user field selections. For instance, the dimensional attribute "date.year" refers to the dimension table "date", and "year" corresponds to the dimensional attribute, intending to avoid confusion in the case of having similar names in different dimensional attributes between the dimension tables.

c) Results table



SUM(count_rentals)	year
15812	2005
158	2006
24	2007
20	2008
30	2009

1-5 of 5 results

Fig. 4. a) Results table, b) Set-based/ nested queries operations and c) Query tabs

Each time a particular operation is applied; a SQL translator automatically processes the query and builds the SQL statement. Once the system has executed the query, it obtains the corresponding output and the results table is then updated with the data retrieved (see Figure 4.a above), showing also the number of rows affected.

d) Set-based and nested queries operations

VISQUE has the ability to tackle multiple queries, using operations that are commonly unusual in other visual query tools. Figure 4.b, shows icons numbered from 1 to 4 representing operations intended to build multiple queries. From left to right, icon one (1) represents the set-based union operation, icon two (2) represents the set-based intersection operation, and icons three (3) and four (4) represent nested queries operations (IN and NOT IN, respectively). These operations provide greater expressiveness and flexibility to build queries in the VISQUE environment, as we will see later.

e) Query tabs

Tabs handle each query independently. Multiple query tabs are displayed when the user operates with set-based or nested queries, showing the query number and the applied operation between queries. Figure 4.c, show the query tabs for a nested-query NOT IN operation ($\not\subset$) corresponding to two existing queries (Query1 $\not\subset$ Query2). Finally, to observe the composite operations, it is necessary to click on the magnifying glass icon located between the two queries, and then the result appears in the operation visualization window that will be explained below.

f) Operation visualization

The operation visualization window displays the visual results of the operation between queries. Moreover, it shows all the field operation icons selected by the user for each query, and the parameters for each operation, that is, selection, filtering, grouping, or ordering. By contrast, it has the disadvantage that it shows too many items onscreen when the number of queries involved in the operation increases significantly (see Figure 8).

3.2. A Worked Use Case

We introduce a use case to briefly describe the process of building queries with VISQUE. To do so, let us introduce the sample database schema shown in Figure 5.

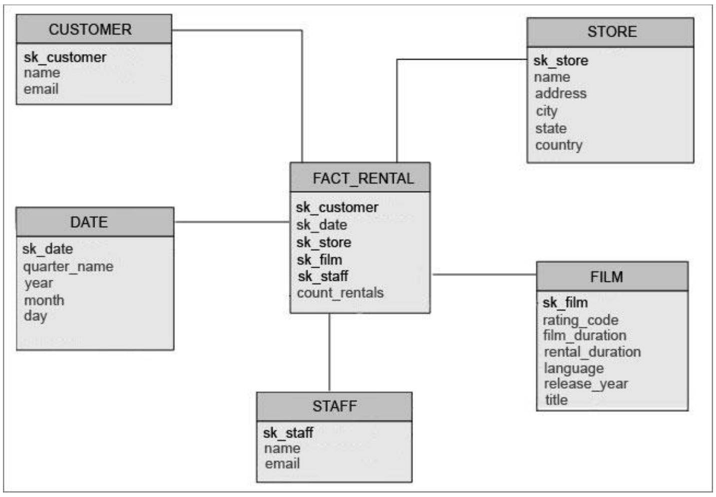


Fig. 5. Star schema to describe the use case including the fact table (FACT_RENTALS) and five dimension fables

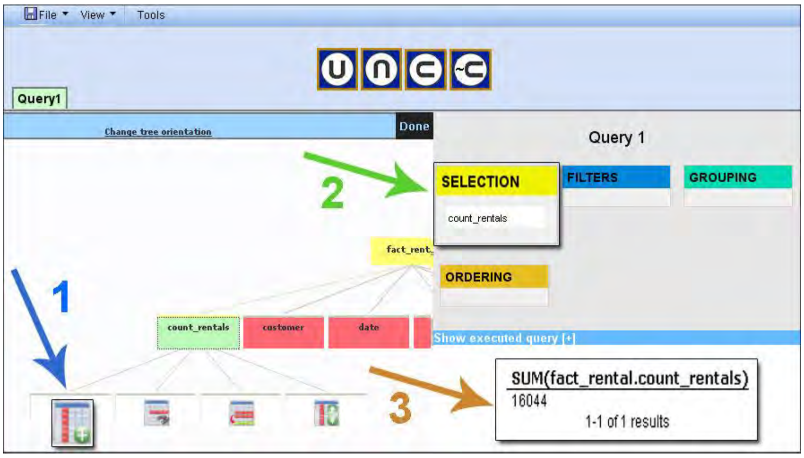
The schema includes a fact table representing movie rentals (FACT_RENTAL) and five dimension tables that represent the films (FILM), the rental date (DATE), the stores in which they have been rented (STORE), the staff who rented each film (STAFF), and the customer (CUSTOMER) who rented each film. Each row in the fact table

(FACT_RENTAL) represents a quantity of rented films (count_rentals). The dimension tables are used to show different views of the measures.

According to the schema shown in Figure 5, let us suppose that the user wants to design a query to obtain the number of films rented in the first quarter of 2007 along with the ones rented during the second quarter of 2008. To do this, two separate queries can be created and linked one another by a union set-based operation. The query can be formally expressed as follows:

$$\left(\Pi_{SUM(count\_rentals)}\left(\sigma_{DATE.year=2007\ \wedge\ DATE.quarte\_name=Q1}\left(FACT\_RENTAL \bowtie_{FACT\_RENTAL.sk_{date}=DATE.sk_{date}} DATE\right)\right)\right)$$
$$\cup$$
$$\left(\Pi_{SUM(count\_rentals)}\left(\sigma_{DATE.year=2008\ \wedge DATE.quarter\_name=Q2}\left(FACT\_RENTAL \bowtie_{FACT\_RENTAL.sk\_date=DATE.sk\_date} DATE\right)\right)\right)$$

Figure 6 shows the sequence of steps to create the first sub-query. Initially, it is necessary to select the field that matches the rented films (count_rentals), clicking on the corresponding element and expanding the options on that field. Then, the user has to click on the operation icon that indicates the simple selection of fields (1), which will be placed in the selection box on the right (2). By default, the tool applies the aggregation function "SUM" when facts are added as part of the selection of fields. However, the user can modify and include any other function by double-clicking the field name in the selection box. Afterward, the tool infers the query and builds a table displaying the query results (3). The results are automatically updated in real time, those being visible in the results box located at the lower right corner that, in this case, displays a single record corresponding to the total number of rented films (i.e., 16044).



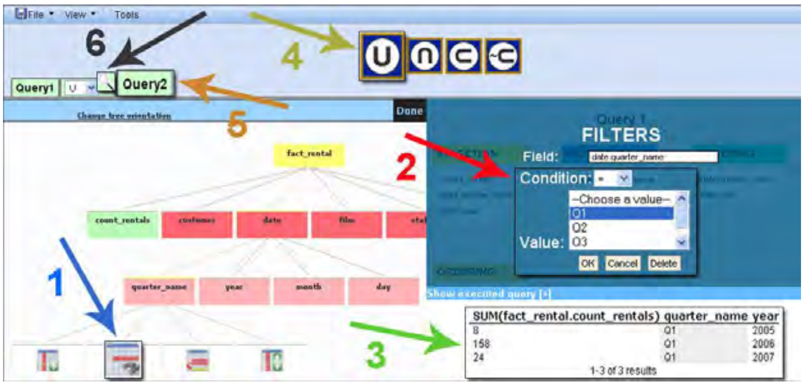
SUM(fact_rental.count_rentals)
16044
1-1 of 1 results

Fig. 6. Fields selection corresponding to the SELECT SQL clause

To carry on with the example, the user can add dimensional attributes by which s/he can observe the facts. In this case, it is necessary to indicate in the query the specific quarter and year.

Both fields come from the table DATE, and they are obtained by using the same selection steps shown before. By default, the tool automatically adds the new selected fields to the grouping box, and therefore the data is grouped by quarter and year. Later on, the user has to indicate the filters to include the results for the first quarter and then the year “2007”.

In Figure 7, the necessary steps to implement these filters can be observed. To carry out this task, the user has to click on the quarter dimensional attribute in table DATE, and then select the icon for the filters (1), which immediately adds the field to the filter box and activates the selection of conditions and values for the filter, suggesting possible values that the user can choose from the dropdown list (2). In this case, the expression "= Q1" is selected. This action automatically updates the results table showing the number of rented films in quarter one, and the data grouped by quarter and year (3).



Field: date.quarter_name
Condition: =
Value: --Choose a value-- / Q1 / Q2 / Q3

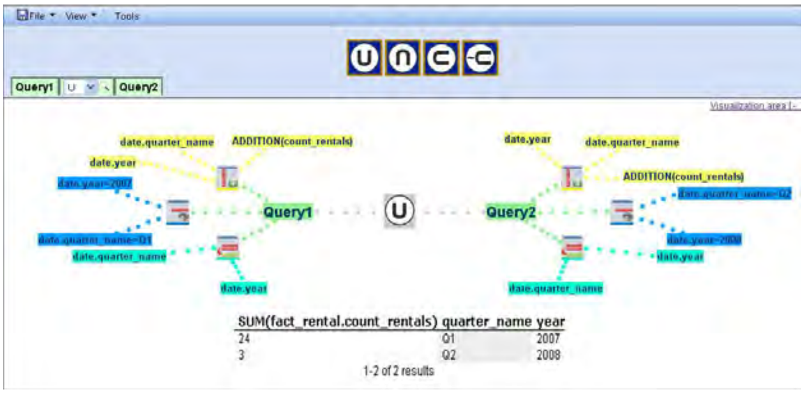
SUM(fact_rental.count_rentals)	quarter_name	year
8	Q1	2005
158	Q1	2006
24	Q1	2007

1-3 of 3 results

Fig. 7. Establishment of conditions for filtering and grouping

Finally, another filter related to the field "year" (in the same dimension table) must be applied, following similar steps as mentioned before for the quarter filter. In this case, the condition "= 2007" provides the first query results referring to the quantity of rented films in the first quarter of 2007. Throughout the process mentioned above, the user completes the first sub-query. Then, the user can create a second sub-query to take into account the rented films in the second quarter of 2008. To do this (see Figure 7), s/he can click on the icon representing the operation UNION (4), allowing the inclusion of a new sub-query (5). Later on, a similar query is built including different conditions and values for the filters (“quarter_name = Q2” and “year = 2008”), in order to establish a result table showing a single record with rented films in the second quarter of 2008. Finally, to see the results based on the union of the two queries together, it is only necessary to click on the icon depicting the magnifying glass that is located between the sub-query tabs one and two (6).

A visual representation of the composite query, as well as the final results, is shown in Figure 8.



SUM(fact_rental.count_rentals)	quarter_name	year
24	Q1	2007
3	Q2	2008

1-2 of 2 results

Fig. 8. Visualization of the query through the UNION operation and the results obtained

For each sub-query, we can see visual elements that indicate the operations that were selected and, in turn, each icon showing the fields that were used, the function that was applied in the selection of facts, the conditions and values through applying information filtering, and the selected field for grouping. Additionally we can observe, at the bottom, an area showing the results, that is, the union between the two data sets representing the main query stated: the number of rented films in the first quarter of 2007 and the second quarter of 2008. The proposed visual language significantly facilitates the visualization of all elements for each query in set-based operations, and enables the possibility to easily construct multiple queries such as $Q_1 \cap Q_2 \cup Q_3 \cap \ldots Q_n$, for instance. This improves the user interaction and the representation of complex queries, which is fairly unusual in other query-by-example tools (as analyzed in Section 2).

In next section, we carry out an initial evaluation of the tool. To do this, a user evaluation is presented, in order to accomplish our objective related to usability.

4. User Evaluation

4.1. The Study

To carry on with our approach, we wanted to have an initial feeling about the general usability of the system in order to improve it further. To perform this task, we have conducted a user test with 12 real users. Specifically, most users were developers having knowledge on programming language: 4 users in PHP programming, 3 users in Java programming, and 1 user on Microsoft .NET platform. On the other hand, another 3 users had knowledge on project management, and the last user serves as manager in an IT department. In particular, there were 11 men and 1 woman aged between 24 and 37, all having general knowledge on databases and related tools, such as Microsoft Access or Excel, but not on Data Warehousing. In addition, only 5 out of 12 users had some basic knowledge on SQL, whereas the rest of them did not have any knowledge on SQL.

The study was focused on the creation of 5 different queries by using VISQUE, representing a notable variety of set-based and nested operation for extracting reliable information from the study. Users were instructed on the tool for about 10 minutes before performing the test, showing them a quick overview of VISQUE, and explaining the purpose of the evaluation. In addition, the evaluator previously commented that the entire session would be recorded, and they could give impressions, questions or doubts along the evaluation. After the training, it was provided a new data model on which users built the selected queries, which consisted on a simple version of the model showing the numeric values they can measure, and the dimensions in which they could give different measure perspectives. After that, the evaluator left them alone until they finished building all the queries. By using the Retrospective Testing, [29] a usability testing technique that records the whole interaction between the end-user and the tool, all the interaction process was recorded on audio and video. This way, the execution of each task performed by the user was studied and analyzed in detail to extract further information about the interaction, so identifying important critical incidents and usual difficulties on selecting specific data collections. The aforementioned technique included the Thinking Aloud protocol [30], which allows analyzing user interaction during the study, with the purpose of obtaining information about their thoughts, feelings and opinions while interacting with VISQUE. This allows observing behavior patterns and/or phrases that may provide clues about the user satisfaction while interacting with our tool.

Once the interactive session with the tool was finished, users were requested to fill in a questionnaire based on the USE survey [31] but including some variations provided by the Purdue Usability Testing [32] and the Perceived Usefulness and Ease of Use [33] questionnaires. The questionnaire had 31 questions classified into the following four variables or dimensions:

1) «Usefulness» – 8 questions
2) «Ease of use» – 10 questions
3) «Ease of learning» – 6 questions
4) «Satisfaction» – 7 questions

Additionally, we included in the questionnaire a set of four open questions intended to have the user's impressions in several aspects of the tool by including explicit issues such as positive/negative aspects, common usage in daily problem-solving activities and so on, as they help contribute to the user's explicit opinion about VISQUE.

4.2. Results

In this subsection, we present the most relevant results. On the one hand, we introduce the main interactive session and Retrospective Testing outcomes. Also, we discuss the results obtained from the questionnaire including the study of the involved variables.

4.2.1 Interactive Session Results

As mentioned before, during the interactive session users were requested to build five queries having incremental levels of difficulty. The data model used for the study sessions was the movie rental use case presented in Section 3 – Figure 5. The queries were selected in order to maximize the expressiveness of the tool, and also to allow users to perform the tasks in different ways, including query constructions with set-based and nested operations. Although users were encouraged to perform the queries without any specific order, the queries were numbered as follows:

- Q1: Number of rentals of each film.
- Q2: Number of rentals of films released in 2006.
- Q3: Number of films rented in the past five years except in 2007.
- Q4: Name and last name of people who rented films with duration between 60 and 90 minutes.
- Q5: Number of films rented in the first quarter of 2006, the second quarter of 2008, and the third quarter of 2009.

Table 1 shows the time spent for users to build each query – i.e., minimal, maximal and average time, the standard deviation, as well as the total number of trial & errors for each query. This information was useful in order to corroborate the level of complexity of each query (effectiveness) and observe the queries in which users may have had specific problems. It is important to point out that the calculated time refers to the user's query-building time (efficiency) and not to the query execution time.

Table 1. Query-building time results and number of trial and errors

Query	MIN Time	MAX Time	Average	Std. Deviation	Trial & Errors
Q1	0.5 min	2.0 min	0.8 min	0.4 min	0 times
Q2	0.5 min	2.5 min	1.1 min	0.6 min	4 times
Q3	1.9 min	5.0 min	2.8 min	1.0 min	9 times
Q4	1.2 min	7.0 min	3.6 min	1.8 min	7 times
Q5	2.0 min	8.0 min	4.0 min	1.4 min	6 times

As presented in Table 1, Q1 can be considered as the simplest query. Users were able to quickly perform the query without major difficulties. Also, Q2 was relatively easy to perform by users, although it resulted a little bit more complex than the first one. Q3 is a bit more variable in terms of the time obtained, mainly because users decided to build it in two different ways, that is, the query can be constructed by using the NOT IN operator (nested-query) or by using filters. However, only half of the users (6 users) decided to use the former facility, whereas the rest of the users preferred the usage of filters. Additionally, Q4 also shows some variability in the construction time as the query could be performed by means of the intersection of two queries (INTERSECT operation), or by applying two different filters. Five users decided to conduct the query with the set-based intersection operation, and the rest (7 users) decided to do it by applying two filters. Finally, Q5 had the highest average construction time, as it required the union of three separate queries (UNION set-based operation), and all the users decided to apply the UNION operation anyway.

In a nutshell, the average time to carry out all the queries was 12 minutes and 30 seconds. The longest time to perform the tasks was 20 minutes and 40 seconds, whereas the shortest time was 6 minutes and 20 seconds. The standard deviation was 3 minutes and 45 seconds. In addition, a total of 6 users (the half of them) decided to utilize nested queries. We particularly noticed that 5 out of 6 users using nested queries in Q3 needed at least 1 trial-and-error attempt in order to accomplish the query. By contrast, only 3 out of 6 users needed 1 trial-and-error attempt to finalize the query when using filters. In consequence, we can say that users using only filters for Q3 were more effective. Similarly, regarding the query involving an intersection operation (Q4), 5 out of the 12 users tended to use this facility, and 4 out of those 5 needed at least 1 trial-and-error attempt to conclude the query. The rest of them (7 users) decided to apply filters, and only 2 of them required at least 1 trial-and-error attempt. Therefore, users applying filters in Q4 were also, in this case, more effective. Finally, the set-based union operation (in Q5) was carried out clearly by all users (12 in total), and 5 users needed at least 1 trial and error attempt to finish the query.

4.2.2 Questionnaire Results

As mentioned above, once the interactive session was finished, users were requested to fill in a satisfaction questionnaire consisting of 31 questions divided into four categories, namely: (a) «Utility», (b) «Ease of use», (c) «Ease of learning», and (d) «Satisfaction».

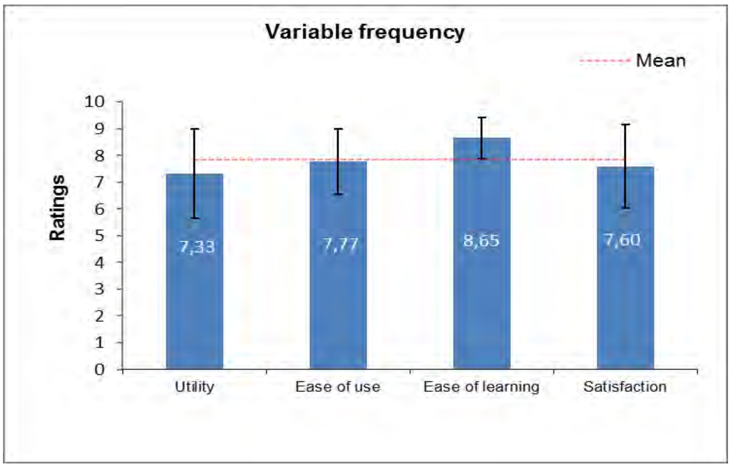


Fig. 9. Mean for each of the four variables, error bars (± σ) and global mean for the four variables.

The users had to respond to each question in a scale ranging from 1 (completely disagree) to 10 (completely agree). Finally, each user was requested to answer four open questions about the general perception of VISQUE, in order to obtain and identify additional comments about strengths and weaknesses to improve the tool.

Initially, to measure the internal consistency of our questionnaire, we calculated the Cronbach's alpha. This measure ranges between 0 and 1. The questionnaire used for this study obtained a reliability value of 94.8% (α = 0.948) for the 31 statement values included. Consequently, we can conclude that the questionnaire had an acceptable level of reliability (an acceptable reliability is considered from 80% onwards).

Next, we proceeded to calculate the mean for each rating provided by users for each question (grouped into the four variables commented before). Furthermore, the standard deviation was also calculated for each variable to determine the dispersion. Figure 9 shows a bar chart including the mean values of all user responses for each grouped variable, as well as an error bar corresponding to the dispersion (± σ) for each one.

The overall mean for all variables is 7.83, with a standard deviation of 1.40. The highest rated variable was «Ease of learning», obtaining an average of 8.65 with a standard deviation of 1.66. This implies that the visual representations are appreciated by users as they reduce the learning burden greatly. The variable related for «Ease of use» obtained an average of 7.76 with standard deviation of 0.77 (the lowest dispersion). Consecutively, «Satisfaction» obtained an average of 7.59 with a standard deviation of 1.55. Finally, «Utility» variable had an average of 7.33 with a standard deviation of 1.66, representing the lowest value among all the variables. The «Utility» value mostly depends on the user's projection about the usage of the tool in (her/his) commonplace context, and so this value usually have a higher deviation. However, in general, the four variables were above 7.0, which is a good indicator for the VISQUE's early usability measurement.

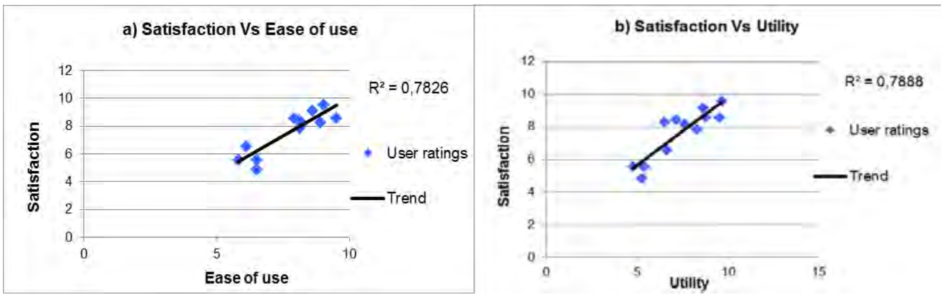


Fig. 10. Correlation between pairs of variables: a) satisfaction vs. ease of use, and b) satisfaction vs. utility

Moreover, we analyzed the correlation degree between all variables (multivariate correlation) by using the Pearson correlation coefficient. This way, we found out a positive correlation between «Ease of use» & «Satisfaction», «Utility» & «Satisfaction» and «Ease of use» & «Utility», having a Pearson coefficient of 0.884, 0.888 and 0.802 respectively. These values are near to 1, indicating a strong correlation. Additionally, we analyzed the correlation degree between those pairs of variables. To do this, we use the average ratings for each variable, comparing the values (one against the other), and calculating a linear regression. Consequently, we corroborated some degree of correlation between such variables.

As depicted in Figure 10, there is a linear relationship between satisfaction, utility and ease of use, which implies that utility and ease of use are probably two of the most important aspects that make users to perceive a high satisfaction. In addition, as shown in the correlation Figure 10.a, there are two clusters of points. The four rating points shown at the bottom of the line correspond to 4 users having knowledge on programming languages (PHP and Java) and basic knowledge on SQL. This indicates that these kinds of user are quite critical about the tool, mainly because of their technical programming and SQL knowledge, therefore their means are lower for this item. Also, Figure 10.b shows a correlation between «Utility» and «Satisfaction» indicating that the users felt more satisfied if the tool covers his/her expectations regarding the work carried in the study.

4.3. Discussion

The measures obtained from the user's interaction with the tool reported interesting results about the user behavior and his/her impression about the application. Specifically, we took into account the following issues:

- Amount of time to build the five queries and the number of trial & errors.
- Audio and video recordings of the interactive sessions.

Regarding the construction of the five queries, simpler constructions were undertaken in a short time without major drawbacks. More complex queries required, in some cases, trial and error to be performed, and thus an extra time to be completed. However, an interesting result was the construction of queries using the UNION set-based operation, as all users tended to use this operation. Generally speaking, this indicates that the proposed visual language met the user's mental model, exploiting expressive facilities especially in the construction of set-based query operations, which is not common in other similar tools. As corroborated by the empirical results, this also denotes a good learning curve.

In addition, 5 users achieved the intersection operation, while the rest decided to use filters. Regarding nested-queries operations, the result is neutral because half of the users used this option and the rest of them used filters. However, according to the trial & error results, the usage of filters seemed more natural and easy in both cases, probably because users already knew how they worked from previous queries (Q2), and also because filter constructions might be closer to the user's mental model than intersection ones. In any case, users were able to perform queries in a way or another, highlighting the expressiveness provided by the tool so that users can successfully build queries in different ways with an acceptable ease of use.

Additionally, the audio and video recorded in the interactive session reported valuable information for improving the tool, analyzing the user's behavioral patterns in the construction of queries. For instance, despite the clear differentiation between measures (in green color) and dimensional attributes (in red color) in the VISQUE interface, we noticed that some users begun to include dimensional attributes firstly and, in consequence, they got empty results since all the information depends on measures. However, VISQUE can be improved in this sense, in order to guide users to select measures firstly. Also, video sessions allowed detecting some minor bugs to be considered in future improvements. However, the tool had an acceptable fault tolerance, giving feedback to the user when an error occurred.

The second phase of the study, related to the usability questionnaire, also provided interesting results about the user's perception of the tool. Some correlations were detected between pairs of variables. Specifically, we found some level of relationship between «Ease of use» and «Satisfaction», and between «Utility» and «Satisfaction».

Additionally, such relationships between variables were corroborated through a multivariate correlation analysis.

According to this first study, used to measure and improve the tool in an iterative and incremental end-user-centered development process, we can affirm that most users found the tool, the visual mechanisms and the interaction method useful for the construction of queries. Also, the most highly rated aspects reported by users in the open questions were, to cite a few: the facility for extracting large amounts of data from a Data Warehouse, the intuitiveness provided, the implicit advantages in the usage of visual set-based operators, the interface look-and-feel and the short time took to understand it, the visual impact of the results in real time, and the expressiveness for building queries, among others.

5. Conclusion

Generally, databases can be considered as the main data source for most information systems. However, the usability of the data application and its user interface can be implicitly affected by data-model design. Consequently, it is necessary to consider aspects of data modeling and user interface design together in order to increase the overall usability [34]. Based on these arguments, the work reported in this paper is an attempt to facilitate database query on Data Warehouses, which are specific database structures helping reduce complex relationships in data modeling and simplify the user's mental model overall. Our work is mainly focused on end-users with some general knowledge on databases and related tools such as Microsoft Access or Excel, but not on Data Warehousing, and with little or inexistent knowledge on SQL.

By means of the competitive analysis carried out on related tools, we missed some features that are difficult to find in most query-building applications, such as an easy way to add aggregation functions, the suggestion of possible values when creating query filters, preventing users from setting relationships between tables, set-based operations, and nested queries, among others. To leverage those common problems that we have identified in existing systems, we have developed VISQUE, which provides more expressive advantages for query building on Data Warehouses, and more specifically on star schemas. Furthermore, we contribute with the following features: a visual language and automatic representation of DW model to avoid dealing with relationships and the data model; an interactive and intuitive mechanism to build queries overcoming the drawbacks previously commented, and the inclusion of set-based operations and nested queries.

Our work is mainly focused on a user-centered approach, so that the work is mainly presented on the basis of a user-centered engineering process, highlighting related aspects such as the VISQUE visual-query language, and the usability features provided. This is also one of the reasons why we have based on star-schema models as they are easier to represent in a visual manner and also they seem to be more natural for end-user understanding. In addition, our research is based on a EUD approach and, more specifically, in instances like PBE and QBE that facilitate the way end-users can manipulate database artifacts easily, avoiding the necessity to deal with programmatic specifications, and helping end-users to learn by example.

As for the future work, first, we will improve our tool further considering the results of the test achieved. We plan to improve on guiding the user to select measures in dimensional attributes, as we have identified this issue as a repetitive problem. Also, VISQUE is limited by the database model it can retrieve/interpret. By now, the tool is only capable of automatically recognizing star-schema models. However, we think that the language can be easily expanded to other DW schema models like the snowflake schema. Also, we consider as a promising future work to investigate on data-model usability in order to obtain new findings on how Data Warehouse structures can affect the overall usability in database applications, in order to envision an explicit user-centered development process focused on both data and data interfaces, and ensuring a tradeoff between expressiveness and ease of use [35] by providing specific analysis, design, implementation and evaluation activities at every step of the software engineering lifecycle.

Acknowledgements. This work has been supported by the founded project TIN2011-24139.

References

1. Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM 13, 377-387. (1970)
2. Inmon, W.:Building the Data Warehouse. Wiley Publishing Inc, New Jersey. (2005)
3. Kimball, R., Ross, M.: The Datawarehouse Toolkit: Second Edition. John Wiley and Sons Inc, New Jersey. (2002)
4. Adamson, C.: Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance. Wiley Publishing Inc., New Jersey. (2006)
5. Giovinazzo, W.: Object-oriented Data Warehouse Design: Building a Star Schema. Prentice Hall, New Jersey. (2000)
6. Borges, C., Macías, J.A.: Interacting with Data Warehouse Schemes through a Visual Web-Based Approach. In: Proceedings of the IADIS International Conference on WWW/Internet. IADIS press, Rome, Italy, pp. 314-318. (2009)
7. Borges, C., Macías, J.A.: Feasible Database Querying Using a Visual End-User Approach. In: Proceedings of the 2nd ACM SIGCHI Symposium on Engineering Interactive Computing Systems. ACM, Berlin, Germany, pp. 187-192. (2010)
8. Lieberman, H., Paternò, F., Wulf, V.: End-User Development. Springer, Dordrecht. (2006)
9. Cypher, A.: Watch what I do: programming by demonstration. MIT Press, Boston. (1993)
10. Lieberman, H.: Your Wish is My Command – Programming by Example. Morgan Kaufmann Publishers, San Francisco. (2001)
11. Zloof, M.M.: Query by example. In: Proceedings of the national computer conference and exposition. ACM, Anaheim, CA, pp. 431-438. (1975)
12. Macías, J.A.,Paternò, F.:Customization of Web Applications through an Intelligent Environment Exploiting Logical Interface Descriptions. Interacting with Computers 20, 29-47. (2008)
13. Catarci, T., Costabile, M., Levialdi, S., Batini, C.:Visual Query Systems for Databases: A Survey. Journal of Visual Languages and Computing 8 (2), 215-260. (1997)
14. Braga, D., Campi, A., Ceri, F.: XQBE (XQuery By Example): A visual interface to the standard XML query language. ACM Transactions on Database Systems (TODS) 30 (2), 398-443. (2005)
15. Özsoyoglu, G., Wang, H.: Example-Based Graphical Database Query Languages. Computer 26 (5), 25-38. (1993)

16. Shneiderman, B. Dynamic queries for visual information seeking. IEEE Software 11 (6), 70-77. (1994)
17. Hundhausen, C., Farley, S., Brown, J.: Can direct manipulation lower the barriers to computer programming and promote transfer of training?: An experimental study. ACM Transactions on Computer-Human Interaction 16 (3). (2009)
18. Microsoft Office Access. Retrieved April 16: from <http://office.microsoft.com/es-es/access/default.aspx>. (2008)
19. Microsoft SQL Query Designer. Retrieved April 23: from <http://www.microsoft.com/sqlserver/2008/en/us/default.aspx>. (2008)
20. Oracle Hyperion Interactive Reporting. Retrieved April 14: from <http://www.oracle.com/technology/products/bi/interactive-reporting/index.html>. (2008)
21. Eclipse BIRT Project. Retrieved April 13: from <http://www.eclipse.org/birt/project/>. (2008)
22. Spago BI QbE Enterprise Scope. Retrieved April 18: from <http://www.spagobi.org/>. (2008)
23. Tableau Desktop. Retrieved May 24: from <http://www.tableausoftware.com/>. (2009)
24. Stolte, C., Tang, D., Hanrahan, P.: Polaris: a system for query, analysis, and visualization of multidimensional databases. Communications of the ACM 51 (11), 75-84. (2008)
25. Ahlberg, C.: Spotfire: an information exploration environment. ACM SIGMOD Record 25 (4), 25-29. (1996)
26. Active Query Builder. Retrieved January 5: from <http://www.activequerybuilder.com/>. (2011)
27. Nielsen, J.: The Usability Engineering Life Cycle. IEEE Computer Journal 25 (3), 12-22. (1992)
28. Rettig, M.: Prototyping for tiny fingers. Communications of the ACM 37 (4), 21-27. (1994)
29. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers, San Francisco. (1993)
30. Nørgaard, M., Hornbæk, K.: What do usability evaluators do in practice?: an explorative study of think-aloud testing. In: Proceedings of the 6th conference on Designing Interactive systems. ACM, University Park, PA, pp. 209-218. (2006)
31. Lund, A.: Measuring Usability with the USE Questionnaire. Usability and User Experience Special Interest Group 8 (2). (2001)
32. Lin, H.,Choong, Y., Salvendy, G.: A Proposed Index of Usability: A Method for Comparing the Relative Usability of Different Software Systems. Behaviour and Information Technology 16 (4/5), 267-278. (1997)
33. Davis, F. D.: Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology.MIS Quarterly 13 (3), 319-340. (1989)
34. Catarci, T., Dix, A., Kimani, S.: User Centered Data Management. Morgan & Claypool, San Francisco. (2010)
35. Macías, J. A.: Enhancing Interaction Design on the Semantic Web: A Case Study. IEEE Transactions on Systems Man and Cybernetics – Part C 42 (6), 1365-1373. (2012)

Clemente R. Borges is a Ph.D. candidate and former researcher in the Autónoma University of Madrid, where he received in 2009 his M.S. Degree in Computer Engineering with a major in Advanced Information Technologies. Most of his research topics are related to Human-Computer Interaction, specifically focused on Visual Database Interaction, Web Interfaces, Content Management Systems, End-User Development and Usability.

José A. Macías received M.S. and Ph.D. degrees in computer science from Madrid Technical University and Autónoma University of Madrid, Spain, in 1999 and 2003, respectively. He is Associate Professor in the Computer Science Department at the Madrid Autónoma University. His principal research area of interest is Human-Computer Interaction (HCI). He appears as author in a great number of publications in prestigious HCI conferences, books and Journals, and he serves as President of AIPO (Spanish HCI Association), also obtaining outstanding academic awards such as the "Outstanding Paper Award" conferred by IEEE Neural Networks Society.

Received: November 30, 2013; Accepted: March 25, 2014

Computer Science and Information Systems 11(2):503–524 DOI: 10.2298/CSIS130610033M

Application and Relational Database Co-Refactoring

Ondrej Macek[1] and Karel Richta[2]

[1] Dept. of Computer Science and Engineering, Faculty of Electrical Engineering,
Czech Technical University in Prague
Karlovo namesti 13
121 35 Praha 2, Czech Republic
macekond@fel.cvut.cz

[2] Dept. of Software Engineering, Faculty of Mathematics and Physics,
Charles University
Malostranske namesti 25
118 00, Praha 1, Czech Republic
karel.richta@mff.cuni.cz

Abstract. A refactoring of application persistent objects affects not only the source code but the stored data as well. The change is usually processed in two steps: refactoring and data migration, which is ineffective and error prone. We provide a formal model for solution which is capable to migrate database according to a refactoring in the application code. The feasibility of the change and its data-secure processing is addressed as well.

Keywords: refactoring, relational schema evolution, application and database co-evolution, formal model

1. Introduction

The Evolution (change) of a software is a common issue during the software development. It occurs for many reasons in all phases of the software lifecycle. The evolution severity usually depends on the number of changes which have to be made and on the number of affected software components. Refactoring [8] is a very popular practice in object-oriented environments for evolving the source code and software architecture. Evolution of database schema and stored data is implemented separately from source code refactoring, although the change of application also affects the database. Object-relational mapping (ORM) frameworks can help with propagation of the evolution from an application to a database. However, these frameworks are usually neither capable of solving complex refactoring cases nor they migrate data properly as it will be shown in Sect. 2.

The problem of application and database evolution is discussed from a developer's point of view and the formal model of application refactoring and its impact is shown in Sect. 3. A developer works with a model of a persistence layer, which can be transformed into a model of a database schema (or into a database schema directly). Changes of the application layer can be represented as a sequence of transformations. These transformations affect the structure of the application layer or the database schema. We show how these transformations can be used not only for a structural change, but for an automatic generation of a data migration script as well. Basic refactoring cases are introduced as well as the complex ones which are created as sequences of the basic ones. Capabilities of the proposed formal models are illustrated in the common refactoring issues in Sect. 5.

2. Refactoring in the Context of ORM

A software implemented by using an object-oriented language, which uses a relational database as a data storage, consists of four main components. There is the application itself, the database schema, stored data and the object-relational mapping. The software can be evolved by adding or removing entities, their properties or associations. These changes affects the database directly. Other kind of evolution is refactoring. Refactoring is a change made to the internal structure of software to make it easier to understand and cheaper to modify without changing its observable behavior [8]. The refactoring may affect not only the application but the database as well. The database has to evolve when the persistence layer of the application changes to fit the object-relational mapping used in the software. The common current solution is based on capabilities of object-relational mapping frameworks which are capable of creating a database schema according to the given source code or model. The process of evolution then proceeds as follows:

1. The code is refactored (usually by using the developer's IDE).
2. The ORM framework generates a new database schema.
3. The data are migrated manually (if needed) from the old to the new database.

The last step is error-prone as it is processed manually and the error probability increases with the complexity of the refactoring. The evolution process requires cooperation of a developer and a database administrator or it requires the developer to have a knowledge of the database used in the software. The knowledge of the ORM is needed in both cases. Remarkable is the fact that the feasibility of data evolution has to be verified for each deployed software instance, because the data can differ. The data and information preservation is a crucial issue of database evolution. The next observation is that the evolution of the software is defined twice for one software - first for the application then for the database.

Example 1 Let us have only two classes *A* and *B* in the application which are not connected by an association and there are corresponding tables *tab_a* and *tab_b* in the database, which contain some data. We decide to merge *A* with *B* during the development. It means (on a structural level) that the result of the merging is a new class *A'*, which contains all properties of old *A* and all properties of *B* and *B* is removed from the application. The database schema is generated by the ORM framework automatically and it contains only the table *tab_a*". The data migration has to be created manually. The developer has to define the evolution twice. The mapping between the data in *tab_a* and *tab_b* (a cartesian product of data in both tables, equality of some columns etc.) has to be provided to merge the stored data correctly. Next the impact of this mapping on the database has to be verified: are there any data which can be lost during inlining and is this loss intentional?

We propose a better solution where the process of database evolution according to the code refactoring is more automatized. The solution is illustrated in Fig. 1. It is based on a change in the evolution process which assumes that the ORM does not change during the evolution:

1. The evolution of the whole software is defined independently of the application or the database.
2. The evolution is interpreted for the application and the database.

3. The evolution is executed.

The process can decrease mistakes in the evolution process, because there is only one source for the evolution and the evolution is automatically interpreted for the application (as a refactoring) and database (as schema and data migration). The existence of the set of all possible software evolutions E is based on a set of evolutionary transformations specific for an application and a database. Each transformation contains conditions of transformation feasibility, thus the feasibility of the evolution can be verified.

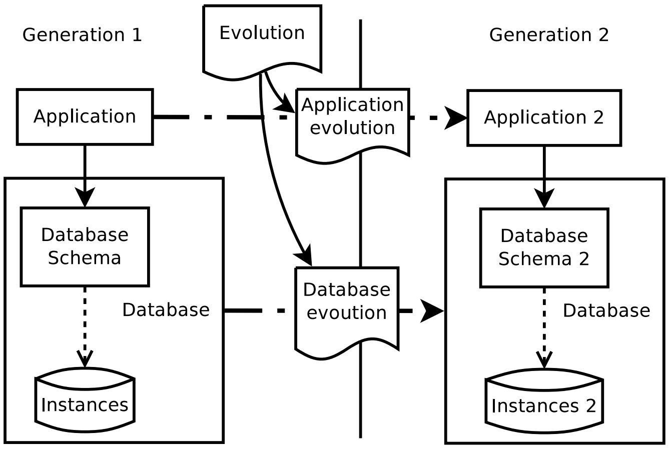


Fig. 1. The evolution of data changes the system on all levels. The figure shows all components of the evolution process.

The set E is created with respect to the needs of a developer, therefore the evolutionary cases are similar to the refactoring cases. Nevertheless the refactoring itself does not provide enough information for the complex migration of database and data, therefore the inputs of some transformation in E are extended beyond normal refactoring inputs. We focus on refactorings (i.e. structural changes) in this paper, therefore some transformations are not mentioned in the paper (e.g. adding data).

Example 2 The situation from Example 1 can be solved more effectively when the *mergeClasses* evolution is defined in E. The *mergeClasses* evolution is then automatically interpreted as merging of application classes (change of structure) and a merge of tables on database level (change of structure and migration of data). All information needed for software evolution is provided as the input of the *mergeClasses* transformation - the inputs are: identificators of both classes to be merged and the mapping between stored data, therefore the script for data migration can be generated. The structural feasibility is verified during the interpretation of the *mergeClasses* evolution and the migration script contains conditions which verify the feasibility on the level of data. (The semantics of the evolution is defined formally later in the paper.)

3. Model of Software Evolution

Model-driven development (MDD) is a good approach to data evolution of various software components [19]. The software is represented by a set of models – concretely, an application model and a database model. The evolution and ORM is represented by a set of model-to-model transformations and the interpretation of transformations from E is a model-to-model transformation again.

3.1. Software Model

The software consists of three important components (as seen in Fig. 1): an application (its persistence layer concretely), a database (consisting of a database schema and stored data) and an ORM, therefore we define a software as a triple consisting of an application and a database which are connected together by an ORM. The software has to be in a consistent state so its users can benefit from its usage. The software is in a consistent state if the application and the database are consistent and the database structure corresponds to the application structure according to the ORM, therefore the software is defined as:

$$software(a, d, \rho) = \begin{cases} consistentSoftware(a, d, \rho) \textbf{ if } a \neq \perp \ \wedge d \neq \perp \\ \wedge \ \rho(a) = d \\ \\ \perp \end{cases} \quad (1)$$

$$a \in Application, d \in Database, \rho \in ORM$$

where the $\perp$ symbol denotes an inconsistent state of a software or its components.

Evolution of a software is a transformation from one consistent state to another one. These states are called generations of the software and the functions which change the state of the software are called transformations.

3.2. Application Model

An application is defined as a set of classes and it creates the context for all structures used in the software persistence layer. For the sake of brevity, we use regular expression as follows: the notation $X = A*$ means X is defined as a sequence of elements from A, $X = (A, B)$ means X is a tuple of pairs from A and B, $X = A|B$ means X is either A or B.

The application is defined as follows:

$$\begin{aligned}
&\textbf{AppType} = APPSTRING \mid APPINTEGER \mid APPBOOLEAN && (2)\\
&\textbf{InheritanceType} = SINGLETABLE && (3)\\
&\textbf{Inheritance} = (Label, InheritanceType) \cup \{OBJECT\} && (4)\\
&\textbf{Property} = (Label, AppType, DefaultValue,\\
&\quad Cardinality, Mandatory) && (5)\\
&\textbf{Class} = (Label, Property*, Association*, Inheritance) && (6)\\
&\textbf{Association} = (Label, Class, StartCardinality, EndCardinality) && (7)\\
&\quad StartCardinality, EndCardinality \in \mathbb{N}_0 \cup \{*\}\\
&\textbf{Application} = Class* && (8)
\end{aligned}$$

Application type Application type ($AppType$) represents primitive types in the application. Programming languages usually provide types such as String, Integer, Boolean etc. The denotation of types begins with ”APP-” prefix to distinguish them from database types. The type casting (i.e. changing type of a property from String to Integer) is not part of transformations defined in this paper, because we focus on structural changes and their impact on data in the first place. However, type casting can be easily integrated into described transformations.

Inheritance The inheritance defines a parent-child relationship between classes. Multiple ancestors are not allowed and of course a class cannot be its own ancestor directly or indirectly. These restrictions are inspired by common programming languages as Java or C#. The *InheritanceType* determines how a inheritance hierarchy is mapped into a database. We consider only one common mappings of inheritance into a relational database for the sake of model abbreviation. We decide to represent an inheritance hierarchy of classes as a single table that has columns for all the fields of the various classes [9]. We assume there is only one type of inheritance type per class hierarchy. There can be several independent inheritance hierarchies in the model. The symbol *OBJECT* represents an universal parent for all classes.

Property Property represents a feature of a class which is represented as a primitive type. A property can be mandatory, can have a default value and according to its cardinality it can represent a single value or a collection of values. The properties of non-primitive types are represented as associations.

Class Class represents a basic structural unit in the application model. It has a unique name, one or more properties and it can be associated to other classes in the application.

Association Association represents a connection between two classes. It has a unique name and the reference is represented by the label of referenced class. The class which owns the association is considered to be the starting class of an association, referenced class is considered to be the ending class of an association. The cardinalities define the multiplicities of both association ends.

3.3. Application Manipulation

The application model defines only the structure of an application's data thus there are defined transformations for adding, altering and removing parts of the model (the set A).

Each transformation has a set of preconditions, which are in case of adding or altering very simple:

1. Name collisions have to be prevented when creating a class in the model or altering its name.
2. All references have to be updated when renaming a class.
3. The existence of the class being referenced has to be verified when creating or altering an association.

The only problem that can occur while deleting classes etc. from an application is that the class can be removed only when it is not associated with other classes. The list of possible transformations from A are in Table 1. The list of altering operations is not complete as transformation for changing the obligation or cardinality of properties and associations are missing. It is because these transformations are not so important for complex refactorings.

Table 1. The content of the set A - transformations for application change and refactoring. The detailed specification can be found in [16].

Type of transformation	Defined transformations
Application Creation	newApplication: $\rightarrow Application$
	addClass: $Class \times Application \rightarrow Application$
	addProperty: $Class \times Property \times Application \rightarrow Application$
	addAssociation: $Class \times Association \times$ $Application \rightarrow Application$
Application Modification	renameProperty: $Class \times Property \times Label$ $Application \rightarrow Application$
	renameAssociation: $Class \times Association \times Label$ $Application \rightarrow Application$
	renameClass: $Class \times Label \times Application \rightarrow Application$
Application Deconstruction	removeProperty: $Class \times Property \times$ $Application \rightarrow Application$
	removeAssociation: $Class \times Association$ $\times Application \rightarrow Application$
	removeClass: $Class \times Application \rightarrow Application$
Inheritance manipulation	addParent: $Class \times Inheritance \times Application \rightarrow Application$
	removeParent: $Class \times Application \rightarrow Application$
	pushDown: $Class \times Property \times Application \rightarrow Application$
	pullUp: $Class \times Property \times Application \rightarrow Application$

3.4. Database Model

A relational database consists of a database schema and data. Database schema defines the structure of the database and data represents stored instances in the software. A database is defined as:

$$\mathbf{DbType} = DBSTRING \mid DBINT \mid DBBOOLEAN \tag{9}$$

$$\mathbf{Constraint} = NOTNULL \mid UNIQUE \tag{10}$$

$$\mathbf{PrimaryKey} = (Label) \tag{11}$$

$$\mathbf{Column} = (Label, DbType, DefaultValue, Constraint*) \tag{12}$$

$$\mathbf{ForeignKey} = (Label, TableSchema, Constraint*) \tag{13}$$

$$\mathbf{TableSchema} = (Label, PrimaryKey, Column*, ForeignKey*) \tag{14}$$

Data Types Database data types $DbType$ represent primitive types in the database. Databases usually provide types such as Varchar, Integer, Boolean etc. We define types for strings, numbers and boolean values which can be further extended according to a specification of a concrete database.

Constraints There are two types of constraints defined in the model. Both constraints are column constraints - the first constraint forces columns to have no non-empty elements, the second constraint requires there have to be unique records in a column or foreign key.

Primary key A primary key is an unambiguous identifier of a record in a table. The primary key is always provided (automatically generated) by the associated sequence s as a non-zero natural number. A new value of a key is obtained by calling the function $next(s)$. The generator of primary keys values is called *Sequence* and there is one sequence per database in the model (see (18)).

A primary key is always defined with constraints $NOTNULL$ and $UNIQUE$.

Column A column defines data values and types which can be part of a table record.

Foreign key A foreign key is a reference to another table's primary key, it has a unique name and it can be constrained. The value of a foreign key is a non-zero natural number or $\varnothing$ if not constrained by $NOTNULL$.

TableSchema A table represents a basic concept of a database schema. It has a unique name, one or more columns and it can be related to other tables in the schema by foreign keys. Rows in the table represent stored data.

Data A database consists not only of a schema but also of data which are represented as rows in a table. A table row in our model consists of value pairs, which represent concrete values of a concrete column or key. Each row contains a reference to a table it belongs to and a primary key's value, which uniquely identifies the row.

$$\mathbf{KeyPair} = (PrimaryKey, Value) \tag{15}$$

$$\mathbf{Pair} = (Column, Value) \mid (ForeignKey, Value) \tag{16}$$

$$\mathbf{TableData} = (Table, KeyPair, Pair*) \tag{17}$$

Database The database is defined by its schema and data it contains. Last important item of a database is a generator of primary key values called *Sequence*.

$$\mathbf{Database} = (TableSchema*, TableData*, Sequence) \tag{18}$$

3.5. Database Manipulation

A database consists of two parts - of a database schema, which defines the structure, and of stored data - hence the transformations from set D have to consider both parts. The transformation for manipulation of structure has similar conditions to the evolution of applications. The transformations for data manipulation are inspired by the SQL language. The basic transformations are in Table 2 and the operations for data manipulation are in Table 3. The set of transformations for database manipulation is limited in contrast to the SQL language. Only transformations necessary for data evolution on database level are introduced.

Table 2. The transformations for database evolution. Transformations can be mapped to SQL intuitively. The detailed specification can be found in [16].

Type of transformation	Defined transformations
Database Creation	newDatabase: $\rightarrow Database$
	addTable: $TableSchema \times Database \rightarrow Database$
	addColumn: $TableSchema \times Column \times Database \rightarrow Database$
	addForeignKey: $TableSchema \times ForeignKey \times Mapping \times Database \rightarrow Database$
Database Modification	alterColumnName: $TableSchema \times Column \times Label \times Database \rightarrow Database$
	alterForeignKeyName: $TableSchema \times ForeignKey \times Label \times Database \rightarrow Database$
	alterTableName: $TableSchema \times Label \times Database \rightarrow Database$
Database Destruction	dropColumn: $TableSchema \times Column \times Database \rightarrow Database$
	dropForeignKey: $TableSchema \times ForeignKey \times Database \rightarrow Database$
	dropTable: $TableSchema \times Database \rightarrow Database$
Copy Structure and Values	copyColumn: $TableSchema \times TableSchema \times Column \times Label \times Mapping \times Database \rightarrow Database$
	copyTable: $TableSchema \times Label \times Database \rightarrow Database$
Data-secure Database Elements Removal	dropEmptyColumn: $TableSchema \times Column \times Database \rightarrow Database$
	dropEmptyForeignKey: $TableSchema \times ForeignKey \times Database \rightarrow Database$
	dropEmptyTable: $TableSchema \times Database \rightarrow Database$

Table 3. The table contains a set of transformations which serve for data manipulation. The detailed specification can be found in [16].

Data Manipulation	selectOne: $TableSchema \times ID \times Database \rightarrow TableData$
	selectAll: $TableSchema \times Database \rightarrow TableData*$
	insertData: $TableData \times Database \rightarrow Database$
	insertValue: $TableData \times Pair \times Database \rightarrow Database$

Copying Database Elements The transformations for copying the structure and values of a column or table serve more as helpers for advanced evolution cases, where they are discussed in detail.

Data-Safe Database Element Removal The transformation that remove elements from the database can have fatal impact on the data preservation, therefore the set of transformations is extended by data safe transformations for removing database elements. These transformations are not part of the SQL standard, although they can be implemented as database functions. These transformations create a safe way to remove elements from the database as they drop empty structural elements only.

3.6. Mapping Between Stored Data

A relation between data from different $TableData$s has to be known during execution of some transformations (e.g. $moveProperty$). The relation is defined as a mapping between $TableData$s. The mapping is defined as follows:

$$mapping : TableData \rightarrow TableData * \cup \{\emptyset\} \tag{19}$$

The mapping has a sequences of $TableData$ in its range set, this allows to define one-to-many and many-to-many relations between data. The $\emptyset$ represents a situation where there is no relation for a given element of the mapping's range. A special case of mapping is an empty mapping denoted as m_e, which is used when there are no $TableData$ in the domain or the range is equal to $\emptyset$ i.e. the transformation takes part on the structural level only. The set of all possible mappings is called $Mapping$.

Each mapping has to fulfill constraints given by the structural definition of its range $TableData$. Concretely: uniqueness of column values:

$$\begin{aligned}
&\forall\, m \in Mapping; x_1, x_2 \in domain(m); p_1 \in pairs(m(x_1)), p_2 \in pairs(m(x_2)) :\\
&x_1 \neq x_2 \land \exists\, c \in Column, UNIQUE \in constraints(c) \land\\
&c \in pairs(columns(range(m))) \implies p_1 \neq p_2
\end{aligned} \tag{20}$$

if the principle of uniqueness is violated then usage of such a mapping leads to an inconsistent database. Next constraint of mappings is the non-emptiness of columns constrained with $NOTNULL$ constraint:

$$\begin{aligned}
&\forall\, m \in Mapping; x \in domain(m) :\\
&\exists\, c \in Column, NOTNULL \in constraints(c) \implies m(x) \neq \emptyset
\end{aligned} \tag{21}$$

if this principle is violated then usage of such a mapping leads to an inconsistent database.

There can occur data loss, when the mapping is a partial function. Usage of such mapping has to be reconsidered before its usage, because it can result in a semantically inconsistent state of the database.

A mapping can be implemented as a nested query in the SQL command representing the transformation. Alternatively, a database view can be implemented to represent such a mapping.

3.7. Object Relational Mapping

ORM is the only fixed point in the software model we use. The mapping is similar to the Hibernate mapping [13], thus a lot of developers should be familiar with it. The main ideas are:

- classes are mapped to tables,
- single properties are mapped to a column or if the property is a collection then the property is mapped as a table,
- associations are mapped to foreign keys or to tables if the association represents a many-to-many relationship,
- primary keys are created automatically for each table,
- names used in application are mapped into the database schema (e.g. because of possible name collision of application classes and names of database schema elements).

3.8. Software Evolution

The evolution of the whole software is described from the developer's point of view, therefore the transformations use the names and elements from application context. Elements have to be transformed into a database context - this is assured by the ORM. In the model we ignore the fact an element's label has to be often transformed as well (e.g. because of collision between the label and label of a database internal table).

Three sets of transformations have to be defined to provide the capabilities described in Example 2: the set E of all possible software transformations:

$$E = \{e | e : ConsistentSoftware \rightarrow Software\} \; , \tag{22}$$

which is limited in this paper to a set of transformations for refactoring, creating and deleting model elements. The list of transformations and their definition is available online [16]. The fact that a transformation produces $Software$ and not $ConsistentSoftware$ supports composition of transformations. If a transformation (refactoring) is applicable on a software it produces consistent software. In contrast if the transformation is not applicable than the software is in the state $\perp$ and no transformation can change it.

Next we define the set of application refactorings:

$$A = \{a | a : \mathit{Application} \rightarrow \mathit{Application}\} \tag{23}$$

and the set of database evolutionary transformations

$$D = \{d | d : Database \rightarrow Database\} \; . \tag{24}$$

Evolution of the software is defined as an interpretation of the transformation for each component of the software:

$$\begin{aligned} &t(s) = software(\Psi(t, application(s)), \Phi(t, database(s)), \rho(s)) \\ &\quad s \in ConsistentSoftware, t \in E \end{aligned} \tag{25}$$

where $\Psi : E \times Application \rightarrow A$ interprets the software evolution cases to the code refactoring and $\Phi : E \times Database \rightarrow D$ interprets to the evolutionary transformation of database. The ORM does not change during the evolution.

The advantage of interpretation is that the semantics of the evolution is defined only once by the E-transformation. This definition contains all necessary information for partial evolutions of all software components. This approach speeds up the work of software developers, because it automatize the process of database evolution.

3.9. Basic Evolutionary Transformations

The evolution of the whole software is based on the atomic transformations specific for each software part, which are defined in Sec. 3.3 and 3.5. This section introduces how those primitives can be used to manipulate the software elements.

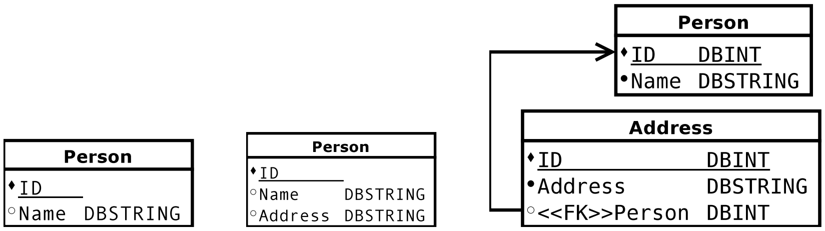
The basic evolutionary transformations of the software are based on basic evolution of an application. These transformations have to respect the ORM, because properties and associations can be mapped as a column (foreign key respectively) as shown in the example of creating a new property:

$$\Phi(newProperty(c,p),d) = \begin{cases} addColumn(ORM(c), ORM(p), d) \\ \textbf{if } cardinality(p) = 1 \\ \\ addForeignKey(ORM(p), fk, m_e, \\ \quad addTable(ORM(p), d)) \\ \textbf{if } cardinality(p) > 1 \\ \textbf{where } fk = (ORM(c), ORM(C), \langle\rangle) \end{cases} \tag{26}$$
$$c \in Class, p \in Property, d \in Database$$

An example of both mappings is in Fig. 2.

A mapping between basic evolutionary transformations is provided in the short in Table 4. The detailed definitions are provided in [16]. The mapping can be implemented as a direct generation of SQL commands from the application transformations or an intermediate database model can be used. The second approach is suitable if there are multiple models affecting database i.e. a model of entities, which is interpreted as database schema and a model of business constraints, which is interpreted as a set of database triggers.

Each basic application transformation is mapped to a set of possible mappings on the database level according to the cardinality. The advanced transformations are explained in detail in Sect. 3.10. Some of the advanced transformations cannot be implemented as a set of simple data manipulations SQL scripts. Rather, advanced constructs such as PL/SQL procedures have to be used, especially when there is a mapping between instances. The



(a) **(b)** **(c)**

Fig. 2. Two possible variants how the $newProperty$ transformation can be interpreted on the database level. The initial state is in Fig. 2a, then a new property called Address is added. In case the Address property has cardinality equal to 1 then the result is in Fig. 2b otherwise the result is in Fig. 2c.

Table 4. The mapping between evolution of the application and database transformations. The advanced transformations are explained in detail in Sect.3.10.

Application Transformation	Database Transformation
newClass	addTable
newProperty	**if** $cardinality \leq 1$ **then** addColumn **else** addTable
newAssociation	**if** $cardinality \leq 1$ **then** addColumn **else** addTable
renameProperty	**if** $cardinality \leq 1$ **then** renameColumn **else** renameTable

renameAssociation	**if** $cardinality \leq 1$ **then** renameForeignKey
	else renameTable

renameClass	renameTable

removeProperty	**if** $cardinality \leq 1$ **then** dropEmptyColumn
	else dropEmptyTable

removeAssociation	**if** $cardinality \leq 1$ **then** dropEmptyForeignKey
	else dropEmptyTable

removeClass	dropEmptyTable

transformations for data-safe removal are used as default removing transformations, although the classic drop- transformations can be used. This should lead to the more careful usage of removing transformations.

3.10. Advanced Evolutionary Transformations

Advanced evolutionary transformations are based on the basic ones, they can be obtained as a concatenation of transformations. This means the complex transformations are limited in the same way as their basic components. All transformations from the set E presented so far have the same input information for both an application and a database, whereas the advanced transformations usually need a mapping between stored data (instances) as their input.

Copy Property The $copyProperty$ creates a duplicate of a property in a given class. If the mapping is not provided the transformation creates only a structural copy of the property, otherwise it copies the values too. Therefore the $copyProperty$ transformation is the simplest way to manipulate the stored data.

$$\begin{aligned}&copyProperty : Class \times Class \times Property \times Mapping \times \\ &\quad ConsistentSoftware \to Software\end{aligned} \tag{27}$$

The $copyProperty$ is the first transformation when an additional information has to be added to the usual code refactoring. It is because the $copyColumn$ transformation needs one more information to succeed - a mapping has to be provided between the source and the target table to assure data information consistency. The transformation is interpreted for both software components - in the application case a new property is added:

$$\begin{aligned}&\Psi(copyProperty(c_s, c_t, p, m, s)) = newProperty(c_t, p, application(s))) \\ &\textbf{if } c_s \neq c_t\end{aligned} \tag{28}$$

In the database case the copy of a column or table is created according the cardinality of the property and then the values are copied:

$$\begin{aligned}&\Phi(copyProperty(c_s, c_t, p, m, d) = \\ &\quad\begin{cases} copyColumn(ORM(c_s), ORM(c_t), ORM(p), m, database(s)) \\ \textbf{if } cardinality(p) = 1 \\ \\ copyPropertyAsTable(ORM(c_s), ORM(c_t), ORM(p), m, database(s)) \\ \textbf{if } cardinality(p) > 1 \end{cases}\end{aligned} \tag{29}$$

Move Property The $moveProperty$ transformation is based on the $copyProperty$ transformation followed by the $removeProperty$ so its construction is easy. On the other hand, special attention has to be paid to the provided mapping of instances, because it can cause loss of data. The ideal case is when the mapping is injective, then the transformation cannot cause loss of data.

$$\begin{aligned}
&moveProperty : Class \times Class \times Property \times Mapping \times \\
&\quad ConsistentSoftware \rightarrow Software \qquad (30)\\
&moveProperty(c_s, c_t, p, m, s) = \\
&\quad removeProperty(c_s, p, copyProperty(c_s, c_t, p, m, s)) \qquad (31)
\end{aligned}$$

Inline and Split Class Inline and split are two opposite transformations. First of them moves data from the source class into the target one and then deletes the source class. Second of them extracts a new class from an existing class. The inline transformations can be composed from already mentioned basic transformations:

$$\begin{aligned}
&inlineClass : Class \times Class \times Mapping \times \\
&\quad ConsistentSoftware \rightarrow Software \qquad (32)
\end{aligned}$$

$$\begin{aligned}
&inlineClass(c_1, c_2, m, s) = removeClass(c_2, moveProperties(c_1, c_2, m, s)) \\
&\ \mathbf{if}\ !isReferenced(c_2, application(s)) \\
&\ \mathbf{where}\ p \in properties(c_2), \\
&\ moveProperties(c_1, c_2, m, s) = \\
&\quad \forall\, p \in properties(c2) : moveProperty(c_1, c_2, p, m, s) \qquad (33)
\end{aligned}$$

A special case of inlining is merge. The $mergeClass$ transformation is used when two classes have the same structure and they should be merged into one.

$$mergeClasses : Class \times Class \times Software \rightarrow Software \qquad (34)$$

The classes are represented as two tables containing different data - these data have to be merged into one table.

The split transformation has to be interpreted:

$$\begin{aligned}
&splitClass : Class \times Label \times Property \times \\
&\quad ConsistentSoftware \rightarrow Software \qquad (35)
\end{aligned}$$

$$\begin{aligned}
&\Psi(splitClass(c, l, p), a) = removeProperty(c, p, newProperty(c_n, p, \\
&\quad addClass(c_n, a))) \\
&\ c_n = class(l, \emptyset, \emptyset, OBJECT) \qquad (36)
\end{aligned}$$

$$\Phi(splitClass(c,l,p),d) =$$
$$\begin{cases} dropColumn(ORM(c), ORM(p), copyColumn(ORM(c), \\ \quad ORM(propToClass(p)), m, addTable(ORM(propToClass(p,l)), d))) \\ \textbf{where } m = \forall\, r \in selectAll(ORM(c), d): \\ \quad m(r) = tableData((ORM(propToClass(p,l)), keyPair(r), \\ \quad\quad pairOfColumn(ORM(p), pairs(r)))) \\ \textbf{if } cardinality(p) = 1 \\ \\ alterTableName(ORM(p), l, dropForeignKey(ORM(p), \\ \quad foreignKeys(ORM(p), d)) \\ \textbf{if } cardinality(p) > 1 \end{cases} \tag{37}$$

The *mapping* is defined as identity in case of *splitClass*, therefore the rows in the new table have the same primary keys' values as in their source class.

3.11. Inheritance Manipulation

Inheritance is an important part of object-oriented world. The impact of change of this relationship to the database depends on the type of inheritance mapping similarly as in the case of *moveProperty*. The *addParent* transformation is described as example:

$$\begin{aligned} &addParent : Class \times Inheritance \times Mapping \times \\ &\quad ConsistentSoftware \rightarrow Software \end{aligned} \tag{38}$$

The interpretation of the $addParent$ transformation for the whole software on the application level uses the transformation $addParent$ defined for the application level (see Tab. 1):

$$\begin{aligned} &\Psi(addParent(c, ih, m), a) = addParent(c, ih, a) \\ &\quad \textbf{if } !\ isReferenced(c) \end{aligned} \tag{39}$$

The interpretation of the $addParent$ transforation on the database level depends on the ORM. We use the single-table mapping in our model, therefore the result is a merge of tables:

$$\begin{aligned} &\Phi(addParent(c, ih, m, s) = dropTable(ORM(c), h(c, ih, p, m, s)) \\ &\quad \textbf{where } h : TableSchema \times Inheritance \times Mapping \times \\ &\quad\quad Software \rightarrow Database \\ &\quad h(c, ih, p, m, s) = \forall\ p \in properties(c): \\ &\quad\quad \Phi(moveProperty(c, class(ih), p, m, s)) \end{aligned} \tag{40}$$

The inverse transformation *removeParent* uses the *splitClass* transformations (this is valid for the simplification when there is only one type of inheritance - *SINGLETABLE*). This does not cover the case when the information from parent are not needed in its child, such a transformation has to be defined by a sequence of steps.

Next transformations connected with inheritance are *pushDown* and *pullUp*. Pull up moves a property from child to parent so the *moveProperty* can be used, the mapping of instances is based on the parent-child relationship. It means the column cannot be constrained with $NOTNULL$ constraint if there are more siblings in the hierarchy. The *pushDown* transformation works in opposite direction, however it moves the property into all children of the parent class. The easiest situation is where there is no sibling in the hierarchy, otherwise we assume there are no stored data in the siblings. It is because the change of parent affects the instances of its children: When a property is moved only into one child, the information consistency is violated, therefore we forbid such transformation, because we cannot anticipate developer's intents.

The next two inheritance-related transformations are *extractParent* and *extractCommonParent* which serve for extracting a parent class based on a given set of property from a class or from a couple of classes.

4. Example of Usage

The examples demonstrate how the transformations help with data evolution in real life example of software evolution. There are two classes in our software *Person* and *LegalParty*, both of them contain information about address (street, city and zip code) as shown in the Fig. 3 where the database tables are shown.

To improve the design of code a class representing Address has to be created. The *Address* class shall be associated with both original classes and shall receive the data already stored in the database.

Id	Surname	Street	City	ZIP
11	Jackson	Central Park St	New York	100 01
12	Clooney	S Orange Ave	Orlando	320 24

(a) Data stored in the table $Person$.

Id	BusinessName	Street	City	ZIP
100	Tools & Machines	Olive ave.	New York	100 01
200	AI Robotics	Pine ave.	LA	900 03

(b) Data stored in the table $LegalParty$.

Fig. 3. The initial state of the example used in the case study. There is a repetition of information structure in the $Person$ and $LegalParty$ class.

The first step is to extract two temporary classes representing addresses of a *Person* or of a *LegalParty*:

$$\begin{aligned} s_1 = splitClass(&Person, "Address\_tmp1", \langle street, city, zip \rangle, \\ &software) \end{aligned} \tag{41}$$

$$s_2 = splitClass(LegalParty, "Address\_tmp2", \langle street, city, zip \rangle, s_1) \tag{42}$$

The *splitClass* transformation moves given properties into the new class, therefore there is no address information in *Person* or *LegalPerson* after these transformations. The temporary classes have to be connected with the origin classes by an association.

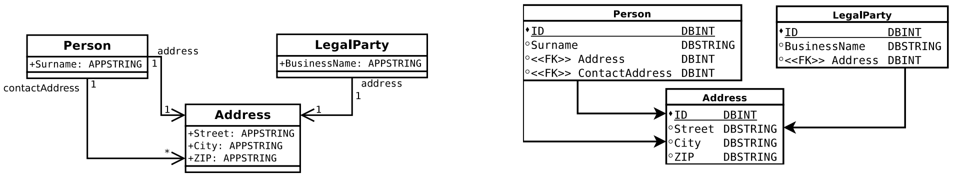
$$\begin{aligned} s_3 = {} & newAssociation(Person, association(\text{"}address\text{"}, \\ & \text{"}Address\_tmp1\text{"}, 1, 1), mapping_1, s_2) \end{aligned} \tag{43}$$

$$\begin{aligned} s_4 = {} & newAssociation(LegalParty, association(\text{"}address\text{"}, \\ & \text{"}Address\_tmp2\text{"}, *, 1), mapping_2, s_3) \end{aligned} \tag{44}$$

The $mapping_1$ and $mapping_2$ can be defined using the equality of primary keys values, because the temporary classes were extracted from the *Person* and *LegalParty*. Finally the temporary class are merged into one:

$$s_5 = mergeClasses(Address\_tmp1, Address\_tmp2, \text{"}Address\text{"}, s_4) \tag{45}$$

The merge of classes is possible because the associations (and corresponding foreign keys) created in the previous step are not constrained. The change of the code continue by adding more addresses to the *Person* class:



(a) The application model.

(b) The database model.

Id	Street	City	ZIP	Person_ContactAddress
1	Olive ave.	New York	100 01	
2	Pine ave.	LA	900 03	
3	Central Park St	New York	100 01	
4	Orange Ave	Orlando	320 24	

(c) Data stored in the table $Address$.

Fig. 4. The final state of the example used in the case study. There is only one table containing all addresses in the system (see 4b), which is referenced by $LegalParty$ class and by the the $Person$ and class twice, which represents regular and contact address.

$$\begin{aligned} s_6 = {} & newAssociation(Person, association(\text{"}contactAddress\text{"}, \\ & Address, 1, 1), m_e, s_5) \end{aligned} \tag{46}$$

The final result of the transformations is in the Fig. 4.

The example shows that the defined transformations are able to perform regular data evolutions. It shows that the usage of the transformations is sometimes not intuitive for the user (e.g. creating temporary classes before merging), therefore the next complex transformations have to be defined based on users' experience reports. The examples show that the column constraints often decide the feasibility of a transformation.

5. Capabilities of Defined Transformations

It is obvious that the set of defined transformation is limited because of focus on data preservation or transformation concatenation. Nevertheless, defined transformations are strong enough to handle a lot of refactoring cases. To verify this we use refactoring cases based on the Eclipse foundation refactoring statistics [7] (choosing only the transformation influencing data) and on Fowler's book [8]. Selected refactorings are:

1. **Rename** is the most used refactoring according to the Eclipse statistic and it is one of the basic refactoring cases introduced in Sec. 3.9.
2. **Move** refactoring is used in Eclipse to move properties from a class to another class within an inheritance hierarchy or to move classes between packages. In contrast, our model does not consider packages, However it is able to move property from one class to another according to the given mapping, which contains not only moves in hierarchy but a lot of other cases too.
3. **Extract Class** is an often used refactoring in Eclipse and it is mentioned in Fowler's book as well. It is introduced as an example of advanced refactorings in Sec. 3.10 as the *splitClass* transformation together with its opposite transformation *mergeClasses*.
4. **Move field** from Fowler's book is introduced as *moveProperty* in Sec. 3.9.
5. **Replace Data Value with Object** is a refactoring case which is not mentioned among the evolutionary transformations. However it can be composed from already defined transformations:

$$\begin{aligned}&replaceDataWithObject(sourceClass, property, newObject,\\&\quad software) = \mathit{removeProperty}(\mathit{sourceClass}, \mathit{property},\\&\quad \mathit{newAssociation}(\mathit{sourceClass}, \mathit{association}(\mathit{label}(\mathit{sourceClass}),\\&\quad \mathit{newObject}, 1, 1), \mathit{mapping}, \mathit{copyProperty}(\mathit{sourceClass}, \mathit{newObject},\\&\quad \mathit{property}, \mathit{mapping}, software)\end{aligned} \tag{47}$$

   where the mapping represents the relation between the new object and the original one - for each instance of the new object there is one instance of the original one.

The transformations proposed in this paper are able to implement all data refactorings supported by the Eclipse IDE. There are some refactorings from Fowler which are not considered in the proposal: replacing array with object, changing unidirectional association to bidirectional (and reverse) and operations which replace types with polymorphism. However, the proposal solves the most used refactorings.

6. Related Work

Research of data and database evolution is not limited to the relational ones only, there is also work in the areas of object databases and XML databases. This work provides solutions specific to concrete types of databases using various ranges of solutions - domain specific languages [14], extensions of existing standards or MDD [15] or formal specification [22]. These solutions are inspiring, however, the domain of the ORM has its specific issues, so a solution from another domain has to be adapted carefully.

Informal definitions The taxonomy of relational database evolution based on the entity-relationship model is proposed in [20]. The evolution is described as a change in the entity-relationship model and a change in a relational database. The semantics change patterns in context of a conceptual schema are described in [25], although its impact on a database schema or data is not described. The main cases of data evolution are defined in both publications. The description is informal. An extensive set of possible database refactorings is provided in [3], where both schema and data evolution is discussed. The refactorings are intended to be used by database administrators, thus it assumes database-first approach to evolution, whereas this proposal is application-first.

Formal frameworks A general formal framework for database evolution is defined in [17]. The framework is based on a set of basic graph transformations which are then extended to transformations of the entity-relationship model. The framework and the defined transformations can be implemented in our proposal too. The contribution of the formal framework is a definition of equivalent structures in relational database schemas. Our proposal is aimed to be used in the domain of object-oriented languages, where a class model is more common than entity-relationship model.

The formal definition of MDD approach to database schema evolution is proposed in [2], where changes of a database conceptual schema are interpreted on both a physical schema and data. In contrast our proposal is aimed to solve the problem of code refactoring and its impact on relational database, next we propose more complex transformations (such as *copyProperty* or inheritance-related transformations) and examine the impact of platform specific constructs (such as foreign keys or constraints) on the evolution.

A categorical framework for the migration of object-oriented systems is proposed in [21]. This framework defines the refactoring of objects, data and methods, which are the main objectives of the framework. The impact of the object change on a relational database is not considered in the paper as it is aimed at object-oriented systems only.

Forward Round-Trip Engineering in Data Evolution A round-trip approach for data evolution was described in [23], which was implemented in the SELF language. This approach proposed a forward-oriented evolution on all application levels in the software application. In contrast with our framework, the round-trip approach does not care about stored data, because it is focused only on transformations for creating, updating and deleting elements of a model.

A meta-model based approach to data evolution is proposed in [1] where a very similar solution for data evolution is proposed, which is based on an extended UML meta-model. The solution provides similar capabilities of change of application and database as does our proposal. In contrast our work is created with the ORM domain in mind and therefore we extended the application meta-model with constructs typical of this domain.

ORM Frameworks There are many object-relational mapping frameworks available for developers, and some of them provide tools for database migration. Hibernate [13] is one of the most popular ORM frameworks in the Java community which is capable of creating a new table or adding a new column according to a change in the application. Active Record [11] is an ORM framework in the Ruby on Rails environment. Since its first version it has contained support for database evolution according to the create-update-delete principle, in the form of so-called migrations [10] which can be extended by adding user SQL commands. Entity Framework [18] is Microsoft's ORM solution for the .NET platform. Its capabilities of data evolution support are similar to those of Active Record. Neither of the frameworks is capable of automatization of complex refactoring cases.

Tools for Database Evolution The MeDEA project [6] offers a tool for evolution of both database schema and stored data based on model-driven approach. The project DB-MAIN [12] provides a MDD approach to data evolution on all the levels of software we do. The project DB-MAIN is well documented formally. The PRISM is a research project for data management under schema evolution [5]. In contrast with our proposal, and with MeDEA, it extends the SQL command set by so-called schema modification operators which implements the schema evolution. A very promising solution is Liquibase [24], a tool for database refactoring and evolution. It is capable of migrating both database schema and stored data. The evolution is described by an XML document which can be interpreted on various databases.

The difference between these frameworks and our proposal is that each has a different focus. All mentioned projects are aimed to be used by database administrators; whereas our focus is on entities evolution, which is then propagated to a database with an emphasis on automatization. Our goal is to hide the entire database level from our users.

The project IMIS [4] follows the same idea of applying MDD into evolution of a whole software, but does not provide a formal model or an overview of capabilities (defined transformations).

7. Conclusion

We discuss the impact of application refactoring on relational database schema and stored data. We introduce basic transformations of application and database, next show how the complex refactorings can be constructed based on basic refactorings. The transformations presented are capable of solving the main refactoring cases. Theirs construction assures structural-safe change of application and data-safe migration of database schema and data.

The proposal is based on the idea of MDD which is implemented by a set of models and transformation rules. This allows to simulate the behavior of an evolution in the platform independent environment.

The main contribution is the new point of view on the application refactoring. The application code and relational database co-evolution can improve the capabilities of IDEs and speed up the work of developers. Next contributions are definitions of impact of advanced refactoring cases on relational database and stored data. We show that the automatic co-refactoring is possible not only in case of basic changes of an application, but even complex refactorings can be processed automatically for the applications' code and database.

Someone can be afraid about applicability in complex scenarios, whether the proposal can really save time in real world scenarios. But due to the fact that hand-crafting of mappings is not necessary in most of the interesting cases (usually an existing association is used as the mapping function), the model proposed has advantages over hand-made migration scripts.

Acknowledgments. The research has been supported by the Czech Technical University in Prague grant No. [SGS12/147/OHK3/2T/13] and partially by the Avast Foundation.

References

1. Aboulsamh, M., Davies, J.: A metamodel-based approach to information systems evolution and data migration. In: Software Engineering Advances (ICSEA), 2010 Fifth International Conference on. pp. 155 –161 (aug 2010)
2. Aboulsamh, M., Davies, J.: A formal modeling approach to information systems evolution and data migration. In: Halpin, T., Nurcan, S., Krogstie, J., Soffer, P., Proper, E., Schmidt, R., Bider, I. (eds.) Enterprise, Business-Process and Information Systems Modeling, Lecture Notes in Business Information Processing, vol. 81, pp. 383–397. Springer Berlin Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-21759-3_28
3. Ambler, S.W., Sadalage, P.J.: Refactoring Databases: Evolutionary Database Design. Addison-Wesley Professional (2006)
4. Bordbar, B., Draheim, D., Horn, M., Schulz, I., Weber, G.: Integrated model-based software development, data access, and data migration. In: Proceedings of the 8th international conference on Model Driven Engineering Languages and Systems. pp. 382–396. MoDELS'05, Springer-Verlag, Berlin, Heidelberg (2005), http://dx.doi.org/10.1007/11557432_28
5. Curino, C.A., Moon, H.J., Zaniolo, C.: Graceful database schema evolution: the prism workbench. Proc. VLDB Endow. 1(1), 761–772 (Aug 2008), http://dl.acm.org/citation.cfm?id=1453856.1453939
6. Domnguez, E., Lloret, J., ngel L. Rubio, Zapata, M.A.: Medea: A database evolution architecture with traceability. Data & Knowledge Engineering 65(3), 419 – 441 (2008), http://www.sciencedirect.com/science/article/pii/S0169023X07002224
7. Eclipse Foundation: Usage Data Collector Results (2013), http://www.eclipse.org/org/usagedata/results.php?kind=command&sort=element[Accesed 19. November 2013]
8. Fowler, M.: Refactoring: Improving the Design of Existing Code. Addison-Wesley, Boston, MA, USA (1999)
9. Fowler, M.: Patterns of Enterprise Application Architecture. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2002)
10. Hansson, D.H., Kemper, J.: ActiveRecord::Migration (2009), http://api.rubyonrails.org/classes/ActiveRecord/Migration.html, Accessed 6. September 2012
11. Hansson, D.H., Kemper, J.: RubyForge:ActiveRecord: Project Info (2009), http://rubyforge.org/projects/activerecord, Accessed 6. September 2012
12. Hick, J.M., Hainaut, J.L.: Database application evolution: A transformational approach. Data & Knowl. Eng. 59(3), 534–558 (2006)
13. JBoss Community: Hibernate (2011), http://www.hibernate.org/, Accessed 6. September 2012
14. Jing, J., Claypool, K., Jin, J., Rundensteiner, E.: SERF: Schema Evolution through an Extensible, Re-usable and Flexible Framework. In: Int. Conf. on Information and Knowl. Management (1998)
15. Lerner, B.S., Habermann, A.N.: Beyond schema evolution to database reorganization. ACM, New York, NY, USA (1990)
16. Macek, O., Richta, K.: Transformations for Application and Database Co-Refactoring (2013), https://github.com/macekond/comsis_transformations_appendix [Accesed 19. November 2013]
17. McBrien, P., Poulovassilis, A.: A General Formal Framework for Schema Transformation. Data and knowledge engineering 28(1), 47–71 (September 1998), http://pubs.doc.ic.ac.uk/formal-framework-transformation/
18. Microsoft: ADO.NET Enitity Framework (2011), http://msdn.microsoft.com/en-us/library/bb399572.aspx, Accessed 6. September 2012

19. Moravec P., Harmanec D, Tarant P., Jezek J.: An practical approach to dealing with evolving models and persisted data. In: Code Generation (2012)
20. Roddick, J.F., Craske, N.G., Richards, T.J.: A taxonomy for schema versioning based on the relational and entity relationship models. In: Proceedings of the 12th International Conference on the Entity-Relationship Approach: Entity-Relationship Approach. pp. 137–148. ER '93, Springer-Verlag, London, UK, UK (1994), http://dl.acm.org/citation.cfm?id=647515.727030
21. Schulz, C., Löwe, M., König, H.: A categorical framework for the transformation of object-oriented systems: Models and data. J. Symb. Comput. 46(3), 316–337 (Mar 2011), http://dx.doi.org/10.1016/j.jsc.2010.09.010
22. Tresch, M.: A framework for schema evolution by meta object manipulation. In: Proc. of the 3d Int. Workshop on Found. of Model. and Lang. for Data and Objects (1991)
23. Van Paesschen, E., De Meuter, W., D'Hondt, M.: Selfsync: a dynamic round-trip engineering environment. Model Driven Eng. Lang. and Syst. pp. 633–647 (2005)
24. Voxland N.: Liquibase (2013), http://www.liquibase.org [Accesed 19. November 2013]
25. Wedemeijer, L.: Semantical change patterns in the conceptual schema. In: Proceedings of the Workshops on Evolution and Change in Data Management, Reverse Engineering in Information Systems, and the World Wide Web and Conceptual Modeling. pp. 122–133. ER '99, Springer-Verlag, London, UK, UK (1999), http://dl.acm.org/citation.cfm?id=647523.728210

Ing. Ondrej Macek. He received his MS in 2008 in study programme Electrical Engineering and Informatics from the Czech Technical University in Prague. He has been an assistant professor at the Czech Technical University since 2010 where he teaches undergraduate courses software analysis, object oriented design and project management. His research interests are model driven development, databases, software engineering and formal specification. He is a member of the Upsilon Pi Epsilon - International Honor Society for the Computing and Information Disciplines and a member of a conference committee of DATESO.

Doc. Ing. Karel Richta, PhD. He graduated from the Czech Technical University in Prague in 1974. Specializes in the area of formal specification and semantics, engaged in programming languages and operating systems, currently focuses mainly on software engineering and database management systems. This topic teaches in the Department of Computer Science & Engineering at Czech Technical University in Prague, and also in the Department of Software Engineering of Charles University in Prague. He has published more than 110 contributions to scientific conferences and in journals. He is a member of numerous program committees, the member of a steering committee of DATAKON conference, and the seminar DATESO. He is the president of the Czech ACM Chapter, the member of TC2 committee of IFIP, the member of Czech Society of Information Systems, the member of Czech Society of Systems Integration, also acts as a permanent reviewer on IEEE Software.

Received: June 10, 2013; Accepted: May 30, 2014.

Computer Science and Information Systems 11(2):525–548 DOI: 10.2298/CSIS130120022D

User-Centric Privacy-Preserving Statistical Analysis of Ubiquitous Health Monitoring Data[⋆]

George Drosatos[1,2] and Pavlos S. Efraimidis[1,2]

[1] Dept. of Electrical and Computer Engineering, Democritus University of Thrace
University Campus, 67100 Xanthi, Greece
{gdrosato,pefraimi}@ee.duth.gr
[2] ATHENA, Research & Innovation Center, University Campus, 67100 Xanthi, Greece

Abstract. In this paper, we propose a user-centric software architecture for managing Ubiquitous Health Monitoring Data (UHMD) generated from wearable sensors in a Ubiquitous Health Monitoring System (UHMS), and examine how these data can be used within privacy-preserving distributed statistical analysis. Two are the main goals of our approach. First, to enhance the privacy of patients. Second, to decongest the Health Monitoring Center (HMC) from the enormous amount of biomedical data generated by the users' wearable sensors. In our solution personal software agents are used to receive and manage the personal medical data of their owners. Moreover, the personal agents can support privacy-preserving distributed statistical analysis of the health data. To this end, we present a cryptographic protocol based on secure multi-party computations that accept as input current or archived values of users' wearable sensors. We describe a prototype implementation that performs a statistical analysis on a community of independent personal agents. Finally, experiments with up to several hundred agents confirm the viability and the effectiveness of our approach.

Keywords: privacy, ubiquitous health data, privacy-preserving statistical analysis, personal software agent, secure multi-party computation.

1. Introduction

The requirement to provide health care to special groups of people who have the need of continuous health monitoring is an integral part of today's society. Moreover, the number of people who need such health monitoring services is increasing. An important reason for this is the aging of the populations, which constitutes a social and economical challenge especially for the developed countries [1]. Related researches which have been carried out both in the European Union [2] and the United States [3] indicate that the number of people over the age of 65 is increasing. A similar increase is expected to take place throughout the developed world. Many elderly people suffer from chronic diseases that require health care and frequent visits to hospitals. For people of this category, it is important to continuously monitor the state of their health. Effective monitoring of the health state can improve the quality of the patients' life or even save their life, while simultaneously reducing the cost of health care [4, 5].

[⋆] Preliminary parts of this work were presented at the 4rd International Conference on PErvasive Technologies Related to Assistive Environments (PETRA 2011) and the 8th International Conference on Trust, Privacy & Security in Digital Business (TrustBus 2011).

The rapid development of the wearable sensors technology led to the appearance and the implementation of prototype Ubiquitous Health Monitoring Systems (UHMS's) [4–6]. Moreover, there is a plethora of researches in the area of ambient assistive living services [7–9] and controlled access to ubiquitous hospital information [10]. The objective of a UHMS is to provide continuous health monitoring, both at home and outdoors. People need to have their health condition under control not only when at home, but wherever they are. One of the main features of a UHMS is to automatically generate alerts to notify the family or the patient's doctor about a possible health emergency so they should rush to their help to him. Examples of the data used for the detection of a possible health incident, as they are reported in [11], are: heart rate, blood pressure, galvanic skin response, skin temperature, heat flux, subject motion, speed and the covered distance.

Although ubiquitous computing is an opportunity for improving the health sector; however, for ubiquitous health monitoring technology to become feasible, a number of challenges are facing its presence [12]. These challenges are related to the deployment of this technology [13] and to issues such as resource constraints, user mobility, cost, heterogeneity of devices, scalability, security and privacy. While in [14] the author believes that challenges associated to sensor technology features also exist, such as Quality of Service (QoS), low power consumption and security of the wireless devices.

Privacy is an important issue of UHMS and health-related applications in general, since health data are sensitive personal data of patients. Privacy-related legislation like the European Data Protection Directive [15] and the HIPAA (Health Insurance Portability and Accountability Act) [16] explicitly define the rules for protecting the privacy of patients. The so far general architecture of a UHMS requires that all personal medical data (such as those reported above) which are produced by the patients' wearable sensors are collected and stored in a central service, specifically at the Health Monitoring Center (HMC) [4, 5]. The HMC is responsible not only for the collection and storage, but also for the control of these critical personal data. However, this technique runs significant risks for the security of the actual data, for the privacy of the monitored people, and, moreover, has an enormous computational and storage cost for the HMC. The distributed architecture that we propose in this work can offer the required scalability to handle large or even huge amounts of personal data.

At the same time, statistics of personal health data can be of high value for medical purposes. For example, the use of statistical methods is an integral part of medical research. A medical statistic may comprise a wide variety of data types, the most common of which are based on vital records (birth, death, marriage), morbidity (incidence of disease in a population) and mortality (the number of people who die of a certain disease in relation with the total number of people). Additional personal data items may needed for other well-known statistical computations, like the demographic distribution of a disease based on geographic, ethnic, and gender criteria, the socioeconomic status and education of health care professionals, and the costs of health care services.

In this work, we deal with the privacy-enhanced management of ubiquitous health monitoring data. Moreover, we describe how this data can be used within privacy-preserving distributed statistical analysis. Regarding the first deal, we suggest the decentralization of the collection of medical data at the users' side. This is achieved by the use of personal agents that will be continuously online and collect the medical data of their owners. In addition to the data that are obtained by wearable sensors, the agents may also have other data, such as demographic elements about the patient and further information about his health records, as well. The additional data can be used to support filtering of the results within distributed computations. Apart from the management of the personal data, the patient agent's automatically monitors the different changes in medical data with a dedicated health component. As soon as the health component detects aberrations in the raw health data, it informs the HMC by giving it access to the user's data so as to decide itself for the danger of the situation. In our approach, the usage of the agents does not block the remote monitoring of the patient's health by an authorized doctor; it only ensures the controlled, user-aware, access to these sensitive data.

For the statistical analysis, we propose a cryptographic protocol based on secure multi-party computations that accept as input current or archived values of users' wearable sensors. This distributed computation is performed by a community of the patients' personal software agents. We design an algorithm for the distributed computation, present a prototype implementation of the proposed solution, and obtain experimental results that confirm the viability and the effectiveness of our approach.

Main Advantages of Our Solution

The personal data management approach proposed in this work achieves a number of advantages in comparison with the existing architecture of a UHMS, and simultaneously enhances the privacy of the patients in such a system. The main advantages are:

- Only controlled access to the health data is provided. Every data access is logged by keeping who retrieved which data items and when this happened.
- The whole history of medical data, including the raw sensors' data, can be kept in the agent, whereas this might not be possible on the HMC for practical reasons. At the same time, decongestion of the HMC from the large amount of data, is achieved. This can make the computational requirements of the central servers more tolerable.
- Less risk of massive theft of personal data since they are distributed at the users' side.
- Option for usability of these data by authorized third independent services or for performing distributed computations.

On the other hand, important advantages of our statistical analysis approach in comparison to traditional statistical analysis techniques are:

- Utilizing valuable, sensitive, up-to-date personal data while ensuring privacy.
- Simplifying the process and significantly reducing the time and cost for conducting a statistical analysis.

A prerequisite for our approach is that each patient must have a personal software agent at his disposal and permanent access to the Internet. The computational requirements for the personal agent can be fulfilled with commodity hardware and hence its cost is not high. Thus, it is plausible to assume that patients with a UHMS can afford the extra cost for such an agent.

Our Contributions

- We present a user-centric software architecture for managing UHMD generated from wearable sensors in a UHMS, that allows controlled access to the health data, decongests the HMC, and enhances the privacy of users.

- We propose the usage of personal software agents for the management of biomedical data at the user side.
- We implement a prototype of the agents for this work.
- We present a privacy-preserving cryptographic protocol for distributed statistical analysis of the health data within agent communities.
- We validate our approach with a set of experiments on generated biomedical data in a community of real software agents.

Outline. The rest of this paper is organized as follows. In Section 2, we describe related work. In Section 3, we introduce the management architecture of our privacy-enhanced UHMS. In Section 4, we propose a system for performing privacy-preserving distributed statistical analysis on ubiquitous health data. Finally, conclusions of this work are given in Section 5.

2. Related Work

Personal data of users are commonly stored in central databases at the service provider's side. In this way, the users have essentially no control over the use of their personal data. The idea that individuals should own their personal information themselves and decide how this information is used, is discussed in [17]. A point made in [18] is that, although considering personal data the owner's private property is a very appealing idea, it would be rather difficult to practically apply it and legally enforce it. The argument that personal data would be safer at the user's side is also examined in [19].

To address privacy concerns, different kinds of frameworks that are related to personal data have recently been proposed. In particular, privacy sensitive management of personal data in ubiquitous computing is discussed in [20], and storing personal data in an individual's mobile device is examined in [21]. Of particular importance for the management of health data in this work is Polis [22], a framework for managing personal data at the owner's side. Polis offers privacy-enhanced management of personal data based on the principle that each individual has absolute control on his personal data, which remain permanently at the side of their owner and only there. Each user of Polis is a unique entity which is represented by a corresponding Polis agent. The Polis agents constitute the backbone of the Polis architecture; they are used to manage the personal data of an entity and provide controlled access at the entity's data. The service providers request personal data items of users from their personal agents. The agents provide the requested data if there is a corresponding policy and/or license agreement. In this work, we extend Polis agents with additional features and adapt the decentralized, agent-based approach of Polis for the management of the patients' personal data. Some work related to Polis has been done within the DISCREET project where a rich but also complicated framework for privacy protection has been proposed [23]. This framework is built on the principle that personal data is kept inside a "Discreet Box", located at the service provider's side. An agent-based solution to address usability issues related to P3P (Platform for Privacy Preferences Project) is presented in [24]. General surveys on privacy enhancing technologies are given in [25, 26].

In the second and main part of this work, we present a solution for distributed privacy-preserving statistical analysis of personal health data. Our approach is based on secure multi-party computations (MPCs). The general model of a MPC was firstly proposed by Yao [27] and later was followed by many others [28, 29]. In general, a MPC problem concerns the calculation of a function with inputs from many parties, where the input of each participant is not disclosed to anyone. The only information that should be disclosed is the output of the computation. The general solution for MPC presented in [27] is powerful but commonly leads to impractical implementations.

A secure two-party computation (S2C) for the calculation of statistics from two separate data sets is presented in [30]. Each data set is owned by a company and is not disclosed during the computation. Similar results are shown in [31], this time focusing on linear regression and classification and without using cryptographic techniques. Some indicative works from the related field of privacy-preserving data mining are [32–34]. A major difference of our work from the above is that in our approach every participant is in control of his health data and that the distributed computation is performed by the community of the personal software agents. Using software agents as building blocks for software systems is an established practice; see for example [35] and for a recent survey [36].

Another approach for statistics on personal data is anonymization, i.e., the sanitization of a data collection by removing identifying information. The data anonymization approach and some of its limitations are discussed for example in [37–39]. Data anonymization applies to data collections in central databases and is not directly comparable to our decentralized approach. Finally, an example of an efficient privacy-preserving distributed computation is given in [40], where personal agents of doctors execute a distributed protocol to identify the nearest doctor to an emergency. The focus of the present work is on privacy-preserving distributed statistical analysis using a massive number of participants.

3. Privacy-Enhanced Management of UHMD

In this section, we describe the proposed architecture for privacy-enhanced management of UHMD and show how it fulfills the goal of protecting the personal data and enhancing the privacy of patients.

3.1. Management Architecture

An overview of the proposed architecture for a UHMS is presented in Figure 1. The emphasis of the description is on the part of personal agents. The biomedical data that are produced by the patients' wearable sensors are wirelessly collected through a local wireless network in the patient's body into a personal mobile device, such as a smart phone. Afterwards, the measured biomedical data are transmitted via multiple complementary wireless networks (GPRS, 3G, Wi-Fi), through the Internet, towards the patient's personal agent. The personal agents that are used for this task are the Polis agents and have been suitably modified for this purpose. The features which have been added to the Polis agents so as to be used in a UHMS are:

1. Ability to collect dynamic personal data, such as the biomedical data of the patients' wearable sensors.
2. Ability to control the values of the biomedical data for the detection of some indicative cases of emergency.

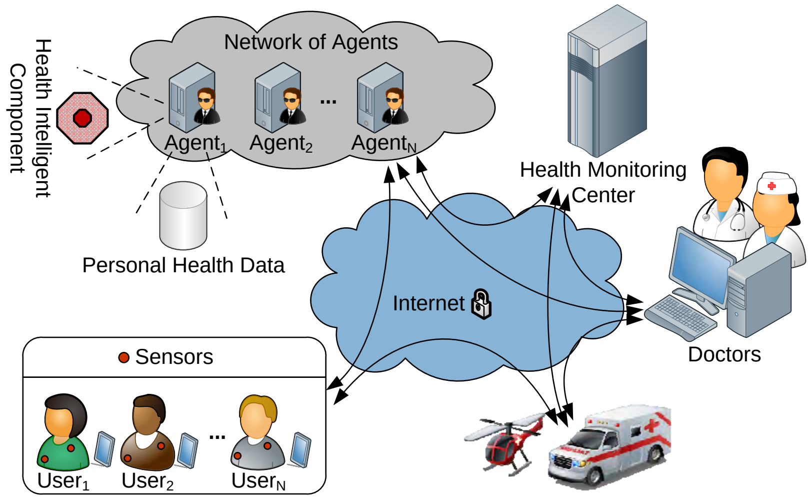


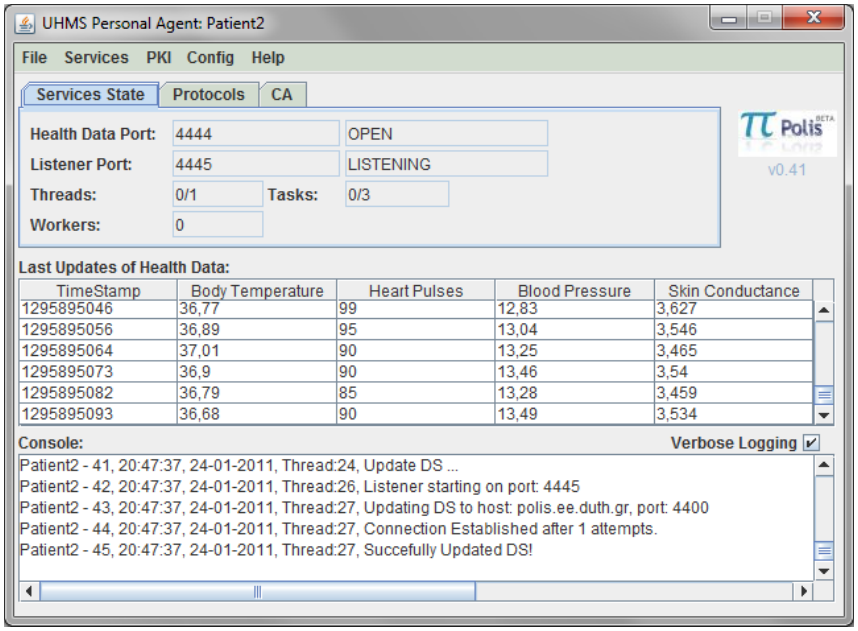
Fig. 1. The proposed management architecture for a UHMS.

A snapshot of a patients' personal agent is shown in Figure 2. On the other hand, the patients' personal agents are self-organized into an appropriate virtual network topology that can provide easy organization and identification of the agents. This network topology can be used as a tool to conduct privacy-preserving distributed computations.

Our architecture can support an intelligent health component which can make a first check of the health data in real time. We provide an overview of the functionality of such a component; a real implementation of such a tool is outside of the scope of this work. The health component of the personal agent checks automatically the incoming vital signs with the purpose to address for further thorough check in HMC if there are indications of an emergency (see Figure 3). An example of rules/decisions that a health component can apply in order to decide about an emergency can be found in [7]. If necessary, the HMC can be consulted by the personal doctor of the patient. The personal doctors are shown as "Doctors" in Figure 1. Depending on the situation, the HMC can coordinate the immediate medical service at the closest or most appropriate local medical facility using the best available transportation service (e.g.: ambulance). Finally, an additional responsibility of the HMC is to inform the family of the patient about his condition so that they could rush to provide their help.

3.2. Benefits of the Architecture

The idea of a decentralized architecture for storage and control of the patients' medical data into their personal agents, as it has already been mentioned provides the advantage of enhanced control on the user's personal data. Moreover, this decentralized approach can also contribute to improved data security, since invaders find large collections of personal data much more inviting than an individual's personal data [19]. The decentralized approach grants to the patient the right to control the disclosure of his health data and mitigates its feeling of being under permanent surveillance. In addition to enhancing privacy,



Health Data Port: 4444 OPEN

Listener Port: 4445 LISTENING

Threads: 0/1 Tasks: 0/3

Workers: 0

Last Updates of Health Data:

TimeStamp	Body Temperature	Heart Pulses	Blood Pressure	Skin Conductance
1295895046	36,77	99	12,83	3,627
1295895056	36,89	95	13,04	3,546
1295895064	37,01	90	13,25	3,465
1295895073	36,9	90	13,46	3,54
1295895082	36,79	85	13,28	3,459
1295895093	36,68	90	13,49	3,534

Console: Verbose Logging ☑

```
Patient2 - 41, 20:47:37, 24-01-2011, Thread:24, Update DS ...
Patient2 - 42, 20:47:37, 24-01-2011, Thread:26, Listener starting on port: 4445
Patient2 - 43, 20:47:37, 24-01-2011, Thread:27, Updating DS to host: polis.ee.duth.gr, port: 4400
Patient2 - 44, 20:47:37, 24-01-2011, Thread:27, Connection Established after 1 attempts.
Patient2 - 45, 20:47:37, 24-01-2011, Thread:27, Succefully Updated DS!
```

Fig. 2. A snapshot of UHMS personal agent. On the top, network configuration parameters of the agent can be seen. In the middle, the latest health data received by the agent are shown. Finally, at the bottom, logging information about the operation of the agents is presented.

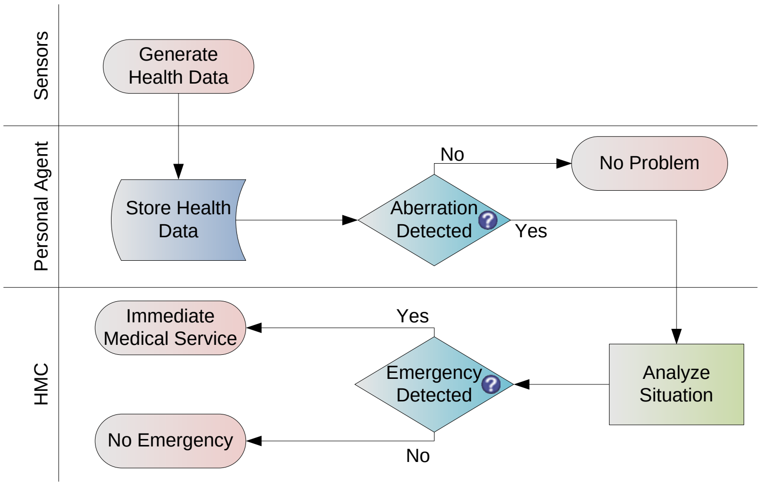


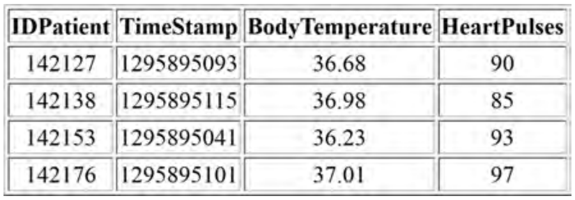
Fig. 3. A system flowchart of the biomedical information.

the decongestion of HMC from the huge amount of data, including raw sensors' data, that would be accepted if the patients sent their data directly to it, is achieved. Even in the case that the data would be collected at the HMC, these would be much less in volume than those that would actually be produced by the sensors, thus the analysis would not be as effective as the one that would be made by the agents themselves by having the complete data. With the proposed health data management approach of this work, the HMC has now to handle only those cases which may be at a certain risk.

Of course, the decentralized architecture holds challenges and issues too. The managing of a personal agent is by definition a critical task, prone to errors and omissions by the user. However, it is possible to mitigate these risks by standarizing or even automating the corresponding procedures of the agent. Furthermore, there are issues about the agent's security; a production-ready agent should satisfy high security levels. We believe that this is a viable task, since the agent has a precise, well-defined functionality and can be operated behind firewalls on a user-controlled computing platform. Also, another issue is what will happen if temporarily the patient's agent has no network connectivity (offline). In this case, the patient's data which are collected by his mobile device could be kept there and later be transmitted to the personal agent as soon as the failure is restored. Moreover, during a failure of the personal agent, health data could also be transmitted directly to the HMC for storage and control. Measures such as the above can ensure fault-tolerance against possible agent failures.

It is noteworthy that storing health data at the patients' side does not exclude the possibility to access the data from a central database as long as the database is entitled to do so. As shown in [22], the personal agents of Polis can be interconnected with mainstream database servers to provide transparent access to the personal data fields. The basic idea is that personal data fields in the central database do not contain the actual data; instead, a ticket represented by an appropriate data object is used to retrieve the data value on the fly. With this approach, which has been tested with an Oracle database server, a query submitted to the database may transparently retrieve – on the fly – personal data items from the associated personal software agents and present the personal data within the recordset (the answer of the database) of the query. An example query and the corresponding recordset are given in Figure 4. The data fields TimeStamp, BodyTemperature and HeartPulses are personal data fields and their content are – transparently for the database user – dynamically retrieved from the corresponding personal agents.

```
SQL> Select IDPatient ,TimeStamp ,BodyTemperature ,
     HeartPulses From CurrentBiomedicalData
     Where IDPatient Between 142120 And 142180;
```



IDPatient	TimeStamp	BodyTemperature	HeartPulses
142127	1295895093	36.68	90
142138	1295895115	36.98	85
142153	1295895041	36.23	93
142176	1295895101	37.01	97

Fig. 4. SQL access to remote health data.

The choice to store the patient's data in an agent enables the possibility to utilize these data for the common wealth. The Nearest Doctor Problem (NDP) [40] is mentioned as a typical example. The NDP is a privacy-preserving protocol, which uses a network of the doctors' agents aiming to find the nearest doctor in case of an emergency, by using dynamic data such as their location. In our case, the data of the patients could be used for a similar distributed computation. Such an example is the monitoring progress/spread of a pandemic in a region. Data such as the location and the body temperature of the patients would be required for this example. Another example is a medical statistical research on the biomedical data of the wearable sensors as well as on the medical records of patients. In the following section we use our distributed data management approach for a distributed statistical analysis application.

4. Privacy-Preserving Statistical Analysis on UHMD

In this section, we present a method for statistical analysis of ubiquitous health monitoring data (UHMD). For the statistical analysis we propose a privacy-preserving distributed computation that is collaboratively executed by the participating personal software agents. We first define the kind of privacy that is achieved and then proceed with the description of the distributed computation.

4.1. Privacy-Preserving Computation

There are two distinct problems that arise in the setting of privacy-preserving statistics/-data mining [41]:

(a) The first is to decide which functions can be safely computed, where safety means that the privacy of individuals is preserved if the result of the computation is disclosed. We will assume that the outcomes of the statistics computations do not violate the privacy of the participating patients and will not further consider this problem in this work.

(b) The second is how, meaning with which algorithms and protocols, to compute the results while minimizing the damage to privacy. For example, it is always possible to pool all of the data in one place and run the computation algorithm on the pooled data. However, this is exactly what we don't want to do (hospitals are not allowed to hand their raw data out, security agencies cannot afford the risk, and governments risk citizen outcry if they do). The focus of our work is on this problem.

Thus, the question we will address is how to achieve privacy of type (b), that is, how to compute the statistic results without pooling the data, and in a way that reveals nothing but the final results of the distributed computation.

4.2. Architecture of the Distributed Computation

Our solution is build on top of the privacy-enhanced UHMS presented earlier in this work. An overview of the architecture of the statistical analysis system including the extra components that are required for a distributed statistical analysis computation, i.e., the Network Community of Personal Agents and the Statistical Analysis Service (SAS), is shown in Figure 5.

The personal agents are organized into a virtual topology, which may be a simple ring topology or a more involved topology for time-critical computations. On the other hand, the SAS is a server that initiates the distributed computation on the users' medical data and collects the aggregate results. Each researcher who wishes to carry out a statistical research and is entitled to do so, can submit his task to the SAS.

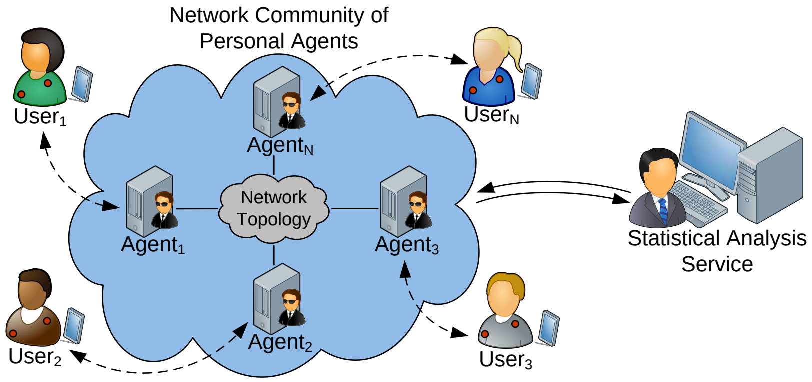

Network Community of Personal Agents
User1
UserN
AgentN
Network Topology
Agent1
Agent3
Agent2
User2
User3
Statistical Analysis Service

Fig. 5. The architecture for performing privacy-preserving statistical analysis.

4.3. The Main Steps of the Distributed Computation

The main steps of the distributed computation that we propose for the statistics calculation are:

- Initially, the researcher submits the request to conduct a specific statistical analysis to the SAS.
- The SAS accepts the request after verifying the credentials of the researcher.
- The SAS picks one of the personal agents to serve as the root-node for the particular computation and submits the request to it.
- The root-node coordinates a distributed computation that calculates the specified statistical function.
- At the end of the distributed computation, the SAS and the researcher will only learn aggregate results of the computation without any additional information of the personal data of individual participants.

4.4. The Secure Distributed Protocol

In this section, we present the main idea of the cryptographic protocol that is used in the statistical computations. The protocol is secure in the Honest-But-Curious (HBC) model (see Section 4.7), where the users' agents participating in the computation follow the protocol steps but may also try to extract additional information. During the calculation the actual users' personal data are not disclosed in any stage of the process but only the aggregate results are revealed at the end. An instance of a statistical computation problem consists of:

- **N patients** $P_1, P_2, \ldots, P_N$ and their personal data.
- **N personal software agents:** The agents of all patients that will participate in the distributed privacy-preserving computation.
  - **Input:** The type of the statistical function and its parameters. In addition, selectivity constraints for the data set may also be specified. Note that more than one statistical functions on the same dataset can be calculated with a single computation.
  - **Output:** The necessary aggregate values (e.g. w_x, u_x, z_{xy} and n, which are defined later) that are needed to calculate the given statistical function.

Consider the following statistical computation instance: Computing the average of the female patients' age in a city. First, we assume that the results of the specific query are not considered a threat against the users' privacy, that is, privacy type (a) of Section 4.1 is preserved. Then, given the computation instance, the SAS chooses a node from the network of the users' agents as the root-node for the particular computation. The SAS sends the type of the requested computation and its parameters to the root-node. The parameters of the computation, i.e., the female gender and the city name, are used to filter the data set. Each personal agent, decides privately to provide data or not to the statistical research.

A simple topology for the personal agents is a virtual ring topology that contains all agents as nodes (Figure 6.b). For time-critical computations, more complex topologies like a virtual tree can be used (Figure 6.a). The tree topology for example has been used in [40]. At the end of the execution, the root-node collects the results of the calculation as an encrypted message and sends it to the SAS. The message is encrypted with the public key of the SAS, which is assumed to be known to all nodes. In this way, the protocol ensures k-anonymity (see Definition 3), where $k = N$ and N is the number of all the nodes in the network.

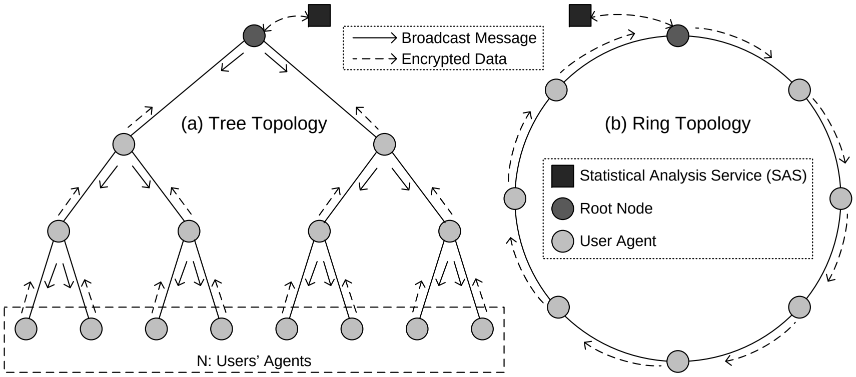


Fig. 6. Possible network topologies.

Cryptographic Tools. The *Paillier cryptosystem* [42] is a probabilistic asymmetric cryptographic algorithm for public key cryptography. The security of Paillier is implied by the Decisional Composite Residuosity Assumption (DCRA). In our cryptographic protocol, we use the additive homomorphic encryption property of the Paillier cryptosystem for calculating aggregate data in a privacy-preserving way.

Definition 1 (Homomorphic Encryption). *Homomorphic encryption [43, 44] is a form of encryption where one can perform a specific algebraic operation on the plaintext by performing a (possibly different) algebraic operation on the ciphertext. Particularly, an encryption algorithm $E()$ is homomorphic if given $E(x_1)$ and $E(x_2)$, one can obtain $E(x_1 \circ x_2)$ without decrypting x_1, x_2, for some operation $\circ$.*

The additive homomorphic encryption property of the Paillier cryptosystem means that multiplication of encrypted values corresponds to addition of decrypted ones, that is,

$$\begin{aligned} E(x_1) \cdot E(x_2) &= (g^{x_1} \cdot r_1^{n_p}) \cdot (g^{x_2} \cdot r_2^{n_p}) \\ &= g^{[x_1 + x_2 \bmod n_p]} \cdot (r_1 r_2)^{n_p} \bmod {n_p}^2 \\ &= E([x_1 + x_2 \bmod n_p]) \ , \end{aligned}$$

where

- x_1 and x_2 are two plain messages such that $x_1, x_2 \in \mathbb{Z}_{n_p}$,
- (n_p, g) is the Paillier public key,
- r_1 and r_2 are two random numbers such that $r_1, r_2 \in \mathbb{Z}^*_{n_p}$, and
- $E(m) = g^m r^{n_p} \bmod {n_p}^2$ is the encryption of message m.

The Paillier cryptosystem is a very popular additively homomorphic cryptosystem. It should be noted, however, that within our proposed distributed computation any other cryptosystem that supports the additive homomorphic property could also be used in place of the Paillier cryptosystem. For example the Benaloh cryptosystem [45] could be used within our solution. Moreover, it would also be possible to use the ElGamal cryptosystem [46], that supports multiplicative homomorphic property, provided that the computations are adapted accordingly. For example in this case one would have to transform the integer x to the group element z^x, for a fixed generator z, before encrypting with ElGamal. Thus, the transformation of multiplicative homomorphic property becomes $E(z^x) \cdot E(z^y) = E(z^x \cdot z^y) = E(z^{x+y})$.

4.5. The Computations

In this section, we use our approach to calculate representative statistical functions with a distributed computation. Wherever it is necessary, the expression of the statistical function is brought to a form that is appropriate for the distributed computation.

Arithmetic Mean. The arithmetic mean of a variable X (with sample space $\{x_1, \ldots, x_n\}$) is given by the following equation:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

We use the additive homomorphic property of Paillier encryption to calculate the value of the terms $u_x = \sum_{i=1}^{n} x_i$ and n. The calculation is privacy-preserving; no single x_i is disclosed. Once the SAS learns the values of the terms u_x and n, it can compute the arithmetic mean. More precisely, using the homomorphic property of Paillier, the two terms u_x and n can be transformed into the following form:

$$E_{pk}(u_x) = \prod_{i=1}^{n} E_{pk}(x_i) \text{ and } E_{pk}(n) = \prod_{i=1}^{n} E_{pk}(1) \ ,$$

where the E_{pk} indicates that the message is encrypted with the current public key of SAS for the specific statistical analysis. Each agent i that participates in the statistical analysis, prepares its own encryptions $E_{pk}(x_i)$ and $E_{pk}(1)$. These encrypted messages are used to calculate the above two global products. Agents that do not participate in the statistical computation (because for example they do not satisfy some selection criterion) multiply each of the above two products with an independent encryption of zero $E_{pk}(0)$.

Frequency Distribution. The frequency distribution is a tabulation of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval; in this way, the table summarizes the distribution of values in the sample. The graphical representation of the frequency distribution is the well known histogram. Figure 7 shows how the frequency distribution would become by using ciphertext as counters in each range, where each ciphertext is given by the following equation:

$$E_{pk}(n_v) = \prod_{i=1}^{n} E_{pk}(m), \text{where } m = \begin{cases} 1, \ x \in [x_{v-1}, x_v) \\ 0, \ x \notin [x_{v-1}, x_v) \end{cases}$$

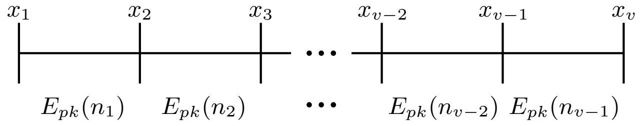


Fig. 7. Representation of a frequency distribution.

Linear Correlation Coefficient. The linear correlation coefficient $corr(X, Y)$ of two random variables X and Y is a measure of the strength and the direction of a linear relationship between two variables and is defined as:

$$corr(X, Y) = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \sqrt{n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2}}$$

The unknown terms that are required to calculate the linear correlation coefficient with the help of the homomorphic property of Paillier are $w_x = \sum_{i=1}^{n} x_i^2$, $u_x = \sum_{i=1}^{n} x_i$, $w_y = \sum_{i=1}^{n} y_i^2$, $u_y = \sum_{i=1}^{n} y_i$, $z_{xy} = \sum_{i=1}^{n} x_i y_i$ and n, by taking the following form:

$$E_{pk}(w_x) = \prod_{i=1}^{n} E_{pk}(x_i^2), \; E_{pk}(u_x) = \prod_{i=1}^{n} E_{pk}(x_i),$$

$$E_{pk}(w_y) = \prod_{i=1}^{n} E_{pk}(y_i^2), \; E_{pk}(u_y) = \prod_{i=1}^{n} E_{pk}(y_i),$$

$$E_{pk}(z_{xy}) = \prod_{i=1}^{n} E_{pk}(x_i y_i) \text{ and } E_{pk}(n) = \prod_{i=1}^{n} E_{pk}(1)$$

Variance. The variance $var(X)$ of a variable X is used as a measure of how far a set of numbers are spread out from each other. The unknown terms that are required to calculate the equation of variance with the help of the homomorphic property of Paillier are w_x, u_x and n. These terms can be calculated as shown earlier in the computations of the *arthmetic mean* and the *linear correlation coefficient*.

Covariance. The covariance $cov(X, Y)$ of two random variables X and Y is a measure of the strength of the correlation between the two variables. The unknown terms that are required to calculate the equation of covariance with the help of the homomorphic property of Paillier are u_x, u_y, z_{xy} and n.

Linear Regression. The linear regression of a dependent variable Y of the regressors X is given by the equation $y = a + bx$, where a and b are parameters. The unknown terms that are required to calculate the parameters of line y with the help of the homomorphic property of Paillier are w_x, u_x, u_y, z_{xy} and n.

From the analysis of the above statistical functions, we conclude that apart from the frequency distribution, all other function can be simultaneously calculated by computing once the required aggregate terms. Moreover, it is clear that the proposed solution can also be used to calculate other statistical functions, such as the polynomial regression and so on. We discuss such issues in the next section.

4.6. Computations with Homomorphic Cryptosystems

In our algorithm, we exploit the additive homomorphic property of Paillier to calculate additive aggregations which are then used to compute the values of statistical functions. We note that the same method can be used for multiplication-based aggregation if a cryptosystem supporting the multiplicative homomorphic property is used in place of Paillier. For example, the ElGamal and the RSA cryptosystems support multiplicative homomorphic encryption. Moreover, there are recent results on “somewhat” homomorphic cryptosystems, i.e., cryptosystems which support a limited number of homomorphic operations including both additive and multiplicative operations. More importantly, during the last years fully homomorphic cryptosystems supporting any number of additions and multiplications have been published, starting with the seminal work of Gentry [44]. Until now, fully homomorphic cryptosystems are not efficient enough to be used in practical applications like ours, though one could probably use somewhat homomorphic cryptosystems for some appropriate functions. A discussion of the efficiency and the practical relevance of current fully homomorphic and somewhat homomorphic cryptosystems [47].

4.7. The Protocol's Security

In this section, we show that the proposed protocol of a distributed statistical analysis in a UHMS does not violate the privacy of participants. The security holds for the model of Honest-But-Curious (HBC) users.

Definition 2 (Honest-But-Curious). *An honest-but-curious party (adversary) [48] follows the prescribed computation protocol properly, but may keep intermediate computation results, e.g. messages exchanged, and try to deduce additional information from them other than the protocol result.*

In the cryptographic protocol described above, the information exchanged by agents is encrypted with the Paillier cryptosystem [42], which is known to offer Semantic Security [49], that is, it is infeasible for a computationally bounded adversary to derive significant information about a message (plaintext) when given only its ciphertext and the corresponding public encryption key. Consequently, assuming honest-but-curious parties and that users' agents do not collude with the SAS party outside of the protocol, our approach is semantically secure. In Section 4.8, we show that the case where some user agents collude with the SAS outside of the protocol can be handled with a threshold decryption model.

From the above, we conclude that the computation with the homomorphic encryptions does not leak personal information of participating individuals (privacy type (b) in Section 4.1). As noted earlier, the (decrypted) outcomes of the statistic computation are also assumed to preserve privacy of type (a). We can now discuss the privacy guarantee of the whole approach. A common criterion for privacy protection is k-anonymity, which requires that data of the outcome cannot be associated with any particular patient.

Definition 3 (k-anonymity). *A simple definition of k-anonymity [50] in the context of this work is that no less than k individual users can be associated with a particular personal data value.*

The proposed solution offers k-anonymity in the sense that the result computed at the end of the protocol cannot be attributed to any of the N participated agents, i.e., $k = N$ even if the list of participating users is known (assuming no background information on specific users is available). In summary, the key security features of our protocol are:

- Each agent that receives a message from the previous node cannot obtain information about the contents of the message, because the ciphertexts are encrypted with the Paillier cryptosystem.
- Each node alters the ciphertexts of the computation. Even the nodes that do not participate in the statistical function multiply the ciphertexts with an encryption of number

"0", which is the neutral element of the additive homomorphic property of Paillier. Thus, the ciphertext is modified at every node, even if the corresponding node does not give any input to the computation.
- At the end of the protocol, only the variables that are needed for a particular statistical function are revealed. As a result, no individual can be associated with the value that he had used in the computation. Consequently, the proposed protocol preserves k-anonymity for $k = N$, where N is the number of all agents in the network.

Another criterion for evaluating privacy protection is the concept of differential privacy. Loosely speaking, the aim of differential privacy is to ensure that the ability of an adversary to inflict harm (or good, for that matter) – of any sort, to any set of people – should be essentially the same, independent of whether any individual opts in to, opts out of, the dataset [51, 52]. The formal definition of differential privacy follows.

Definition 4 (ϵ-Differential privacy [52]). *A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$, the following holds:*

$$Pr[\mathcal{K}(D_1) \in S] \leq exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S]$$

The probability is taken is over the coin tosses of $\mathcal{K}$.

If privacy of type (a) (Section 4.1) is preserved, for example, no queries or sequences of queries addressing a very small number of individuals are permitted etc., then it is plausible to assume that our approach achieves a satisfactory level of differential privacy. Note that the outcomes of the statistical computations are sums or aggregate results computed from a large number of sensor measurements and demographic values of a large population. One may also consider of adding Laplace noise [53] to the statistical results in order to further enhance the differential privacy criterion, even though there is some recent criticism of such an approach [54].

4.8. Security Discussion

In this section, we identify some representative threats against our application and discuss how they are or can be addressed within our approach. The threats concern either the correctness of the aggregated results or the privacy of the involved participants.

- *Incorrect sensor measurements.* This case refers to the case where one or more sensors generate erroneous data of values large enough to significantly influence the aggregate result. Such incidents could disrupt a statistical analysis and would be difficult to be noticed in the statistical results. However, such incorrect measurements could be detected by the intelligent health component or some dedicated filter of the patient's agent and excluded from the current statistical analysis. This solution is acceptable in the HBC model. Moreover, even for the case where such incorrect measurements could be maliciously submitted in order to skew the statistical result, we could use more advance techniques of the area of electronic voting [55]. In this case, each node would have to run a zero-knowledge proof with its predecessor/s with purpose to verify that the measurements are within an acceptable range.

- *Dedicated queries with purpose to reveal personal biomedical data of a particular patient.* One query or a set of queries may be chosen and submitted to target specific patients, by using background information on the set of participating individuals. Such dedicated queries may cause leakage of personal data of the selected patients. As noted earlier, such an attack is a threat against privacy of type (a) and the participants have to be protected with respect to such attacks. The problem is well known in the area of statistical databases [56] and it is not something new. A possible solution could be to use a second authority which will check if there are enough patients who cover the query's criteria before the SAS performs the specific statistical analysis.
- *Collusion among some patients and the SAS.* In this case, the SAS will try to collaborate with at least two patients (in the simple ring topology) with purpose to reveal the private values of a patient. These two patients have to be the predecessor and the successor of the particular patient. More specifically, the colluding predecessor creates neutral ciphertexts and forwards them to the intermediate node. This node would then encrypt its private values and forward the result to its colluding successor (according to the topology). The successor would then immediately return the values to the SAS which now gets to decrypt these private values. Such malicious behaviors can be effectively handled by deploying threshold decryption model [57] for the decryption of the encrypted values. Threshold decryption model requires a number of designated parties exceeding an appropriate threshold to cooperate for the decryption to be possible.

4.9. Experimental Results

To evaluate our solution, we developed a prototype that carries out distributed statistical analysis on medical data. The application is implemented in Java and for the cryptographic primitives the Bouncycastle [58] library is used. The personal agents of the Polis platform [22] are used as the personal data management agents of the patients. For this approach, the Polis agents were suitably modified so as to be able to manage both health records and health data that would actually be collected through a secure communication channel by the patients' wearable sensors. The community of the personal agents is organized as a Peer-to-Peer network. At this stage of development of the prototype, the backbone of the topology is a virtual ring topology. The ring offers a simple and reliable solution for the interconnection of the agents. For time-critical calculations of statistics a more involved topology like a virtual tree should be used.

The personal agents use production-ready cryptographic libraries and employ 1024 bits RSA X.509 certificates. The communication between agents is performed over secure sockets (SSL/TLS) with both client and server authentication. Below we describe an experiment of a distributed statistical analysis with 6 agents and the SAS. The requested statistic is:

- *The arithmetic mean of the current body temperature of patients who are aged between 55 and 65 years old and their gender is female.*

For the needs of the experiment, each agent generates random values for the age, the gender, and the current body temperature. We assume that the selectivity of the query criteria is high enough to preserve privacy of type (a). Then, in brief, the statistical computation works as follows. Initially, the SAS randomly chooses a node from the agents' network, in this case agent ‘Patient2’, as the root-node, and forwards the description of the statistical computation to it. The values of each agent which are related to the computation are shown in Table 1. The last two columns show the aggregate values that are encrypted after the corresponding agent applies its values to the results. Since the homomorphic property of Paillier applies to integers, decimal values like the body temperature have also to be represented with integers. In our example, the temperature is rounded to a number with at most two decimal digits and then multiplied by 100 to become an integer.

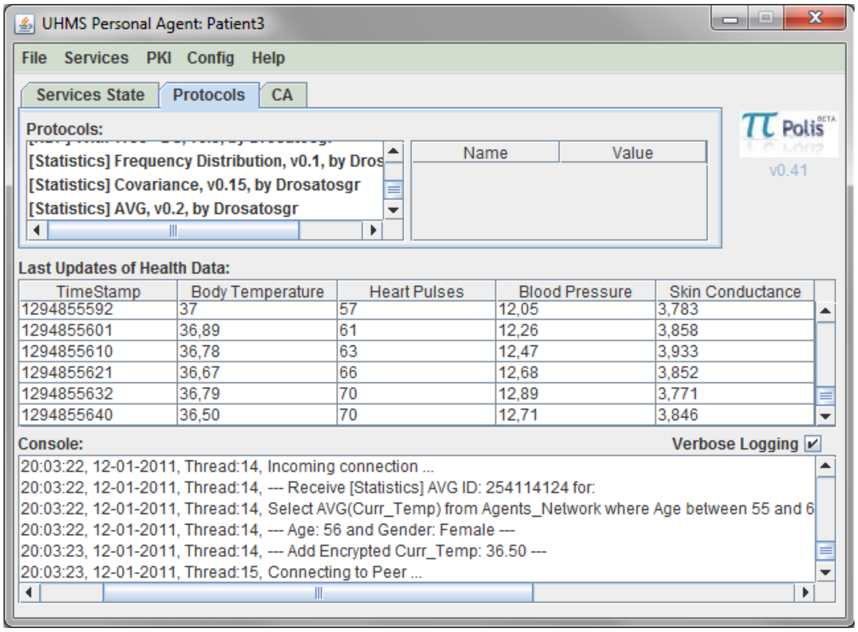
Table 1. Example of computation, where the agents in gray rows did not take part in computation.

Agent	Curr. Temp.	Age	Gender	$E_{pk}(u_x) = \prod_{i=1}^{n} E_{pk}(x_i)$	$E_{pk}(n) = \prod_{i=1}^{n} E_{pk}(1)$
Patient2	36.68 ^oC	51	Female	$E(0)$	$E(0)$
Patient3	36.50 ^oC	56	Female	$E(3650)$	$E(1)$
Patient4	37.70 ^oC	60	Female	$E(7420)$	$E(2)$
Patient5	38.10 ^oC	65	Female	$E(11230)$	$E(3)$
Patient6	37.12 ^oC	59	Male	$E(11230)$	$E(3)$
Patient1	36.20 ^oC	63	Female	$E(14850)$	$E(4)$

At the end of the computation, the agent ‘Patient2’ as the root-node collects the results and sends them back to the SAS. Finally, the SAS decrypts the results and finds that the average of the question which was submitted is $37.125\ ^oC$. A snapshot of the application during the execution of the experiment is shown in Figure 8.

We also performed a set of large-scale experiments with up to 300 agents. More precisely, we evaluated the efficiency of our solution with a series of experiments on a gradually increasing number of up to 300 agents. For this experiment, a network of 30 computer workstations with Intel Core 2 Quad Q8300 CPU’s at 2.5 GHz, 2 GB RAM and a 100 Mbps network, were used. The workstations were running a 32-bit operating system and the agents were executed in 32-bit Java virtual machines. Each computer was shared by at most 10 agents, to ensure an even workload distribution and avoid single overloaded workstations; an overloaded workstation would become a bottleneck that could significantly delay the execution of the whole protocol.

The running times of our experiments are shown in Figure 9. In this figure, we present the execution times for the computation of the arithmetic mean, the variance and the frequency distribution (for 10 subintervals) functions. As expected, the execution times depend practically linearly on the number of agents which take part in the computation and on the number of encryptions and multiplications in every statistical function. The overall running time is more than satisfactory for batch execution of statistical computations. In case of large numbers of statistical computations, the rate of computations can be substantially improved by using a pipeline of independent computations. For cases where the run-time of the computations is important, the distributed computation can be executed on a virtual tree or some other – low depth – topology, instead of the ring topology. In this case one would expect, and we actually have such preliminary measurements albeit within a different context [40], that the total running time will depend only logarithmically on the total number of nodes.



Protocols:

[Statistics] Frequency Distribution, v0.1, by Dros
[Statistics] Covariance, v0.15, by Drosatosgr
[Statistics] AVG, v0.2, by Drosatosgr

Last Updates of Health Data:

TimeStamp	Body Temperature	Heart Pulses	Blood Pressure	Skin Conductance
1294855592	37	57	12,05	3,783
1294855601	36,89	61	12,26	3,858
1294855610	36,78	63	12,47	3,933
1294855621	36,67	66	12,68	3,852
1294855632	36,79	70	12,89	3,771
1294855640	36,50	70	12,71	3,846

Console:

```
20:03:22, 12-01-2011, Thread:14, Incoming connection ...
20:03:22, 12-01-2011, Thread:14, --- Receive [Statistics] AVG ID: 254114124 for:
20:03:22, 12-01-2011, Thread:14, Select AVG(Curr_Temp) from Agents_Network where Age between 55 and 6
20:03:22, 12-01-2011, Thread:14, --- Age: 56 and Gender: Female ---
20:03:23, 12-01-2011, Thread:14, --- Add Encrypted Curr_Temp: 36.50 ---
20:03:23, 12-01-2011, Thread:15, Connecting to Peer ...
```

Fig. 8. A snapshot of the agent ‘Patient3’.

Finally, the execution times of the computations could be significantly reduced by simply using 64-bit Java virtual machines for running the experiments. This change would greatly improve the execution times especially of the heavy encryption operations which involve BigInteger[3] variables. In a comparable, independent, experiment we noticed an almost four-times improvement of the execution times when 64-bit Java was used in place of 32-bit Java. The use of the 64-bit virtual machine seems to effectively exploit the bigger registers of the AMD64 architecture for the cryptographic operations.

5. Conclusions

The tendency of the society towards increasing numbers of elderly people and generally people who need continuous health monitoring makes the need of Ubiquitous Health Monitoring Systems (UHMS) imperative. At the same time the concerns of the public about privacy are also rising. In this work, we presented a software architecture for privacy-enhanced UHMS and proposed the use of the ubiquitous health data that are obtained by the wearable sensors in a UHMS for caring out statistical researches. The proposed architecture allows the patients to have enhanced control over their personal data, so as not to have the feeling of being continuously under surveillance. The enhanced control on their personal data was achieved by using personal software agents for the management of the patients’ personal data. Putting personal agents in charge of personal health data can open the way for the definition and implementation of new services which utilize

[3] BigInteger is an immutable arbitrary-precision integer.

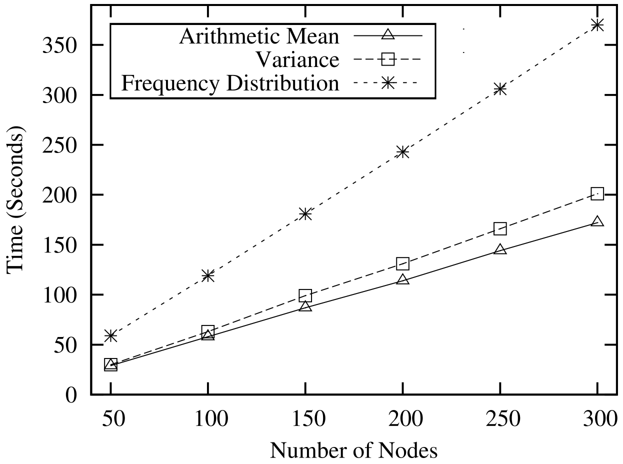


Fig. 9. Computation times of arithmetic mean, variance and frequency distribution (for 10 subintervals) statistical functions with respect to the number of agents.

personal data to contribute to public well being, while at the same ensuring the privacy of the involved individuals.

In this direction, we designed an algorithm for the distributed computation and, based on this algorithm and cryptographic primitives, we presented a solution for privacy-preserving statistical analysis on ubiquitous health data. The protection of privacy is achieved by using cryptographic techniques and performing a distributed computation within a network of patients' personal agents. We described how representative statistical functions can be executed distributedly by using the proposed cryptographic protocol. Finally, we developed a prototype implementation and performed an experimental evaluation that confirmed the viability and the efficiency of our approach.

Overall, our work demonstrates the feasibility of decentralized, scalable, privacy-enhanced management of Ubiquitous Health Monitoring Data (UHMD), and, most importantly, presents how privacy-preserving statistical analysis can be efficiently performed on such an architecture.

Acknowledgments. The present work was partially funded by the project ATLAS (Advanced Tourism Planning), GSRT/CO-OPERATION/11SYN-10-1730.

References

1. Commissioned by Philips. Healthcare strategies for an ageing society. In *The fourth report in a series of four from the Economist Intelligence Unit*, pages 1–31, UK, December 2009. The Economist. `http://graphics.eiu.com/upload/eb/Philips_Healthcare_ageing_3011WEB.pdf`.
2. Asghar Zaidi. Features and challenges of population ageing: The european perspective. In *This Policy Brief is derived from the presentation made at the Social and Economic Council of Spain (CONSEJO ECONOMICOY SOCIAL, CES, Madrid), as their keynote speaker in the*

conference "Ageing of Population". European Centre for Social Welfare Policy and Research, 2008. http://www.euro.centre.org/data/1204800003_27721.pdf.

3. G.T. Huang. Monitoring mom: As population matures, so do assisted-living technologies. In *Technical Review 20*, July 2003.
4. Chris Otto, Aleksandar Milenkovic, Corey Sanders, and Emil Jovanov. System Architecture of a Wireless Body Area Sensor Network for Ubiquitous Health Monitoring. *Journal of Mobile Multimedia*, 1:307–326, January 2006.
5. Arjan Durresi, Arben Merkoci, Mimoza Durresi, and Leonard Barolli. Integrated biomedical system for ubiquitous health monitoring. In *Proceedings of the 1st international conference on Network-based information systems*, NBiS'07, pages 397–405, Berlin, Heidelberg, 2007. Springer-Verlag.
6. Akira Yamazaki, Akio Koyama, Junpei Arai, and Leonard Barolli. Design and implementation of a ubiquitous health monitoring system. *Int. J. Web Grid Serv.*, 5:339–355, December 2009.
7. Dimitrios D. Vergados. Service personalization for assistive living in a mobile ambient healthcare-networked environment. *Personal Ubiquitous Comput.*, 14:575–590, September 2010.
8. A. Wood, G. Virone, T. Doan, Q. Cao, L. Selavo, Y. Wu, L. Fang, Z. He, S. Lin, and J. Stankovic. Alarm-net: Wireless sensor networks for assisted-living and residential monitoring. Technical report, Department of Computer Science, University of Virginia, 2006.
9. L. Atallah, B. Lo, Guang-Zhong Yang, and F. Siegemund. Wirelessly accessible sensor populations (wasp) for elderly care monitoring. In *Proceedings of the 2nd International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '08, pages 2 –7, 2008.
10. Xuan Hung Le, Sungyoung Lee, Young-Koo Lee, Heejo Lee, Murad Khalid, and Ravi Sankar. Activity-oriented access control to ubiquitous hospital information and services. *Information Sciences*, 180(16):2979 – 2990, 2010.
11. Fabrice Camous, Dónall McCann, and Mark Roantree. Capturing personal health data from wearable sensors. In *Proceedings of the 2008 International Symposium on Applications and the Internet*, pages 153–156, Washington, DC, USA, 2008. IEEE Computer Society.
12. M. Elkhodr, S. Shahrestani, and H. Cheung. Ubiquitous health monitoring systems: Addressing security concerns. *J. Comput. Sci.*, 7(10):1465–1473, 2011.
13. Shinyoung Lim, Tae Hwan Oh, Y.B. Choi, and T. Lakshman. Security issues on wireless body area network for remote healthcare monitoring. In *Proceedings of IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC)*, pages 327–332, 2010.
14. T. Znati. On the challenges and opportunities of pervasive and ubiquitous computing in health care. In *Proceedings of the 3rd IEEE International Conference on Pervasive Computing and Communications (PerCom '05)*, pages 396–396, 2005.
15. European Parliament. Directive 95/46/EC. In *Official Journal L 281*, pages 0031–0050. 24 October 1995. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML.
16. 104th U.S. Congress. Health insurance portability and accountability act. In *Public Law 104-191*. Aug. 21 1996.
17. K.C. Laudon. Markets and privacy. *Commun. ACM*, 39(9):92–104, 1996.
18. P. Samuelson. Privacy as intellectual property? *Stanford Law Review*, 52:1125, 2000.
19. Deirdre Mulligan and Ari Schwartz. Your place or mine?: privacy concerns and solutions for server and client-side storage of personal information. In *Proceedings of the tenth conference on Computers, freedom and privacy: challenging the assumptions (CFP '00)*, pages 81–84, New York, NY, USA, 2000. ACM.
20. J.I. Hong. *An Architecture for Privacy-Sensitive Ubiquitous Computing*. PhD thesis, University of California at Berkeley, Computer Science Division, Berkeley, 2005.
21. P. Jäppinen. *ME - Mobile Electronic Personality*. PhD thesis, Lappeenranta University of Technology, Finland, 2004.

22. Pavlos S. Efraimidis, Georgios Drosatos, Fotis Nalbadis, and Aimilia Tasidou. Towards privacy in personal data management. *Journal on Information Management & Computer Security*, 17(4):311–329, 2009.
23. G.V. Lioudakis, E.A. Koutsoloukas, N.L. Dellas, N. Tselikas, S. Kapellaki, G.N. Prezerakos, D.I. Kaklamani, and I.S. Venieris. A middleware architecture for privacy protection. *Comput. Networks*, 51(16):4679–4696, 2007.
24. Hsu-Hui Lee and Mark Stamp. An agent-based privacy-enhancing model. *Information Management & Computer Security*, 16(3):305–319, 2008.
25. Ian Goldberg. Privacy-enhancing technologies for the internet iii: Ten years later. In A. Acquisti, S. Gritzalis, C. Lambrinoudakis, and S. di Vimercati, editors, *Chapter 1 of Digital Privacy: Theory, Technologies, and Practices*. Auerbach, December 2007.
26. S. Gritzalis. Enhancing web privacy and anonymity in the digital era. *Information Management and Computer Security*, 12(3):255–287, 2004.
27. Andrew Chi-Chih Yao. Protocols for secure computations (extended abstract). In *Proceedings of Twenty-third IEEE Symposium on Foundations of Computer Science*, pages 160–164. Chicago, Illinois, November 1982.
28. Li Shundong, Wang Daoshun, Dai Yiqi, and Luo Ping. Symmetric cryptographic solution to yaos millionaires problem and an evaluation of secure multiparty computations. *Information Sciences*, 178(1):244 – 255, 2008.
29. Kun Peng, Colin Boyd, Ed Dawson, and Byoungcheon Lee. Ciphertext comparison, a new solution to the millionaire problem. In *Proceedings of the 7th International Conference on Information and Communications Security (ICICS 2005)*, volume 3783 of *LNCS*, pages 84–96. Springer, 2005.
30. W. Du and M. Atallah. Privacy-preserving cooperative statistical analysis. In *Proceedings of the 17th Annual Computer Security Applications Conference*, pages 102–112, Washington, DC, USA, 2001. IEEE Computer Society.
31. Wenliang Du, Shigang Chen, and Yunghsiang S. Han. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In *Proceedings of the 4th SIAM International Conference on Data Mining*, pages 222–233, 2004.
32. Murat Kantarcioglu and Onur Kardes. Privacy-preserving data mining in the malicious model. *International Journal of Information and Computer Security*, 2:353–375, January 2008.
33. Yitao Duan, NetEase Youdao, John Canny, and Justin Z. Zhan. P4P: practical large-scale privacy-preserving distributed computation robust against malicious users. In *USENIX Security Symposium*, pages 207–222, 2010.
34. Nissim Matatov, Lior Rokach, and Oded Maimon. Privacy-preserving data mining: A feature set partitioning approach. *Information Sciences*, 180(14):2696 – 2720, 2010.
35. Michael R. Genesereth and Steven P. Ketchpel. Software agents. *Commun. ACM*, 37(7):48–ff., July 1994.
36. Costin Badica, Zoran Budimac, Hans-Dieter Burkhard, and Mirjana Ivanovic. Software agents: Languages, tools, platforms. *Comput. Sci. Inf. Syst.*, 8(2):255–298, 2011.
37. Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, VLDB '05, pages 901–909. VLDB Endowment, 2005.
38. Victor Muntés-Mulero and Jordi Nin. Privacy and anonymization for very large datasets. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 2117–2118, New York, NY, USA, 2009. ACM.
39. Sheng Zhong, Zhiqiang Yang, and Tingting Chen. k-anonymous data collection. *Information Sciences*, 179(17):2948 – 2963, 2009.
40. George Drosatos and Pavlos S. Efraimidis. An efficient privacy-preserving solution for finding the nearest doctor. *Personal and Ubiquitous Computing*, 18(1):75–90, 2014.
41. Yehuda Lindell and Benny Pinkas. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality*, 1:59–98, 2009.

42. P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *Advances in Cryptology–EUROCRYPT 99*, pages 223–238. Springer Verlag LNCS 1592, 1999.
43. R. Rivest, L. Adleman, and M. Dertouzos. On data banks and privacy homomorphisms. In *Foundations of Secure Computation*, pages 169–177. Academic Press, 1978.
44. Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41st annual ACM symposium on Theory of computing (STOC '09)*, pages 169–178, New York, NY, USA, 2009. ACM.
45. Josh Benaloh. Dense probabilistic encryption. In *Proceedings of the workshop on selected areas of cryptography*, pages 120–128, 1994.
46. T. Elgamal. A public key cryptosystem and a signature scheme based on discrete logarithms. *Information Theory, IEEE Transactions on*, 31(4):469 – 472, July 1985.
47. Michael Naehrig, Kristin Lauter, and Vinod Vaikuntanathan. Can homomorphic encryption be practical? In *Proceedings of the ACM workshop on Cloud computing security workshop (CCSW '11)*, pages 113–124, New York, NY, USA, 2011. ACM.
48. A. Acquisti, S. Gritzalis, C. Lambrinoudakis, and S. De Capitani di Vimercati. *Digital privacy*. Auerbach Publications, Taylor & Francis Group, 6000 Broken Sound ParkWay NW, 2008.
49. Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28(2):270 – 299, 1984.
50. V. Ciriani, S. Capitani di Vimercati, S. Foresti, and P. Samarati. κ-anonymity. In Ting Yu and Sushil Jajodia, editors, *Secure Data Management in Decentralized Systems*, volume 33 of *Advances in Information Security*, pages 323–353. Springer US, 2007.
51. Cynthia Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54:86–95, January 2011.
52. Cynthia Dwork. Differential privacy: a survey of results. In *Proceedings of the 5th international conference on Theory and applications of models of computation*, TAMC'08, pages 1–19, Berlin, Heidelberg, 2008. Springer-Verlag.
53. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.
54. Rathindra Sarathy and Krishnamurty Muralidhar. Evaluating laplace noise addition to satisfy differential privacy for numeric data. *Transactions on Data Privacy*, 4(1):1–17, April 2011.
55. Steve Kremer, Mark Ryan, and Ben Smyth. Election verifiability in electronic voting protocols. In *ESORICS 2010*, pages 389–404, Heidelberg, 2010. Springer.
56. Nabil R. Adam and John C. Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Comput. Surv.*, 21:515–556, December 1989.
57. Ivan Damgård and Mats Jurik. A generalisation, a simplification and some applications of paillier's probabilistic public-key system. In *Proceedings of the 4th International Workshop on Practice and Theory in Public Key Cryptography: Public Key Cryptography (PKC '01)*, pages 119–136, London, UK, 2001. Springer-Verlag.
58. Bouncycastle. Legion of the bouncy castle, January 2011. `http://www.bouncycastle.org/`.

George Drosatos is a Post-Doctoral Researcher at the Athena Research and Innovation Center, branch of Xanthi (Greece). He received his diploma thesis in Electrical and Computer Engineering from Democritus University of Thrace (Greece) in 2006. In addition, he acquired a Master's degree (March 2010, advised by Prof. Alexandros Karakos) and a PhD degree (December 2013, advised by Assistant Prof. Pavlos S. Efraimidis) both from the Dept. of Electrical and Computer Engineering of the Democritus University of Thrace. His research interests are in the field of privacy and in particular privacy in ubiquitous computing.

Pavlos S. Efraimidis is an Assistant Professor at the Dept. of Electrical and Computer Engineering of the Democritus University of Thrace (Greece). He graduated from the Dept. of Computer Engineering and Informatics of the University of Patras (Greece) in 1996 and obtained a PhD in Informatics in 2000 from the University of Patras under the supervision of Paul Spirakis. His research interests are in the fields of algorithms and privacy. He is a member of ACM, IEEE and EATCS.

Received: January 20, 2013; Accepted: March 20, 2014.

Computer Science and Information Systems 11(2):549–567 DOI: 10.2298/CSIS130823024S

Iris: A Decentralized Approach to Backend Messaging Middlewares

Péter Szilágyi[1,2]

[1] Eötvös Loránd University
1053 Budapest, Hungary
[2] Babeş-Bolyai University
400084 Cluj-Napoca, Romania
peterke@gmail.com

Abstract. In this work we introduce the design and internal workings of the Iris decentralized messaging framework. Iris takes a midway approach between the two prevalent messaging middleware models: the centralized one represented by the AMQP family and the socket queuing one represented by ZeroMQ; by turning towards peer-to-peer overlays as the internal transport for message distribution and delivery. A novel concept is introduced, whereby a distributed service is composed not of individual application instances, but rather clusters of instances responsible for the same sub-service. Supporting this new model, a collection of higher level messaging patterns have been identified and successfully implemented: broadcast, request/reply, publish/subscribe and tunnel. This conceptual model and supporting primitives allow a much simpler way to specify, design and implement distributed, cloud based services. Furthermore, the proposed system achieves a significant switching speed, which – given its decentralized nature – can be scaled better than existing messaging frameworks, whilst incurring zero configuration costs.

Keywords: peer-to-peer, decentralized, message oriented middleware.

1. Introduction

A message-oriented middleware (MOM) is either a hardware or software infrastructure component, with the sole purpose of removing the complexity of communication from a distributed system, allowing individual network components to focus on their specific tasks [2].

The core concept behind MOMs is based on the observation that classical, stream-based communication introduces a very tight coupling between network components, requiring significant efforts to develop and maintain distributed systems. Instead, in MOM based systems the unit of communication is a message, *an undividable, self-contained block of data propagating from sender to recipient(s)*. As the transferred data is self contained, the network can evolve dynamically without costly protocol setups and teardowns.

The second core concept is the middleware part, whereby all distributed clients are in contact only with the MOM, but not each other. The most important implication is that the middleware is responsible for transferring application messages from origin to destination, whilst clients remain ignorant of the routing complexity. This decoupling grants the whole system both simplicity and greater flexibility in terms of scalability and environmental heterogeneity [8].

The last (optional) concept in MOMs are message queues, which provide further decoupling by allowing applications to communicate asynchronously [23], not requiring the communicating peer to be accessible at the time of messaging. However, for distributed systems targeted by this paper (i.e. dynamically scaling back-end services), persistent messaging is not a requirement.

Due to the significant popularity of cloud computing [3], messaging middlewares gained an even bigger role in distributed system infrastructures. Whilst previously internal back-end systems could have used arbitrary network topologies, with the advent of clouds, system designers are forced to think in more general – and many times, less reliable – solutions. Messaging middlewares provide the extra simplification of distributed systems to keep focus on the problem and prevent it from shifting towards cloud communication.

Even though the operational environment of cloud providers, as well as operational requirements of cloud services are drastically different from *pre-cloud* ones, the same underlying messaging models are continued to be used. We argue, that although these models are perfectly feasible, not taking advantage of cloud specifics leads to suboptimal distributed systems. Essentially, cloud platforms encourage the *load-based elastic evolution of hosted services*, where deployed applications must be ready to scale at a short notice. This is the weakness of current messaging models, as they were not designed with constant scaling in mind. To prepare for such scenarios, all non-scaling concepts need to be eliminated: location and cardinality.

This paper presents a MOM model, design and algorithms that significantly simplify the development of distributed back-end services. Firstly, a novel networking abstraction is introduced, whereby the smallest unit of composition in a distributed system is a *cluster of instances*, opposed to individual instances in previous literature. Secondly, four core communication patterns are defined, which are essential for back-end services and are also fully compatible with the *clustered* compositional model: request/reply, broadcast, tunnel and publish/subscribe. Finally, the feasibility of the model and messaging schemes are demonstrated through peer-to-peer networks, fully developing a P2P overlay supporting the required operations, and at the same time requiring zero configuration and maintenance. The P2P related challenges of decentralized bootstrapping and peer-to-peer security have already been covered in two previous papers [21,22], the present one focusing solely on decentralized routing.

The paper starts out by presenting the existing solutions, continuing with the core communication patterns and reliability guarantees identified as essential for back-end services. Afterwards, the implementing models, algorithms and optimizations are presented. Finally, the proposed solution is validated through a series of benchmarks, confirming the model's feasibility.

2. State of the art

Historically, hard requirements were placed on these messaging middlewares: the guarantee of data security and integrity in any operational environment; the guarantee that no messages are lost in the face of any network, software or hardware failure; yet to still achieve a significant switching throughput.

The most prevalent technology to have met these criteria is the Advanced Message Queuing Protocol (AMQP) [24,14], with RabbitMQ[3] being one of the leading implementations of the specs. However, the reliability guarantees AMQP undertook blew up the protocol complexity enormously, leading to centralized solutions that are hard to scale [10].

iMatix took a new approach to messaging – diverging from AMQP – by removing the concept of message brokers and instead, placed the message queues directly into the client processes with their ZeroMQ library[4], coining the term *"sockets on steroids"*, arguing that all messaging should happen at the endpoints [11,19]. The main issue with iMatix's approach is that they reintroduced the tight coupling that MOMs originally set out to remove, and although ZMQ provides higher level communication primitives, these are bound to socket level, forcing the user of the library to define and implement the needed network topology. This works well in static environments, but with the prevalence of clouds (massive distributions and massive failures), the administrative overhead of a custom, user-managed topology becomes a significant cost [18].

A different initiative is the Data Distribution Service for Real-Time Systems (DDS) a standardized MOM specification [15,17], with RTI Connext DDS[5] being one of the most advanced implementation of it [12]. As its name suggests, the primary goal of DDS is modeling complex data flows through which data can be disseminated to interested parties in a large network of nodes. This is achieved using a publish/subscribe model, where the physicality of the network is hidden from participating nodes and all addressing is done through so called topics. Although very powerful, the publish/subscribe abstraction models only information flows, but does not cater for other messaging patters – such as request/reply or load balancing – which are essential for implementing back-end services.

A final MOM needed mentioning is the eXtensible Messaging and Presence Protocol (XMPP), an IETF standardized protocol[6] focusing on data exchanges relating to instant messaging, presence management and social collaboration. It can be considered a generalized routing protocol for XML data. However, its goals are very different from the back-end service ones targeted here.

3. Proposed abstractions

To retain as much flexibility as possible, our messaging model assumes only a bare-bone cloud environment. Such a cloud service-model is commonly called *Infrastructure as a Service* (IaaS), and provides the consumer with the capability to provision fundamental computing resources (e.g. processing, storage, network, etc.) and run arbitrary software on them [13].

3.1. Distribution model and unit of composition

Originally, back-end services used vertical distribution: the logically distinguishable parts of a complex service were split into individual component applications, each hosted either together, or separately based on their resource consumption.

[3] `http://www.rabbitmq.com/`

[4] `http://zeromq.org/`

[5] `http://www.rti.com/products/dds/index.html`

[6] `http://xmpp.org/xmpp-protocols/rfcs/`

This solution has however quickly shown its weaknesses to failures, as well as its limitations under high load. System designers hence extended the vertical model with the concept of horizontal distribution: multiple instances of the same service components are run simultaneously on different machines with tasks distributed between them according to some load metric.

Whilst theoretically sound, the extended model incurs significant administrative costs, since each entity must be able to contact its own servicing components and load-balance between them. The addition or removal of instances further complicates overall component logic and maintenance.

We introduce the high level abstractions required to retain the simplicity of the vertical distribution, whilst obtaining the flexibility of the horizontal distribution. This is done by turning towards peer-to-peer networks as the base communication model (Fig. 1).

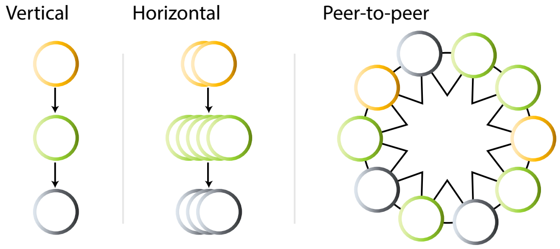


Fig. 1. The vertical model is the theoretical target, whilst the horizontal one is the operational goal. The paper presents a solution based on the last model, achieving the best of both worlds.

State of the art distribution models consider an application *instance* (e.g. web-server, database, etc) the smallest unit. We argue that this is the most significant design flaw responsible for the increased complexity of distributed systems: distributed communication is too complex to keep track of individual instances.

We propose a novel abstraction, where the *cluster* is at the lowest level: a *group of applications responsible for the same sub-task* of the service (e.g. collection of databases). At any point in time there can be any number of instances belonging to the same cluster (including zero); it is all the same from the consumer's point of view: a reply either arrives or it does not.

These clusters can then be assembled into a *service*, where any component can query another cluster for some sub-service, without having to know neither where, nor how many instances can handle its requests. Such a *service* is considered the unit of security [22].

Finally, multiple such services can be composed into a *federation* or a *cloud*, crossing security boundaries and allowing inter-service communication (shortly expanded in the future works section).

3.2. Communication patterns

In order to support the proposed distribution and composition model, four communication patters have been identified: *request/reply*, *broadcast*, *tunnel* and *publish/subscribe* (Fig. 2). These are refinements over author's previously identified ones in [20].

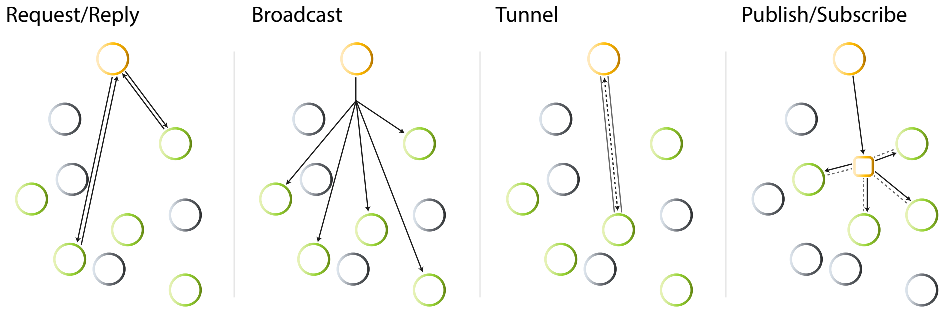


Fig. 2. Core messaging patterns to support the model. The first three accomplish standard back-end communication, whilst the last one enables more specialized operations.

The first pattern is the *request/reply*, a natural communication pattern for the target use-case of interconnected services: a consumer sends a request to a sub-service, which responds with a reply. The proposed model adds a twist, which is a direct consequence of the *cluster* being the smallest unit of composition: whenever a request is made, the target is not a single entity, but rather the cluster of entities serving the same purpose. It is the responsibility of a messaging middleware implementing this model to make sure that an appropriate node gets the request, taking load balancing into consideration too. The consumer should be aware of neither the serving cluster's size, nor individual component's whereabouts, but the name/type of the cluster alone (e.g. *'databases'*) must be enough to process the request.

The second pattern is the *broadcast*, a supporting scheme besides the *request/reply*. The difference between the two is, that whilst a request is delivered to a single member of the cluster, a broadcast is forwarded to all of the participants. This is the only way to contact every instance within a group. But since there is no concept of member count, there is no possibility for individual responses, as the recipient would never know how many replies to wait for. Hence broadcasting is an asynchronous, one-way operation.

The third pattern is the *tunnel*, which itself is another supporting scheme of the cluster communication. Its goal is to solve the challenge of stream communication and/or stateful protocols, where an operation consists of multiple data exchanges (e.g. a database transaction). A *tunnel* establishes a communication stream between a client and a member of a cluster, with ordered and throttled message delivery, persisting until either side closes the connection.

The messaging patters enumerated above solve most of the communication requirements of a back-end service. The last scheme, the *publish/subscribe*, is a very specialized one, as it allows breaking cluster boundaries. The underlying concept is well known in the literature [5]. Any instance within the network may subscribe to a *topic* (any number of them), effectively forming temporary clusters. Any node in the network can then publish events to these clusters the same way as *broadcast* does.

3.3. Reliability guarantees

The reliability guarantees of a distributed messaging model are *the assurances it can always satisfy about message delivery in the presence of software, system or network failures*. From this point of view, the proposed model takes a significantly more modest standpoint compared to previous messaging models (e.g. AMQP family). Most of such requirements cannot be handled at the messaging level, and trials usually result in complexity explosions [10]. This paper instead focuses on the robustness of the messaging layer, providing only those guarantees that are essential and natural to the respective level of abstraction.

The *request/reply* pattern has two possible points of failure. The first scenario of them is that the request is lost during transit or that there are no available instances to respond. This is impossible to circumvent in a loosely coupled environment, since any hardware failure will result in message loss. The second scenario is when the request is successfully serviced, but the reply is the one lost. The reason this problem is impossible to solve at the messaging layer is twofold. Per the Two Generals' Paradox [1], both of the endpoints of a request/reply exchange can never be sure of the operation's success, hence a reply cannot be cached for resending. Secondly, without the guarantee of idempotence, neither can the request be cached for automatic resending. This means, that transit failures can be detected at the messaging layer (through timeout mechanisms for example), but their handling must be delegated up to the application layer.

The *broadcast* and *publish/subscribe* schemes are sensitive to network partitioning, where some of the addressed instances receive a message, but not all of them (possibly even none). Although unfortunate, this issue persists in any distributed environment, and can be solved solely using a central tracker. We argue, that reliance on such guarantees is a fundamental design flaw in most distributed systems, and as such, the proposed communication model makes no attempt to provide it at the messaging layer. Of course, a high-level client application can easily implement such behavior on top of the core schemes with the exact guarantees needed (bearing the necessary sacrifices).

Finally, opposed to the previous patterns, the *tunnel* is a high reliability messaging primitive. It provides the very same guarantees that a transport layer TCP connection does: as long as the connection is alive, the messages are guaranteed to be delivered eventually and in the same order as sent. Besides ordering, the data flow is also throttled to prevent overloading the recipient.

4. Overlay network

The Iris overlay adopts a fully decentralized peer-to-peer model, where – given the one messaging node per host constraint – each participant has equal responsibilities. To achieve the desired messaging patterns, it builds application *clusters* based on the multi-cast trees formed by Scribe [7], which in turn are based on the routing paths of Pastry [16].

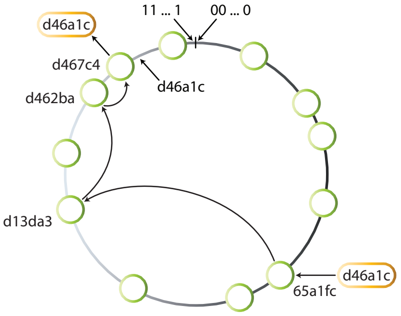
4.1. Foundations: Pastry and Scribe

Pastry is a generic, efficient, scalable, fault resilient, and self-organizing peer-to-peer substrate [6], forming a highly robust distributed hash table (DHT). The goal of the Pastry DHT is that, given a message and an associated routing key, to deliver the message to that specific network participant, whose node identifier is numerically closest to the key itself.

To achieve this, each participating node is assigned a unique, uniform random node id, taken from a circular, 128-bit identifier space. Each node is then considered responsible for the slice of the id space circularly the closest to it. Furthermore, each node maintains a local routing state: Leaf-set (numerically closest peers), Routing table (peers with matching prefixes of $1..N$ digits) and Neighborhood (physically closest peers).

Given the above routing state information, at each routing step a node checks the common prefix (length l) of the message's routing key with its own node id, and then selects a node from its routing table which shares at least $l + 1$ digits with the message. If such a node is not present within the routing table, the leaf-set is consulted and the numerically closest peer receives the message. If according to the leaf-set, the current node is the closest, the message is delivered upstream for higher level processing (Fig. 3).

Of course, Pastry achieves self-organization by dynamically maintaining the routing state within each node, these synchronizing between each other whenever a churn event is detected (a node joins or leaves). For such fine grained details about the protocol we refer the reader to the original paper [16].



```
func route(dest, message) {
  l := prefix(self, dest)
  d := digit(dest, l + 1)
  if peer := routes[l][d]; peer != nil {
    send(peer, dest, message)
  } else if leaves[0] < dest < leaves[last] {
    peer := closest(leaves, dest)
    if peer != self {
      send(peer, dest, message)
    } else {
      deliver(dest, message)
    }
  }
}
```

Fig. 3. Pastry routing algorithm and example. A message with the routing key *d46a1c* is forwarded from node *65a1fc* towards its destination. The message is prefix routed, decreasing the distance from its destination exponentially in each step. At the very last step no longer matching prefix is found, so the leaf-set is consulted and the message delivered to its final destination.

Pastry was originally conceived for internet-scale networks. However, in our proposed messaging scenario we assume that a service is running in its entirety within one data center[7], due to which two modifications were made. Firstly, the concepts of proximity and associated neighborhood set were removed, since within a single cloud the latency/bandwidth differences between links are not significant enough to warrant the complexity. Secondly, with the reduced node count of back-end services – in the range of hundreds – an ID space of 128-bits leaves the routing tables mostly empty. It should thus

[7] It will be the responsibility of a *federation* to span data centers.

be reduced as much as possible, while keeping the probability of a random ID collision insignificant. According to the birthday paradox [9], a 40-bit ID space S for 10K nodes N results in a collision probability $p(N, S)$ of 4.5483×10^{-5}.

$$p(N, S) = 1 - e^{-N^2/(2S)}$$

Scribe is a decentralized, scalable application-level multicast infrastructure, built on top of the Pastry overlay [7]. The goal of Scribe is to allow nodes to create *groups*, which can then be joined/subscribed by any peer within the network. Messages sent to these groups are delivered to all members.

To achieve this, whenever a node wishes to create a new group, the name of the group is mapped to a Pastry identifier, after which Scribe uses Pastry to route a *create* message to the node responsible for that specific ID, which implicitly assigns the group to the specific node. This node will assume responsibility for it, becoming the group's *rendez-vous point* (Fig. 4a).

After group creation, any participant of the Scribe network can become a member of the group by routing a *join* message towards the group's *rendez-vous point*. Opposed to the *create* message, the *join* is not forwarded blindly between nodes, but instead each node along the path to the group owner maintains a local subscription list, and if the current node is already part of the group, then the new node will be attached to it. If the traversed node is not a member of the group, it will create a new local subscription, terminate the arrived join, and initiate a new join itself. This procedure will create a tree structure for every group, rooted at the *rendez-vous point* (Fig. 4b).

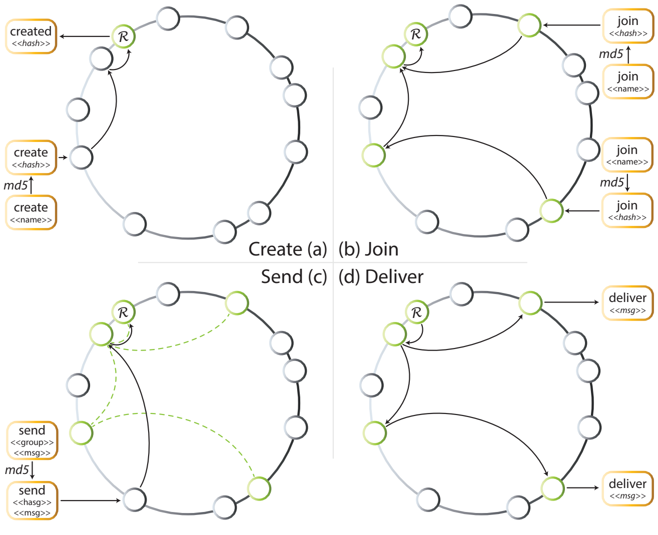


Fig. 4. Scribe group operations.

Given the freshly built subscription tree, whenever a message is to be disseminated to the group members, it is first routed to the group's *rendez-vous point*, which will initiate a distribution along the edges of the tree, from parent to child, each internal node of the tree in its turn also forwarding the message to its children (Fig. 4c & 4d).

Since Scribe was conceived for public peer-to-peer networks, it contains an underlying credential system to limit group membership and communication to authorized entities only. The proposed system uses a different authentication mechanism [22], whereby a successful connection already pertains the granting of full privileges. Removing the credential sub-system makes group creation redundant, since group joining will already entail all the necessary steps. Due to the same reason, during content distribution, the messages do not necessarily have to reach the root node before dissemination can begin: any subscriber receiving such a message can immediately begin distributing to both children and parent, saving valuable network hops.

4.2. Iris overlay

Iris is a decentralized, application-scope, group communication platform, built on top of modified Scribe and Pastry protocols. The goal of Iris is to implement a batch of high level messaging primitives to simplify the distributed communication among nodes in a back-end service.

Cluster formation and teardown As defined in section 3.1, the *cluster* is the smallest unit of composition in the Iris platform. Whenever an application wishes to communicate through Iris, it must first become a member of a cluster. Hence the first step in the life of an Iris application is joining – possibly by creating – such a cluster.

Cluster formation and membership management are based on the Scribe multi-cast groups. Whenever an application requests to join a cluster with a given name, Iris maps that cluster name to a Scribe group name by prepending a tag to it. The goal of the tag is to differentiate between other types of Scribe groups. After tagging the name, a Scribe *join* is initiated, turning the node a member of the requested group, implicitly creating the group if non-existent.

When the application finishes using the Iris network, it issues a close request, leaving the joined cluster. After the same cluster to group mapping, a Scribe *leave* is initiated, removing the application from the group subscription tree. If a node's local subscriptions all terminate, the node itself leaves the group, cascading until either a live member remains or the whole group is torn down.

Messaging primitives The simplest communication pattern is the cluster *broadcast*, specifically because Scribe already provides it out of the box. When an application broadcasts a message to a cluster, Iris first resolves the Scribe group name of the cluster, after which it delegates message distribution to the Scribe protocol according to its multi-cast subscription tree (Fig. 4c & 4d).

Request/replies are a little more tricky. The core concept is the same as with *broadcast*: Iris resolves the Scribe group name and issues a Scribe delivery. But opposed to the previous scenario, the message must be delivered to a single recipient. To achieve this, the message is routed towards the group rendez-vous point, but after entering the multicast tree it is not simply distributed to all sub-trees, but rather at every group member, a decision is made. Each member maintains the load balancing state of itself and all local subtrees. An outstanding request is either delivered to a locally available cluster member, or routed on a single path to a subtree (details at the end of the section). When the request is finally delivered to a node and processed, the reply is routed back to the sender using the Pastry DHT directly.

The *tunnel* primitive is assembled through a combination of the previous *request/reply* pattern and direct usage of the Pastry network. When a tunneling request is made to a member of a cluster, that request is delivered the same way as a normal request. However, upon arrival, the remote side replies with its own Pastry address. At this point both parties have obtained each other's Pastry identifier and can communicate directly through the DHT. The tunnel is assembled using the same concepts as a TCP stream: each side is assigned a tunnel id (incremental throughout the lifetime of the node) and each packet assigned a sequence id. With these two information, any packet can be reliably delivered to the correct tunnel endpoint in the correct order. To prevent overloading a peer, each tunnel has a throttling mechanism, whereby each packet requires an acknowledgment from the remote side, allowing only a limited number of un-acked packages to be sent.

The last pattern, the *publish/subscribe*, is based on the same exact principle as the *broadcast*. The difference is that topic names are mapped to a separate set of Scribe groups than clusters by using a different prepended tag. This ensures that cluster and topic names will never collide with each other. Additionally, opposed to clusters, applications have to manually subscribe to topics, but can be subscribed to an unlimited number of then.

Split clustering Although the Scribe multi-cast trees supporting the Iris clusters are decentralized, since all inbound messages target the root of the tree for distribution, this root node can become a significant bottleneck.

Iris circumvents this issue by introducing *split clustering*. Instead of relying on a single Scribe multi-cast tree, every cluster and topic internally uses a number of Scribe groups simultaneously (simply adding a second tag to the name). Whenever a message enters the system, a sub-cluster is chosen using round-robin for delivery, ensuring that consecutive messages take different paths (Fig. 5).

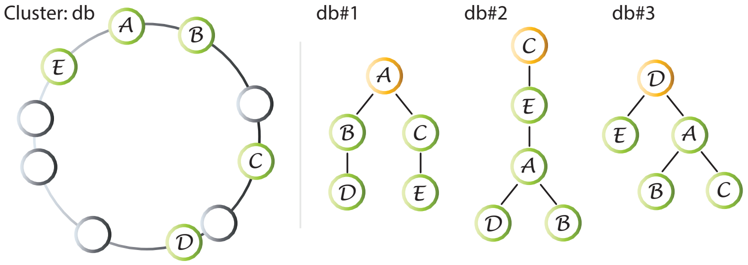


Fig. 5. A single Iris cluster split into multiple Scribe groups. On the left, the Pastry DHT can be seen with the green circles being part of an Iris group. On the right, these same few nodes can form a number of different Scribe multi-cast trees, each time with a different group rendez-vous point. Note, additional intermediary nodes may be part of these trees, but were not displayed to prevent clutter.

Load balancing As mentioned in the previous section, using the *request/reply* scheme, the framework is expected to load balance messages between all possible destinations that could handle them. In order to see how this might be accomplished, it is important to observe the internal state of the system, more specifically, that of an Iris cluster.

An Iris cluster – implemented by a Scribe group – is a rooted multi-cast tree, where each node within the tree has knowledge only about its direct children and parent. This conceptual separation, that some neighboring nodes are children and one is the parent, is essential for maintaining the tree structure, but irrelevant otherwise. Thus, from a load balancing standpoint, each node can be considered the root of a tree, where each child is a subtree of the original cluster (Fig. 6). This abstraction permits a uniform balancing algorithm for all nodes, without needing specialized logic for different parts of a tree.

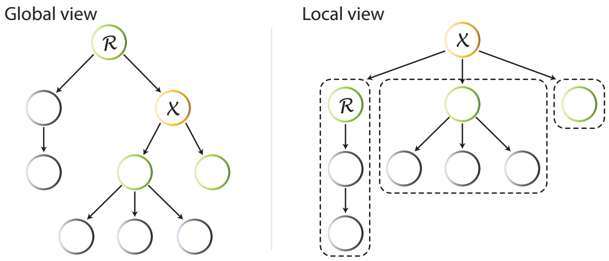


Fig. 6. Different views of the cluster. The global view is how the multi-cast tree really looks like in the network, but the local view is how an internal node sees the tree from a balancing standpoint.

With the above abstraction, the load balancing mechanism boils down to a greedy algorithm, where each node, upon receiving a request, decides whether to deliver it to a locally attached client application, or route it down to one of its local subtrees (except the one from which the message originates).

The logic which decides on the exact subtree to forward a request to is based on the *dynamic weighed round-robin* algorithm. The core concept of the round-robin algorithm is to issue every request to a different entity; but since these entities – subtrees in the current scenario – are wildly heterogeneous, specific weights are assigned to each, ensuring that the number of requests are proportional to the total available computing capacity of the subtree. Additionally, the capacities of each subtree can vary over time due to new nodes arriving, old ones leaving, or simply by experiencing an unrelated load on certain members. To cater for these dynamic scenarios, the nodes periodically exchange capacity information between neighbors, ensuring that each node has a local estimate of the capabilities of each of its neighboring subtrees.

Although this mechanism is very robust, it suffers from two weaknesses. Firstly, requests are passed on each link in both directions due to the greedy decision making. But each request matched by another – flowing in the opposite direction – could have been avoided in the first place. A theoretical solution would be a predictive algorithm, but in the presence of sporadic bursts, such an algorithm causes massive congestions. Furthermore, the effect of this weakness is only visible if the transfer of the data is more expensive that it's processing.

The other weakness is in the capacity estimation, which at the moment is based on the number of requests processed at the present load since the last periodic capacity exchange. If the time required to process a request comes close to or goes over this cycle, the load balancer will become random. Still, given the use case of back-end services, this should occur rarely and only during serious system overloads.

5. Preliminary evaluation

To analyze the performance of the suggested Iris overlay, a prototype[8] was implemented in the Go programming language and benchmarks executed in two different environments:

- For computational performance measurements, a single machine with dual Intel Xeon E5-2687W processors and 64GB of memory was used, with all messaging passing running through localhost.
- For scalability measurement, a small physical cluster of 12 machines, each containing two dual code AMD Opteron 275 processors and 4GB system memory was used, with all messaging passing through a 1 Gigabit network.

5.1. Request/reply performance and scalability

The computational broadcast benchmarks were run by starting 4 Iris nodes on the Xeon machine, with one client application attached to each. The clients formed two clusters, two in each. Furthermore, each client was a simple echo service (i.e. responds with the request itself).

To measure the performance, each cluster started issuing requests of a given size to the other, loading the system until saturation. At that point the load was kept up for 30 seconds and the throughput measured. Afterwards the system stepped to the next message size and repeated the procedure. The results were plotted on Fig. 7.

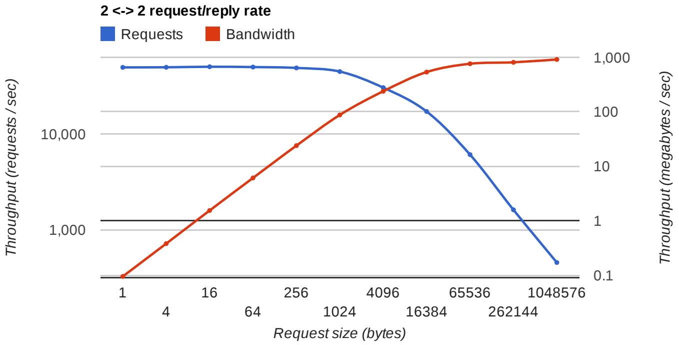


Fig. 7. Request/reply performance evaluation. Blue designates the requests served by the whole system while the red the payload bandwidth.

[8] Publicly available at `http://iris.karalabe.com`

As seen from the above chart, the system reaches a remarkable switching capacity, capping at around 50 thousand requests per second (100 thousand messages counting the replies). Looking at the bandwidth, the useful data throughput caps at over 900MB/s. It is important to emphasize that this performance is achieved with 128-bit AES encryption included between nodes [22].

For the scalability benchmark the same echo service model was used, but each machine ran a single client application and one Iris node. This benchmark was ran for systems of gradually increasing sizes from 2 to 12 hosts, and the same metrics collected.

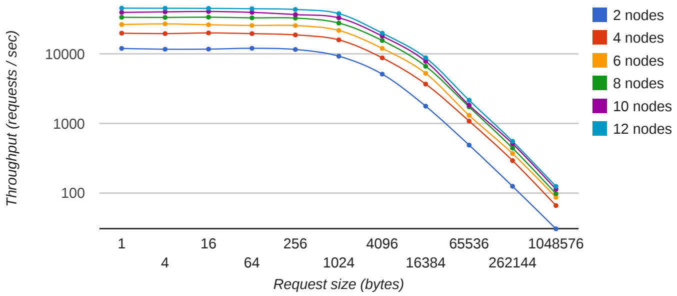


Fig. 8. Globally served requests.

The scalability results look very promising, with additional nodes linearly increasing the overall system throughput (Fig. 8): two nodes have capped at around 12 thousand requests per second and each additional pair increased this by 6K, reaching 46 thousand requests (92 thousand messages) for the full 12 nodes.

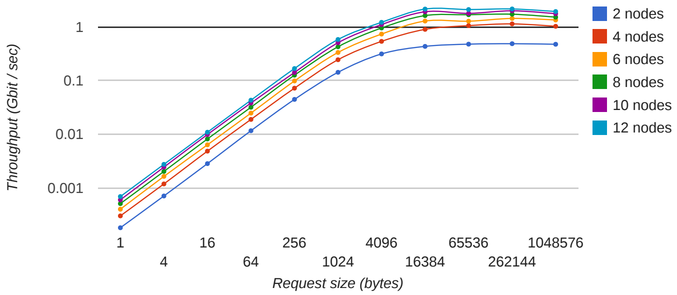


Fig. 9. Globally useful bandwidth (request + reply).

Looking at the data-rate chart (Fig. 9), the most interesting part is the bandwidth cap, which for 12 nodes is at around 2.1Gbits/sec. Since each node had a 1 Gigabit network connection, the theoretical maximum throughput is 6Gbits/sec. This means that – ignoring system messages and headers – each data packet traversed an average 2.85 network hops.

This hop count is not caused by the overlay network, but rather the inferior load balancer currently implemented.

5.2. Broadcast performance and scalability

To benchmark the broadcast operation, the same environmental setup was used as in the request tests: a number of clients form the Iris network, half joining one cluster and the other half another. The clients then started broadcasting all of the members of the opposite group.

As previously, the performance benchmarks were ran on the Xeon machine and the scalability benchmarks on the AMD cluster, plotting the delivered message and payload data-rates on Fig. 10 and Figs. 11, 12 respectively.

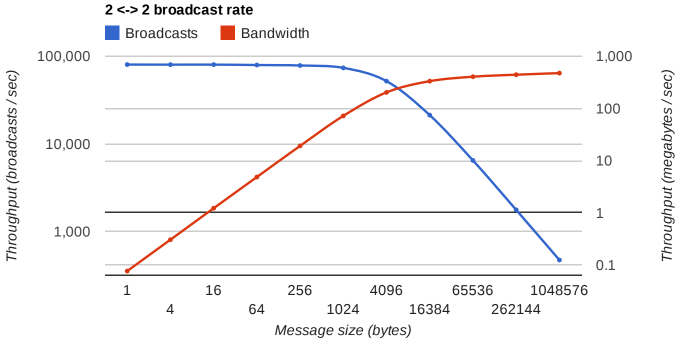


Fig. 10. Broadcast performance evaluation. Blue designates the delivered broadcasts by the whole system while the red the payload bandwidth. Note, these are twice as many as initiated since each has two recipients.

An important observation is a regression in both maximal request throughput as well as bandwidth cap. The system was able to deliver 80 thousand messages only (20% lower than using requests), and has also peaked at approx. 475MB/s throughput (down from 900MB/s).

We consider two explanations probable: since the network is very small – consisting of only 4 nodes – requests almost always get delivered to the right place, but broadcasts need to traverse the Scribe multi-cast tree. One additional hop per broadcast would be enough to halve the maximal data-rate. Another possible explanations is within the implementation details of the broadcast, namely that the handling of broadcast messages has higher memory costs both size and operation wise. Since the single machine is saturated already, extra memory allocations and copies could have an adverse effect on performance.

Looking at the scalability charts (Fig. 11), the most pleasant thing to notice, is that the total message switching capacity of the system is the same as in the request/reply benchmarks. This leads to the presumption that the above noticed regression may be a benchmark anomaly and not a real issue. Still, further experiments are needed to confirm one or the other.

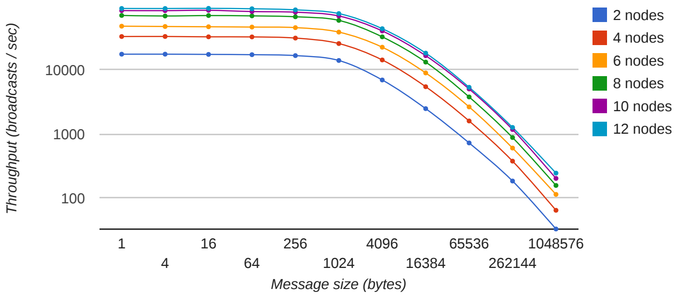


Fig. 11. Gl

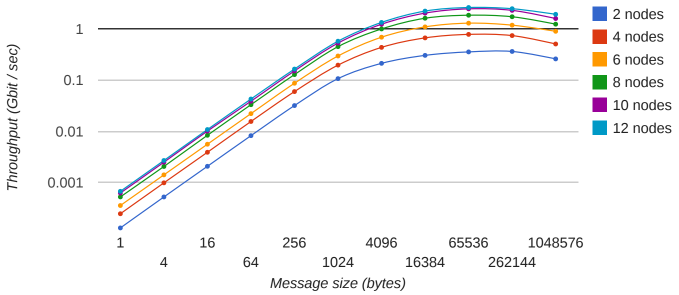


Fig. 12. Globally useful bandwidth.

On the data-rate side (Fig. 12) again we have positive results: not only did the anomaly noticed earlier disappear, but the system reached an overall data throughput of over 2.6 Gbits/s, reducing the average hop count to 2.3 per message.

At closer look, there is a small inflection on the data-rate chart when broadcasting large messages. This is deemed to be caused by a limitation of the benchmark and not the system itself and is only seen due to a combination of the AMD machines reaching their processing limits and the network being saturated simultaneously. The net effect is pulsating broadcasts, alternating between overload and starvation.

5.3. Tunnel performance and scalability

The tunnel benchmarks use the same environmental setup: a number of clients form the Iris network, out of which half join one cluster and the other half the other. Each node will then establish a tunnel into the opposite cluster and will stream messages.

It must be noted however, that this benchmark is not accurate. The problem is that these tunnels will not be distributed evenly between the nodes forming the two clusters, and because of that, overloaded nodes will produce significant delays. This could be avoided with more sophisticated benchmarks, but that would require a production quality overlay implementation.

As previously, the performance benchmarks were executed on the Xeon machine (Fig. 13). Skipping the wobbly effect cause by the aforementioned anomaly, we can observe a staggering throughput regression compared to previous messaging patterns. The highest rate achieved was 36–40 thousand messages per second, half of previous speeds.

The cause however lies within the implementation of the tunnel: to achieve ordered and throttled delivery, all messages need to be acked, which in effect matches each data packet with a system packet. The result is, that for 40K data message, the system passes 80K messages in total. This means, that for small messages, a tunnel is limited by the system noise.

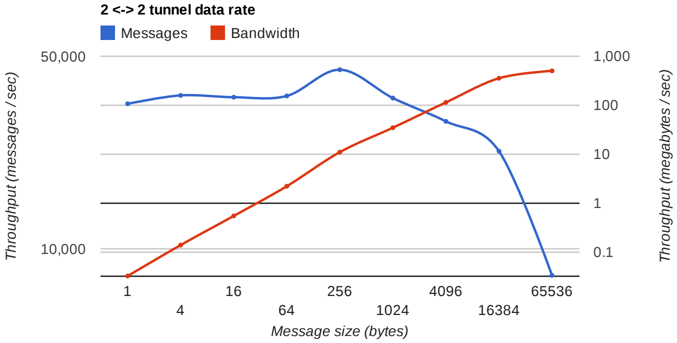


Fig. 13. Tunnel performance evaluation. The blue line is the total number of messages transferred through tunnels whilst the red is the useful payload bandwidth.

Looking at larger messages, we can see this issue alleviated a bit, with only 30% performance loss compared to the asynchronous message patterns, capping at around 500MB/s compared to the requests' 762MB/s.

The message throughput plot on the scalability charts (Fig. 14) presents both good and bad news. The good is, that – as expected – the tunnel primitive too scales linearly with the number of added nodes. The bad however, that the benchmark anomaly is getting worse with each added node.

Plotting the scalability data throughput (Fig. 15) tells a similar story to the performance benchmarks, that tunnels suffer a 36% performance loss compared to the *requests*, with the total useful bandwidth reaching 1.6Gbits/sec for 12 nodes compared to 2.2Gbits/sec.

The moral of the tunnel benchmarks is that the current *ack*-ing implementation has a serious hit on total system performance. The proposed exploration points are prioritized system messages, the splitting of larger messages into smaller manageable chunks to prevent clogging up network links and even a completely new tunnel implementation based on Iris requests and lower level direct TCP links.

The final scheme, the *publish/subscribe* has the exact same implementation internally as the broadcasts (from the message passing point of view) and hence are expected to perform identically. Due to space limitations, these have not been separately plotted and analyzed.

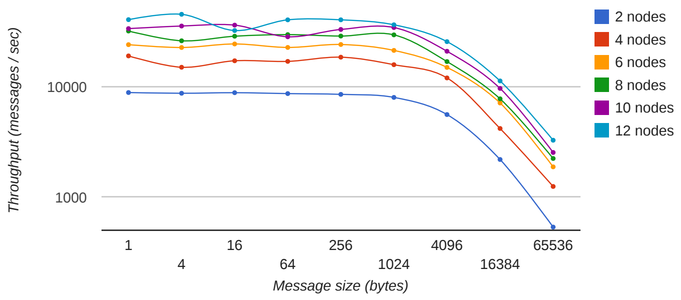


Fig. 14. Gl

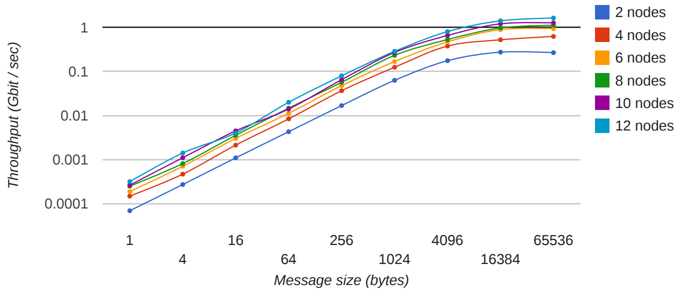


Fig. 15. Globally useful bandwidth.

6. Conclusions

This paper presented the design of the Iris decentralized group communication infrastructure based on the Scribe multi-cast protocol and the Pastry distributed hash table. The proposed overlay introduced a novel concept where *clusters* of nodes are considered the unit of composition, supporting four communication primitives: request/reply, broadcast, tunnel and publish/subscribe. These, combined with the framework's zero configuration nature, result in a significantly simplified approach to implementing decentralized back-end services.

To support the design, a preliminary evaluation was executed using a prototype implementation. Through the obtained results we conclude, that the Iris decentralization model looks very promising, reaching a simplicity yet scalability beyond previous messaging attempts. A further case study has also been carried out, assembling a decentralized ray-tracer using the Iris system. Due to space limitations, this experiment can be found as supplementary material on the Iris project's website[9].

The model presented in the paper supports a single decentralized, back-end service. It is important to emphasize that these distributed systems can be composed to create an even larger ecosystem of intercommunicating services. Depending on the desired goal, these

[9] `http://iris.karalabe.com/papers`

ecosystems can be achieved using two models. Tightly coupled services could be federated across data centers for availability and location considerations using direct links between services and having gateway nodes relay messages between them. Or more loosely coupled ones could be assembled into a cloud of services using a super-peer architecture [4], creating a Platform-as-a-Service middleware. Each of these models, however, poses significant new challenges and thus are the next steps in the Iris research project.

Acknowledgments. The research was partially carried out as part of the EITKIC_12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group.
This research is partially supported from the project POSDRU/88/ 1.5/S/60185 Innovative doctoral studies in a Knowledge Based Society PhD scholarship, Project co-financed by the SECTORAL OPERATIONAL PROGRAM FOR HUMAN RESOURCES DEVELOPMENT 2007 - 2013, Babeş-Bolyai University, Cluj-Napoca, Romania.

References

1. Akkoyunlu, E., Ekanadham, K., Huber, R.: Some constraints and tradeoffs in the design of network communications. ACM SIGOPS Operating ... pp. 67–74 (1975), http://dl.acm.org/citation.cfm?id=806523
2. Banavar, G., Chandra, T., Strom, R., Sturman, D.: A case for message oriented middleware. Lecture Notes in Computer Science, Distributed Computing 1693, 1–17 (1999), http://link.springer.com/chapter/10.1007/3-540-48169-9_1
3. Berman, S., Kesterson-Townes, L., Marshall, A., Srivathsa, R.: The power of cloud: Driving business model innovation. Tech. rep., IBM Global Business Services (2011), http://www.ibm.com/cloud-computing/us/en/assets/power-of-cloud-for-bus-model-innovation.pdf
4. Beverly Yang, B., Garcia-Molina, H.: Designing a super-peer network. In: Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405). pp. 49–60. IEEE (2003), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1260781
5. Birman, K.P., Joseph, T.A.: Exploiting virtual synchrony in distributed systems. In: 11th ACM Symposium on Operating systems principles. pp. 123–138. ACM New York, NY, USA (1987), http://dl.acm.org/citation.cfm?id=37515
6. Castro, M., Druschel, P., Hu, Y.C., Rowstron, A.: Exploiting network proximity in peer-to-peer overlay networks. Tech. Rep. MSR-TR-2002-82, Microsoft Research (2002), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.7902&rep=rep1&type=pdf
7. Castro, M., Druschel, P., Kermarrec, A.M., Rowstron, A.: SCRIBE: A large-scale and decentralized application-level multicast infrastructure. Selected Areas in Communications 20(8), 100–110 (2002), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1038579
8. Curry, E.: Message-oriented middleware. In: Middleware for communications, chap. 1. John Wiley & Sons, Ltd, Chichester, UK (2005)
9. Girault, M., Cohen, R., Campana, M.: A generalized birthday attack. Advances in Cryptology - Eurocrypt '88 pp. 129–156 (1988), http://link.springer.com/chapter/10.1007/3-540-45961-8_12
10. Hintjens, P.: What is wrong with AMQP (and how to fix it). Tech. rep. (2008), http://www.imatix.com/articles:whats-wrong-with-amqp

11. IMatix Corporation: Multithreaded Magic with ØMQ. Tech. rep. (2010), http://zeromq.wdfiles.com/local--files/whitepapers%3Amultithreading-magic/imatix-multithreaded-magic.pdf
12. Innovations, R.T.: RTI Connext (2012)
13. Mell, P., Grance, T.: The NIST Definition of Cloud Computing, Recommendations of the National Institute of Standards and Technolog. National Institute of Standards and Technology (2011), http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf
14. Oasis: Advanced Message Queuing Protocol Specification - Version 1.0. Tech. Rep. October (2012), http://docs.oasis-open.org/amqp/core/v1.0/amqp-core-complete-v1.0.pdf
15. OMG: Data distribution service for real-time systems - Version 1.2. Tech. Rep. January (2007), http://kurser.iha.dk/ee-ict-master/timico/slides/2012_Fischer_DDS_Slides.pdf
16. Rowstron, A., Druschel, P.: Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. Middleware 2001 (November 2001) (2001), http://link.springer.com/chapter/10.1007/3-540-45518-3_18
17. Schneider, S., Farabaugh, B.: Is DDS for You? A Whitepaper by Real-Time Innovations (April) (2009), http://omg.org/news/whitepapers/IsDDS4U.pdf
18. Sústrik, M.: Broker vs. Brokerless. Tech. rep., iMatix (2008), http://zeromq.org/whitepapers:brokerless
19. Sústrik, M.: ØMQ: The Theoretical Foundation (2011), http://250bpm.com/concepts
20. Szilágyi, P.: ErlHop: Erlang Hosting Platform. Achieving Dynamic Scalability and Load Balancing on a Cloud Architecture. Masters, Babes-Bolyai University (2010), https://dl.dropboxusercontent.com/u/10435909/Documents/Theses/MSc-ErlangHostingPlatform.pdf
21. Szilágyi, P.: Decentralized bootstrapping in clouds. In: 10th Jubilee International Symposium on Intelligent Systems and Informatics. pp. 277–281. IEEE, Subotica (Sep 2012), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6339529
22. Szilágyi, P.: Securing communication in a peer-to-peer messaging middleware. In: 15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing. IEEE, Timisoara, Romania (2013)
23. Tanenbaum, A.S., van Steen, M.: Distributed Systems: Principles and Paradigms (2nd Edition). Prentice Hall, 2 edn. (2006)
24. Vinoski, S.: Advanced message queuing protocol. Internet Computing, IEEE (December), 87–89 (2006), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4012603

Péter Szilágyi is currently finalizing his PhD studies at the Eötvös Loránd University in Budapest and Babeş-Bolyai University in Cluj, while also being enrolled as a business student of the European Institute for Innovation and Technology. At the moment he's focused on simplifying the development of backend services through the Iris project, and the launch of a public demo service on top of it, called RegionRank.

Received: August 23, 2013; Accepted: January 10, 2014.

Computer Science and Information Systems 11(2):569–599 DOI:10.2298/CSIS130523034K

A Direct Approach to Physical Data Vault Design

Dragoljub Krneta[1], Vladan Jovanović[2], and
Zoran Marjanović[3]

[1] Lanaco Information Technologies, Knjaza Milosa 15,
78000 Banja Luka, Bosnia and Herzegovina
dragoljub.krneta@lanaco.com
[2] Georgia Southern University, 1500 IT Drive,
Statesboro, Georgia 30458, USA
vladan@georgiasouthern.edu
[3] Faculty of Organizational Sciences, Jove Ilica 154,
11000 Belgrade, Serbia
zoran.marjanovic@fon.rs

Abstract. The paper presents a novel agile approach to large scale design of enterprise data warehouses based on a Data Vault model. An original, simple and direct algorithm is defined for the incremental design of physical Data Vault type enterprise data warehouses, using source data meta-model and rules, and used in developing a prototype case tool for Data Vault design. This approach solves primary requirements for a system of record, that is, preservation of all source information, and fully addresses flexibility and scalability expectations. Our approach benefits from Data Vault dependencies minimizations and rapid loads opportunities enabling greatly simplified ETL transformations in a way not possible with traditional (i.e. non data vault based) data warehouse designs. The approach is illustrated using a realistic example from the healthcare domain.

Keywords: Enterprise data warehouse, system of records, design automation, Data Vault model.

1. Introduction

In this section we delineate the scope of the work, introduce baseline terminology, present selected requirements for design automation of a Data Vault (DV) type data warehouse, and explain the organization of this paper. The scope of the work is the design of DV based Enterprise Data Warehouse (EDW) / Data Mart (DM) content. The paper does not addresses storage and distribution opportunities i.e. recent advances such as (Hadop, NoSQL, etc.).

Data warehouse (DW) is a subject oriented, nonvolatile, integrated, time variant collection of data in support of management's decisions [1]. A DW may be used to support permanent systems of records and information management (compliance and improved data governance). Data marts are small data warehouses that contain only a subset of the EDW [2]. The data mart provides the platform for Online Analytical Processing (OLAP) analysis. Therefore, OLAP is a natural extension of the data warehouse. The results from OLAP analysis can be presented visually, which enables improved comprehension [3]. The data mart is a part of data storage, but usually contains summarized data [4]. The Extract, Transform, Load (ETL) process involves fetching data from transactional systems, cleaning the data, transforming data into appropriate formats and loading the result to a warehouse [5]. In the ETL process, data from data sources is extracted by extraction routines. Data are then propagated to the Data Staging area where they are transformed and cleaned before being loaded to the data warehouse [6], [7]. An Important element of the DW is metadata, including definitions and rules for creating data [8]. Metadata play an important role in data warehousing. Before a data warehouse can be accessed efficiently, it is necessary to understand what data is available in the warehouse and where the data is located [2].

The design of data warehouses is a difficult task. There are several problems designers have to tackle. First of all, they have to come up with semantic reconciliation of the information lying in the sources and the production of an enterprise model for the data warehouse [2]. Data warehouses can be distinguished by the type of architecture. Bill Inmon [9], [1], [10] proposed the CIF (Corporate Information Factory) as an integrated data warehouse, i.e. database in the third normal form (3NF), from which multidimensional data marts are to be derived. The second option is bus architecture, defined by Ralph Kimball [11], [12], [13] where a data warehouse is just a collection of data marts with conformant dimensions. Data Warehousing 2.0 (DW 2.0) is a second-generation attempt to define a standard Data Warehouse architecture. One of the advantages introduced in DW 2.0 is its ability to support changes of data over time [10].

Data modeling techniques for the data warehouse differ from the modeling techniques used for operational systems and for data marts. This is due to the unique set of requirements, variables and constraints related to the modern data warehouse layer. Some of these include the need for an integrated, non-volatile, time-variant, subject oriented, auditable, agile, and complete store of data. To address these needs several new modeling approaches have been introduced within the Data Warehouse/Business Intelligence industry. Among these are Data Vault modeling and Anchor Modeling [14].

The Data Vault approach is introduced by Daniel Linstedt [15], [16], [17], to solve the problems of flexibility and performance, enabling maintenance of a permanent system of records [18]. Data Vault (DV) model is recognized by the modeling style using Hub, Link and Satellite entities. In terms of entity relationship data models, a Hub entity holds identities, typically business keys (or their combinations), Link entities are representation of foreign key references (typically used to represent transactions between two or more business components i.e. Hubs). A Satellite entity shows context information i.e. attributes of a Hub or a Link. We can split out satellites by: rate of change, type of data and source system. [19]. See the example of a Data Vault in figure 1. The main difference from traditional ER data modeling style is in representing attributes, namely Hub/Link entities which have relationship/identify Satellite entities (representing traditional attributes in time).

Anchor Modeling is a database modeling technique built on the premise that the environment surrounding a data warehouse is in a constant state of change, and furthermore that a large change on the outside of the model should result in only a small change on the inside of the model [20], [21]. The goal of using the anchor modeling technique is to “achieve a highly decomposed implementation that can efficiently handle growth of the data warehouse without having to redo any previous work” [21], mimicking the ideas of isolated semantic temporal data put forth in DW 2.0 architecture [22].

Data Vault and Anchor models are characterized by strong normalized data and insensitivity to changes in the business environment, and adapted to frequent changes, as well as the modeling of small parts of the data model, without the need for redesign.

Sixth Normal Form (6NF) is a term used in relational database theory by Christopher Date to describe databases which decompose relational variables to irreducible elements. While this form may be unimportant for non-temporal data, it is certainly important when maintaining data containing temporal variables of a point-in-time or interval nature [22], [23]. In [23], Chris Date points out several shortcomings when attempting to model fully temporalized data using standard Fifth Normal Form (5NF) principles and introduces Sixth Normal Form (6NF) as "A relvar (table) R is in 6NF if and only if R satisfies no nontrivial join dependencies at all, in which case R is said to be irreducible" [23].

The Data Vault model, in the extreme case where a satellite table consists of one attribute, becomes a 6NF design.

According to [16], [24], a Data Vault makes most sense in the case of distributed data sources. To ensure traceability of data, the Data Vault is not transforming data from different sources before they are loaded into the warehouse, thus enabling permanent system of records (i.e. Data Vault). This differs from the CIF where data are consolidated up front.

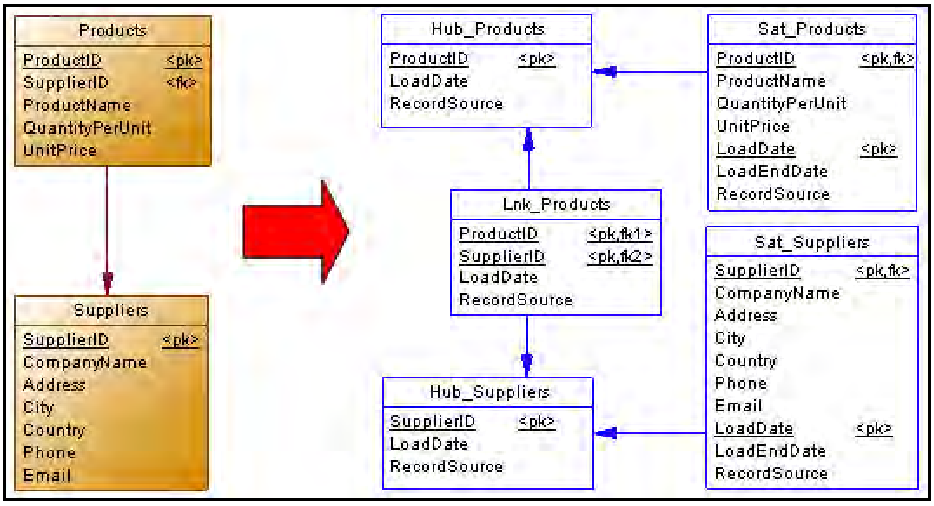


Fig. 1. An example of mapping a relational to the data vault physical model

A system theoretic block diagram representation, figure 2 from [25], shows explicit separation of data warehouse data into permanent original data (DW staging or to put it more precisely, a Data Vault) and derived data in data marts. The DB mimics system's state at the last updated moment in time while DW tracks the system's state over time.

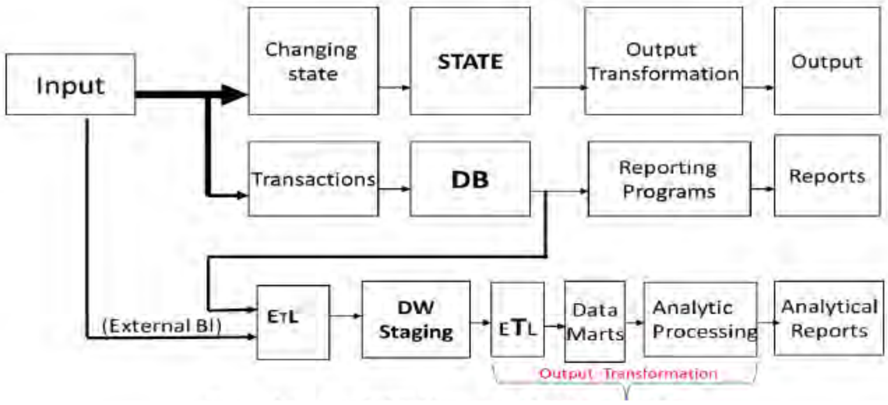


Fig. 2. DW State history (Data Vault) and DW Reporting Area

Methodologies of data warehouse design, such as ([1], [13], [26], [27], [19]) are not fully formalized. Based on the starting point to be used in determining the DW content, the following four classes of approaches to DW design can be recognized:

1. Data-driven (or supply-driven) [9], [28], [29], [30], [31] is based on an analysis of the data, the database schema and the identification of available data.

2. Demand-driven (also user-requirements driven) [11], [28], [30], [31] starts with the determination of requests for information from the user, but mapping that to available data remains a problem.

3. Goal-driven [32], [33], [30] is based on a business strategy i.e. begins with an analysis of key business processes, but according [28], [34] is a subset of Demand-driven approach.

4. The Hybrid approach combines user requirements with analysis of management and operational data. It distinguishes the interleaved and sequential access. For sequential access, the data-driven and demand-driven paradigms are applied independently of each other, one after the other, while in the interleaved hybrid approach both paradigms are applied in parallel [34].

This work is part of a larger research program, the DW21 dealing with the next generation DW design technology [35] where Data Vault is a baseline approach to DW design and most of the requirements are derived from a developer's standpoint as shown in figure 3.

One practical problem for organizations building data warehouses using different methodological approaches is that the process of designing a data warehouse including a process of identifying and/or generating measures and dimensions (or Hub, Link and Satellite concepts in case of Data Vault approach) is not sufficiently automated.

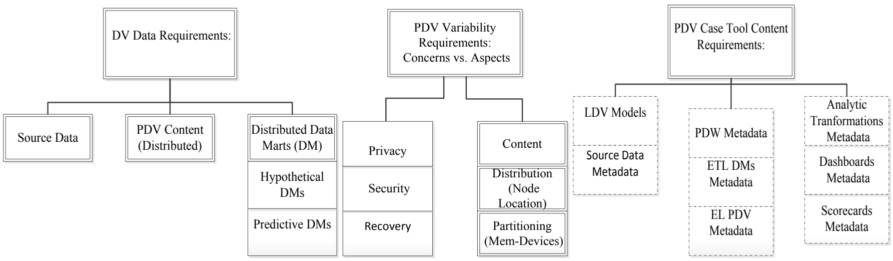

DV Data Requirements:
Source Data
PDV Content (Distributed)
Distributed Data Marts (DM)
Hypothetical DMs
Predictive DMs
PDV Variability Requirements: Concerns vs. Aspects
Privacy
Security
Recovery
Content
Distribution (Node Location)
Partitioning (Mem-Devices)
PDV Case Tool Content Requirements:
LDV Models
Source Data Metadata
PDW Metadata
ETL DMs Metadata
EL PDV Metadata
Analytic Tranformations Metadata
Dashboards Metadata
Scorecards Metadata

Fig. 3. Some Requirements of DW 2.1 Research Program

The rest of the paper is structured as follows: section 2 reviews work on data warehouse design automation, section 3 presents our approach to physical Data Vault design, section 4 illustrates related Physical Data Vault (PDV) design tool we developed to automate prototyping and implementation of data vaults based data warehouses, section 5 illustrates the use of the tool on a real example for distributed data base sources, and section 6 outlines future work.

2. A Review of Data Warehouse Design Automation

The starting point for review is selection of relevant works in the field of design automation for DW content (to narrow the scope somewhat, we did not emphasize a large body of work dealing primarily with ETL). Our choices are based on perceived contributions to the theory and practice of design, confirmed by the number of citations (at www.harzing.com).

A semiautomatic sequential hybrid approach of conceptual modeling of a data warehouse that starts with E/R model is proposed in [36]. After the analysis and design of the conceptual/logical model, the identification of facts is followed by semi-automatic creation of attribute trees in accordance with the requirements. The next step is to identify the dimensions, measures and aggregate functions after which logical and physical schema of data warehouse is performed. Improvement of the procedure is given in a textbook form in the [26].

The Hybrid approach in designing multidimensional database warehouse on the basis of transactional normalized systems is proposed in [37]. In this approach an ER diagram is transformed into a Structured Entity Relationship Models (SERM) diagram.

Phipps and Davis [38] start with a data-driven approach, and later uses the demand-driven approach (sequential hybrid approach). The initial scheme is obtained using a multidimensional E/R model. This paper proposes an algorithm for obtaining a conceptual scheme of transactional schemas. The algorithm uses the numeric fields and relationships between entities as the basis for the creation of Multidimensional E/R schemes. The algorithm is applied to each data model where data can be divided into numeric, date/time, and text data types. In addition, this approach can be used to evaluate candidates for the conceptual schema using custom queries.

Peralta (et al.) [39] represents a step forward in the automation of data warehouse based on the rules applying existing knowledge in the data warehouse design. The proposed rules are built into the design strategy that is run in accordance with the requirements and data extraction based on the transactional database schema. The framework consists of: rules, guidelines for the design based on non-functional requirements, the mapping scheme of the original database to a conceptual scheme, and a number of schema transformations. This approach allows the use of existing design techniques, with improved productivity.

Romero and Abello [40] propose a semi-automatic method for recognizing business multidimensional concepts from domain ontology representing different and potentially heterogeneous data. With the use of ontologies and tools it can search for multi-dimensional patterns. Simulations of real cases to verify the algorithm are performed as well as theoretical analysis of the algorithm. This approach is able to integrate data from heterogeneous sources that describe their domains through ontologies. One of the most promising areas where the method can be applied is the semantic Web, which contributes to the extraction and integration of external data from the Web.

Zepeda (et al) [41] introduces a semi-automatic process to build a DW based on MDA (Model Driven Architecture). It starts by collecting and consolidating user requirements. Other steps include recognizing structural information from the original database schema supported by an automation technique. This approach uses a tool to guide designers toward effective solutions in line with customer requirements.

Nazri (et al) [42] is based on a model of data source and proposes a semi-automatic tool that uses lexical ontology as the knowledge base. Once you identify the facts and dimensions, the generation of a multidimensional model is made with minimal involvement of a user. The method is illustrated using a semantic lexicon WorldNet as knowledge domain, to identify measures and dimensions.

Zekri (et al) [43] introduces a semi-automatic approach for the design of multidimensional diagrams. Conceptual Data Model (CDM) is first translated into the multivariate pattern, and then refined with user requirements. Both steps use graphs as a formalism for the representation of the data model for decision making.

Design automation is not an easy process because, in addition to the identification of the original database schema, it attempts to formalize the requirements of end users. However, certain steps must be performed manually, for example, the identification of business keys and measures. Table 1 provides a comparative overview of the various approaches, including a new one based on the Data Vault. Note that the public domain literature concerning the design of Data Vault systems [15], [16], [24], [44], [45], [46] does not elaborate on actual design process automation.

Our direct approach to automatic generation of Data Vault physical models is based on metadata schemas of structured data sources (transactional databases) taking into account some semi-structured and unstructured sources. In addition, compared to the alternatives (Table 1), our direct approach to the physical design automation of data warehouse is achieved through the use of rules. It should be noted that majority of listed approaches are academic and that only Golfarelli [26] has a long tradition of industrial use which is also a characteristic of the Data Vault approach itself [15], [16], [44]. A direct physical design for data warehouses is practical only in the case of a data vault type of data warehouses and is made possible by separating storage of identities (permanent keys) from evolving relationships and characteristics (attributes).

Table 1. Comparison of data warehouse design approaches

Approach	Generating			Data Source		Using Rules
	Conceptual DW model	Physical DW model	Data Vault	Structured	Semi-structured	
[36]	X	X		X		
[37]	X			X		
[38]	X			X		
[39]	X			X		X
[40]	X	X		X	X	
[41]	X			X		
[43]	X	X		X		
[42]		X		X	X	
Direct Physical DV		X	X	X	X	X

An important contribution of our approach is the realization of a possibility to directly derive Data Vault schema from source RDBMS schemas. The direct approach to physical Data Vault design is now possible thanks to the feature of the Data Vault models i.e. the separation of unchangeable identities of entities in real systems (Hubs) from time variant relationships (Links) and the characteristics of such entities and relationship (Satellites).

There are many approaches to modeling and/or generating ETL process and code for example [45, 46, 47, 48, 49, 50, 51 and 52]. In Muñoz [51] an approach to the automatic code generation of ETL processes is given. The modeling of ETL processes used MDA (Model Driven Architecture) with the formal definition of a set of QVT (Query, View, Transformation) transformation. The problem of defining ETL activities and securing their formal conceptual representation is given in Simitsis and Vassiliadis [48]. The proposed conceptual model provides custom attributes for monitoring and appropriate ETL activities in the early stages of project data warehouse, and flexibility, extensibility and reuse patterns for ETL activities. Jovanovic (et al) [53] presents a tool GEM that from a given set of business requirements and data source descriptions semi-automatically produces multidimensional and ETL conceptual designs. In addition, many commercial ETL design tools exists [54, 55, 56, 57, etc.]. Nevertheless none of the before mentioned approaches and tools consider the ETL process in a Data Vault based data warehouse context that is for a second generation of data warehousing models.

Sources listed in Table 1 above mainly represent academic contributions. Of interest to practitioners and to our research are also several additional sources illustrating the development and/or use of tools to automate the design of data warehouses.

Golfarelli (et al) [58] presented the prototype CASE tool WAND. This tool assists the designer in structuring a data mart, carries out conceptual design in a semi-automatic fashion, allows for a core workload to be defined on the conceptual scheme and carries out logical design to produce the data mart scheme.

QBX is a CASE tool for data mart design resulting from a close collaboration between academy and industry. It supports designers during conceptual design, logical design, and deployment of ROLAP data marts in the form of star/snowflake schemata, and it can also be used by business users to interactively explore project-related knowledge at different levels of abstraction. [59].

BIReady's Model-driven Data Warehouse generation engine [60] is one extension of CA Erwin tool that allegedly uses best practices and a stable architecture so that one can easily manage and evolve data warehouse. BIReady is able to generate a Data Warehouse automatically from a business model. BIReady even generates the ETL and loads data. BIReady builds data architectures from the ground up based on best-practices. For example, the Corporate Information Factory (CIF) pattern by Bill Inmon, the father of Data Warehousing, and Dan Linstedt's Data Vault modelling approach. Even BIReady's datamarts are real Kimball Star-schemas.

The Birst [61] tool automatically compiles a logical dimensional model into a star schema design and then generates and maintains fact and dimension tables. Logical measures are automatically analyzed for calculation grain, while logical dimensions are analyzed for levels requiring persistence. This tool generates and manages all key relationships, including surrogate keys where necessary and provides data connectivity and extract options for a wide variety of databases, both flat and structured files. The Birst also supports scheduled data extraction from all major relational database management systems, including Oracle, DB2, SQLServer, MySQL, Sybase, and PostgreSQL.

Quipu [62] is an open source data warehouse generation system that creates data warehouses and supports Data Vault architectures. Quipu automates the data warehouse data model design and generates the ETL load code to fill the data warehouse from source systems. With Quipu you can simply and quickly generate and implement a source driven Data Vault, often referred to as source or raw Data Vault. Additional business value can be achieved by implementing a business Data Vault, where source data is combined in a Data Vault implementation of a single business model. Quipu supports both Data Vault architectures. Quipu's repository holds all relevant metadata of the source data elements and generates a data warehouse model based on the Data Vault modeling methodology. Quipu provides the functions to build a Data Vault data warehouse quickly and reliably. Quipu fills the gap in the tool sets available today to implement the data warehouse architecture by generating, maintaining and populating (database) structures and code to capture changes in data, both transactional and reference [62]. According to a white paper available on [62], functionality to assist the construction of Data Marts is not yet available in the Version 1.1 release of Quipu.

Pentaho Data Integration (PDI, also called Kettle) [46] is the component of Pentaho responsible for the ETL processes. Though ETL tools are most frequently used in data warehouses environments, PDI can also be used for other purposes: migrating data between applications or databases, exporting data from databases to flat files, parallel loading of data into databases, data cleansing, integrating applications. PDI can be used as a standalone application, or it can be used as part of the larger Pentaho Suite. As an ETL tool, it is the most popular open source tool available Moreover, the transformation capabilities of PDI allow you to manipulate data with very few limitations [46]. The ETL process is fully automated, but data vault and data mart processes are not part of the framework [63].

Table 2 summarizes information about relevant tools.

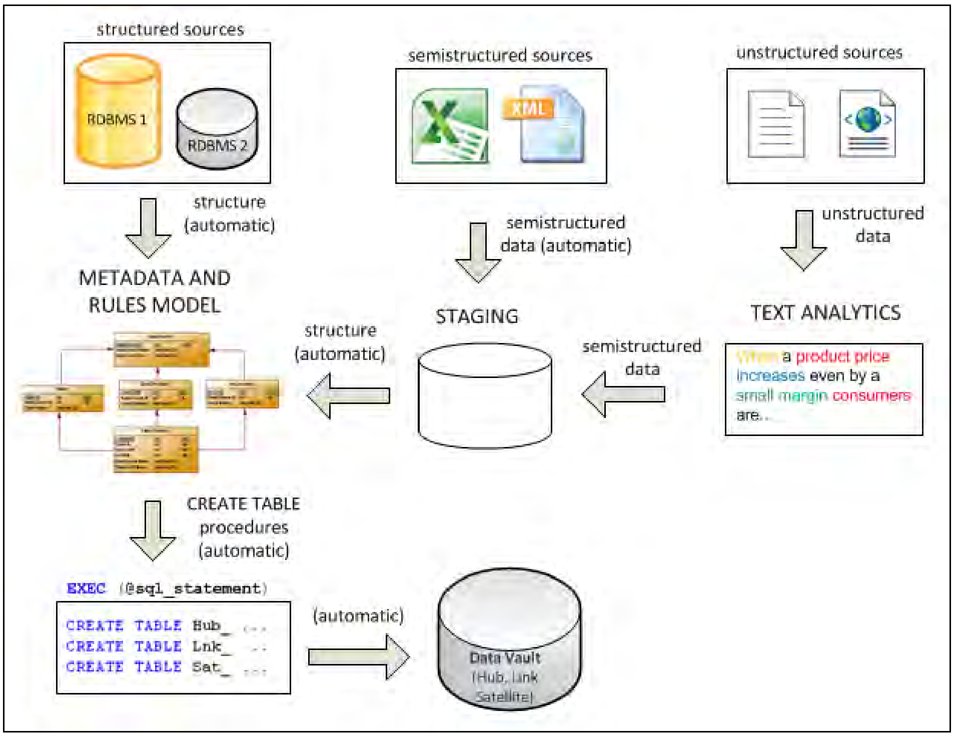
Table 2. Comparison of data warehouse design tools

	Type of DW tool	Physical DW model	Data Vault design	Generate ETL code	Data Mart design
WAND	Academic	Automatic	Not supported	Automatic	Automatic
QBX	Academic	Automatic	Not supported	Automatic	Automatic
Quipu	Open source	Automatic	Automatic	Automatic	Manual
Pentaho Kettle	Open source	Manual	Manual	Automatic	Manual
BIReady	Commercial	Automatic	Manual	Automatic	Automatic
Birst	Commercial	Automatic	Not supported	Automatic	Automatic
Direct PDV	Prototype	Automatic	Automatic	Automatic (in progress)	Automatic (in progress)

The advantage of our approach is the availability of a graphical view of the physical diagrams for Data Vaults and data marts during the design process. Furthermore, except for WAND and QBX, these tools do not present formalized and publically available information in detail.

3. Direct Physical Data Vault Design

The aim of physical design is to assist in implementing a DW on a selected database management platform. In the process of storing data, ones take large amounts of data from variety of sources: ERP (Enterprise Resource Planning) systems, relational databases, Excel files, DBF files, TXT and XML files or data from the Web. The data in databases are called structured as they are defined with the data schemas. Information contained in file systems is considered unstructured and exists in two basic forms: as textual and as non-textual. Existing technology allows for easy loading of textual data in the staging database or data warehouse, which is not the case with non-textual. Some textual data (documents) can be structured and are known as semi-structured.



When a product price increases even by a small margin consumers are..

```
EXEC (@sql_statement)
CREATE TABLE Hub_ ...
CREATE TABLE Lnk_ ...
CREATE TABLE Sat_ ...
```

Fig. 4. Overview of proposed approach

We propose a direct approach to the design of physical Data Vault data warehouses based on the historical preservation of the original data. An important methodological contribution of our approach is that conceptually, at the metadata level, data sources (DS) of all types are presented with relational model schema equivalents. The relational model is treated as the common factor for Data Vault implementation and all the source data and therefore all types of sources (structured and some semi-structured) are the first to be abstracted via meta-model into relational model equivalents. The term direct approach means that an approach directly leads to Data Vault models based on a physical model of sources in a relational form. The approach is facilitated by the fact that, no matter how many sources of operational data are drawn from, it is the aim of each information system to integrate basic information of wider significance and establish control using only one or few primary database(s) that include most of the data (and these databases, in most cases, reside in relational database engines). On the other hand, theoretical analysis [64] considers relational data models as general models, in which all data can be represented and so reduced. This work covers only the first stage from Figure 5 of the full scope of our research on Data Vault approach automation. The phases of initial Data Vault design automation (for the first stage) are shown in Figure 6.

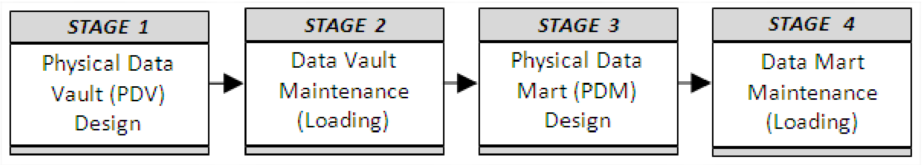

STAGE 1
Physical Data Vault (PDV) Design
STAGE 2
Data Vault Maintenance (Loading)
STAGE 3
Physical Data Mart (PDM) Design
STAGE 4
Data Mart Maintenance (Loading)

Fig. 5. Stages of Data Vault data warehouse design and exploitation

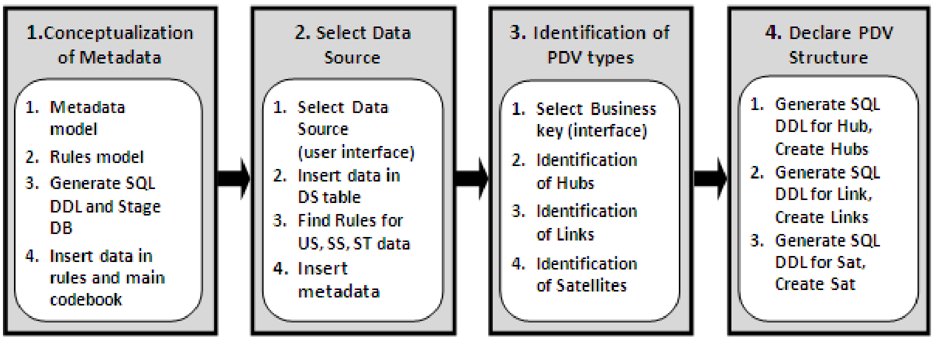

1.Conceptualization of Metadata
1. Metadata model
2. Rules model
3. Generate SQL DDL and Stage DB
4. Insert data in rules and main codebook
2. Select Data Source
1. Select Data Source (user interface)
2. Insert data in DS table
3. Find Rules for US, SS, ST data
4. Insert metadata
3. Identification of PDV types
1. Select Business key (interface)
2. Identification of Hubs
3. Identification of Links
4. Identification of Satellites
4. Declare PDV Structure
1. Generate SQL DDL for Hub, Create Hubs
2. Generate SQL DDL for Link, Create Links
3. Generate SQL DDL for Sat, Create Sat

Fig. 6. Phases of initial Physical Data Vault design automation

The direct approach to Physical Data Vault (PDV) design automation of a data warehouse presented in this paper includes the use of rules for data warehouse design. To understand the detailed rules we first outline a general algorithm for recognition of Hub, Link and Satellite relations based on mapped original data (equivalent relational model tables). In a nutshell the algorithm consists of the following:

```
I- For each source:
      a) For each Table: the user confirms Hubs by
         selecting a Business Key
      b) For each Table, if not already selected as
         a Hub, create Link
      c) For each Table:
            i. for each FK create  Link
      d) for each non-key attribute: create
         Satellite
II-   For each Hub (PK)
   a) For each source
    i.   For each Table (search for matching PK): if
         found, create Link (to Hub ad II; this only
         illustrate possible integration).
```

The effect of the algorithm is to create a Link for each FK (except for tables whose PK are completely composed of FKs as all such tables will became Links in step I- b- if not confirmed Hubs in step I- a). The proof of completeness of the algorithm follows from the assumption that, no matter the source, every table must have a PK, and the fact that tables are originally valid i.e. have been made from real data so that no contradictions in terms of circular dependencies on other tables could prevent their formation.

1. Every PK is either composed (partially or fully) from the PKs from other tables or is a separate PK (possibly, but rarely, composed of several fields concatenated or not. See [19]).
2. Tables with separate PKs are designated as Hubs
3. Tables with fully composed PKs are designated as Links.
4. Remove all Links tables whose PK are not used as FK
5. Remove all Hubs tables whose PK are not part of any PK in any remaining table
6. Represent all remaining FK references with separate Links tables (each of those Links will now have two FK references); this is a crucial step separating mutually dependent tables
7. Represent what is left, treating it as a directed graph expressing FK references (as precedence relations) in a matrix form
   a. Nodes are tables (including Links created in step 6),
   b. Branches are FK references whose direction follows the FK reference i.e. branches are pointing to tables whose PK is referenced by each FK. The result is a connected directed graph.
8. We can fully reduce the matrix using Warshall's algorithm following precedence relations (as adjacencies) on a connected directed acyclic graph (in a finite number of steps) where the number of steps determines the length of the longest precedence chain.
9. The key for a proof that the above procedure finishes is that the graph obtained by carving initial, i.e. root, Hubs and Links must be acyclic (proof by contradiction)
   a. If the precedence (FK path) forms a cycle the table has a PK that depends on itself but this contradicts the initial requirement that tables can exists in extension, meaning that records in the tables exist i.e. not only can be formed with complete PK in the time of schema creation but also effectively populated in the time of record

Next figure shows how introduction of Links breaks cycle in ER models.

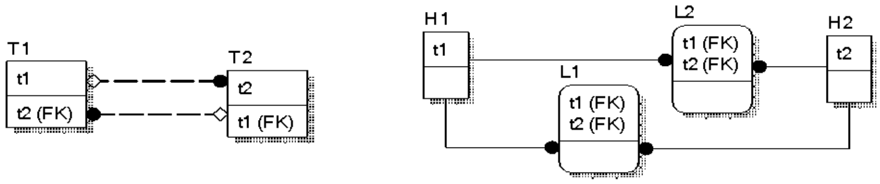


Fig. 7. Links for FK relationship

3.1. Conceptualization of Metadata – Phase 1

The main phase is the Conceptualization of Metadata (general Physical Data Vault structure) resulting in a physical data model that can be used in designing individual Data Vault data warehouses.

The first two steps of this stage assume building a model of the metadata (Figure 8) and of the modeling rules (Figure 9). When making the meta model and rules we distinguish between part of the model that is independent of the data source (source-independent design) and part of the model, which depends on the source (source-specific design). The independent part refers to the metadata tables and rules, and the dependent part involves procedures related to different data sources. The purpose and a method of how to update the metadata table are shown in Table 3**.**

The RuleTypes table contains information about the types of policies, such as rules for different data sources, rules for identifying Hub, Link and Satellite tables and rules for creating Hub, Link and Satellite tables. The Rules Table contains information about the rules for a particular type of rules and actions to be performed when a certain condition is satisfied. Rules in column Rules allow easier execution of commands that are stored in the column RuleAction. The following example is given to show the SQL (Structured Query Language) rule to describe a set of constraints that are applied to identify candidate Hub based on the value of the column BusinessKey.

```
UPDATE TableColumns SET TableColumns.HubCandidate = 1
WHERE TableColumns.BusinessKey = 1
```

The third step in this phase is generating code to create a staging database and creating tables and stored procedures from the model (Figure 8 and 9). From this presentation of the physical model we can generate a script to create a table in the staging database and insert initial data into tables StructureType and SourceType. The fourth step is to insert data into the rules table and basic codebook (reference data).

After the first stage and creation of the tables and appropriate procedures conditions are created for the development of an application that allows one to automate the design of the physical model of Data Vault data warehouse, with minimal user interaction. Second, third and fourth phase of Figure 6 needs to be done for each specific system.

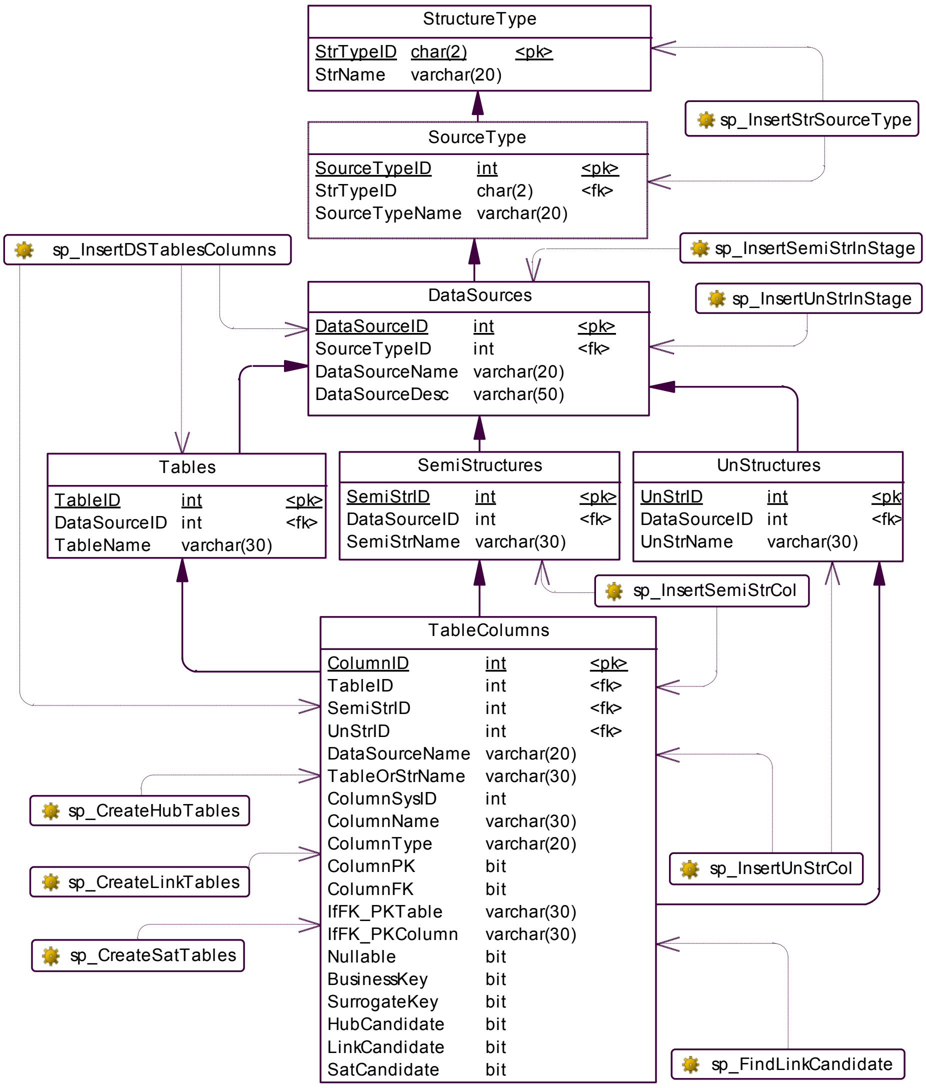


Fig. 8. Physical metadata model

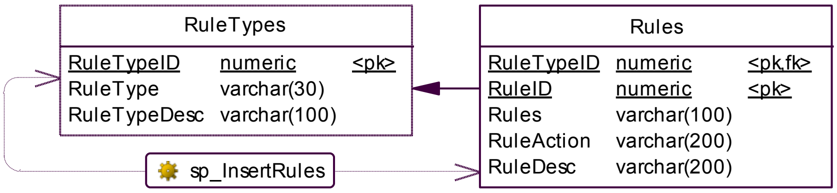


Fig. 9. Model of Rules

Table 3. Purpose of the generated tables

Table	Purpose	Insert method
StructureType	Codebook data: structured (ST), semi-structured (SS) or unstructured (US)	From model, example: INSERT INTO StructureType (StrTypeID, StrName) VALUES ('ST', 'Structured')
SourceType	Data sources codebook (example: MS SQL Server, Oracle, xls, XML, txt, etc.)	From model, example: INSERT INTO SourceType (SourceTypeID, StrTypeID, SourceTypeName) VALUES (1, 'ST', 'MS SQL Server')
DataSource	Data source that will be used by data warehouse	Automatically after identification all data sources through the user interface
Tables	Table from operational data	Automatically after identification all data sources through the user interface
SemiStructures	Semi-structured data source (xls, XML, etc.)	Semi-automatic, after importing and structuring (if possible) in the staging database
UnStructures	Unstructured data source (txt, etc.)	Semi-automatic, after textual analysis and importing and structuring in the staging DB
TableColumns	Columns of data tables from databases and other sources that can by structured	Automatically, except BusinessKey fields that will be selected through the user interface
RuleTypes	Types of rules for creating a data warehouse	From model
Rules	Contain rules and appropriate action if condition is satisfied	From model

3.2. Identification of Data Sources – Phase 2

This phase is applied to each individual DW (further development of CASE tools will take into account experiences with the prototype). The first step involves the identification of specific applications and systems (databases, structured and unstructured data sources) for data warehouse by a user selecting specific data sources (through the user interface). The second step in this phase is to insert data into table data sources based on the selected data source.

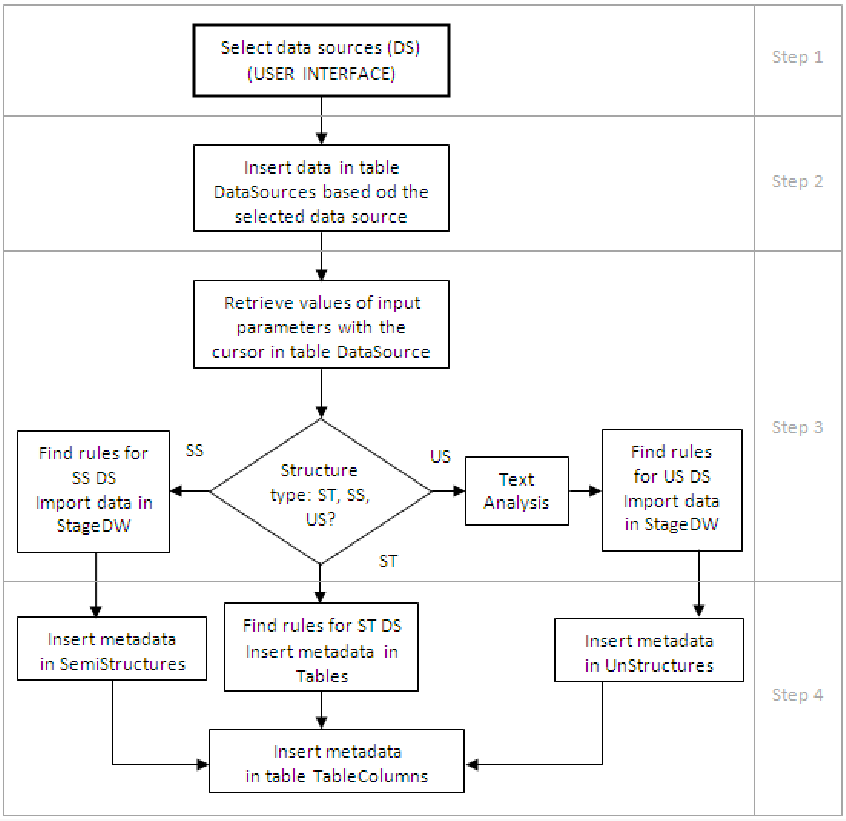

Select data sources (DS)
(USER INTERFACE)
Step 1
Insert data in table DataSources based od the selected data source
Step 2
Retrieve values of input parameters with the cursor in table DataSource
Step 3
Find rules for SS DS Import data in StageDW
SS
Structure type: ST, SS, US?
US
Text Analysis
Find rules for US DS Import data in StageDW
ST
Insert metadata in SemiStructures
Find rules for ST DS Insert metadata in Tables
Insert metadata in UnStructures
Step 4
Insert metadata in table TableColumns

Fig. 10. Flowchart of the second phase

The third step defines rules for inserting metadata for structured, semi-structured and unstructured data sources into a staging database. The rules defined in the model are stored in a table Rules. For identification of the rules, it is necessary to find data type (ST, SS, US) from all sources of data based on input parameters. These parameters will be obtained through the mechanism of the cursor through tables from data sources. The cursor is introduced as a set of records which is attached to the pointer at current row. Commands in SQL statements include moving the cursor to work with the current row. If it is semi-structured data sources, in the absence of a clear structure, their operational data need to be imported into the staging database. Semi-structured data such as Excel files can be imported into a database using ETL processes [13].

Import metadata in the staging database will be prepared using a stored procedure as indicated in the model *sp_InsertSemiStrInStage* (with appropriate parameter). In case of unstructured data sources, it is necessary to first identify their metadata. Unstructured data can be placed into the database in the traditional way by using the files metadata, the path to the file or URL, and attributes and links between files placed in the database. The newer method of storing unstructured data in the database is known as text analytics. It is the process of converting unstructured documents into structured documents by analysis of the structure of the text in the document [65]. Text analytics is the process of enabling a computer to extract meaning from text and is performed as a series of iterative process prior to loading into the database [66]. Some of the procedures are simple editing, stop-word removal, replacement of synonym or concatenation, homographic resolution, thematic clustering, external glossary/taxonomy overlay, stemming, alternate spelling resolution, foreign language accommodation, search direct and indirect support. A description of each of the procedures is given in [67].

The fourth step in this phase is to fill the tables Tables, SemiStructures, UnStructures TableColumns and information about the tables and their columns on transactional databases. This step will be automatically generated from the procedure for importing data by using data from the target system schema repository containing the tables, columns, indexes, constraints and relationships for each specific system for database management. Inserting metadata into the table TableColumns looks more complex due to the large number of metadata and a number of columns. A similar approach is taken with other structured data sources that import metadata into tables and TableColumns. Tables will be prepared using a stored procedure as indicated in the model *sp_InsertDSTablesColumns*. Based on data sources, this procedure will be transmitted as a parameter for a specific structured data source (MS SQL Server, Oracle, IBM DB2 and others), on the basis of the rules in the table of rules.

In a nutshell, the algorithm for this phase consists of the following:

```
1. Loop over data sources (incrementally or in sets of DBs
   as sources): Select data source and Insert metadata in
   table DataSources
2. For each row in table DataSources: find structure type,
   source type and rules
      a) For each structured type:
           i.Insert metadata in Tables
          ii.For each row in table Tables: insert metadata
                in table TableColumns
      b) For each semistructured type:
           i.Extract and load data in database StageDW
          ii.Insert metadata in table SemiStructures
         iii.For each row in table SemiStructures: insert
                metadata in TableColumns
      c) For each unstructured type:
           i.Text analytics process
          ii.Extract and load data in database StageDW
         iii.Insert metadata in table UnStructures
          iv.For each row in table UnStructures: insert
                metadata in TableColumns
```

Identification of PDV types (Hubs, Links, Satellites) for each DW is initially reduced to the identification of a business key until the Hub, Link and Satellite are automatically generated. On the model in Figure 8, the meaning of the columns in the tables StructureType, SourceType, DataSource, Tables, SemiStructures, UnStructures clear,only the structure and purpose of the specific TableColumns column (Table 4) will be explained.

This table should contain metadata from databases, semi-structured metadata and unstructured data sources that can be appropriately structured. Our approach provides the automatic loading of metadata from all TableColumns structured data sources and partly from some unstructured and semi-structured data sources. In addition to the data that will be filled on the basis of a database schema, this table will contain a column indicator that will give us information as to whether it is a business key, surrogate key, or if the table Hub, Link or Satellite candidate.

Table 4. TableColumns Table structure

Column	Type	Description	Load automatic	Load through UI
ColumnID	int	Primary key	✓	
TableID	int	Table ID	✓	
SemiStrID	int	ID from semistructured data source	✓	
UnStrID	int	ID from unstructured data source	✓	
DataSourceName	varchar	Data source name	✓	
TableOrStrName	varchar	Table name or semi/un struct. file name	✓	
ColumnSysId	integer	Sys column ID	✓	
ColumnName	varchar	Column name	✓	
ColumnType	varchar	Data type	✓	
ColumnPK	bit	PK column?	✓	
ColumnFK	bit	FK column?	✓	
Nullable	bit	Nullable column?	✓	
IfFK_PKTable	varchar	Table name on PK side (if column FK)	✓	
IfFK_PKColumn	varchar	Column name on PK side (if column FK)	✓	
BusinessKey	bit	Business key column?		✓
SurrogateKey	bit	Is column Surrogat for corresponding Business key?	✓	
HubCandidate	bit	Hub candidate?	✓	
LinkCandidate	bit	Link candidate?	✓	
SatCandidate	bit	Satellite candidate?	✓	

This phase involves the following steps:

1. Identification of business key through a user interface. Through the appropriate user interface, based on the data, the user should check the business keys. Based on the business key, the column with the appropriate BusinessKey value (0 or 1) will be filled. Filling in this section can be automatically based on rules stored in the table Rules, according to [15], [16], [24], [44], and based on BusinessKey the column SurrogateKey will be filled.

2. Identification of hubs. The Hub entity table contains a single list of business keys. These are the keys that organizations use in their daily operations, such as customer number, code of the employee, account number, and so on [15]. According to [15], [16], [24], Hub candidates can be identified using the filled BusinessKey and SurrogateKey on the basis of rules for identifying hubs which are found in the tables RuleTypes and Rules.

3. Identification of links. Link is the physical representation of references, foreign keys and many-to-many relationships in third normal form [15]. Links can be identified using the procedure *sp_FindLinkCandidate* (which is an integral part of the model and that is called from the table Rules and RuleTypes) that includes the following steps:

a) Find many-to-many table

b) Set to true LinkCandidate field in many-to-many table

c) Set 1 value in the column LinkCandidate for tables that have a foreign key

4. Identification of satellites. The Satellite entity contains context data of hub and contains attributes that are not primary or foreign keys [15]. Satellites are identified on the rules in tables RuleTypes and Rules, according to [15], [16], [24], [46].

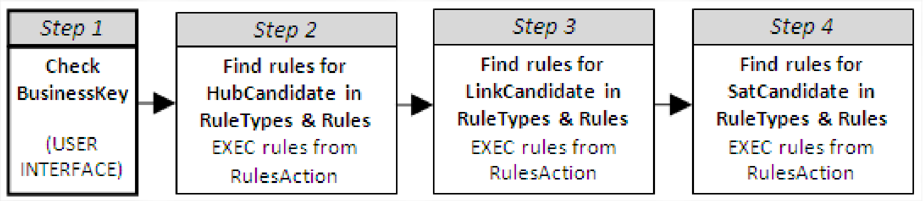

Step 1
Check BusinessKey
(USER INTERFACE)
Step 2
Find rules for HubCandidate in RuleTypes & Rules
EXEC rules from RulesAction
Step 3
Find rules for LinkCandidate in RuleTypes & Rules
EXEC rules from RulesAction
Step 4
Find rules for SatCandidate in RuleTypes & Rules
EXEC rules from RulesAction

Fig. 11. Flowchart of the third phase

3.4. Initial Declaration of PDV Structure- Phase 4

The last phase in the process of automation of the physical design of a data warehouse is the initial declaration of the PDV structure. This phase include following steps:

1. Generate and execute the script to create a hub table. When we have the business key, appropriate tables can be identified (by setting the value 1 in column HubCandidate). It is possible to generate a script to create a hub table in a Data Vault, based on the rules in the Rules table. This step provides the following:

a) Forming a cursor to go through the TableColumns Where HubCandidate=1

b) Retrieve data from a table and assign variables

c) In each iteration, use dynamic SQL to supplement sql_statement

d) Perform an sql_statement, creating a hub table

2. Generate and execute the script to create the link table. The link table is used to represent relationships or transactions between two or more business components (two or more hubs). We identified the appropriate link tables containing value 1 in column

LinkCandidate which enables the generation script to automatically generate a link table, based on the rules in the Rules table. It is possible to generate a script to create a link table in a data warehouse, based on the rules in Table Rules. This step involves:

a) Forming a cursor to go through the table TableColumns Where LinkCandidate=1
b) Retrieve data from a table and assign variable
c) In each iteration, use dynamic SQL to supplement sql_statement
d) Performing an sql_statement, creating a link table

3. Generate and execute script to create the satellite table. Satellite entity shows how context hub data. Satellite table created for each hub table will contain non-key attributes in a transactional database. We identified the appropriate Satellite tables by setting the value 1 in column SatCandidate which enables the generation script to automatically generate the satellite table based on the rules in the Rules table. In this case, one table is generated for each hub table (or link table, if the transaction has only a link table). This step involves:

a) Forming a cursor to go through the table TableColumns Where SatCandidate=1
c) Retrieving data from a table and assign variable
d) In each iteration, use dynamic SQL to supplement sql_statement
e) Performing an sql_statement, thus creating satellite tables

If the source of simple unstructured data, (Excel or TXT file), the data on clients who are not in a transactional database (or other contact information, information about market position, business data, etc.), then an additional Satellite table will be created that will include data from Excel or TXT file.

Procedures described in the fourth phase make it possible to automatically create a complete data warehouse based on Data Vault concepts. Among meta-requirements for any design approach and its automation at least the following four are obvious: performance, scalability, flexibility and agility. In order to fully appreciate potential of our PDV approach one have to first realize what DV as such (comparing to traditional alternatives namely normalized EDW and dimensional Data Marts) contributes to satisfying such meta-requirements. The DV separation of Identities and Links from attributes (Satellites) by design creates a scale-free network [17] and thus greatly reduces stress of incremental expansion (scalability) this supports expansion of scope. The structural changes are also additions (no deletes) so flexibility of designs is assured. The DV by design foster higher levels of performance by (decoupling dependencies and) allowing parallel data loads all the time from all source systems. By requiring input without any irreversible data alterations (ELT as opposed to ETL) from a data source into a raw data vault (as a permanent fully auditable system of records) Linstedt suggest typical loading speed of 100K rows per minute is normal (as a benchmark). Nature of the DV model preserves scalability and flexibility as well as performance of the data vault designs weather manual or automated.

4. Physical Data Vault Design Tool

The goal of this section is to present the prototype CASE tool the authors have implemented to support their methodology. The PDV (Physical Data Vault) design tool assists the designer in structuring a data vault, carries out physical design in a semi-automatic fashion, and allows for a core workload to be defined on the physical scheme.

The tool was developed using Larman's [68] method. Specifications of requirements can be presented as verbal descriptions of the model and use cases. Verbal Description: Need to make an application that will provide support to the process of designing a data warehouse. The data warehouse should provide an analysis of data from structured data sources and simple unstructured and semi-structured (xls, txt). Structured data should not be loaded in a separate staging database, but directly loaded into the data warehouse. Semi-structured data has to be further structured in the staging database and then loaded into the data warehouse. Unstructured data has to go through the process of textual analysis, and then loaded into staging database for some structuring, and then loaded into the data warehouse. The Figure 12 shows the observed use-cases for the PDV tool.

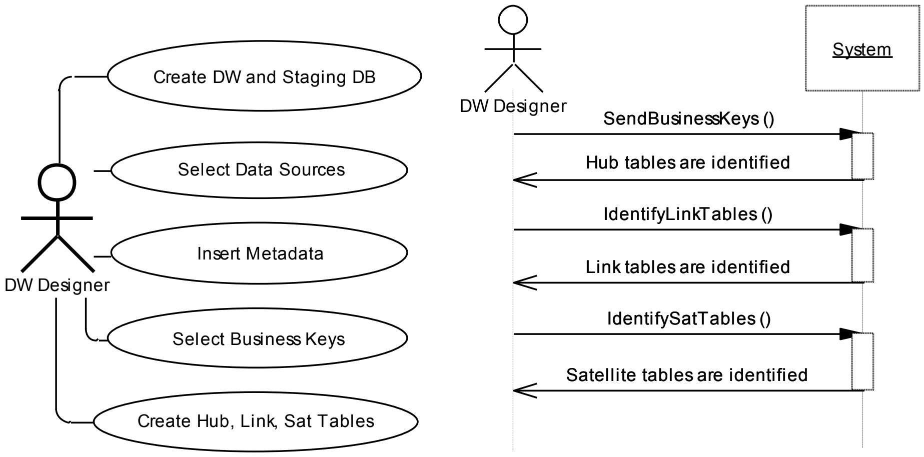


Fig. 12. Use-case diagram and Sequence diagram for the use-case select Business Keys

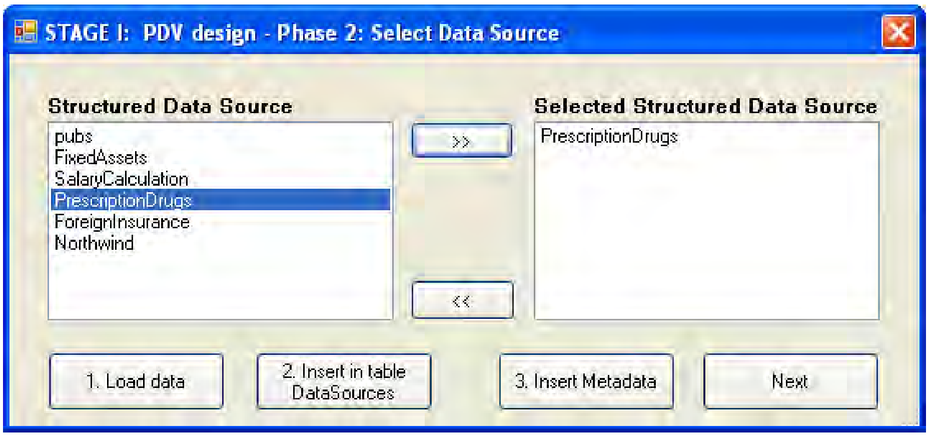
In the analysis phase, the behavior of software systems is determined by using system sequence diagram. For each use-case and for each scenario, a sequence diagram was created. The example in Figure 12 shows the sequence diagram for the use case Select Business Keys. Sequence diagram for use-case Identification of business keys, baseline scenario:

1. Designer sends business keys to system
2. The system returns the information based on business keys and identified Hub tables
3. Designer call the system to identify the Link tables
4. The system returns the information of identified Link table
5. Designer calls the system to identify the Sat tables
6. The system returns the information of identified Satellite tables

The architectural design includes the design of application logic, the user interface and the internal metadata model database. The tool is built in three layer architecture with the database layer, the user interface and business logic layer. The staging database was designed on the basis of the physical model shown in Figures 8 and 9. We used the Database management system Microsoft SQL Server 2008 R2. Designing the user interface included designing Windows forms. Some examples of the forms are given in the next section. The tool is implemented using Microsoft Visual Studio.NET environment using the C# programming language. The Microsoft's NET Framework was selected as it has a consistent programming model for building diverse applications [69].

5. Experimental verification of research results based on a prototype application

Experimental verification of research results was done in the area of health insurance using a prototype tool PDV for a data vault data warehouse design.


STAGE I: PDV design - Phase 2: Select Data Source
Structured Data Source
>>
<<
Selected Structured Data Source
1. Load data
2. Insert in table DataSources
3. Insert Metadata
Next

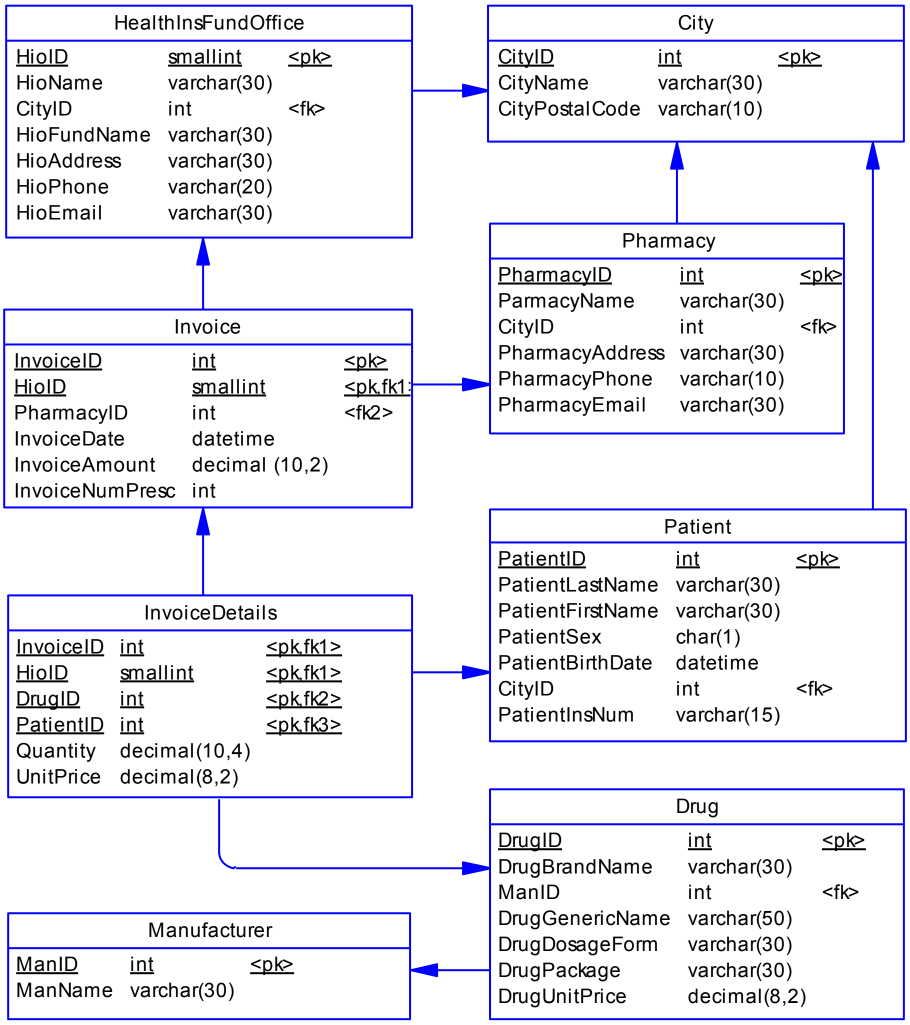
Structured Data Source:
pubs
FixedAssets
SalaryCalculation
PrescriptionDrugs
ForeignInsurance
Northwind

Selected Structured Data Source:
PrescriptionDrugs

Fig. 13. Form for selection of structured data sources

Health Insurance Fund annually enters into contracts with pharmacies to issue prescription drugs to insured patients. Prescription drugs are prescribed in health institution (family doctor’s office or hospital). Every fifteen or thirty days (depending on the contract), the pharmacy sends to Fund an invoice for drugs issued on prescriptions. Invoice consists of a header and items. The header contains the name of pharmacy, the date of the invoice, the total amount and the number of receipts. Items invoices contain information about drug, quantity and amount, and patient’s information. Invoices are to be submitted electronically. Health insurance Fund distributes databases in offices in each of several regions. The number of records in the table invoice items annually reaches several millions. To reduce the load on the transactional system, view reports (which often changes the format and appearance), to meet the requirements of users (in terms of reports with different grouping and diagrams), it was decided to implement a data warehouse and business intelligence system. When choosing a data warehouse architecture, the Data Vault approach was chosen (emphasizing the need to leave a trail from where and when the information originated from the databases). Moreover, a Data Vault is designed to model data that can be easily changed following rapid changes in the business environment.

After the first stage (initialized data warehouse and staging databases) associated procedures based on the model are derived from the second stage, the selection of data sources.



HealthInsFundOffice

HioID	smallint	<pk>
HioName	varchar(30)	
CityID	int	<fk>
HioFundName	varchar(30)	
HioAddress	varchar(30)	
HioPhone	varchar(20)	
HioEmail	varchar(30)	

City

CityID	int	<pk>
CityName	varchar(30)	
CityPostalCode	varchar(10)	

Pharmacy

PharmacyID	int	<pk>
ParmacyName	varchar(30)	
CityID	int	<fk>
PharmacyAddress	varchar(30)	
PharmacyPhone	varchar(10)	
PharmacyEmail	varchar(30)	

Invoice

InvoiceID	int	<pk>
HioID	smallint	<pk,fk1:
PharmacyID	int	<fk2>
InvoiceDate	datetime	
InvoiceAmount	decimal (10,2)	
InvoiceNumPresc	int	

Patient

PatientID	int	<pk>
PatientLastName	varchar(30)	
PatientFirstName	varchar(30)	
PatientSex	char(1)	
PatientBirthDate	datetime	
CityID	int	<fk>
PatientInsNum	varchar(15)	

InvoiceDetails

InvoiceID	int	<pk,fk1>
HioID	smallint	<pk,fk1>
DrugID	int	<pk,fk2>
PatientID	int	<pk,fk3>
Quantity	decimal(10,4)	
UnitPrice	decimal(8,2)	

Drug

DrugID	int	<pk>
DrugBrandName	varchar(30)	
ManID	int	<fk>
DrugGenericName	varchar(50)	
DrugDosageForm	varchar(30)	
DrugPackage	varchar(30)	
DrugUnitPrice	decimal(8,2)	

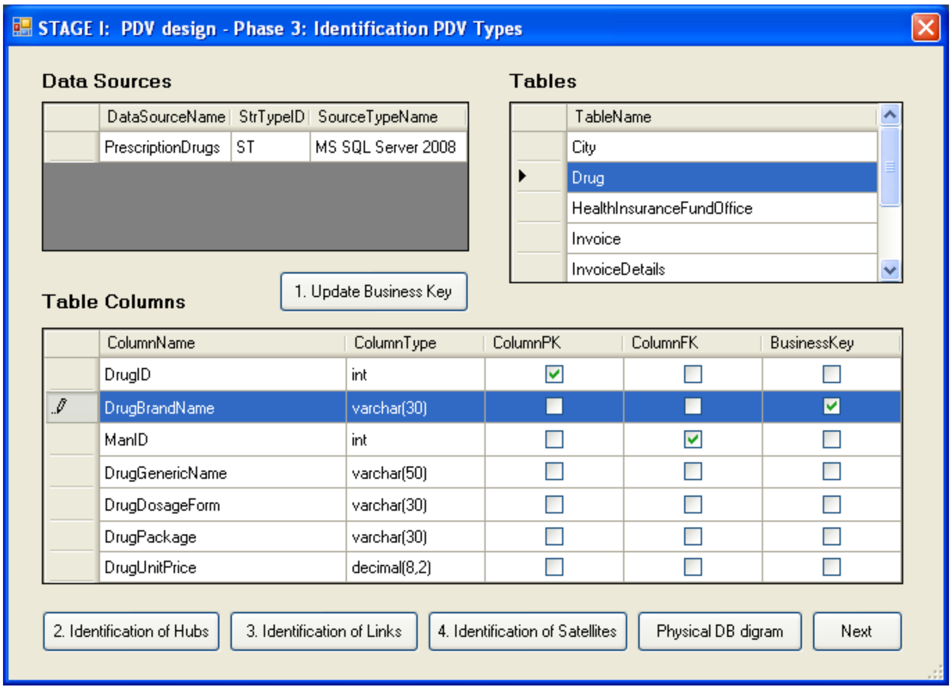
Manufacturer

ManID	int	<pk>
ManName	varchar(30)	

Fig. 14. Physical model of transactional database PrescriptionDrugs

After the selection of data sources, the user starts uploading the metadata of selected data sources into the staging database. Part of the physical model of transactional database PrescriptionDrugs is given in figure 14.

In the following form (Fig.15), users simultaneously (on three DataGridView controls) see selected data sources, data source tables and columns. The third DataGridView enables the checking of business keys. On the basis of the business keys and corresponding rules for Hub tables, the system identifies a Hub table. After that, based on the hub tables and rules, the system identifies link tables. At the end of this phase, system identifies Satellite tables. After clicking the Next button, a form opens to declare a PDV structure as shown in the figure 16.



Data Sources

DataSourceName	StrTypeID	SourceTypeName
PrescriptionDrugs	ST	MS SQL Server 2008

Tables

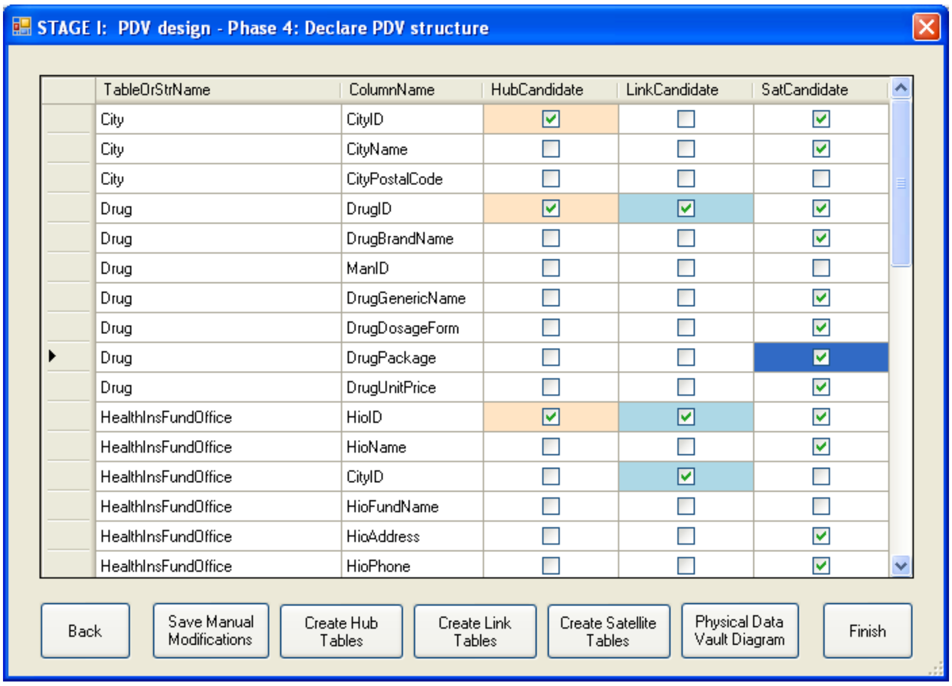
TableName
City
Drug
HealthInsuranceFundOffice
Invoice
InvoiceDetails

Table Columns

ColumnName	ColumnType	ColumnPK	ColumnFK	BusinessKey
DrugID	int	☑	□	□
DrugBrandName	varchar(30)	□	□	☑
ManID	int	□	☑	□
DrugGenericName	varchar(50)	□	□	□
DrugDosageForm	varchar(30)	□	□	□
DrugPackage	varchar(30)	□	□	□
DrugUnitPrice	decimal(8,2)	□	□	□

2. Identification of Hubs
3. Identification of Links
4. Identification of Satellites
Physical DB digram
Next

Fig. 15. Form for identification PDV types


STAGE I: PDV design - Phase 4: Declare PDV structure
Back
Save Manual Modifications
Create Hub Tables
Create Link Tables
Create Satellite Tables
Physical Data Vault Diagram
Finish

	TableOrStrName	ColumnName	HubCandidate	LinkCandidate	SatCandidate
	City	CityID	☑	□	☑
	City	CityName	□	□	☑
	City	CityPostalCode	□	□	□
	Drug	DrugID	☑	☑	☑
	Drug	DrugBrandName	□	□	☑
	Drug	ManID	□	□	□
	Drug	DrugGenericName	□	□	☑
	Drug	DrugDosageForm	□	□	☑
▸	Drug	DrugPackage	□	□	☑
	Drug	DrugUnitPrice	□	□	☑
	HealthInsFundOffice	HioID	☑	☑	☑
	HealthInsFundOffice	HioName	□	□	☑
	HealthInsFundOffice	CityID	□	☑	□
	HealthInsFundOffice	HioFundName	□	□	□
	HealthInsFundOffice	HioAddress	□	□	☑
	HealthInsFundOffice	HioPhone	□	□	☑

Fig. 16. Form for declare PDV structure

This form allows for manual modification of the proposed Hub, Link and Satellite table. After manual modifications, the designer starts the process of creating Hub, Link and Satellite tables that are just being identified. This step completes the process of creating a data warehouse based on the Data Vault concept. The next button is clicked to enable visualization of Data Vault data warehouse structure. The created Data Vault Physical model is shown in figure 17. According [15], in the tables can be generate a surrogate key - optional component, possibly smart key or sequential number, if the composite primary key might cause performance problems.

In our trials and experiments with the use of the PDV design tool (as it was evolving as a prototype itself) it was easy to create a complete data warehouses based on Data Vault concepts. In addition, the tool excelled when used to develop prototypes of data warehouses. In fact, only when a tangible DW prototype was completed, users become more interested in participating and frequently stated new requirements. Automating design for a data warehouse significantly accelerated the development of a robust system by allowing prototyping in the early stages of contact with customers, and customers were more interested in providing information.

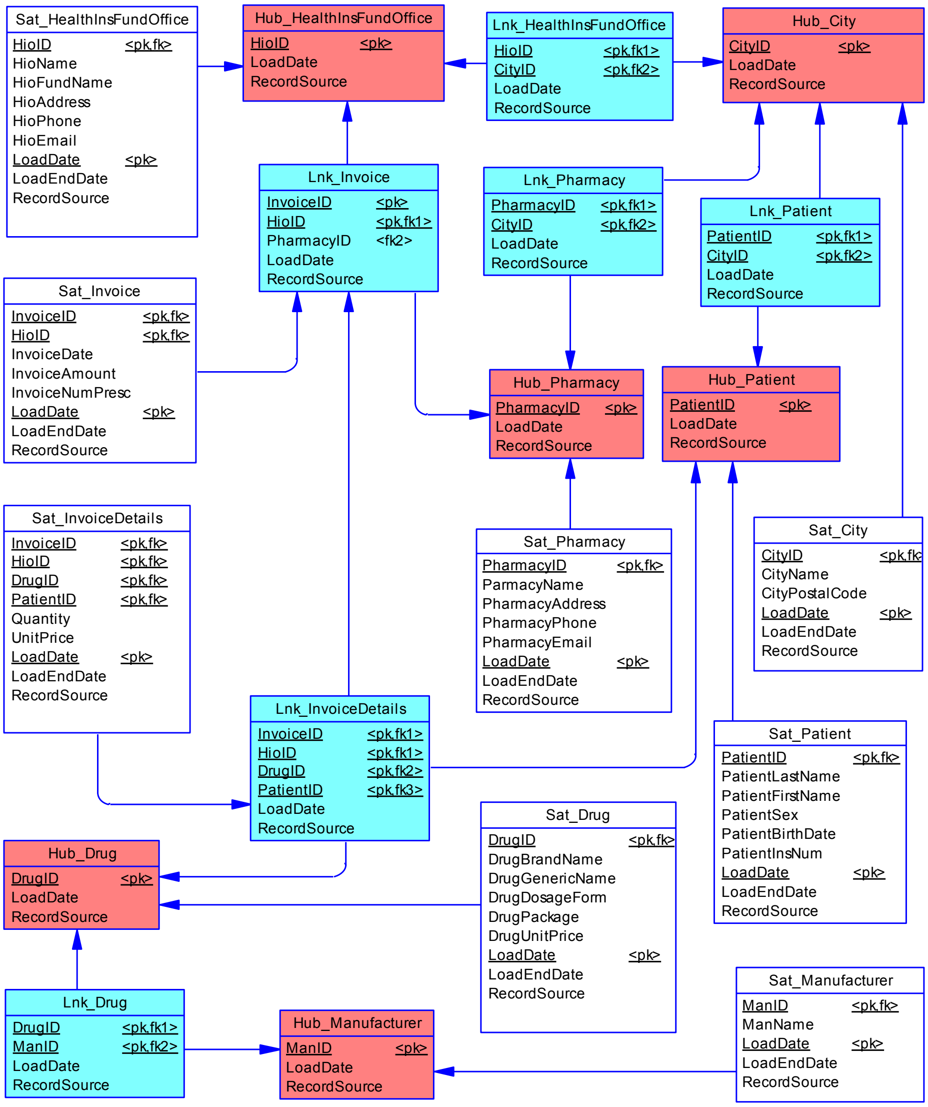


Fig. 17. Model of Physical Data Vault data warehouse

6. Conclusion

This paper presents the basic algorithm for the initial physical design stage of the Data Vault types of enterprise data warehouses i.e. integrated data warehouses as systems of records not open to end user reporting. The approach is based on the incremental expansion of data warehouse adding new data sources in sets or one at a time. The algorithm utilizes metadata model and rules for the design starting with existing (mainly) transactional data sources. Relations between entities in transactional systems and rules for the development of a data warehouse based on Data Vault concept are crucial for physical design automation of a data warehouse model. The conceptualization of metadata presented a physical model that can be used in the design of individual data warehouses, and this became the basis for development of a tool. The most important contribution of this paper is realization of Data Vault schema directly from RDBMS schemas. Such a direct approach was possible thanks to the feature of the Data Vault models i.e. separation of unchangeable identities of entities in real systems (Hubs) from time variant relationships among such entities (represented by Links) and the characteristics of such entities and their relationships (represented by Satellites). Traditional approaches to integration used pruning of data from the source and other forms of derivation i.e. consolidation that requires much intervention by experts (due to creative and semantically rich transformations). Data Vaults provide the unique ability to integrate data incrementally by adding links (of 'same as' type essentially 1:1 mappings) between initial and added hubs, while preserving all data in satellites, links and hubs without any reconstructions and deletes (guaranteeing preservation of information necessary for an enterprise size system of records). The subject of ongoing research is detailed specifications of dynamic expansion of Hub, Link (and their Satellite tables) and their additional linking for possible cases of merging Data Vault schemes in operation. Within the achieved scope, work in progress is focused on code generation for the initial loading of the created Data Vault enterprise data warehouses, as well as code generation for the Data Vault updating with new values and/or updates of the code to update (all without slowing down the original system).

The PDV approach is based on available relational schema and this satisfies meta-requirements stated earlier. Loading a schema and transforming it following preprogrammed rules certainly supports design performance, scalability and agility (user intervention is minimal but necessary, and is mainly focused on recognizing major permanent business keys). We claim that any indirect DV design driven by a conceptual or a logical data model (derived from the existing data sources) even when supported by some automation, is less flexible than direct PDV. Furthermore, it increases a danger of losing data from the source, potentially invalidating the central DV EDW purpose - to maintain a system of unaltered records.

Future work relate to the remaining three stages of Figure 5. First is the automation of the ETL process from data sources to feed a Data Vault. The Data Vault type data warehouse is a solution for integrating, storing and protecting data. However, it is not intended, nor suitable, for intensive queries or reports. That is why the data warehouse architecture with a Data Vault layer (persistent system of records with full history of changes) also contains a data mart layer using the star schemas for reporting data access [46]. According to [70] the dimensions of the star schema result from the Hub and Satellite tables, and the fact tables from a Satellite and Link tables. The next item of research is addressing the output area (data marts) with the following steps: Create a model of DMs and DMs materialization code and create metadata for Analytics Tracking (Dashboards/Scorecards) and standard reporting.

The process of designing an enterprise data warehouse based on the Data Vault model can be formalized, generalized and to some extent based on the automated physical model for structured, semi-structured and simple unstructured data sources, including transactional database. Our direct approach integrates elements of the physical design of enterprise data warehouses based on a data vault model as a system of records. This paper also illustrated the development of a tool for automation of design for data vault based enterprise data warehouses. The tool has been implemented and used on a real case in the field of healthcare and medical insurance and provided satisfactory results.

The paper and the approach presented so far do not address design of data marts, data virtualization, data warehouse schema evolution, master data management, mappings to NoSql data stores or hybrid databases, nor fully elaborates on ELT/ETL transformation automatization as those issues are part of a lager ongoing research program.

References

1. Inmon H. W.: Building the Data Warehouse. Wiley Computer Publishing. (1992)
2. Jarke M., Lenzerini M., Vassiliou Y, Vassiliadis P.: Fundamentals of Data Warehouses. Springer-Verlag. (2003)
3. Ćamilović D., Bečejski-Vujaklija D., Gospić N.: A Call Detail Records Data Mart: Data Modelling and OLAP Analysis. Computer Science and Information Systems, Issue: 12, page 87-110. (2009).
4. Krneta D., Radosav D., Radulovic B.: Realization business intelligence in commerce using Microsoft Business Intelligence. 6th International Symposium on Intelligent Systems and Informatics (SISY), pp. 1–6. (2008)
5. Larson B.: Delivering Business Intelligence with MSSQL Server 2008. Mc Graw Hill (2009)
6. Vassiliadis P., Simitsis A., Skiadopoulos S.: Conceptual Modeling for ETL Processes. In Proc. 5th International Workshop on Data Warehousing and OLAP, VA, USA. (2002)
7. Vassiliadis P., Simitsis A, Baikousi E.: A Taxonomy of ETL Activities. InProc. ACM 12th International Workshop on Data Warehousing and OLAP (DOLAP 2009), Hong Kong
8. Adelman S., Moss L., Abai M.: Data Strategy, Addison Wesley, New York. (2005)
9. Inmon W.H.: Building the Data Warehouse: Gettiing Started, White Paper BillInmon.com. (2000)
10. Inmon H.W., Strauss D., Neushloss G.: DW 2.0 The Arcitecture for the Next Generation of Data Warehousing, Morgan Kaufman. (2008)
11. Kimball R.: The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, Willey. (1996)
12. Kimball R., Ross M.: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, Wiley. (2002)
13. Mundy J., Thornthwaite W., Kimball R.: The Microsoft Data Warehouse Toolkit, Second edition, Wiley. (2008)
14. Rönnbäck L., Hultgren H.: Comparing Anchor Modeling with Data Vault Modeling, Modeling for the modern Data Warehouse, White paper. (2013)
15. http://www.tdan.com/view-articles, Data Vault Series 1-5 by Dan E. Linstedt. (2002)
16. Linstedt D.: Data Vault Modeling & Methodology, http://www.learndatavault.com. (2011)
17. Linstedt D.: Super Charge your Data Warehouse, Kindle Edition. (2010)
18. Jovanović V., Bojičić I.: Conceptual Data Vault Model, Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA. (March, 2012)

19. Hultgreen H.: Modeling the Agile Data Warehouse with Data Vault. New Hamilton. (2012)
20. Rönnbäck L.: Anchor Modeling – A Technique for Information Under Evolution, GSE Nordic Conference, Stockholm, Sweden. (2001)
21. Regardt, O., Rönnbäck, L., Bergholtz, M., Johannesson, P., Wohed, P.: Analysis of Normal Forms for Anchor Tables. (2010)
22. Knowles C.: 6NF Conceptual Models and Data Warehouses 2.0. Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA. (March 2012)
23. Date, C.J., Darwen, H., Lorentzos, N.: Temporal data and the relational model: A detailed investigation into the application of interval and relation theory to the problem of temporal database management. Morgan Kaufmann Publishers. (Amsterdam, 2002)
24. Graziano K.: Introduction to Data Vault Modeling. True BridgeResources, White paper. (2011)
25. Jovanović V., Bojičić I., Knowles C., Pavlić M.: Persistent staging area models for Data Warehouses. Issues in Information Systems, Volume 13, Issue 1, pp. 121-132. (2012)
26. Golfarelli M., Rizzi S.: Data Warehouse Design Modern Principles and Methodologies. McGraw-Hill. (2009)
27. Corr L., Stagnitto J.: Agile Data Warehouse Design. DecisionOne Press, Leeds. (2011)
28. Winter, R., Strauch, B.: A Method for Demand-Driven Information Requirements Analysis in DW Projects. In Proceedings of 36th Annual Hawaii International Conference on System Sciences. (2003)
29. Jensen, M., Holmgren, T., Pedersen T.: Discovering Multidimensional Structure in Relational Data. In Proceedings of DaWaK. (2004)
30. Guo Y., Tang S., Tong Y., Yang D.: Triple-Driven Data Modeling Methodology in Data Warehousing A Case Study. DOLAP '06 Proceedings of the 9th ACM international workshop on Data warehousing and OLAP. (2006)
31. Golfarelli M.: Data warehouse life-cycle and design. White Paper, DEIS – University of Bologna. (2008)
32. Boehnlein, M., Ulbrich v.E, A.: Business Process Oriented Development of Data Warehouse Structures. Proceedings of Data Warehousing 2000, Physica Verlag. (2000)
33. Westerman P.: Data Warehousing. Morgan Kaufman Publishers. (2001)
34. Romero O., Abello A.: A Survey of Multidimensional Modeling Methodologies. International Journal of Data Warehousing & Mining, 5(2), 1-23. (April-June 2009)
35. Jovanovic V., Marjanovic Z.: DW21 Research Program-Expectations, FON/Breza software engineering, White paper. (February 2013)
36. Golfarelli M., Mario D., Rizzi S.: Conceptual Design of Data Warehouses from E/R Schemes. System Sciences. (1998)
37. Boehnlein M., Ulbrich-vom Ende A.: Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems. 2nd Int. Workshop on Data Warehousing and OLAP. (1999)
38. Phipps C., Davis K. C.: Automating Data Warehouse Conceptual Schema Design and Evaluation. 4th International Workshop on Design and Management of DW. (2002)
39. Peralta V., Marotta A., Ruggia R.: Towards the Automation of Data Warehouse Design. 15th Conference on Advanced Information Systems Engineering, Velden, Austria. (2003)
40. Romero O., Abello A.: Automating Multidimensional Design from Ontologies, DOLAP'07. (November 2007)
41. Zepeda L., Ceceña E., Quintero R., Zatarain R., Vega L. , Mora Z., Clemente G.: A MDA Tool for Data Warehouse. International Conference on Computational Science and Its Applications. (2010)

42. Nazri M.N.M., Noah S.A., Hamid Z.: Using lexical ontology forsemi-automatic logical data warehouse design. 5th international conference on Rough set and knowledge technology. (2010)
43. Zekri M., Marsit I., Adellatif A.: A new data warehouse approach using graph. Eight IEEE International Conference on e-Business Engineering. (2011)
44. Damhof R.: The next generation EDW. Database Magazine (Netherlands). (2008)
45. Hammergren T. , Simon A.: Data Warehousing for Dummies. 2nd edition. (2009)
46. Casters M., Bouman R., van Dongen J.: Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration. Wiley Publishing. (2010)
47. Luján-Mora, S., Vassiliadis, P., Trujillo, J.: Data Mapping Diagrams for Data Warehouse Design with UML. (2004)
48. Vassiliadis, P., Simitsis, A., Skiadopoulos, S.: Conceptual Modeling for ETL Processes. In DOLAP. 2002.
49. Kimball, R., Caserta, J.: The Data Warehouse ETL Toolkit Wiley Publishing, Inc. (2004)
50. Simitsis A., Vassiliadis P., A method for the mapping of conceptual designs to logical blueprints for ETL processes. Decision Support Systems, vol. 45, no. 1. pp. 22–40. (2008)
51. Muñoz L., Maz n J., Trujillo J.: Systematic review and comparison of modeling ETL processes in data warehouse, Inf. Syst. Technol. (CISTI), 5th Iberian Conference. (2010)
52. Oliveira B., Santos V., Belo O.: Pattern-Based ETL Conceptual Modelling. International Conference on Model & Data Engineering, Calabria, Italy. (2013)
53. Jovanovic P., Romero O., Simitsis A., Abell A., Requirement-driven creation and deployment of multidimensional and ETL designs, Advanced Concept. Model. (2012)
54. http://technet.microsoft.com/en-us/library/ms141026.aspx
55. http://www.oracle.com/us/products/middleware/data-integration/enterprise-edition/overview/index.html
56. http://www-03.ibm.com/software/products/en/ibminfodata
57. http://www.informatica.com/us/products/data-integration/enterprise/powercenter/
58. Golfarelli M., Rizzi S., Saltarelli E.: WAND: A CASE Tool for Workload-Based Design of a Data Mart. SEBD. 2002.
59. Battaglia A., Golfarelli M., Rizzi S., QBX: A CASE tool for Data Mart design, O. De Troyer et al. (Eds.): ER 2011 Workshops, LNCS 6999, pp. 358–363, Springer-Verlag Berlin Heidelberg. (2011)
60. http://biready.com
61. http://www.birst.com/product/technology/data-warehouse-automation
62. http://www.datawarehousemanagement.org/Quipu.aspx
63. http://www.bi-podium.nl/mediaFiles/upload/DWHgen/Pentaho_en_DV_-_KdG.pdf
64. Date, C, Darwen, H., Databases, types and the relational model: the third manifesto (3rd ed.). Reading, MA: Addison-Wesley. (2006)
65. Rainardi V.: Building The Data Warehouse With Examples in SQL Server, Springer, New York. (2008)
66. Nicola M., Sommerlandt M., Zeidenstein K.: From text analytics to data warehousing. http://www.ibm.com/developerworks/data/library/techarticle/dm-0804nicola/ (2008)
67. Inmon W.H., Krishnan K.: Building the Unstructured Data Warehouse, Technic Publications LLC. (2012)
68. Larman Craig: Applying UML and Patterns: An introduction to object-oriented analysis and design, Second edition, Prentice Hall, New Jersey. (1998)
69. http://www.microsoft.com/net
70. Govindarajan S.: Data Vault Modeling The Next Generation DW Approach. http://www.globytes.com. (2010)

Dragoljub Krneta received the M.Sc. from Technical Faculty "Mihajlo Pupin" University of Novi Sad. Currently he is a PhD student in Faculty of Organizational Sciences, University of Belgrade, Serbia. He lives in Banja Luka, Bosnia and Herzegovina, where he works as a Software Development Manager in a company Lanaco Information technologies. Throughout his career in IT that lasts more than twenty years, he was actively involved in the design and development of information systems as a programmer, systems analyst, database designer, system architect or project manager. He has published 12 papers on international scientific conferences and journals. His research interests include information systems development, databases, data warehouses and business intelligence systems. He is a member of IEEE.

Vladan Jovanović, born in Belgrade Serbia, received all his degrees from the University of Belgrade: a) B. Eng. in Cybernetics Information Systems in 1975, b) M. Sci. in Computer Information Systems in 1978, and c) Ph.D. in Organizational Sciences-Software Engineering in 1982. Since 2001 works as a professor of Computer Sciences at the Georgia Southern University's College of Engineering and IT. Research interests: models, methodology and/or processes of data and software intensive systems design. Professional contributions: a) standardization within the IEEE Systems and Software Engineering Standardization areas of software architecture and design, SQA, and SW Processes, and b) curriculum development in IT, IS, CS, Systems Assurance and Software Engineering, as well as participating in CS and SE program accreditation, as an ABET Program Evaluator.

Zoran Marjanović is a full professor at Faculty of Organizational Sciences (FOS), University of Belgrade, Serbia and a president of Breza Software Engineering company. His research interests are IS development methodologies, databases, semantic enterprise application interoperability, and information systems development. Prof Marjanovic is a lead on several on-going projects with government and commercial entities that address design, deployment, and testing of ERP and other business systems. He received his MS and PhD degrees in Information Systems from University of Belgrade. He is a member of ACM and IEEE. E-mail: zoran.marjanovic@brezasoftware.com

Received: May 23, 2013; Accepted: June 05, 2014

Computer Science and Information Systems 11(2):601–622 DOI: 10.2298/CSIS130415035A

Flow-Based Anomaly Intrusion Detection System Using Two Neural Network Stages

Yousef Abuadlla[1], Goran Kvascev[2], Slavko Gajin[3], and Zoran Jovanović[3]

[1] School of Electrical Engineering, University of Belgrade,
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
abouadlla@gmail.com
[2] School of Electrical Engineering, University of Belgrade,
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
kvascev@etf.bg.ac.rs
[3] School of Electrical Engineering, University of Belgrade,
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
{slavko.gajin, zoran}@rcub.bg.ac.rs

Abstract. Computer systems and networks suffer due to rapid increase of attacks, and in order to keep them safe from malicious activities or policy violations, there is need for effective security monitoring systems, such as Intrusion Detection Systems (IDS). Many researchers concentrate their efforts on this area using different approaches to build reliable intrusion detection systems. Flow-based intrusion detection systems are one of these approaches that rely on aggregated flow statistics of network traffic. Their main advantages are host independence and usability on high speed networks, since the metrics may be collected by network device hardware or standalone probes. In this paper, an intrusion detection system using two neural network stages based on flow-data is proposed for detecting and classifying attacks in network traffic. The first stage detects significant changes in the traffic that could be a potential attack, while the second stage defines if there is a known attack and in that case classifies the type of attack. The first stage is crucial for selecting time windows where attacks, known or unknown, are more probable. Two different neural network structures have been used, multilayer and radial basis function networks, with the objective to compare performance, memory consumption and the time required for network training. The experimental results demonstrate that the designed models are promising in terms of accuracy and computational time, with low probability of false alarms.

Keywords: Intrusion Detection system, Anomaly detection system, Neural Network, NetFlow.

1. Introduction

With the rapid growth of the Internet and due to increase in number of attacks, computer security has become a crucial issue for computer systems. Intrusion Detection

Systems (IDS) technology is an effective approach in dealing with the problems of network security.

Intrusion Detection Systems (IDS) have become increasingly important in recent years to reveal the growing number of attacks. They need to be able to adapt to the rise in the amount of traffic as well as the increase in line speed [1]. However, researchers assess the payload-based IDSs processing capability to lie between 100 Mbps and 200 Mbps when commodity hardware is used [2, 3], and close to 1 Gbps when dedicated hardware is employed [4, 5]. Well-known systems like Snort [6] and Bro [7] exhibit high resource consumption when confronted with the overwhelming amount of data found in high-speed networks [8]. In addition, the spread of encrypted protocols poses a new challenge to payload-based systems. In addition to that, the constant increase in network traffic and the fast introduction of high speed (tens of Gbps) network equipment [9] make it hard to preserve traditional packet based intrusion detection systems. Such systems rely on deep packet inspection, which does not scale well.

Having this in mind, flow-based approaches seem to be a promising candidate for research in the area of IDS. A flow is defined as a unidirectional stream of packets that share common characteristics, such as source and destination addresses, ports and protocol type. Additionally a flow includes aggregated information about the number of packets and bytes belonging to the stream, as well as its duration. Flows data are often used for network monitoring, allowing us to obtain a real time overview of the network traffic. Common tools for this purpose are Nfsen [10] and Flowscan [11], while the *de facto* standard technology in this field is Cisco NetFlow, particularly its versions 5 and 9 [12], [13]. This technology is now becoming a formal standard through the work of the IETF IPFIX working group [14].

Network flows are monitored by specialized accounting modules usually placed in network routers. These modules are responsible for calculating flow statistics and exporting these statistics (flow-data) to external collectors. Flow-based IDSs analyze these flows data to detect anomaly and alarm possible attacks. Compared to traditional IDSs based on deep packet inspection, flow-based IDSs have to handle a considerably lower amount of data. Another important motivation for using flow-data in our research lays in the fact that flow-data is easily collected from network routers (several network nodes or just one central router) using standard protocols (such as Cisco NetFlow, Juniper Jflow, IETF IPFIX), without need to install additional software and collect data from every single computer on the network.

Therefore we have developed our approach by focusing on network flows. Approaches that rely on aggregated traffic metrics, such as *flow-based* approaches, show better scalability and therefore seem to be more promising.

In general, the techniques for Intrusion Detection (ID) fall into two major categories depending on the modeling methods used: misuse detection and anomaly detection. Misuse detection compares the usage patterns with known techniques of compromising computer security. Although misuse detection is effective against known intrusion types, it cannot detect new attacks that were not predefined. Anomaly detection, on the other hand, approaches the problem by attempting to find deviations from the established patterns of usage. Anomaly detection may be able to detect new attacks. However, it may also have a significant number of false alarms because the normal behavior varies widely and obtaining complete description of normal behaviors is often difficult. Architecturally, an intrusion detection system can be categorized into three types: host based IDS, network based IDS and hybrid IDS [15], [16]. A host based intrusion detection system uses the audit trails of the operating system as a primary data source. Network based intrusion detection systems, on the other hand, use network traffic information as their main data source. Hybrid intrusion detection systems use both methods [17].

In recent years, Neural Networks (NN) have been successfully used in the context of network intrusion detection. They have been extensively used in discriminating normal behavior from abnormal behavior in a variety of contexts. Neural networks have become a very useful technique to reduce information overload and improve decision making by extracting and refining useful information through a process of searching for patterns from the extensive collected data. Classification is a very common neural network task. In classification (section 3), we need to examine the features of newly presented objects and try to assign it to one of the predefined sets of classes. Supervised learning methods are applied to solve classification problems. Multilayer Feedforward NN (MLNN), and radial basis function (RBF) are representative supervised learning methods that can be applied to classification problems.

In this paper, flow-based anomaly IDS is implemented using two neural network stages. In many previous studies [18],[8],[19] the implemented system is a neural network based on DARPA [20] or KDD [21] dataset with the capability of detecting normal or abnormal traffic. In our study the existence of attacks and the attack type is classified by using extracted data from the labeled DARPA dataset. The restriction for the extracted data was that it should be limited to the type of data that can also be extracted from the router NetFlow data. This labeled data in the following text will be named labeled NetFlow dataset or simply NetFlow dataset, and the training procedure for neural networks is based on it.

This paper is organized as follows, section 2 present an overview of some of the previous works, section 3 provides a brief introduction about classification methods, section 4 explains the proposed system and feature selection, section 5 evaluates the proposed system, and section 6 discusses the experimental results followed by conclusions and future work plans.

2. Previous Work

Due to the increase in network speed, flow-based techniques attracted the interest of researchers, especially in analysis of high-speed networks. Day to day increase in network usage and load, have clearly pointed out that *scalability* is a growing problem. In this context, flow based solutions to monitor and, moreover, to detect intrusions help to solve the problem. They achieve, indeed, *data and processing time reduction*, opening the way to high-speed detection on large infrastructures. Gao and Chen [22] designed and developed a flow-based intrusion detection system. Karasaridis et al. [23], Shahrestani et al. [24] and Livadas et al. [25] proposed a concept for the detection of botnets in network flows. Sperotto et al. [1] provided a comprehensive survey on current research in the domain of flow-based network intrusion detection. A sound evaluation of a neural network based IDS requires high-quality training and testing datasets. The de facto standard is still the DARPA dataset created by Lippmann et al. [26]. Despite its severe weaknesses and the critique published by McHugh [25], it is still used. The KDD-99 [21] dataset can be regarded as another popular dataset. All these datasets were prepared for deep packet inspection and preprocessing was necessary to prepare flows. Sperotto et al. [28] created the first labelled flow based dataset intended for evaluating and training flow based network intrusion detection systems.

Several Neural Network approaches were implemented for Intrusion Detection systems based on NetFlow and DARPA [20] dataset. Muna Mhammad T. Jawhar [29] used Neural Networks and Fuzzy C-Mean (FCM) clustering algorithms. Rodrigo Braga [30] used OpenFlow and the SOM unsupervised neural network. Vallipuram and Robert [31] used back-propagation Neural Networks having all features of KDD (Knowledge Discovery in Databases) data [21]. Tie and Li [32] used the back propagation (BP) network with Genetic Algorithms (GAs) to enhance BP, for selected attacks and some features of the KDD dataset as input. Mukkamala, Andrew, and Ajith [33] used Back Propagation Neural Network with many types of learning algorithm. Jimmy and Heidar [34] used Neural Network for classification of unknown attacks. Dima, Roman and Leon [35] used MLP and Radial Based Function (RBF) Neural Network for classification of five types of attacks. Iftikhar, Sami and Sajjad [36] used Resilient Back propagation algorithm for detecting network intrusion attacks in a precise way by using the power of RPROP (Resilient Backpropagation) learning algorithm.

3. Detection and Classification Methods

Neural Networks (NNs) have recently attracted more attention compared to other techniques due to their strong discrimination and generalization abilities, when utilized for classification purposes [37], especially in the case of large amounts of data. An increasing amount of research in the last few years has investigated the application of Neural Networks to intrusion detection. If properly designed and implemented, Neural Networks have the potential to address many of the problems encountered by rule-based approaches. Neural Networks were specifically proposed to learn the typical characteristics of system's users and identify statistically significant variations from their established behavior. In order to apply this approach to Intrusion Detection, several steps have been taken:

Learning of Neural network: introducing data representing attacks and normal network flow to the Neural Networks to adjust the coefficients of these Networks automatically during the training phase. In other words, it will be necessary to collect data representing normal and abnormal behavior to train the Neural Networks.

Testing: after training has been accomplished, a certain number of performance tests with real network traffic and attacks have been conducted [38]. Instead of processing sequentially, Neural Network based models simultaneously explore several hypotheses through the use of several computational interconnected elements (neurons); this parallel processing may imply time savings in malicious traffic analysis [39]. In our study and based on prior research, two different neural network methods have been used for our intrusion detection system: Multilayer Feedforward neural network (MLFF), and Radial Basis Function Network (RBFN), [40]. NetFlow data has been used for training of these neural networks, while in many previous studies [18], [8], [19]; the implemented system is a neural network with the use of other features from the DARPA dataset that don't exist in NetFlow data [20].

3.1. Multilayer Feedforward NN

A multilayer feedforward NN is a structure that maps sets of input data onto a set of appropriate outputs. An MLFF consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one as shown in Figure 1. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLFF utilizes a supervised learning technique called backpropagation for training the network [41], [42] and has the ability to classify data that is not linearly separable [43].

The training algorithm rule repetitively calculates an *error function* for each input and backpropagates the error from one layer to the previous one. The weights for a particular node (w_{ij}) are adjusted in direct proportion to the error in the units to which it is connected.

The error function for pattern p is defined as to be proportional to the square of the difference of desired output d_{pj} (j-th node) and actual output y_{pj} for all nodes in output layer ($j=1,...,n$). The backpropagation algorithm implements weight changes that follow the path of steepest descent on a surface in weight space. The height of any point on this surface is equal to the error measure E_p. This can be presented by showing that the derivative of the error measure with respect to the fact that each weight is proportional to the weight change dictated by the delta rule, with a negative constant of proportionality, i.e.,

(1)

The most used training algorithm is back propagation algorithm gradient descent (GDA) with the disadvantage of slow training. In other hand Levenberg-Marquardt [44], [45] is one of the accurate algorithms and faster than GDA, but consumes more memory space.

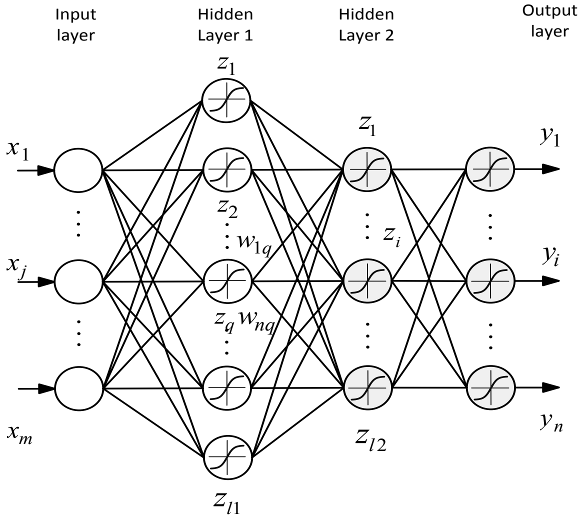


Fig. 1. Multilayer Feedforward NN, *m* – inputs (x_j), *l1,l2* – neurons in hidden layers (z_q), *n* – outputs (y_i), with different activation functions *f* for hidden and output layer

3.2. Radial Basis Function Network

The radial basis function network (RBFN) [46],[40] has the architecture of the instar-outstar model (Figure 2) and uses the hybrid unsupervised and supervised learning scheme, unsupervised learning in the input layer and supervised learning in the output layer.

The purpose of the RBFN is to pave the input space with overlapping receptive fields. For an input vector **x** lying somewhere in the input space, the receptive fields with centres close to it will be appreciably activated. The output of the RBFN is then the weighted sum of the activations of these receptive fields. The RBFN is designed to perform input-output mapping trained by examples, pairs of inputs and outputs (***x***, ***y***). The hidden nodes in the RBFN have normalized Gaussian activation function.

$$z_q = g_q(\mathbf{x}) = \frac{R_q(\mathbf{x})}{\sum_k R_k(\mathbf{x})} = \frac{\exp\left(-\frac{|\mathbf{x}-\boldsymbol{m}_q|^2}{2\sigma_q^2}\right)}{\sum_k \exp\left(-\frac{|\mathbf{x}-\boldsymbol{m}_k|^2}{2\sigma_k^2}\right)} \tag{2}$$

where **x** is the input vector and z_q output of hidden layer. m_q and σ_q are the mean (an *m*-dimensional vector) and variance of the *q*-th Gaussian function in hidden layer.

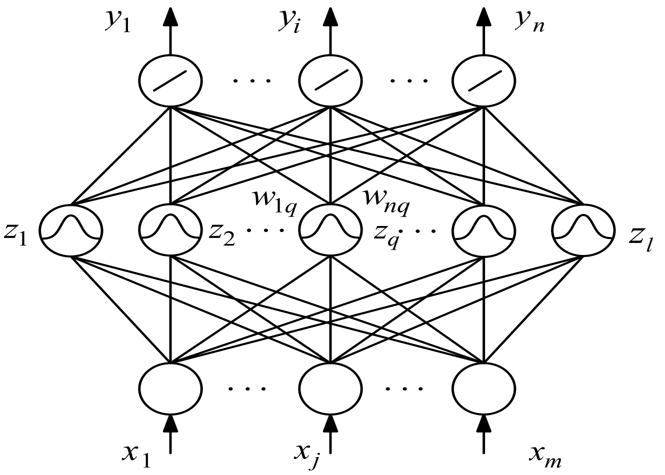


Fig. 2. Radial basis function network structure, with *m* inputs, *l* nodes in hidden layer, and *n* outputs

The output of the RBFN is simply the weighted sum of the hidden node output:

$$y_i = f_i\left(\sum_{q=1}^{l} w_{iq} z_q + \theta_i\right) \tag{3}$$

where $f_i(.)$ is the output activation function, generally linear function, and θ_i is the threshold value.

The weights in the output layer can be updated simply by using the delta learning rule (supervised learning). The unsupervised part of the learning involves the determination of the receptive field centres m_q and widths σ_q, $q = 1, 2... l$. The proper centres m_q can be found by unsupervised learning rules such as the vector quantization approach, competitive learning rules, or simply the Kohonen learning rule. Another learning rule for the RBFN with node-growing capability is based on the orthogonal least squares learning algorithm [47]. This procedure chooses the centres of radial basis functions one by one in a rational way until an adequate network has been constructed, or maximal number of nodes is reached.

The RBFN offers a viable alternative to the two-layer neural network in many applications of signal processing, decision making algorithms, pattern recognition, control, and function approximation. It has been shown that the RBFN can fit an arbitrary function with just one hidden layer [48], but they cannot quite achieve the accuracy of the back-propagation network. Although, RBFN can be trained several orders of magnitude faster than the back-propagation network, and this is a very important advantage in real or semi real time applications.

4. The Proposed NN Based Two Stages System

Our proposed system for intrusion detection and classification (Figure 3) consists of the following five main modules:

- Flow collector module.
- Feature preparation module.
- Anomaly detection module (NN stage one).
- Detection and classification module (NN stage two).
- Alert module.

NetFlow enabled routers are considered as external devices which permanently monitor network traffic, account statistics, and export flow-data to our system according to Cisco NetFlow [12], [13] or similar protocols.

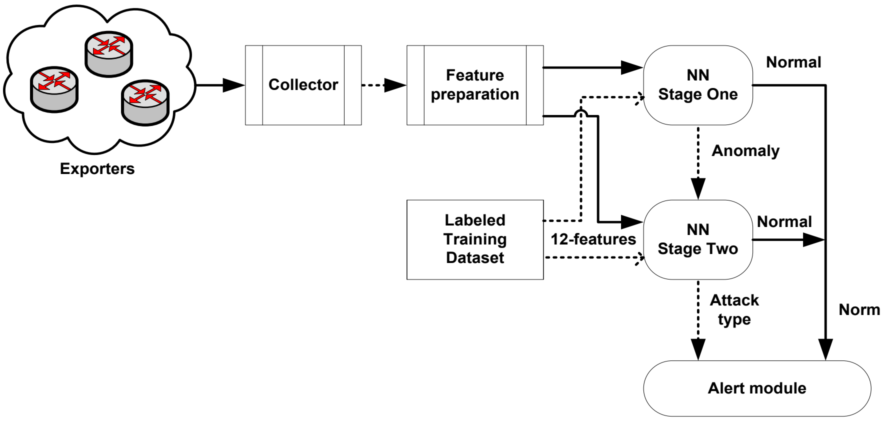


Fig. 3. Implemented two stages NN based system

4.1. Flow Collector Module

The operation of this module is to collect flow-data exported from one or several exporters. The received data need to be recognized by protocol and version (for instance NetFlow version 5 or 9, J-flow or IPFIX) and transform into an internal format. These data are constantly being sent to the Feature preparation module.

4.2. Feature Preparation Module

The Feature preparation module receives and processes flow-data sent from the Flow collection module. The main function is to prepare the features that are important for anomaly intrusion detection and classification modules. The features are combined in two groups: 7-*tuples and 12-tuples* that are passed to stage one and stage two detection modules respectively. Section 4.2.1 gives a more detailed explanation of the stage one features and section 4.2.2 demonstrate the added features for stage two modules. The selected features for both stages are suitable only for the selected attack in our study, and many other attacks have deviations of these features. Preprocessing must be done on all selected features before passing them to the detection modules; this phase involves normalizing all features by mapping all the different values for each feature to [0, 1] range.

4.3. Stage one Features

In stage one there are features that have most influence on the decision whether a traffic flow is normal or abnormal (an attack). It can also have the role to warn the network administrators that there is unusual traffic. A more detailed description of the vector of features is as following:

Average Flow Size: it provides a useful hint for anomalous events, such as port scan, and it is typically very small in order to increase the efficiency of attacks.

Average Packet Size: another factor is the size of each packet in the flow; low average size can be a sign of anomaly. For example, in TCP flooding attacks, packets of 120 bytes are typically sent.

Average Packet Number: one of the main features of DoS attacks is the source IP spoofing, which makes the task of tracing the attacker's true source very difficult. A side effect is the generation of flows with a small number of packets, i.e. about 3 packets per flow. This differs from normal traffic that usually involves a higher number of packets per flow.

Number of different flows to the same Destination IP: This feature counts the number of flows to the same destination IP address. A high number of flows could mean a flood attack or a port scan attack.

Number of flows to different Destination Ports: also it has influence on detecting attacks. An abnormally large number of different destination ports means that the system is probably under attack (port scan attack).

Land: this feature is responsible for checking whether there is a land attack in the network or not.(i.e.SrcIP=DestIP,SrcPort=DestPort)

SYN - SYN/ACK: this feature was used by many researchers [49] to detect DoS Attack, by comparing the numbers of SYN and SYN/ACK packets that a host receives and returns respectively. Under normal conditions, the two numbers should be balanced since every SYN packet is answered by a SYN/ACK packet. Consequently, a high number of unanswered SYN packed is an indication of ongoing SYN flood.

4.4. Stage Two features

In addition to the stage one feature there are five additional features that can significantly improve both the detection rate and the classification of the type of attack. Warnings and alerts to network administrators are created from this stage. The classification done by the system can be verified by inspecting complete corresponding flows when new types of attacks appear or when the network administrator wants to verify the classification done by stage 2. The additional five features are:

Number of flows from the same source IP: attacker can send for example ICMP ping packets to every possible address within a subset and wait to see which machine respond.

Number of flows from different source IP: IP spoofing is widely used by attackers to attack the networks. A high number of different IP addresses to the same destination address within a short period of time is a strong sign for attack (DoS/ DDoS attack).

Number of flows to the same Destination Port: in some cases the attacker sends GET request to some ports only (ex. Port 80) to crash the server.

Number of flows from different source Port: As IP spoofing is generated by DDoS attack; ports can also be changed during an attack at random.

Protocol type (TCP, UDP, and ICMP): knowing the protocol type in combination to the all previous features can help to determine the type of attack.

All added features have an important role in improving detection and classification of attacks. In order to have a full picture of what's going on in the network, a full history of the source address, used ports, and protocol type should be considered in many cases.

4.5. Anomaly Detection Module (Stage one NN1)

Anomalies in our system are defined as unusual activities in the network. The purpose of this module is to find out such activities using a small number of features extracted from NetFlow raw data. For the neural network that used in stage one the algorithm below is a simplified general description of the detection process.

```
Algorithm: Anomaly detection module

Loop
Read 7-tuple inputs from the feature preparation module.
Feed parameters to the NN1
If the data is "normal", then
Assign "01" to the output of the NN1 as normal traffic
      Else
Assign "10" to the output of NN as anomaly activity.
Call stage two Procedure.
End Loop
```

The number of input nodes of the NN1 corresponds to the number of the selected features of the NetFlow dataset for the first stage (7 Features). The implemented NN1 includes one input layer, one hidden layer and an output layer of 2 nodes (01 as normal traffic, and 10 as anomaly traffic). The number of nodes in the hidden layers has been determined based on the back propagation (BP) computation process and the process of trial and error.

4.6. Attack Detection and Classification Module (Stage two NN2)

In our work, neural networks were used for attack classification. It was crucial to successfully train the network with reliable data gained from classified attacks and then to test the trained network. The result from the network is classified into one of five possible categories. Table 1 maps these categories to the actual outputs from Neural Network module NN2.

Table 1. Neural Network Classified Categories

No	Category	NN2 outputs
1	DoS/ DDoS Attack	10000
2	Port Scan Attack	01000
3	Land Attack	00100
4	Other/unknown Attack	00010
5	Normal	00001

The number of input nodes to the NN2 corresponds to the number of the selected features from NetFlow dataset for the second stage NN2 (12 Features). The implemented neural network includes one input layer, one hidden layer and an output layer of 5 nodes (Table 1 contains the descriptions of the outputs). The numbers of nodes in the hidden layers has been determined based on the back propagation (BP) computation process and the process of trial and error. Table 2 describes the detection and classification procedure.

Table 2. Detection and Classification Procedure

```
Stage two Procedure

Begin
Read corresponding 12/tuple inputs for
NN2
If the data is "normal", then
Assign "00001" to the output
of NN2
Else
Assign appropriate attack category to
the NN2 outputs according to Table 1.
End
```

4.7. Alert Module

This is the final stage of the proposed system. This stage involves identifying the events that occurred. It also presents status of the observed network to the administrator and creates alarms when appropriate.

5. Experimental Results

The experiments were performed in MATLAB, using neural network toolbox which implements several training algorithms including Resilient Backpropagation, Radial Basis Function net, and Levenberg-Marquardt.

Considering the fact that the previous works commonly use DARPA dataset as a trusted labeled dataset for intrusion detection research, we built our NetFlow dataset as a subset of DARPA dataset. Since DARPA dataset is in form of TCP dump data, therefore we created flows from the raw DARPA dataset using a modified version of softflowd [50]. In our dataset, a flow closely follows the NetFlow v5 definition and has the following form:

$$F= \{IPsrc,IPdst,Psrc,Pdst,Pckts,Octs,Flags,Protcl,Tstart,Tend\}$$

It represents the unidirectional communication from the source IP addresses IPsrc and port number Psrc to the destination IP address IPdst and port number Pdst, using protocol type Protcl. The Pckts and Octs give the total number of packets and octets transferred during this communication. The field Flags is related to the TCP header flags which are computed as a binary OR of TCP flags in all packets of the flow. The start and end time of the flow are given by Tstart and Tend respectively, in millisecond resolution. The extracted flows are labeled according to the log file of DARPA dataset and used to prepare all selected features, which used to train NN1 and NN2.

In the experimental stages we have used different number of iterations and hidden layers to determine the level of training. This test has been done to find out when the neural network was trained properly to detect attacks. This test has also provided the background for choosing the number of hidden layers and iterations for the training of the neural network for the last experiments.

The experiments show that Levenberg-Marquardt is the best training algorithm because it takes less time, low number of epochs and has good performance and high accuracy. The Detection Rate (DR) and False Positive rate (FP) have been calculated for different scenarios. The considered scenarios in our experiments are as follows:

1. Anomaly Detection Phase: This case includes detection of attacks by deciding whether flows are normal or abnormal. The training and testing have been performed with seven selected features. One input layer, one hidden layer, and one output layer have been used in NN1 module.

The experiments have three phases: a training phase, a validation phase and a testing phase. All experiments in this stage were done with 96852 records of attack traffic, and 48556 of normal traffic. 21816 records were used for testing the neural network and it contains 14527 records of attack traffic, and 7289 records of normal traffic. Table 3 and Figure 4 show the performance of anomaly detection module. Detection rate and false positive rate has been calculated according to the following formulas:

$$Detection\ Rate = \frac{number\ of\ detected\ attacks}{total\ number\ of\ attacks} \times 100[\%] \tag{4}$$

$$False\ positive\ Rate = \frac{number\ of\ normal\ classifed\ as\ attacks}{total\ number\ ofnormal\ traffic} \times 100[\%] \tag{5}$$

Table 3. The Results of Anomaly Detection phase

Training Algorithm	Resilient Backpropagation Test 1	Levenberg-Marquardt Test2	Radial Basis Function Net Test 3
Parameters			
Training dataset	101806		
Validation Data	21816		
Testing set	21816		
Hidden Layer	50	50	20
Number of detected attacks(14527)	13468	13684	13234
Number of detected traffic as normal(7289)	6757	6866	6640
Detection Rate	92.7%	94.2%	91.1%
False positive Rate	3.6%	3.4%	5.1%

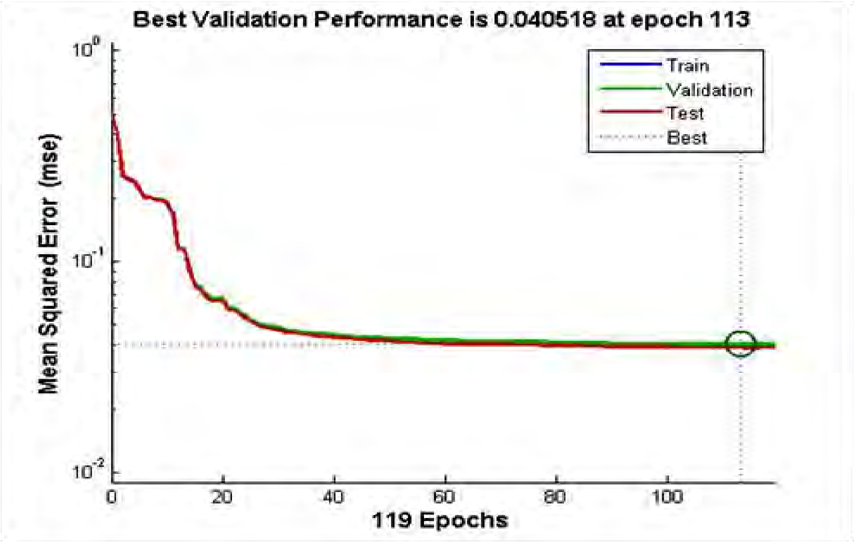


Fig. 4. Performance of the anomaly detection module (stage 1)

2. Detection and Classification Phase: This scenario includes detection of packets and their classification as normal with a larger number of inputs, or one of the four main attack types (DoS/DDoS, Port Scan, Land attack, or other/unknown attack). 12 selected features were used for training and testing the neural network. One input layer, one hidden layer, and output layer. All experiments in this stage have been done with 96852 records of attack traffic, and 48556 of normal traffic. 21816 records were used for testing the neural network and it contains 14527 records of attack traffic, and 7289 records of normal traffic. Table 4 and Figure 5 show the performance of the detection and classification module. Detection rate and false positive rate was also calculated. During the testing phase, the classification rate of each attack types was calculated according to the following formula:

$$Classification\ Rate = \frac{number\ classifed\ \ attacks}{total\ number\ of attack\ type} \times 100[\%] \quad (6)$$

The best result of the classification module during the test phase is shown in table 5.

Table 4. Results of detection and classification module

Training Algorithm Parameters	Resilient Backpropagation Test 1	Levenberg-Marquardt Test2	Radial Basis Function Net Test 3
Training dataset	96852		
Validation Data	21816		
Testing set	21816		
Hidden Layer	80	80	40
Number of detected attacks(14527)	14439	14443	13858
Number of detected traffic as normal(7289)	7145	7246	6953
Detection Rate	99.4%	99.42 %	95.4%
False positive Rate	0.3%	0.32%	2.6%

Table 5. Classification results of Stage 2

Attack name	Total number of attacks	Number of classified attacks	Classification rate
DoS	4490	4490	100%
Port scan	9929	9919	99.9%
Land	85	85	100%
Unknown	23	18	78%

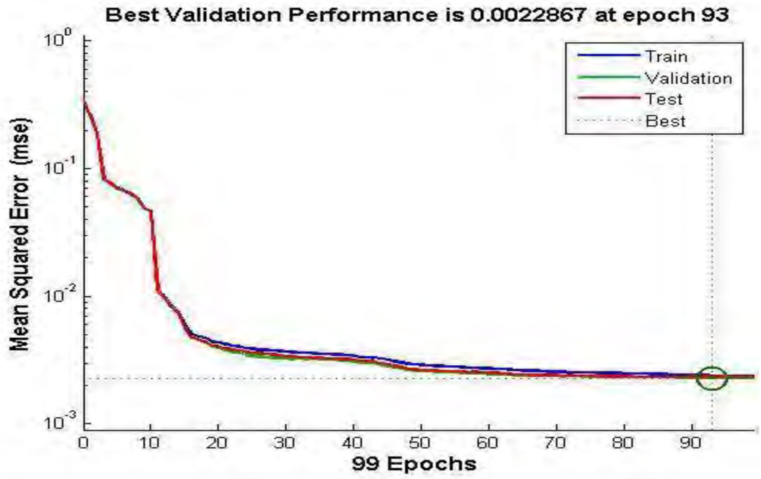


Fig. 5. Training performance of the detection and classification module (stage 2)

6. Discussion of Results

Two stages of Anomaly detection systems using neural networks and based on NetFlow dataset have been proposed and tested. Three different training algorithms (Resilient Backpropagation, Radial Basis Function net, and Levenberg-Marquardt) were used for training of both neural network stages. Anomaly detection stage (NN1) was trained until the best validation performance 0.0405 was met at epoch 113 as shown in Figure 4. The results in Table 3 show that the detection rate is 94.2% with false positive of 5.8%. Results from detection and classification stage (NN2), show significantly larger improvement of prediction accuracy than the Anomaly detection phase. Figure 5 shows that, the best validation performance 0.0022 was met at epoch 93.Table 4 shows that the detection rate is relatively high at 99.42% for MLP, and 95.4% for RBF detection algorithm. The false alarms were as low as 0.58% in MLP neural network and 4.6% in RBF neural network. Table 5 shows that, 100% of DoS attack, 99.9% of port scan attack, 100% of land attack, and 78% of unknown attack were detected and classified correctly by using stage two neural networks. The analysis of both stages results shows that, MLP with Levenberg-Marquardt is found to be fast compared to Resilient Backpropagation, has low memory consumption compared to Radial Basis Function, and has a lower false alarm rates.

6.1. Comparison of Results

In this section, we compare our results with the other researcher's results available in the literature. Vallipuram and Robert [31] used backpropagation neural network based on KDD'99 dataset, the detection rate was 86% with high false alarm rate at 14%.

Mukkamalaa [51] used backpropagation neural networks with the use of DARPA dataset, and the detection rate was 97.04% and false alarm rate of 2.06%. Dima, Roman, and Leon [35] used both MLP and RBF neural network with KDD'99 as a dataset, their results was 93.2% for RBF, and 92.2% for MLP with 7.2% as false alarm. Muna Mohammad [29] used MLP AND Fuzzy-clustering algorithm with the use of DARPA dataset, the detection rate was 99.9% and low false alarm rate 0.1%. Rodrigo Braga [30] used unsupervised neural network with flow dataset and the results for detection rate was 99.11% with false alarm rate of 0.99%. Govindarajan and Chandrasekaran [53] used neural based hybrid classification methods and they used flow dataset, their results were 96.67% for abnormal traffic, and 96.54% for normal traffic. Prasanta, Bhattacharyya, Borah and Jugal [54] used both supervised and unsupervised neural network with the use of flow dataset and KDD'99 dataset, the detection rate were 99.1% for flow dataset and 92.26% for KDD'99 dataset with false rate of 0.9%.

In our research with two neural network stages based on extracted NetFlow dataset, we have achieved the detection rate at 99.4% for MLP, and 94.6% for RBF neural network with low false alarm rate at 0.6%. Figure 6 shows that our proposed system is greatly competitive and performs significantly better Detection Rate (DR). From Table 6, we observe and conclude that our system with two neural network stages based on flow dataset and the use of a small number of extracted features can effectively and efficiently detect and classify both known and unknown attacks. The obtained false alarm rate is low compared to other methods that use different techniques and different datasets.

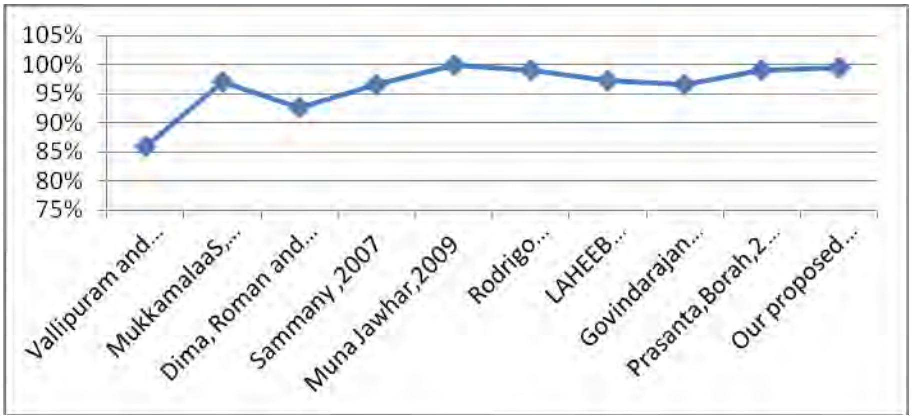


Fig. 6. Detection Rate on Different Datasets for IDSs

Table 6. Comparison of Intrusion Detection Systems Using NN.

Research	NN type	Dataset used	Detection Rate (%)	False Alarm Rate
Vallipuram and Robert,2004	Backpropagation	KDD-99	86% for normal traffic	14%
[51] Mukkamalaa S., 2005	Backpropagation	DARPA	97.04%	2.06%
Dima, Roman and Leon,2006	MLP and RBF	KDD-99	93.2% using RBF and 92.2% using MLP	8.8%
[52] Sammany M,2007	2 hidden layers MLP	DARPA	96.65%	3.35%
Muna Mhammad T. Jawhar,2009	MLP and Fuzzy C-Mean (FCM) clustering algorithms	DARPA	99.9%	0.1%
Rodrigo Braga,2010	SOM	Open flow dataset	99.11%	0.99%
Laheeb Mohamad Ibrahim,2010	Distributed Time-Delay Neural Network	KDD-99	97.24%	2.76%
[53] Govindarajan , Chandrasekaran,2011	hybrid classification methods	Flow data set	96.67% for abnormal traffic, and 96.54% for normal traffic	3.33%
[54] Prasanta Gogoi, Bhattacharyya,Borah and Jugal Kalita,2013	Supervised and unsupervised neural network	Packet Level and Flow Level dataset, KDD-99	99.1% for packet/flow level data, and 92.26% for KDD	0.9%
Our proposed IDSs	Two stage neural network	NetFlow dataset	94.2% for stage one NN, and 99.4% for stage two NN	0.6%

7. Conclusion and Future Work

In this paper we presented flow based intrusion detection and classification method using two neural networks for separate tasks. One neural network detects traffic anomalies that can be attacks and the other one classifies attacks if they exist. This system can easily be extended, configured, and/or modified by replacing some features or adding new features for new types of attacks.

The training of the NNs modules requires a very large amount of NetFlow data with known types of attacks and considerable time to ensure that the results from the NNs are accurate. The changes in patterns of usage of the network should not be undetected, but at the same time, these changes are isolated to NN1. Appearance of new patterns of attack affects only classification in NN2, which is the main reason to have two stage neural networks instead of one. Consequently, the events that require retraining for the two networks are completely independent. Experiments with different NNs were crucial to define the NN which yields the best classification and training speed results for both NN stages.

The experimental results of the proposed method prove that the use of NetFlow dataset and extracting only features that significantly contribute to intrusion detection gives promising results. The obtained detection rate (94.2% for anomaly detection at stage one, and 99.4% for classification at stage two) is remarkably good compared to other approaches, which use larger training sets [20]. These results are comparable to the best researches that are based on a similar approach using the same type of training dataset.(Table 6).

The multilayer Feedforward neural network has a better classification ability compared to RBFN, but memory and time consumption is 3-5 times greater. Otherwise, RBFN has a simple architecture and hybrid learning algorithm which leads to less time/memory consumption and it is better for working in real-time and for retraining with new data.

Our future research will be directed towards developing a more accurate model that can be used in real-time for detecting and classifying anomaly with minimum features and less time for training.

References

1. A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, and B. Stiller." Anoverview of ip flow-based intrusion detection". IEEE Communications Surveys &Tutorials, 12(3):343–356, (2010).
2. H. Lai, S. Cai, H. Huang, J. Xie, and H. Li. A parallel intrusion detection system for high-speed networks. In Proc. of the 2nd Int. Conf. Applied Cryptography and Network Security, pages 439–451,(May 2004).
3. M. Gao, K. Zhang, and J. Lu. Efficient packet matching for gigabit network intrusion detection using TCAMs. In Proc. of 20th Int. Conf. on Advanced Information Networking and Applications (AINA'06), pages 249–254, (2006).

4. W. de Bruijn, A. Slowinska, K. van Reeuwijk, T. Hruby, L. Xu, and H. Bos. Safe- Card: A Gigabit IPS on the Network Card. In Proc. of the 9th Int. Symp. on Recent Advances in Intrusion Detection (RAID '06), pages 311–330,(2006).
5. G. Vasiliadis, S. Antonatos, M. Polychronakis, E. P. Markatos, and S.Ioannidis.Gnort: High Performance Network Intrusion Detection Using Graphics Processors.In Proc. of the 11th Int. Symp. on Recent Advances in Intrusion Detection, pages 116–134, (2008).
6. M. Roesch. Snort, intrusion detection system. http://www.snort.org, Sept. 2010.
7. V. Paxson. Bro: a system for detecting network intruders in real-time. Computer Networks, 31(23–24):2435–2463, (1999).
8. J. Ryan, M. Lin, and R. Miikkulainen, "Intrusion Detection with Neural Networks," *AI Approaches to Fraud Detection and Risk Management: Papers from the 1997 AAAI Workshop,* Providence, RI, pp. 72-79, (1997).
9. Internet2 NetFlow: Weekly Reports. NetFlow.internet2.edu/weekly, (April 2008).
10. Net flow sensor: nfsen.sourceforge.net
11. D. Plonka. Flowscan. Available at: www.caida.org/tools/utilities/flowscan/, (April 2008).
12. B. Claise. Cisco Systems NetFlow Services Export Version 9. Request for Comments: 3954, (October 2004).
13. Cisco IOS NetFlow Configuration Guide. Available at: www.cisco.com, (April 2008).
14. IP Flow Information Export Working Group. Available at: www.ietf.org/html.charters/ipfix-charter.html, (April 2008).
15. J., Muna. M. and Mehrotra M., "*Intrusion Detection System : A design perspective"*, 2rd International Conference On Data Management, IMT Ghaziabad, India. (2009).
16. M. Panda, and M. Patra, "*Building an efficient network intrusion detection model using Self Organizing Maps*", proceeding of world academy of science, engineering and technology, Vol. 38.(2009).
17. M. Khattab Ali, W. Venus, and M. Suleiman Al Rababaa, "*The Affect of Fuzzification on Neural Networks Intrusion Detection System"*, IEEE computer society,(2009).
18. James Cannady, "Artificial neural networks for misuse detection," Proceedings of the 1998 National Information Systems Security Conference, Arlington, VA,(1998).
19. Srinivas Mukkamala, "Intrusion detection using neural networks and support vector machine," *Proceedings of the 2002 IEEE International* Honolulu, HI, (2002)
20. DARPA1998. Available at: http://www.ll.mit.edu/IST/ideval/docs/1998.
21. KDDCup1999. Available at: http://kdd.ics.uci.edu/databases.
22. Y. Gao, Z. Li, and Y. Chen, "A DoS Resilient Flow-level Intrusion Detection Approach for High-speed Networks," in *Proc. of the 26th IEEE International* Conference on Distributed Computing Systems, Washington, USA, p.39,(2006).
23. A. Karasaridis, B. Rexroad, and D. Hoeflin, "Wide-scale botnet detection and characterization," in Proc. of the first conference on Hot Topics in Understanding *Botnets (HotBots '07)*, Berkeley, CA, USA, p.7,(2007).
24. A. Shahrestani, M. Feily, R. Ahmad, and S. Ramadass, "Architecture for Applying Data Mining and Visualization on Network Flow for Botnet Traffic Detection," in Proc. of the International Conference on Computer Technology and Development, Washington, DC, USA, pp. 33 – 37,(2009).
25. C. Livadas, R. Walsh, D. Lapsley, and W. T. Strayer, "Using Machine Learning Techniques to Identify Botnet Traffic," in *2nd IEEE LCN Workshop on Network Security*, pp. 967 – 974,(2006).
26. R. P. Lippmann *et al.*, "Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation," in *Proc. of the DARPA Information* Survivability Conference and Exposition, pp.12 – 26,(2000).
27. J. McHugh, "Testing Intrusion detection systems: a critique of the 1998 and 1999

DARPA intrusion detection system evaluations as performed by Lincoln Laboratory," ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262 – 294, (2000).

28. A. Sperotto, R. Sadre, F. Vliet, and A. Pras, "A Labeled Data Set for Flow-Based Intrusion Detection," in Proc. of the 9th IEEE International Workshop on IP *Operations and Management,* Berlin, Heidelberg, pp. 39 – 50, (2009).
29. Muna Mhammad T. Jawhar," Design Network Intrusion Detection System using hybrid Fuzzy- Neural Network", International Journal of Computer Science and Security, Volume (4),(2009)
30. Rodrigo Braga," Lightweight DDoS Flooding Attack Detection Using NOX/OpenFlow", 35th Annual IEEE Conference on Local Computer Networks LCN 2010, Denver, Colorado 6.(2010)
31. M. Vallipuram and B. Robert, *"An Intelligent Intrusion Detection System based on Neural Network",* IADIS International Conference Applied Computing.(2004).
32. T. Zhou and LI Yang, *"The Research of Intrusion Detection Based on Genetic Neural Network"*, Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition, Hong Kong, IEEE.(2008).
33. S. Mukkamala, H. Andrew Sung, and A. Abraham, *"Intrusion detection using an ensemble of intelligent paradigms"*, Journal of Network and Computer Applications 28. pp167–2,(2005).
34. S. Jimmy and A. Heidar, "*Network Intrusion Detection System using Neural Networks",* IEEE computer society.Vol.05,pp 242-246,(2008).
35. D. Novikov, V. Roman Yampolskiy, and L. Reznik, *"Anomaly Detection Based Intrusion Detection"*, IEEE Third International Conference on Communication, Networking & Broadcasting. pp 420-425,(2006).
36. I. Ahmad, S. Ullah Swati and S. Mohsin, "*Intrusions Detection Mechanism by Resilient Back Propagation (RPROP)"*, European Journal of Scientific Research ISSN 1450-216X Vol.17 No.4, pp.523-531. (2007).
37. M. Al-Subaie, "*The power of sequential learning in anomaly intrusion detection",* degree master thesis, Queen University, Canada. (2006).
38. D. Novikov, V. Roman Yampolskiy, and L. Reznik, *"Artificial Intelligence Approaches for Intrusion Detection"*, IEEE Long Island Systems Applications and Technology Conference. pp 1-8, (2006).
39. S. Lília de Sá, C. Adriana Ferrari dos Santos, S. Demisio da Silva, and A. Montes, *"A Neural Network Application for Attack Detection in Computer Networks",* Instituto Nacional de Pesquisas Espaciais – INPE, BRAZIL.(2004).
40. Chin-Teng Lin, C. S. George Lee: Neural fuzzy systems: a neuro-fuzzy synergism to intelligent, Prentice-Hall, Inc. Upper Saddle River, (1996).
41. Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, (1961).
42. Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: MIT Press, (1986).
43. Cybenko, G. "Approximation by super positions of a sigmoidal function", *Mathematics of Control, Signals, and Systems*, 2(4), 303–314,(1989).
44. Hagan, M.T., H.B. Demuth, and M.H. Beale, Neural Network Design, Boston, MA: PWS Publishing, (1996).
45. Hagan, M.T., and M. Menhaj, "Training feed-forward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, Vol. 5, 6, pp. 989–993, (1994).
46. Moody, J., and Darken C.: "Fast learning in networks of locally-tuned processing units", Neural Comput. MIT Press Cambridge, MA, USA, 1:281-294. (1989).

47. Chen, C. H., ed. Neural Networks in Pattern Recognition and Their Applications. River Edge, NJ: World Scientific, (1991)
48. E. J. Hartman, J. D. Keeler, and J. M. Kowalski. Layered neural networks with Gaussian hidden units as universal approximations. Neural Computation, 2:210–215, (1990)
49. H. Wang, D. Zhang, and K. G. Shin, .SYN-dog: Sniffing SYN Flooding Sources, In Proc. of 22nd International Conference on Distributed Computing Systems, Vienna, Austria, (2002).
50. Softflowd. Available at:: https://code.google.com/p/softflowd/.
51. Mukkamala, S.; and Sung, A.H. Feature selection for intrusion detection using neural networks and support vector machines. *Transportation Research Record*, 1822, 33-39. (2003).
52. Sammany, M.; Sharawi, M.; El-Beltagy, M.; and Saroit, I. Artificial neural networks architecture for intrusion detection systems and classification of attacks. *Accepted for publication in the 5th international conference INFO2007*, Cairo University.(2007).
53. M.Govindarajan, and RM.Chandrasekaran," Intrusion detection using neural based hybrid classification methods", computer networks journal. Volume 55, issue 8, pp.1662-1671, (2011).
54. Prasanta Gogoi, D.K. Bhattacharyya, B. Borah, and Jugal K. Kalita "MLH-IDS: A Multi-Level Hybrid Intrusion Detection Method", the Computer Journal Advance Access published May 12, (2013).

Yousef Abuadlla Received his BSc degree in Computer Engineering from University of Tripoli, Libya in 1993, and his MSc degree in Computer Engineering from University of Belgrade, Serbia in 2003. Currently he is a PhD candidate in Computer Engineering at the School of of Electrical Engineering, University of Belgrade. His research interests include network security, intrusion detection system and data mining.

Goran Kvascev was born in 1975. He received the Engineer, Magister and Ph.D. degrees from School of Electrical Engineering, University of Belgrade, Serbia, in 2000, 2005, and 2012, respectively. Since 2000, he has been with the Control Department, School of Electrical Engineering, where he was a research and teaching assistant, and became an Assistant Professor in 2013. His main areas of research interest are real-time and industrial process control, neural networks, robust system identification, thermal power plant control.

Slavko Gajin received dipl. Eng., MS and PhD degrees from University of Belgrade, School of Electrical Engineering, Serbia, in 1993, 1999, and 2007, respectively. He is a director of Belgrade University Computer Center, where he started working as a network engineer since he received bachelor's degree. He is a professor at the Department of Computer Engineering and Computer Science at the School of Electrical Engineering, University of Belgrade. His current research interests include the modelling and performance analysis of communication systems, routing algorithms in NoC and interconnection networks, computer network management, monitoring and performances.

Zoran Jovanović received his dipl. eng., MS and PhD degrees from University of Belgrade, School of Electrical Engineering in 1978, 1982, and 1988, respectively. He is the Director of the National Research and Education Network of Serbia (AMRES) and professor at the Department of Computer Engineering and Computer Science at School of Electrical Engineering, University of Belgrade. His current research interests include parallel computing, networking, concurrent and distributed programming.

Received: April 15, 2013; Accepted: May 25, 2014

Computer Science and Information Systems 11(2):623–644 DOI:10.2298/CSIS130429023A

Building ontologies for different natural languages

Emhimed Salem Alatrish[1], Dušan Tošić[2], and Nikola Milenković[3]

[1] Faculty of Mathematics, Studentski Trg 16,
11000 Belgrade, Serbia
emhimed_alatrish@yahoo.com
[2] Faculty of Mathematics, Studentski Trg 16,
11000 Belgrade, Serbia
dtosic@matf.bg.ac.rs
[3] Faculty of Mathematics, Studentski Trg 16,
11000 Belgrade, Serbia
nikola.milenkovic@live.com

Abstract. Ontology construction of a certain domain is an important step in applying the Semantic web. A number of software tools adapted for building domain ontologies of most wide–spread natural languages are available, but accomplishing that for any given natural language presents a challenge. Here we propose a semi-automatic procedure to create ontologies for different natural languages. Our approach utilizes various software tools available on the Internet most notably DODDLE-OWL - a domain ontology development tool implemented for English and Japanese languages. By using this tool, WordNet, Protégé and XSLT transformations, we propose a general procedure to construct domain ontology for any natural language.

Keywords: Semantic Web, Ontology, Natural language, DODDLE-OWL.

1. Introduction

Semantic Web has lately been a popular and prolific field of research with numerous scientific papers published on the topic so far. Ontology is an important component of the Semantic Web and a lot of papers about applying ontology in specific fields have been published (see [27, 28]). Ontologies are closely connected to Natural Language Processing (NLP) - a field of artificial intelligence, computer science and linguistics. As such, NLP is related to the area of human–computer interaction. Ontologies provide an explicit and formal way for data interpretation, integration and sharing, helping to understand natural (human) language. Understanding of natural language is not an aim per se, but it is useful in different fields, such as: Information Extraction (IE), Machine Translation (MT), Question Answering (QA), etc. (Fig. 1.). Because of that the production of software tools to support ontology and Semantic web has accelerated. A number of these tools are free and available on the Internet (see: [2, 5, 13]).

Unfortunately, most of them are made to work with only a small set of widely used languages such as: English, Spanish, French etc. Some natural languages are not represented in these tools and it is a challenge to create domain ontologies for text written in these languages.

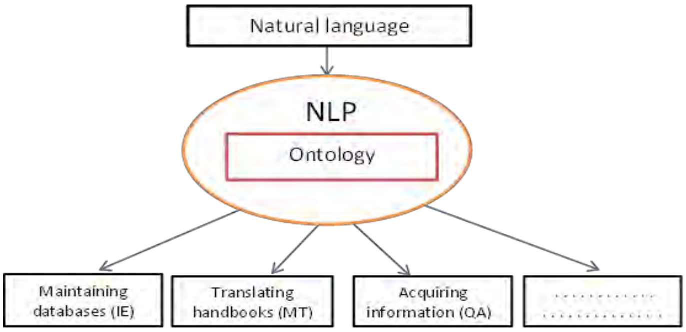

Natural language
NLP
Ontology
Maintaining databases (IE)
Translating handbooks (MT)
Acquiring information (QA)
....

Fig. 1. Relation between natural language, NLP and ontology

Our idea is to combine different accessible software tools for the purpose of semi-automatic construction of Natural Language Ontologies (NLOs) from specific domains. This approach aims to be general and applicable for any natural language. The proposed approach is very ambitious because the problem NLO construction is very general and difficult. Understandably, certain adaptations and constraints are necessary depending on the features of the natural language in question. The domain of applying ontology is important too. Usually, we should make text classification before ontology construction. A good methodology for text classification is proposed in [14].

2. Related work

The problem of NLO construction became apparent immediately after ontologies appeared in computer sciences and it is still present today. This problem is considered in some books (for example, [3], [21]) and in many articles ([15], [29], [8], [30]). In [15], an automatic ontology building method is proposed. The authors described a system which starts from small ontology kernel and constructs the ontology by automatically understanding the text. This system is implemented in the project named Hasti and applied to Persian (Farsi) texts. Paper [8] contains a project description where ontologies are part of the reasoning process used for information management and for the presentation of information. Both accessing and presenting information are mediated via natural language. In [1], an automatic construction of ontology from Arabic texts is proposed, by using statistical techniques to extract elements of ontology. In this work initialization of the ontology is started manually and it is difficult to describe it as a fully automatic process. An approach to converting hierarchical classifications (whose nodes are assigned natural language labels) into lightweight ontologies is proposed in [11]. In paper [12] a model of a Conversational Recommendation Agent (CoRA) is described. It is a domain-specific dialogue system, which implements an ontology-based Natural Language Processing system for shopping situations. The problem of content determination in natural language generation (NLG) is considered in [16]. The authors try to answer the question "What is an A?" where A is a that building ontologies for different natural languages is currently a challenging problem. In recent years, CNL (Controlled Natural Language) has received much attention with regard to ontology ([20], [5]). CNLs, as subsets of natural languages, can eliminate ambiguity of natural languages and successfully apply in ontology construction. Introducing CNLs, authors impose restrictions on used natural languages.

By developing the area of ontology construction, a lot of new problems are raised. The growing number of ontologies available in different natural languages leads to an interoperability problem. This problem is considered in [9] and a new architecture for a multilingual ontology matching service is proposed. A Framework for merging the heterogeneous ontologies based on WordNet is described in [4]. In fact, a new methodology for merging the different ontologies is introduced. Casual users often use large database. For these purposes it convenient to use Natural Language Interfaces (NLIs) (often referred as closed-domain Question Answering (QA) systems). In [6] system FREyA (Natural Language Interfaces to ontologies) is presented.

3. Semi-automatic creation of NLO

Even though automatic creation of domain NLO has been attempted (see [15]), it is still a difficult task in general. It is particularly challenging to do so for the texts written in different natural languages and related to some domain. In this case the domain ontology structure depends in some aspects on human users. Because of that, it is convenient to provide refined semi-automatic software tool for building NLOs. Those kind of tools are available on the Web ([24], [5]) and one of them is DODDLE-OWL (see: [18] and [26]). DODDLE-OWL is an interactive domain ontology development environment created for Japanese and English language. We adopted this environment for any natural language (that has a dictionary on WordNet) by applying translation of original text into English text and transforming the obtained English ontology. Since DODDLE-OWL is an essential tool in our approach, we are going to describe it in more detail.

3.1. DODDLE-OWL Overview

DODDLE-OWL (a Domain Ontology rapiD DeveLopment Environment - OWL extension) is a domain ontology development tool for the Semantic Web. It is written in Java language. According to [7], “DODDLE-OWL reuses existing ontologies such as WordNet and EDR as general ontologies to construct taxonomic relationships (defined as classes) and other relationships (defined as properties and their domains and ranges) for concepts”. An initial concept hierarchy is constructed as a (is-a) hierarchy of terms. Here, it is assumed that there are one or more domain specific documents and that the user can select important terms needed to construct domain ontology. DODDLE-OWL has the following six main modules: Ontology Selection Module, Input Module, Construction Module, Refinement Module, Visualization Module, and Translation Module. We assume that there are one or more domain specific documents, and we also assume that the user can select important terms needed to construct domain ontology (Fig. 2, see [18] and [26]).

First, as an input to DODDLE-OWL, the user selects several concepts in Input Module. In Construction Module, DODDLE-OWL generates the basis of the ontology, an initial concept hierarchy and set of concept pairs, by referring to appropriate reference ontologies and documents. In Refinement Module the initial ontology, generated by Construction Module, is refined by the user through interactive support by DODDLE-OWL. The ontology constructed by DODDLE-OWL can be exported with the representation of OWL. Finally, Visualization Module (MR3) (described in [18]) is connected with DODDLE-OWL and works with a graphical editor ([26]).

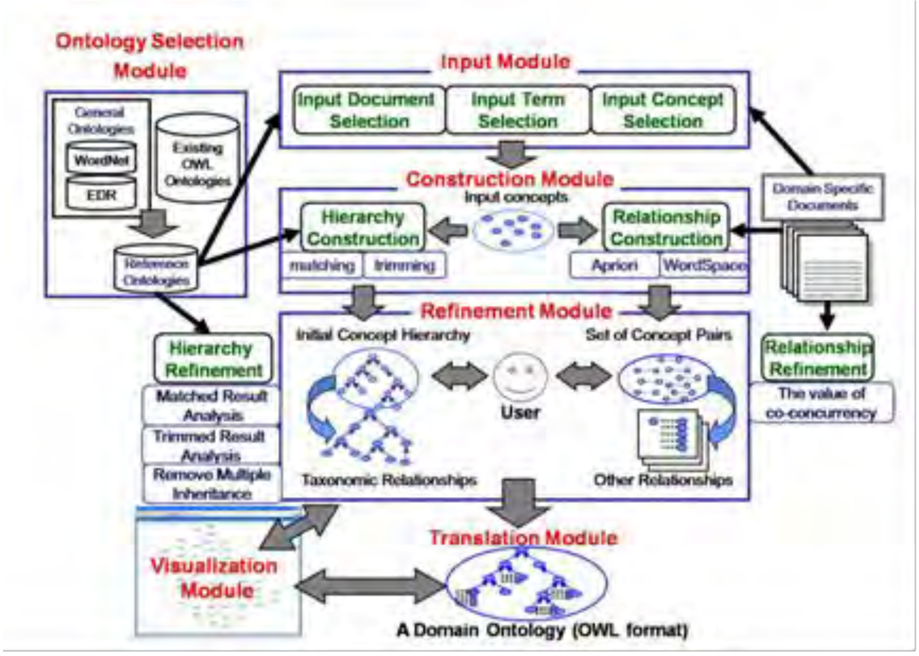


Fig. 2. Overview of DODDLE-OWL

3.2. Construction of domain NLO for different languages

In order to construct the NLO for different languages, text document from any natural language is translated to English language. English ontology is built by using DODDLE-OWL. In DODDLE-OWL the following steps are executed:

1. In the Ontology Selection Module, user selects reference ontologies on WordNet, EDR (general vocabulary dictionary or technical terminology dictionary), and existing OWL ontologies in the ontology selection as shown in Fig. 3.

2. In the Input Document Selection Module, user selects domain specific documents described in English. In this step, some words in the documents are extracted. During the same phase, user can select a part of speech (POS) for extraction of words from the documents. For example, if noun or verb words are extracted, checkbox "Noun" or "Verb" should be checked as shown in Fig. 3.

3. In the Input Term Selection Module, a list of extracted terms is formed. This list includes (for more details see [3] and [14]): compound words, part of speech (POS), Term Frequency (TF of term t in document d is defined as the number of times that t occurs in d), Inverse Document Frequency (IDF estimate the rarity of a term in the whole document collection - if a term occurs in all the documents of the collection, its

IDF is zero) and TF-IDF in the documents (TF-IDF is weight of a term - the product of its TF weight and its IDF weight). Domain specific documents contain many significant compound words. Therefore, accurate extraction of compound words is necessary to construct domain ontologies. At this step, while considering part of speech (POS), TF, and so on, the user selects input terms which are significant terms for the domain. For certain domains, important terms do not occur in the documents. In such case, the Input Term Selection Module has a function, allowing the manual addition of important terms as input terms by the user. In order to prevent the leakage of the selection of input terms from the documents, the Input Term Selection Module maintains the relationships between the extracted terms and the terms in the documents as shown in Fig. 3.

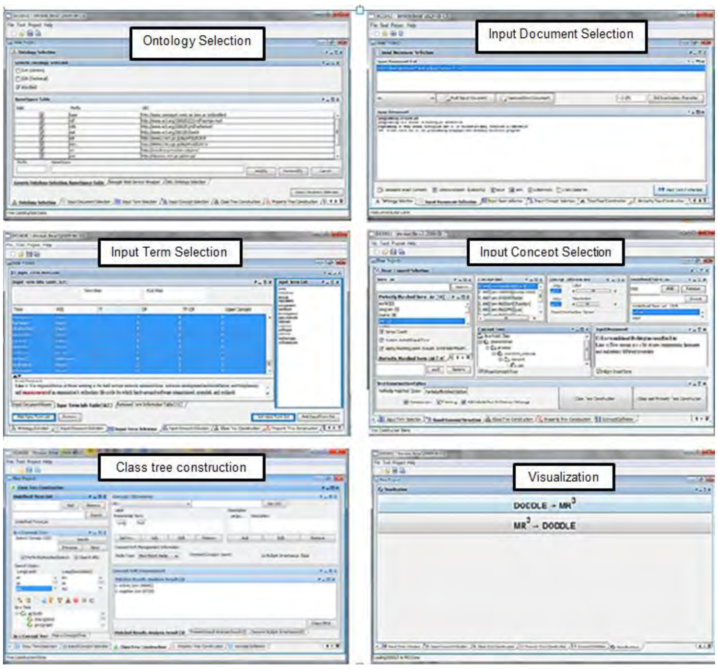


Fig. 3. Typical usage of DODDLE-OWL

4. In the Input Concept Selection Module, the user identifies the word sense of input terms in order to map those terms to the concepts in the reference ontologies selected with the Ontology Selection Module. A particular single term may have many word senses. Therefore, there may be many concepts corresponding to a word. The Input Concept Selection Module has a function enabling automatic word disambiguation. This function shows the list of concepts, ordered by some criteria, corresponding to the selected input term. Input term not corresponding to the labels of concepts in the reference ontologies is marked as undefined. The input terms are also undefined if the concept exists, but there are no appropriate concepts in the reference ontologies. The user defines the undefined terms manually in the refinement module, as shown in Figure 3.

5. The Hierarchy Construction Module automatically generates the basis of ontology, an initial concept hierarchy (by referring to reference ontologies) and documents. An initial concept hierarchy is constructed as a taxonomic relationship.

6. DODDLE-OWL uses MR3: Meta-Model Management based on RDFs Revision Reflection [18] as the Visualization Module. Figure 4. shows the product of MR3 as RDFs description and graphical representation. Finally, through the translation module, we can export the constructed domain ontology described in RDFs. For example, a portion of the obtained English Ontology OWL code is presented in the following document (1):

```
<rdf:Description rdf:about="use">
  <rdfs:subClassOf rdf:resource="activity"/>
  <rdfs:comment xml:lang="en">the act of using; "he warned
against the use of narcotic drugs"; "skilled (1) in the
utilization of computers"</rdfs:comment>
  <rdfs:label xml:lang="en">use</rdfs:label>
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
</rdf:Description>
```

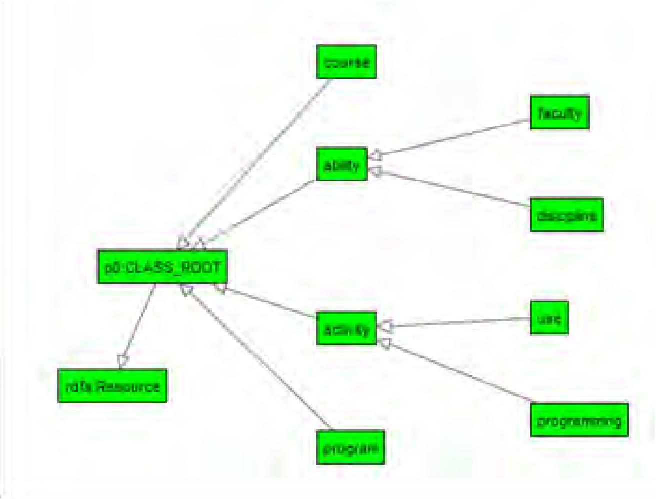


Fig. 4. Products of MR3

7. To build ontology represented by OWL, we use Protégé editor ([23]). Protégé has plugin to enhance ontology development, such as the OWL plugin (see: [10]). We use this possibility to get the OWL document. For example, the document (1) is transformed in the following text.

```
<owl:Class rdf:ID="use">
  <rdfs:label xml:lang="en">use</rdfs:label>
  <rdfs:comment xml:lang="en">
    the act of using; "he warned against the …
  </rdfs:comment>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="activity"/>
  </rdfs:subClassOf>
</owl:Class>
```

The similar ontology graph for the related English words could be generated as in Fig. 4. by using Protégé editor.

8. Very important step in localization process is translation of the ontology recognized in a source language into target language by using XSLT. We are looking for all tags `<rdfs:label>`, `<owl:Class>` (this includes `rdf:ID` and `rdf:about`) and `<rdfs:subClassOf>` (this includes attributes `rdf:about` and `rdf:resource`) and duplicates them. XSLT Translation script is included in this paper and also publically available online[1].

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#">
  <xsl:output method="xml" indent="yes"/>
  <xsl:variable name="dict"
                 select="document('dictionary.xml')/*" />
  <xsl:variable name="sourceLanguage" select="$dict/@from" />
  <xsl:variable name="targetLanguage" select="$dict/@to" />

  <xsl:template match="@* | node()">
    <xsl:copy>
      <xsl:choose>
         <xsl:when test=".=rdfs:label">
          <xsl:apply-templates select="rdfs:label"/>
        </xsl:when>
         <xsl:when test=".=rdfs:subClassOf ">
          <xsl:apply-templates
                             select="rdfs:subClassOf "/>
          </xsl:when>
         <xsl:otherwise>
          <xsl:apply-templates select="@* | node()"/>
        </xsl:otherwise>
      </xsl:choose>
    </xsl:copy>
  </xsl:template>

  <xsl:template match="rdfs:label">
    <xsl:variable name="word" select="." />
```

[1] https://github.com/nikolamilenkovic/doddle-owl-rdf-translator

```
      <rdfs:label xml:lang="{$sourceLanguage}">
       <xsl:value-of select="$word"/>
     </rdfs:label>
      <rdfs:label xml:lang="{$targetLanguage}">
       <xsl:value-of select="$dict/word[@name=$word]"/>
     </rdfs:label>
  </xsl:template>

  <xsl:template match="owl:Class">
    <xsl:if test="@rdf:ID">
      <xsl:variable name="word" select="@rdf:ID" />
      <xsl:variable name="translated_word">
        <xsl:value-of select="$dict/word[@name=$word]"/>
      </xsl:variable>

     <owl:Class rdf:ID="{$translated_word}">
       <xsl:apply-templates />
     </owl:Class>
    </xsl:if>

    <xsl:if test="@rdf:about">
      <xsl:variable name="word"
                  select="substring(@rdf:about, 2)" />
      <xsl:variable name="translated_word">
        <xsl:value-of select="$dict/word[@name=$word]"/>
      </xsl:variable>

      <owl:Class rdf:about="{concat('#',$translated_word)}">
       <xsl:apply-templates />
     </owl:Class>
    </xsl:if>
  </xsl:template>

  <xsl:template match="rdfs:subClassOf">
    <xsl:choose>
      <xsl:when test="@rdf:resource">
        <xsl:variable name="word"
                     select="substring(@rdf:resource, 2)" />
        <xsl:variable name="translated_word">
          <xsl:value-of select="$dict/word[@name=$word]"/>
        </xsl:variable>
        <rdfs:subClassOf
             rdf:resource="{concat('#',$translated_word)}" />
      </xsl:when>
      <xsl:otherwise>
        <xsl:copy>
          <xsl:apply-templates select="@* | node()"/>
        </xsl:copy>
      </xsl:otherwise>
    </xsl:choose>
  </xsl:template>
</xsl:stylesheet>
```

For example, if the input document includes these tags:

```
<owl:Class rdf:ID="use">
  <rdfs:label xml:lang="en">use</rdfs:label>
  <rdfs:comment xml:lang="en">
    the act of using; "he warned against the …
  </rdfs:comment>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="activity"/>
  </rdfs:subClassOf>
</owl:Class>
```

and the target language is Serbian-Cyrillic, we will get document like this one:

```
<owl:Class rdf:ID="користити">
  <rdfs:label xml:lang="sr">користити</rdfs:label>
  <rdfs:comment xml:lang="en">
    the act of using; "he warned against the …
  </rdfs:comment>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="активност"/>
  </rdfs:subClassOf>
</owl:Class>
```

After that, we can generate an ontology graph for related Serbian words from a specified text (Fig. 5.).

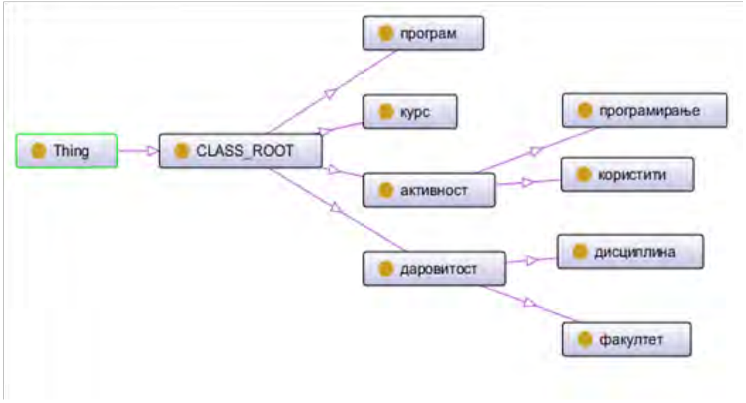


Fig. 5. Ontology graph for related Serbian words

The whole process could be graphically described as in Fig. 6.

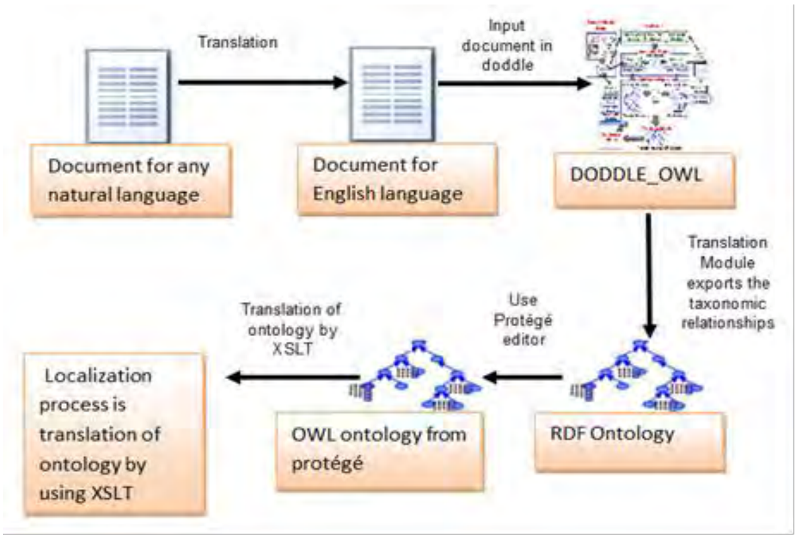

Translation
Input document in doddle
Document for any natural language
Document for English language
DODDLE_OWL
Translation Module exports the taxonomic relationships
Translation of ontology by XSLT
Use Protégé editor
Localization process is translation of ontology by using XSLT
OWL ontology from protégé
RDF Ontology

Fig. 6. Building of NLO

3.3. Creating dictionary for the translation process

By applying DODDLE-OWL, after finishing step 4, the project is completed and saved. When the project is saved, a file named “InputWordSet” is automatically created. This file contains all English words selected from Input Term Info Table. This file consequently contains all relevant words for the target ontology. These words should be translated into corresponding words in the target language. Here we utilize MyMemory online translation service [19]. MyMemory is the world's largest Translation Memory (TM). It has been created collecting TMs from the European Union, United Nations and aligning the best domain-specific multilingual websites.

MyMemory’s translation service is accessible over the Internet via their translation API. We wrote WordTranslator[2] console application in .NET Framework 4.5 which utilizes their translation API. Performing translation of a single word is done by calling WordTranslator with the following arguments: word to be translated, language of the provided word and the desired language for the translation. Program outputs a single translated word. This greatly simplifies the translation process, since the complexity of

[2] https://github.com/nikolamilenkovic/word-translator

natural language translation is reduced to a single console command execution. (If some words are not translated, then we use Google translator or any available tool for translation.)

Since WordTranslator translates a single word at a time, we also created a script[3] which automatically iterates over each word in the input word list and translates each word individually. Script is written in Batch shell scripting language and its source code is:

```
@ECHO OFF & SetLocal EnableExtensions EnableDelayedExpansion
::Input file name
SET in=%1
::Input language (for example: en)
SET inLang=%2
::Output language (for example: ar)
SET outLang=%3
::Output file name (for example: dictionary.xml)
SET out=%4
::Check if output file exists
IF EXIST %out% (
  ECHO %out% already exists! Deleting...
    DEL %out%
    ECHO Deleted!
)

ECHO ^<?xml version="1.0" encoding="UTF-8"?^> >> %out%
ECHO ^<dictionary from="%inLang%" to="%outLang%"^> >> %out%

FOR /F "tokens=*" %%l in (%in%) do call :TRANSLATE %%l
ECHO ^</dictionary^> >> %out%
ECHO Dictionary created. Perform manual check before usage.

GOTO :EOF

:: ----------------- PROCEDURE TRANSLATE ------------------------
:TRANSLATE
  SET inputWord=%1
  FOR /f "delims=" %%a in ('WordTranslator.exe --word %inputWord% --
inputLanguage %inLang% --outputLanguage %outLang%') DO SET
outputWord=%%a
```

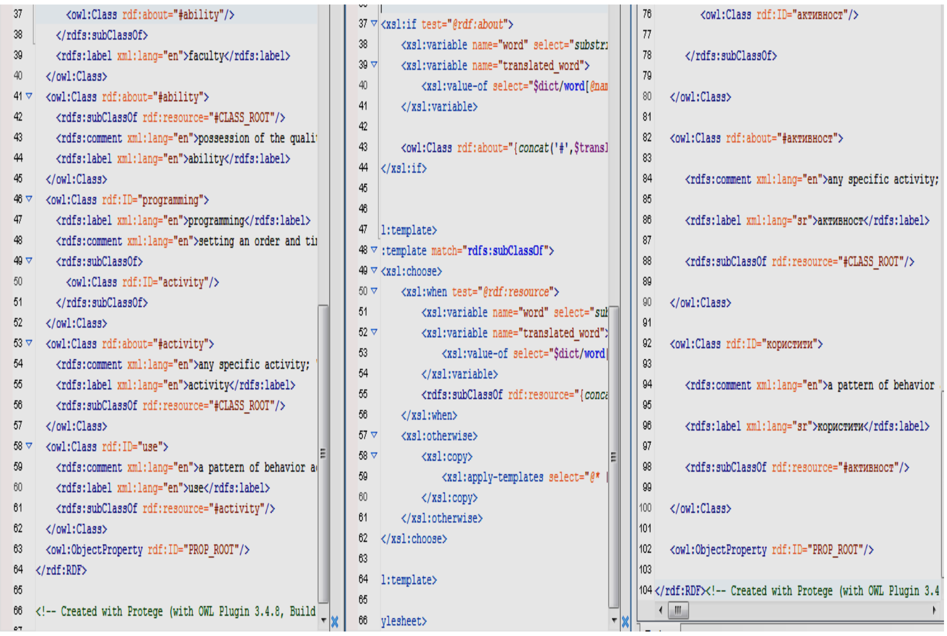
[3] https://gist.github.com/nikolamilenkovic/9169523

```
    ECHO Translated %inputWord% to %outputWord%
    ECHO ^<word name="%inputWord%"^>%outputWord%^</word^> >> %out%
    EXIT /b
:: ------------- END OF PROCEDURE: TRANSLATE --------------------
```

WordTranslator can be used for any combination of input/output languages. Since we are starting with English ontologies, all our input words will be in English, and output in the target language. In the Batch script we use outputs generated by WordTranslator to construct dictionary used by XSLT transformation. Format of the dictionary as follows (mapping English to Serbian words):

```
<?xml version="1.0" encoding="utf-8"?>
<dictionary from="en" to="sr">
   <word name="activity">активност</word>
   <word name="abstraction">апстракција</word>
   <word name="group">група</word>
   <word name="creative_activity">креативна_активност</word>
   <word name="sun">сунце</word>
</dictionary>
```

After building the English OWL representation of our text in step 5, we transform this document into Serbian OWL representation by using XSLT transformer. For this purpose we use an XML editor. There are several available XML editors (see: [17] and [22]). For example, by using Oxygen XML Editor [10], we get the workspace organized as in Fig. 7.



```
      <owl:Class rdf:about="#ability"/>
    </rdfs:subClassOf>
    <rdfs:label xml:lang="en">faculty</rdfs:label>
  </owl:Class>
  <owl:Class rdf:about="#ability">
    <rdfs:subClassOf rdf:resource="#CLASS_ROOT"/>
    <rdfs:comment xml:lang="en">possession of the quali
    <rdfs:label xml:lang="en">ability</rdfs:label>
  </owl:Class>
  <owl:Class rdf:ID="programming">
    <rdfs:label xml:lang="en">programming</rdfs:label>
    <rdfs:comment xml:lang="en">setting an order and ti
    <rdfs:subClassOf>
      <owl:Class rdf:ID="activity"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="#activity">
    <rdfs:comment xml:lang="en">any specific activity;
    <rdfs:label xml:lang="en">activity</rdfs:label>
    <rdfs:subClassOf rdf:resource="#CLASS_ROOT"/>
  </owl:Class>
  <owl:Class rdf:ID="use">
    <rdfs:comment xml:lang="en">a pattern of behavior a
    <rdfs:label xml:lang="en">use</rdfs:label>
    <rdfs:subClassOf rdf:resource="#activity"/>
  </owl:Class>
  <owl:ObjectProperty rdf:ID="PROP_ROOT"/>
</rdf:RDF>

<!-- Created with Protege (with OWL Plugin 3.4.8, Build
```

```

<xsl:if test="@rdf:about">
    <xsl:variable name="word" select="substr
    <xsl:variable name="translated_word">
        <xsl:value-of select="$dict/word[@nam
    </xsl:variable>

    <owl:Class rdf:about="{concat('#',$transl
</xsl:if>


l:template>
:template match="rdfs:subClassOf">
<xsl:choose>
    <xsl:when test="@rdf:resource">
        <xsl:variable name="word" select="sub
        <xsl:variable name="translated_word">
            <xsl:value-of select="$dict/word[
        </xsl:variable>
        <rdfs:subClassOf rdf:resource="{conca
    </xsl:when>
    <xsl:otherwise>
        <xsl:copy>
            <xsl:apply-templates select="@*
        </xsl:copy>
    </xsl:otherwise>
</xsl:choose>

l:template>

ylesheet>
```

```
      <owl:Class rdf:ID="активност"/>

    </rdfs:subClassOf>

  </owl:Class>

  <owl:Class rdf:about="#активност">

    <rdfs:comment xml:lang="en">any specific activity;

    <rdfs:label xml:lang="sr">активност</rdfs:label>

    <rdfs:subClassOf rdf:resource="#CLASS_ROOT"/>

  </owl:Class>

  <owl:Class rdf:ID="користити">

    <rdfs:comment xml:lang="en">a pattern of behavior

    <rdfs:label xml:lang="sr">користити</rdfs:label>

    <rdfs:subClassOf rdf:resource="#активност"/>

  </owl:Class>

  <owl:ObjectProperty rdf:ID="PROP_ROOT"/>

</rdf:RDF><!-- Created with Protege (with OWL Plugin 3.4
```

Fig. 7. XSLT transformer applied to Oxygen XML Editor

4. Examples

In this section we present two examples of applying our procedure for Arabic and French. We started with the following Arabic text:

البرمجة هي فن جميل. البرمجة هي المقرر التعليمي في كلية الرياضيات. البرمجة هو الانضباط فعال للغاية
ويتم تعليمها في العديد من الكليات في العالم. الآن يمكننا استخدام الكثير من لغات البرمجة الجديدة وجعل العديد من
البرامج المختلفة

The above Arabic text is translated into English as follows:

"Programming is a beautiful art. Programming is a course in the Faculty of Mathematics. Programming is a discipline very effective and it is learned in many colleges in the world. Now we can use a lot of new programming languages and make many different programs."

After applying the DODDLE-OWL, the obtained document is used by Protégé editor to get the ontology for the above-mentioned English text. The obtained ontology document is very long and here we present only a part of this document that is related to the notions: “discipline” and “course”:

```
</owl:Class>
  <owl:Class rdf:about="discipline">
    <rdfs:subClassOf>
      <owl:Class rdf:about="activity"/>
    </rdfs:subClassOf>
    <rdfs:label xml:lang="en">discipline</rdfs:label>
    <rdfs:comment xml:lang="en">training to improve strength
     or self-control</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:about="course">
    <rdfs:label xml:lang="en">course</rdfs:label>
    <rdfs:comment xml:lang="en">a line or route along which
     something travels or moves; "the hurricane demolished
     houses in its path"; "the track of an animal"; "the
     course of the river"</rdfs:comment>
    <rdfs:subClassOf>
```

By using the Protégé editor, we can generate ontology graph for related English words from specified text (Fig. 8.).

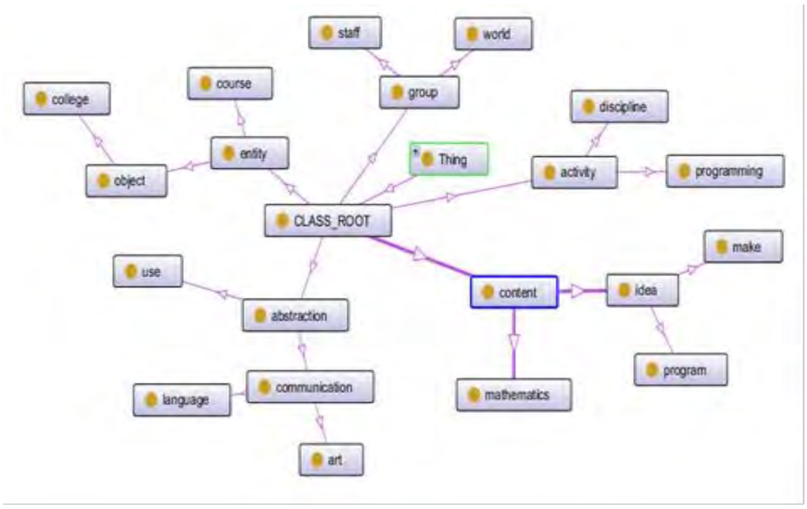
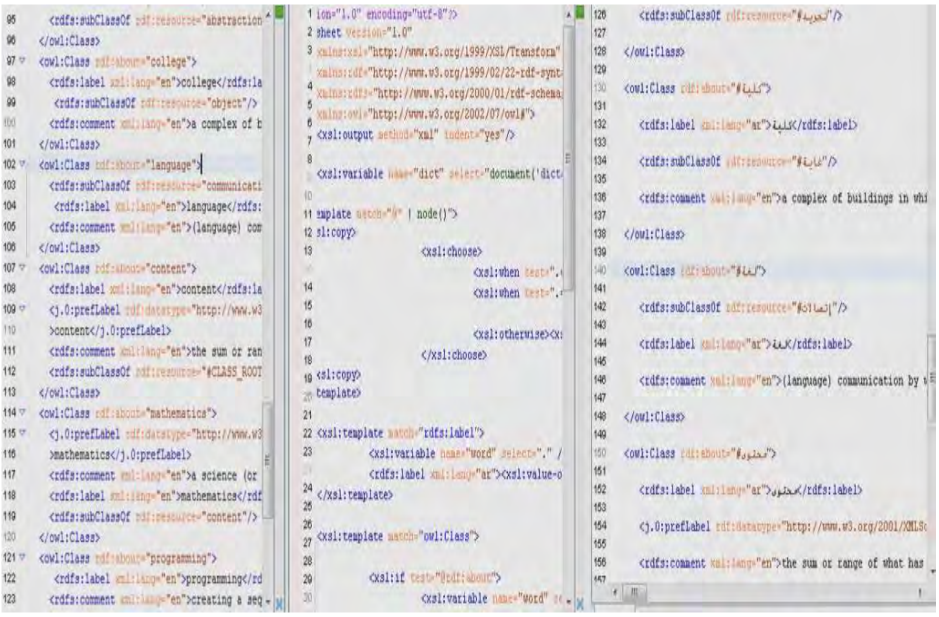


Fig. 8. Ontology graph for English words

A part of XSLT transformation for this text has the form as in Fig. 9.



```
  <rdfs:subClassOf rdf:resource="abstraction
</owl:Class>
<owl:Class rdf:about="college">
  <rdfs:label xml:lang="en">college</rdfs:la
   <rdfs:subClassOf rdf:resource="object"/>
  <rdfs:comment xml:lang="en">a complex of b
</owl:Class>
<owl:Class rdf:about="language">
  <rdfs:subClassOf rdf:resource="communicati
   <rdfs:label xml:lang="en">language</rdfs:
  <rdfs:comment xml:lang="en">(language) com
</owl:Class>
<owl:Class rdf:about="content">
  <rdfs:label xml:lang="en">content</rdfs:la
  <j.0:prefLabel rdf:datatype="http://www.w3
  >content</j.0:prefLabel>
  <rdfs:comment xml:lang="en">the sum or ran
  <rdfs:subClassOf rdf:resource="#CLASS_ROOT
</owl:Class>
<owl:Class rdf:about="mathematics">
  <j.0:prefLabel rdf:datatype="http://www.w3
  >mathematics</j.0:prefLabel>
  <rdfs:comment xml:lang="en">a science (or
  <rdfs:label xml:lang="en">mathematics</rdf
  <rdfs:subClassOf rdf:resource="content"/>
</owl:Class>
<owl:Class rdf:about="programming">
   <rdfs:label xml:lang="en">programming</rd
  <rdfs:comment xml:lang="en">creating a seq
```

```
ion="1.0" encoding="utf-8"?>
sheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-synt
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema
xmlns:owl="http://www.w3.org/2002/07/owl#">
<xsl:output method="xml" indent="yes"/>

<xsl:variable name="dict" select="document('dict

emplate match="/" | node()">
sl:copy>
                    <xsl:choose>
                              <xsl:when test=".
                              <xsl:when test=".

                              <xsl:otherwise><x
                    </xsl:choose>
<sl:copy>
template>

<xsl:template match="rdfs:label">
          <xsl:variable name="word" select="." /
          <rdfs:label xml:lang="ar"><xsl:value-o
</xsl:template>

<xsl:template match="owl:Class">

          <xsl:if test="@rdf:about">
                    <xsl:variable name="word" se
```

```
   <rdfs:subClassOf rdf:resource="#تجريد"/>

</owl:Class>

<owl:Class rdf:about="#كلية">

   <rdfs:label xml:lang="ar">كلية</rdfs:label>

   <rdfs:subClassOf rdf:resource="#بناية"/>

   <rdfs:comment xml:lang="en">a complex of buildings in whi

</owl:Class>

<owl:Class rdf:about="#لغة">

   <rdfs:subClassOf rdf:resource="#اتصالة"/>

   <rdfs:label xml:lang="ar">لغة</rdfs:label>

   <rdfs:comment xml:lang="en">(language) communication by 

</owl:Class>

<owl:Class rdf:about="#محتوى">

   <rdfs:label xml:lang="ar">محتوى</rdfs:label>

   <j.0:prefLabel rdf:datatype="http://www.w3.org/2001/XMLSc

   <rdfs:comment xml:lang="en">the sum or range of what has
```

Fig. 9. XSLT transformer applied Arabian text

A corresponding piece of the obtained ontology for the previous Arabic text (the entire document is too long and will therefore not be presented in its entirety) looks like the following one:

```
</owl:Class>
  <owl:Class rdf:about="#الانضباط">
```

```
    <rdfs:subClassOf>
      <owl:Class rdf:about="#نشاط"/>
    </rdfs:subClassOf>
    <rdfs:label xml:lang="ar">الانضباط</rdfs:label>
    <rdfs:comment xml:lang="en">training to improve strength
    or self-control</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:about="#المقرر_التعليمي">
  <rdfs:label xml:lang="ar">المقرر التعليمي</rdfs:label>
    <rdfs:comment xml:lang="en">a line or route along which
     something travels or moves; "the hurricane demolished
     houses in its path"; "the track of an animal"; "the
     course of the river"</rdfs:comment>
  <rdfs:subClassOf>
```

By applying the Protégé editor to the obtained document, we can generate an ontology graph for related Arabic word from a specified text (Fig.10.). From this graph we can see the relations between concepts in given text.

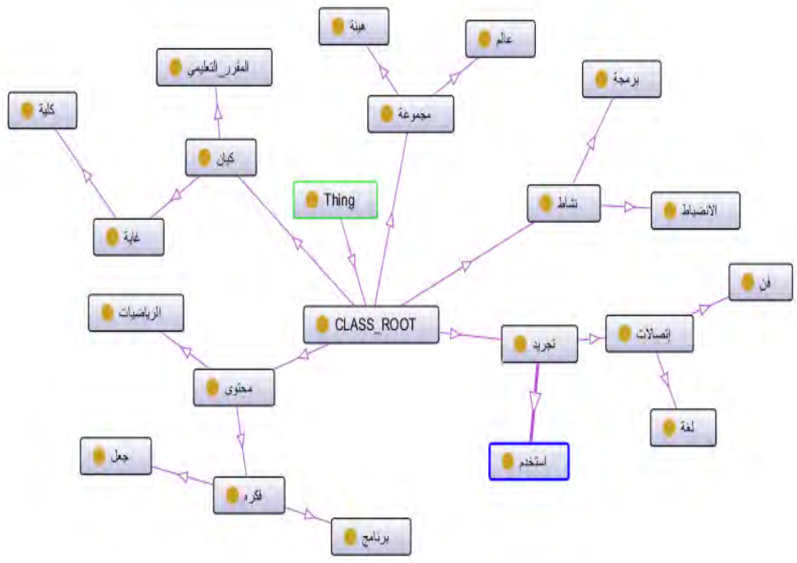


Fig. 10. Ontology graph for Arabic words

The second example is related to French language. For the following text:

Un thésaurus est constitué d'un ensemble organisé de termes, choisis pour leur capacité à faciliter la description d'un domaine et à harmoniser la communication et le traitement de l'information. Les termes d'un thésaurus sont reliés entre eux par des relations sémantiques (hiérarchique, équivalence, etc.).

A fraction the OWL document is below:

```
<owl:Class rdf:about="description">
  <rdfs:subClassOf>
```

```
    <owl:Class rdf:about="knowledge"/>
  </rdfs:subClassOf>
  <rdfs:label xml:lang="en">description</rdfs:label>
  <rdfs:comment xml:lang="en">sort or variety; "every
description of book was there"</rdfs:comment>
</owl:Class>
```

By applying the Protégé editor, as in previous example, we get an ontology graph for an English text (Fig. 11.).

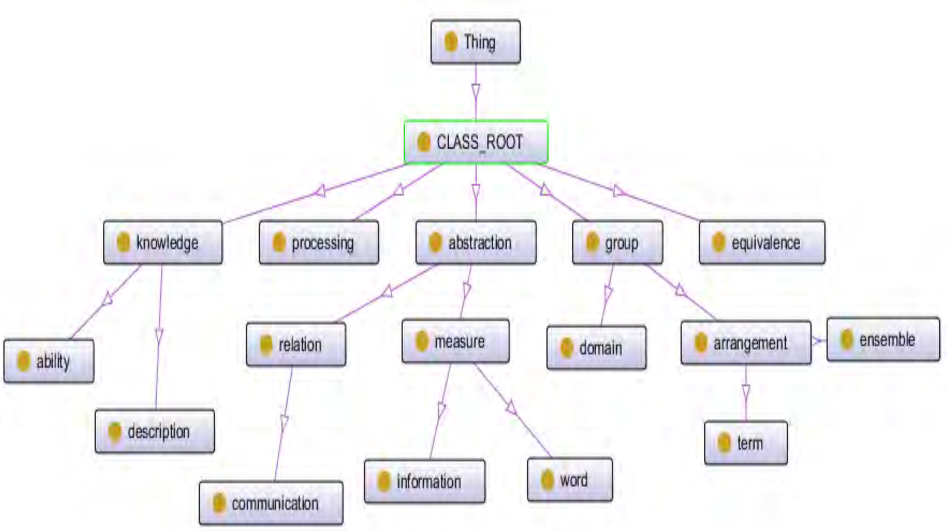
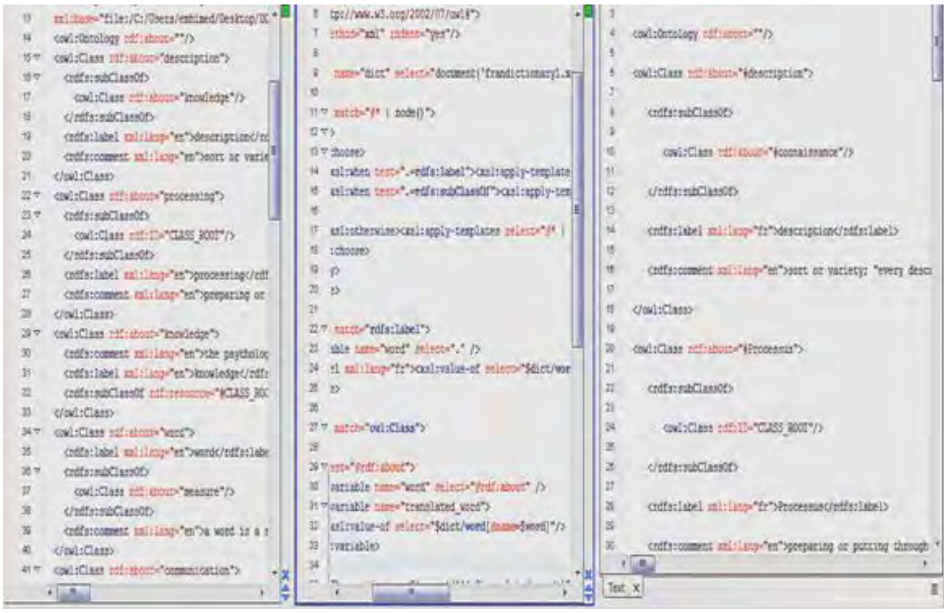


Fig. 11. Ontology graph for English words

By using the XSLT transformer (Fig. 12.), we generate a corresponding OWL document for French language.

```
<owl:Class rdf:about="#description">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#connaissance"/>
  </rdfs:subClassOf>
  <rdfs:label xml:lang="fr">description</rdfs:label>
  <rdfs:comment xml:lang="en">sort or variety; "every
  description of book was there"</rdfs:comment>
</owl:Class>
```



```
xml:base="file:/C:/Users/emhimed/Desktop/DC
<owl:Ontology rdf:about=""/>
<owl:Class rdf:about="description">
  <rdfs:subClassOf>
    <owl:Class rdf:about="knowledge"/>
  </rdfs:subClassOf>
  <rdfs:label xml:lang="en">description</rd
  <rdfs:comment xml:lang="en">sort or varie
</owl:Class>
<owl:Class rdf:about="processing">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="CLASS_ROOT"/>
  </rdfs:subClassOf>
  <rdfs:label xml:lang="en">processing</rdf
  <rdfs:comment xml:lang="en">preparing or
</owl:Class>
<owl:Class rdf:about="knowledge">
  <rdfs:comment xml:lang="en">the psycholog
  <rdfs:label xml:lang="en">knowledge</rdfs
  <rdfs:subClassOf rdf:resource="#CLASS_ROC
</owl:Class>
<owl:Class rdf:about="word">
  <rdfs:label xml:lang="en">word</rdfs:labe
  <rdfs:subClassOf>
    <owl:Class rdf:about="measure"/>
  </rdfs:subClassOf>
  <rdfs:comment xml:lang="en">a word is a r
</owl:Class>
<owl:Class rdf:about="communication">
```

```
tp://www.w3.org/2002/07/owl#">
ethod="xml" indent="yes"/>

 name="dict" select="document('frandictionary1.x

match="/ | node()">
>
choose>
xsl:when test=".=rdfs:label"><xsl:apply-template
xsl:when test=".=rdfs:subClassOf"><xsl:apply-tem

xsl:otherwise><xsl:apply-templates select="@* |
:choose>
y>
e>

match="rdfs:label">
able name="word" select="." />
el xml:lang="fr"><xsl:value-of select="$dict/wor
e>

match="owl:Class">

ute="@rdf:about">
variable name="word" select="@rdf:about" />
variable name="translated_word">
xsl:value-of select="$dict/word[@name=$word]"/>
:variable>
```

```
<owl:Ontology rdf:about=""/>

<owl:Class rdf:about="#description">

  <rdfs:subClassOf>

    <owl:Class rdf:about="#connaissance"/>

  </rdfs:subClassOf>

  <rdfs:label xml:lang="fr">description</rdfs:label>

  <rdfs:comment xml:lang="en">sort or variety; "every desc

</owl:Class>

<owl:Class rdf:about="#Processus">

  <rdfs:subClassOf>

    <owl:Class rdf:ID="CLASS_ROOT"/>

  </rdfs:subClassOf>

  <rdfs:label xml:lang="fr">Processus</rdfs:label>

  <rdfs:comment xml:lang="en">preparing or putting through
```

Fig. 12. Generating ontology for French language by applying XSLT transformer

The corresponding ontology graph is presented in Fig. 13.

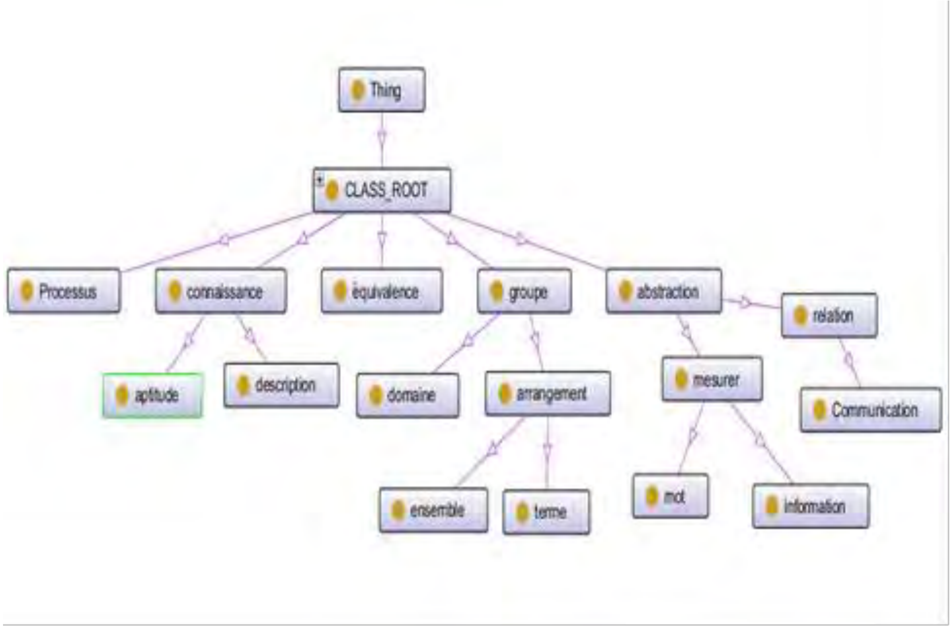


Fig. 13. Ontology graph for French words

Consider another more complex example. Let us have the following French text:

L'informatique est le domaine d'activité scientifique, technique et industriel concernant le traitement automatique de l'information via l'exécution de programmes informatiques par des machines : des systèmes embarqués, des ordinateurs, des robots, des automates, etc.

Ces champs d'application peuvent être séparés en deux branches, l'une, de nature théorique, qui concerne la définition de concepts et modèles, et l'autre, de nature pratique, qui s'intéresse aux techniques concrètes d'implantation et de mise en œuvre sur le terrain. Certains domaines de l'informatique peuvent être très abstraits, comme la complexité algorithmique, et d'autres peuvent être plus proches d'un public profane. Ainsi, la théorie des langages demeure un domaine davantage accessible aux professionnels formés (description des ordinateurs et méthodes de programmation), tandis que les métiers liés aux interfaces homme-machine sont accessibles à un plus large public.

Le terme « informatique » résulte de la combinaison des trois premières syllabes du terme « information » et des deux dernières syllabes du terme « automatique » ; il désigne à l'origine l'ensemble des activités liées à la conception et à l'emploi des ordinateurs pour traiter des informations. Dans le vocabulaire universitaire américain, il désigne surtout l'informatique théorique : un ensemble de sciences formelles qui ont pour objet d'étude la notion d'information et des procédés de traitement automatique de celle-ci, l'algorithmique. Par extension, la mise en application de méthodes informatiques peut concerner des problématiques annexes telles que le traitement du signal, la calculabilité ou la théorie de l'information.

After applying our method, the following ontology graph is obtained (we omit intermediate ontology graph for English words and parts of XSLT transformer):

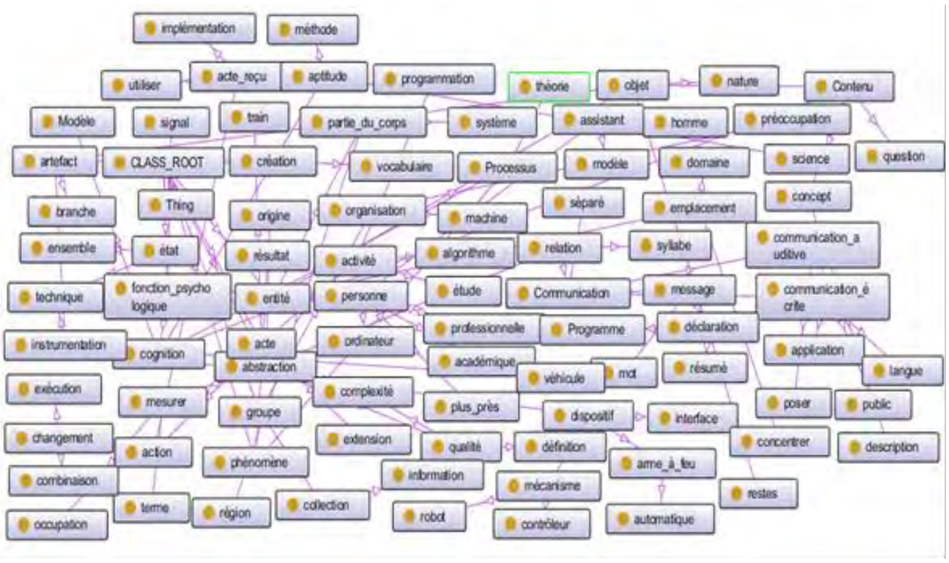


Fig. 14. Ontology graph for French words

5. Discussion

The strong point of our approach is its generality, i.e. the possibility to apply it in same way for any natural language. Our starting goal was to create automatic method for building domain ontologies related to any natural language. Moreover, we conceive that it is not possible to do in this moment. So, we include expert in generating of ontology and make a semi-automatic approach. Participating of an expert (in selection some words) is probably the weakest point of our method. Also, the problems could appear during translation of some text into English language. A lot of new questions arise. For example, let we have a text in the natural language NL1 (denote it with t.NL1) and translate this text into natural language NL2 (denote it with t.NL2). After applying of our method to t.NL1 and t.NL2 will we get same ontology graph, at least will we get similar ontology graph. Special problem is how to measure the similarity of graphs. These problems could be subject of further research.

6. Conclusion and future work

Ontologies are very important in different scientific fields such as: knowledge engineering and representation, information retrieval and extraction, knowledge management, agent systems, and so on. We can say that ontologies represent the backbone of the semantic web. The possibility to create ontology for any natural language gives us an opportunity to work with information that can be processed by both humans and computers in a natural way which is, unfortunately, still difficult to do that automatically. However, semi-automatic implementation of this process, including a human expert, is possible. We described our approach for discovering taxonomic conceptual relations from text facility ontology by using open source software tool DODDLE-OWL. The main challenge we faced is that this software is available only for Japanese and English languages. To address that, we proposed the procedure where DODDLE-OWL is used as an auxiliary tool to build an ontology from the given text for any natural language (referred in our paper as target language). For this approach the other auxiliary tools are necessary as well as an existing WordNet database for the target language, Protégé semantic web editor and Oxygen XML Editor. The main contribution of this paper is the integration of different software tools, which gives new quality and provides the building of ontologies for different natural languages. We plan to perform further analysis of the results and compare the obtained ontology trees using different natural languages with the same input text. We will try to improve the proposed approach by integrating additional software tools and making certain steps simpler.

Acknowledgment. Research was partially supported by the Ministry of Education and Sciences Republic of Serbia, through project no. OI 174010 Mathemaical models and optimization methods of large systems.

References

1. Ahmed, C, M., Hassina, A., Zaia, A.: Automatic Construction of Ontology from Arabic Texts. ICWIT 2012: 193-202. (2012)
2. Alatrish S.E.: Coparison of Ontology Editors, eRAF Journal on Computing, Vol. 4, 23-38, 2012.
3. Antoniou G. and van Harmelen F.: Semantic Web Primer, The MIT Press, 2008.
4. Bajwa S. I.: A Framework for Ontology Creation and Management for Semantic Web, International Journal of Innovation, Management and Technology, Vol. 2, No. 2, April 2011.
5. Bernstein A. and Kaufmann E.: GINO – A Guided Input Natural Language Ontology Editor, Lecture Notes in Computer Science Volume 4273, pp 144-157, 2006.
6. Damljanovic D., Agatonovic M. and Cunningham H.: Natural Language Interfaces to Ontologies:Combining Syntactic Analysis and Ontology-based Lookup through the User Interaction, in: Proc. of the 7th Extended Semantic Web Conference (ESWC) (ed. Lora Aroyo and al.), Springer Verlag, 106-120, 2010.
7. DODDLE-OWL,a Domain Ontology rapiD DeveLopment Environment - OWL extension, [Online]. Available: http://doddle-owl.sourceforge.net/en/.
8. Dominique, E., Chris, N., Andrew, Z.: Towards Ontology-based Natural Language Processing". RDF/RDFS and OWL in Language Technology: 4th Workshop on NLP and XML (NLPXML-2004), ACL 2004, Barcelona, Spain, [Online]. Available: www.aclweb.org/anthology-new/W/W04/W04-0609.pdf. (2004)
9. Haytham A-F., Schafermeier R. and Paschke P.: An Inter-lingual Reference Approach For Multi-Lingual Ontology Matching, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, pp. 497- 503, 2013.
10. Holger, K., Ray, F., Natalya, N., Mark, M.: The Protégé OWL plugin: An open development environment for semantic web applications, The Semantic Web–ISWC 2004, 229-243. (2004)
11. lya, Z., Lei, S., Fausto, G., Wei, P., Qi, J., Mingmin, C., Xuanjing, H.: From Web Directories To Ontologies: Natural Language Processing Challenges, Technical Report # DIT-07-029. May (2007)
12. Janzen, S., Maass, W.: Ontology-based Natural Language Processing for In-store Shopping Situations, Third IEEE International Conference on Semantic Computing (ICSC 2009), Berkeley, California. USA (2009)
13. Kapoor B. and Sharma S.: A Comparative Study Ontology Building Tools for Semantic Web Applications, International Journal of Web & Semantic Technology , vol. 1, no. 3, pp. 1-13, 2010.
14. Karanikolas N. and Skourlas C.: A parametric methodology for text classification, Journal of Information Science, Vol. 36 (4), pp. 421-442, 2010**.**
15. Mehrnoush, S., Ahmad, A.: Learning Ontologies from Natural Language Texts, International Journal of Human-Computer Studies, Volume 60, Issue 1, January, Pages 17-63. (2004)
16. Mellish, C., Pan, J.: Natural Language Directed Inference from Ontologies, Artificial Intelligence 172(10): 1285-1315. (2008)
17. Microsoft Core XML Services (MSXML) 6.0, [Online]. Available: http://www.microsoft.com/enus/download/details.aspx?id=3988.
18. Morita, T., Izumi, N., Fukuta, N.: A Graphical RDF-Based Meta-Model Management Tool. IEICE Transactions 89-D(4): 1368-1377. (2006).
19. MyMemory: next generation Translation Memory technology, [Online]. Available: http://mymemory.translated.net/doc/.
20. Namgoong H. and Kim H-G.: Ontology-Based Controlled Natural Language Editor Using CFG with Lexical Dependency, ISWC/ASWC, 353-366, 2007.
21. Nitin, I., Fred, J, D (editors).: Handbook of Natural Language Processing, Second Edition, Taylor & Francis. (2010)

22. Oxygen XML Editor 14.2. [Online]. Available: http://www.oxygenxml.com/doc/Editor-UserManual.pdf
23. Protégé ontology editor developed by Stanford Medical Informatics, Stanford University School of Medicine, further information, [Online]. Available: http://protege.stanford.edu/.
24. Szulman S., Charlet J., Aussenac-Gilles N., Nazarenko A., Sardet E. and Téguiak H.V.: DAFOE: An Ontology Building Platform - From Texts or Thesauri, KEOD 2009: 372-375
25. Takeshi, M., Naoki, F., Noriaki I., Takahira, Y.: DODDLE-OWL: Interactive Domain Ontology Development with Open Source Software in Java. IEICE Transactions 91-D(4): 945-958. (2008)
26. Takeshi, M., Naoki, F., Noriaki I., Takahira, Y.: DODDLE-OWL: A Domain Ontology Construction Tool with OWL. ASWC 2006: 537-551. (2006)
27. Vesin, B., Ivanović, M., Klašnja-Milićević, A., Budimac, Z.: Ontology-Based Architecture with Recommendation Strategy in Java Tutoring System. Computer Science and Information Systems, Vol. 10, No. 1, 237-261. (2013)
28. Wei, M., Xu, J., Yun, H., Xu, L.: Ontology-Based Home Service Model. Computer Science and Information Systems, Vol. 9, No. 2, 813-838. (2012)
29. Wilson, W., Liu, W., and Bennamoun, M. 2012. Ontology learning from text: A look back and into the future, ACM Comput. Surv. 44, 4, Article 20 (August 2012), 36 pages, DOI=10.1145/2333112.2333115 http://doi.acm.org/10.1145/2333112.2333115
30. Wolf, F. and Bernhard, B.: Combining Ontologies And Natural, Language, [Online]. Available: krr.meraka.org.za/~aow2010/Fischer-etal.pdf. (2010)

Emhimed Salem Alatrish was born in 03/09/1972 at Yefren in Libya. He finished primary and secondary school in Yefren, and then graduated from the Faculty of Science and Arts in 1996. After that he started master study in Cultural University in Istanbul and ended it in 2005/2006 school year. From 2008 year he is enrolled in the doctoral program at the Faculty of Mathematics in Belgrade.In the period from 1997 to 2001 year he worked as an assistant professor at the Department of Physics, Faculty of Science and Art in Yefren in Libya.

Dušan Tošić received his B.Sc. degree in mathematics 1972, M.Sc. in mathematic 1977 and PhD in mathematics 1984 from University of Belgrade, Faculty of Mathematics. Since 2003 he is a full professor of computer science at Faculty of Mathematics, University of Belgrade. Professor Tošić is an associate editor for the Journal ComSIS (Computer Science and Information Systems) and reviewer of Zentarblat. His publication list contains more than 100 titles of articles and books. He has supervised about 20 candidates for PhD and M.Sc. degree. On the beginning his research interest was numerical solving of differential equations. After that he was oriented toward parallel algorithms, optimizations, genetic algorithms and teaching of computer sciences. He is Fellow of Serbian Mathematical Society and member of Committee Serbian Society of Computer Sciences.

Nikola Milenković was born in 11/18/1987 in Belgrade. He finished primary school in Barič and secondary school “Nikola Tesla” in Belgrade, and then graduated at the Faculty of Mathematics in 2009, Computer Science course. After that he started master studies at the Faculty of Mathematics and ended it in 2011. On the same year he enrolled doctoral studies at the Faculty of Mathematics. Since 2009 he worked at a various IT companies as a software developer, team leader and project manager. In 2013 he cofounded software development company in Belgrade where he currently holds position of director.

Received: April 29, 2013; Accepted: March 30, 2014

Computer Science and Information Systems 11(2):645–664 DOI: 10.2298/CSIS130824025P

A Question-Based Design Pattern Advisement Approach

Luka Pavlič, Vili Podgorelec, and Marjan Heričko

Faculty of Electrical Engineering and Computer Science, University of Maribor
2000 Maribor, Slovenia
{luka.pavlic, vili.podgorelec, marjan.hericko}@uni-mb.si

Abstract. Design patterns are a proven way to build flexible software architectures. But the selection of an appropriate design pattern is a difficult task in practice, particularly for less experienced developers. In this paper, a question-based design pattern advisement approach will be proposed. This approach primarily assists developers in identifying and selecting the most suitable design pattern for a given problem. We will also propose certain extensions to the existing Object-Oriented Design Ontology (ODOL). In addition to the advisement procedure, a new design pattern advisement ontology will be defined. We have also developed a tool that supports the proposed ontology and question-based advisement (OQBA) approach. The conducted controlled experiment and two surveys have shown that the proposed approach is beneficial to all software developers, especially to those who have less experience with design patterns.

Keywords: design patterns, pattern selection, ontology, semantic web, selection algorithm.

1. Introduction

One of the basic characteristics of any engineering discipline is that new systems are built and developed from existing, already proven reusable elements using well known approaches and best practices. Reuse has become an essential and important strategy - also in the area of software and information systems development. The reuse of concrete assets and software elements such as functions, classes and components has already been well established and continues to be practiced on a daily basis. Attention should also be placed to reuse at higher levels of abstraction i.e. to software patterns. A pattern is a form of knowledge for capturing a recurring successful practice [1]. Design patterns capture the best practices for solving recurring software design problems. They explicitly capture knowledge that experienced developers understand implicitly and facilitate training and knowledge transfer to new developers [2]. A survey conducted by the Microsoft Patterns and Practice Group [3] indicated a low adoption of design patterns among practitioners: respondents estimated that no more than half of the developers and architects in their organization use software patterns. Therefore, bridging the gap between expert design pattern communities and the typical design pattern user is critical for achieving the full benefits of design patterns [4].

In [5] we can find an observation that it might be difficult to find a suitable design pattern even in a catalogue such as the GoF (Gang Of Four) Design Pattern Catalogue – with no more than 24 patterns. Several hundred software patterns have already been published. The Pattern Almanac [6] published in 2000, provided a list of over 700 previously published patterns organized into 70 categories. Since then the number of software patterns and catalogues has increased significantly. Consequently, software developers have been experiencing more and more problems identifying appropriate catalogues and finding patterns that match their design problems. Useful patterns might easily be overlooked. That is why a manual identification and application of patterns is not efficient enough. For the efficient identification and selection of suitable design patterns, automated support is of critical importance.

There were some efforts within the research community to automate the application of design patterns [7,8,9]. Mainly these efforts were focused on code generation and the identification of design patterns in existing designs and/or source code and less on the selection of suitable design patterns. Although it is not reasonable to believe that the responsibility for selecting design patterns will be completely delegated to tools, they could provide at least some advice for design pattern(s), which may be potentially useful in a given situation. In order to be able to develop and provide efficient automated tools, we need an adequate design pattern description approach. An appropriate knowledge representation technique should be computer-readable, based on standard technologies, extendable and relatively simple for developers to work with. Consequently, the main focus and aim of this work is twofold:

- To define a means for capturing information and knowledge on design patterns in a form that would enable a computer to process and use it in a more intelligent way whilst also keeping it readable for humans.
- To provide assistance for software developers searching for an appropriate design pattern for a given design problem.

The main idea of the proposed question-based approach originates from our undergraduate course on software design patterns. Our students are taught how to identify suitable design patterns and/or choose the most appropriate one when they are unsure about two or more design patterns (e.g. whether to use the Adapter or Façade pattern). In order to select the most appropriate design pattern some questions -- such as those presented in Table 1 -- have proven to be helpful.

Table 1. Sample "recipe" on how to differentiate the applicability of two design patterns

Question		
Could the necessary functionality be found in the existing classes?	Yes	Yes
Do we need a simpler interface for existing classes?	Yes	No
Is there a predefined interface that a class under development should be compliant with?	No	Yes
Are class objects expected to demonstrate polymorphic behaviour?	No	Probably
Do we want to change the way of how existing functionalities are used?	Yes	No
Possible solution	Facade	Adapter

Figure 1 shows the holistic view of the proposed Ontology and Question-Based Advisement (OQBA) approach. An ontology-based technique is used to represent knowledge on design patterns, catalogues, pattern containers and pairs of questions/answers used for reasoning regarding a suitable pattern for a particular design problem. This design-pattern expert knowledge, gathered in the Ontology-Based Design Pattern Repository, is then used for guiding the question/answer interaction with the developer in order to identify and propose a suitable design pattern, as well as to verify its applicability for a given situation.

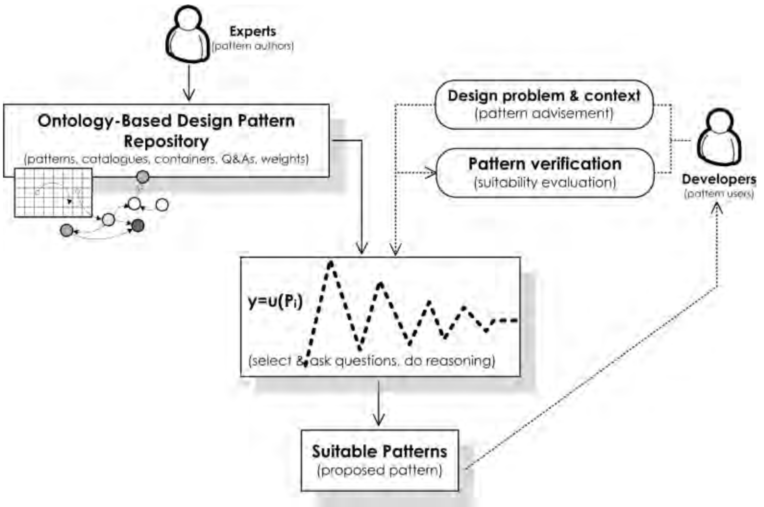


Fig. 1. The holistic view of the proposed OQBA approach

The approach presented in this paper is focused on object-oriented design patterns. To show the benefits, we experimented with the GoF design patterns catalogue, since even catalogues with few patterns proved to be problematic when identifying the most suitable pattern[5]. At the moment we cannot claim the approach to be general to all software patterns. However, it is our belief that a primary idea of question-based advisement approach could eventually be used also for other software patterns. However for that it would be necessary to formalize additional concepts and relationships that are specific to other software patterns groups.

This paper is organized as follows: An overview of related work and research is given in the second section. The third section discusses problems related to design pattern descriptions and presents our extensions to an existing ODOL ontology (Object-Oriented Design Ontology). It also introduces a new ontology named DPAO (Design Pattern Advisement Ontology). Section four gives a general description of the proposed question-based approach and presents algorithms for assessing the usability of patterns. In addition, the advisement procedure as it is implemented in the DPEX (Design Pattern Expert) tool is introduced. The description of the conducted surveys and experiment with their results are given in the fifth section. Finally, some concluding remarks and directions for future work will be presented.

2. Related Work

Kung et al. [4] presented a five-step methodology for constructing an expert system is presented that suggests design patterns to solve a design problem. The focus is on the collection and analysis of knowledge on patterns in order to formulate questions, threshold values and rules to be used by an expert shell. A prototype of the Rule-Based tool (for a subset of GoF patterns) was developed. It selects the design pattern through dialogue with the software designer to narrow down the choices. A preliminary evaluation of the proposed methodology was done on a group of ten students. The focus of the experiment was not on evaluating the efficiency of a tool in finding a solution for a given problem situation. Rather it was aimed at proving that, with a known design pattern, a tool would lead the user to a suitable suggestion or solution. Therefore, the focus was on verifying that a tool would confirm the suitability of a particular design pattern. Our approach is different: we want to suggest which design pattern might be useful in given context.

Gomes et al. [10] introduced similar approach that is based on Case-Based Reasoning (CBR) and lexical database WordNet [10]. The approach is based on the idea that a system can learn to select and to apply design patterns if it can store and reuse knowledge on pattern usage experience. An application of a specific pattern to a specific software design is represented in the form of a design pattern application (DPA) case. A DPA case describes a specific situation where a software design pattern was applied to a class diagram. Our approach does not require a developer to prepare a model first. The information needed to identify a suitable solution is gathered using a guided dialogue, where a developer provides information on problem characteristics by choosing items from the lists of available answers.

Birukuo et al. [11] proposed a multi-agent system that supports collaboration between developers in order to chose the suitable design pattern for a given problem. Suggestions are made using experiences from a group of developers who were previously faced with a similar problem. The problem for which an appropriate pattern is searched for is described using a “bag of words” approach as a sequence of terms. Having a vector that maps these terms to a vocabulary of all terms, a search for patterns can be initiated at all agent nodes. The approach assumes that the developers will provide a textual description of the problem. Their work could be seen as a complementary approach to ours – in the case that our approach would not result in a suitable recommendation, a term vector could be automatically formed and used as an input to their system.

Ontologies have already been successfully used to describe design patterns within the scope of the ”Web of Patterns“ (WoP) project [8]. The main goals of the WoP project were (1) to define an ontology with which object-oriented models could be described, (2) to describe design patterns, anti-patterns and refactorings and (3) to develop tools based on this ontology which would be useful for software engineers. The main focus of the project was on finding pattern instances in Java projects as well as on refactoring and anti-patterns. The project is not directly related to our work. However, one of the artifacts of the WoP project is the Object Oriented Software Design Ontology (ODOL), which contains a set of concepts and relationships used in all of the more important object-oriented (OO) languages. ODOL includes basic OO concepts like classes and methods, OO design concepts and high-level concepts like design patterns, pattern categories, aggregations etc. ODOL (available at [12]) is an open ontology so it is possible to extend it with concepts that are currently not present in the ontology. During our research we extended already existing ontology ODOL with proposing capabilities as described later in paper. It was extended with additional concepts and a new ontology that was needed to provide an infrastructure for the application of the question-based approach.

3. Using Ontology for Describing Knowledge of Design Patterns

3.1. Towards a More Suitable Design Pattern Description

In order to develop a tool that would assist and advise the developer on the most appropriate design patterns, or a combination of patterns, for a given design problem, knowledge and experiences on design patterns should be gathered and described. Since 1994, when design patterns were introduced, many different approaches have been used to document them. In general there are three main categories of descriptions: (1) informal representations, (2) semiformal representations based on graphical notations such as UML, and (3) various formal representations which also include notations using semantic web technologies. Design patterns are traditionally described using natural language and published in printed catalogues [5]. These documents are loosely structured in a canonical form which consists of a series of fields: name, intent, applicability, structure, participants, consequences, implementation etc. Because of its loose structure, this kind of representation is less suitable for knowledge management and sharing [13]. Informal representations based on a canonical form do not enable the desired and necessary level of design pattern identification and application. For this we need more structured representation forms. This has led some researchers to devise more formal presentations, mainly by using existing mechanisms of UML or by extending UML specifications [14,15]). Design patterns are usually described using class diagrams and interaction diagrams – primarily parameterized communication diagrams. The main drawback of these approaches is an over-reliance on visual specifications with UML diagrams and limited support for the behavioural aspects of design patterns [16]. They are efficient for a basic understanding of patterns since they cover their structural elements. But they do not provide information and knowledge on high level aspects such as intent, usability and consequences. The introduction of automated support requires a formal approach to design pattern description.

3.2. How Can Ontologies and Semantic Web Technologies Help?

The semantic web enables knowledge to be expressed in a way that enables machine processing and its use in web environments by both intelligent agents and human users

[17]. It is considered to provide an efficient way to present data, information and knowledge on the internet or in the scope of a global interconnected database. Since many semantic web technologies have reached high community consolidation and have become W3C standards (including RDF - Resource Description Framework and OWL - Ontology Web Language) it can also be considered as a long-term platform for intelligent services based on a common knowledge base [18].

One of the enabling approaches used in the semantic web is metadata. It is supported with the concept of ontologies and has a foundation in W3C standards. Ontology describes the subject domain using the notions of concepts, instances, attributes, relations and axioms. Concepts are typically organized into taxonomies through which inheritance mechanisms can be used in an ontology. Ontologies build on description languages such as RDF(S) and OWL, and add semantics to the model representation. Their formal, explicit and shared nature makes them an ideal object repository for catalogues.

With the presented facts we also justify our decision to use ontologies as well as other semantic web technologies to provide a basis not only for a design pattern description but also for providing additional information, relations and rules, needed to define and implement the OQBA approach:

- Ontology-based design pattern descriptions are computer readable and therefore suitable for automated (computer) processing.
- If design pattern descriptions are provided in the form of ontological definitions, they can be presented in textual or graphical form as well as transformed to presentation forms customized for developers (developer friendly representation).
- Ontology and related technologies are well established, recognized and also extendable, whereas the semantic web is becoming an enabling factor for better knowledge management and the management of a high volume of data that still has to be inter- and cross-linked.
- Ontologies enable the establishing and revising of a knowledge base on design patterns based on common, accepted standards and technologies.
- Navigation based on relationships between patterns helps with a better understanding of a pattern space [19]; consequently ontology-based descriptions are ideal because navigation-based relationship investigation capabilities are inherent to ontologies and semantic web technologies.
- Ontologies enable the exchangeability of design pattern descriptions and are extendable.
- Several knowledge sources can easily be integrated using relatively simple transformations.
- The knowledge base can be distributed (on the web or in closed networks).
- Third-party ontology (OWL)–enabled tools are available, which can extend the use of an ontology-enabled knowledge base.
- The behaviour of developed system can be improved simply by changing the ontology and/or data without changing the system itself.
- Even generic search engines can be impacted to retrieve more reasonable results by using ontology-rich data.

These were the main reasons to use the results of the WoP project and ODOL ontology as a starting point for defining a novel approach and tools that are based on ontologies and other semantic web technologies. Alongside the definition of the question-based advisement approach we have developed a new ontology, called DPAO (Design Pattern Advisement Ontology) that represents the foundation for the proposed question-based approach.

The WoP project was oriented toward design pattern instance identification and not on the selection of suitable design patterns. Consequently, ODOL ontology does not provide and/or enable one to capture all the information needed to introduce and apply the OQBA approach. We needed additional information that would direct and guide the advisement procedure as well as the process of verification so that the selected design pattern actually represents a reasonable solution for a given problem. That is why we extended the ODOL ontology with additional concepts that would enable the efficient grouping of design patterns. In addition, we developed a new ontology, named DPAO, which provides the infrastructure for two main aspects of the advisement:

- Advisement on an appropriate design pattern for a given problem situation: Developers are faced with a problem for which they presume that the solution in the form of pattern already exists, but they have no idea which pattern it is or which pattern group or catalogue they should look at.
- Advisement on the applicability of a particular design pattern for a given problem situation: Developers believe they know which are suitable design patterns within a given context, however they would like to verify their choice and/or evaluate the suitability of the identified design pattern(s).

3.3. Extending ODOL with New Concepts and Relations - Pattern Container, Related, Alternative and Composed Patterns

The existing ODOL ontology has two concepts for identifying and classifying design patterns: Pattern and PatternCatalog. This two-level hierarchy becomes insufficient when the number of design patterns stored in the catalogue starts to rise. PatternCatalog can only inlcude Patterns, not other also PatternCatalogs. For these reason we have defined a new concept: PatternContainer. As the name suggests, it should serve as a container for patterns and/or other pattern containers. PatternContainer provides a means to group an arbitrary set of patterns based on selected aspects. In addition, the same design pattern can be an element of many different groups/categories. Even more: design patterns can be related to each other while being part of different containers/catalogues. As can be seen in Figure 2, PatternContainer becomes the main class for pattern grouping. In Figure 2, the original ODOL concepts are shown as shadowed and/or written in italics whereas new concepts are in bold. The existing PatternCatalog concept then becomes just a type of PatternContainer. In order to provide efficient advisement some additional concepts were added to ODOL. They enable us to connect an existing design pattern with possible related patterns such as:

- Alternative design pattern that could represent a reasonable alternative to a particular pattern – e.g. Façade to Adapter, Visitor to Observer, Decorator to Adapter, Builder to Abstract factory.

- Related design pattern as a pattern for which it is very likely to be used in combination with a selected pattern, e.g. Chain of Responsibility together with Composite, Memento together with Command, Abstract Factory implemented with Singleton, Composite processed with Iterator.
- Composite pattern that provides higher granularity of interrelated patterns, e.g. the MVC (Model-View-Controller) pattern is a combination of patterns and incorporates Composite, Observer and Strategy.

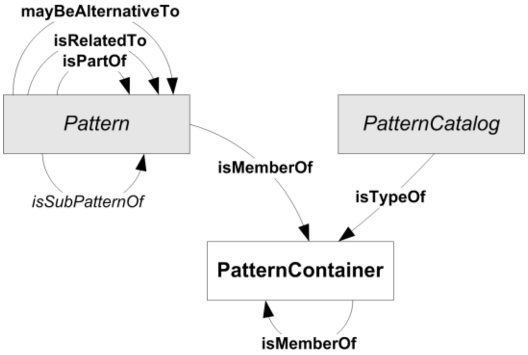


Fig. 2. Extensions (in bold) of the ODOL ontology

With these extensions, the descriptions of GoF design patterns in ODOL ontology can be updated, for instance, with definitions for the three GoF pattern categories, namely: creational, structural and behavioural patterns. Existing definitions of GoF patterns could then be expanded with categories as well as information on related and alternative patterns. ODOL ontology has been extended with a few additional relations that are of crucial importance for efficient automation and support in the process of suitable design pattern identification and selection.

3.4. Design Pattern Advisement Ontology (DPAO)

In addition to the knowledge on patterns, a precondition for successfully advising on patterns is the tool. It should be able to ask the right questions and according to the answers lead the dialogue with a developer until enough certainty is reached to propose a certain design pattern. We need an ontology that provides the basis for: (1) the selection of appropriate patterns and (2) the verification of candidates that are appropriate for a particular design problem. The developed DPAO ontology meets both goals.

Figure 3 shows the concepts that cover the aspect of selecting the most appropriate design pattern for a given design problem in a particular context. AdviceArea class represents a set of matrices (AdviceMatrix) which are connected in a graph. The advice process starts with the initial matrix. The initial matrix typically holds the knowledge necessary to make decisions about general context whereas latter matrices can also hold some context-specific knowledge. AdviceMatrix is three-dimensional structure and is constructed from questions, answers and pattern/pattern container candidates. A candidate is modeled with the class AdviceMatrixNode and can be a terminal one (AdviceMatrixLeafNode leading to particular Pattern or PatternContainer) or a link to another AdviceMatrix (AdviceMatrixNodeMatrix). The matrix content is modelled as AdviceMatrixCell, holding in CellAssessment the weight value of how relevant a candidate can be if a particular answer for a given question is selected. Since assessments can be given by more than one expert, assessments are grouped into AssesmentSets. Expert knowledge can be gathered using simple in-house developed tools. Weights, questions, answers and their relations to patterns and pattern containers are freely inserted by experts. For example data and meaning during the dialogues please see section 4.1.

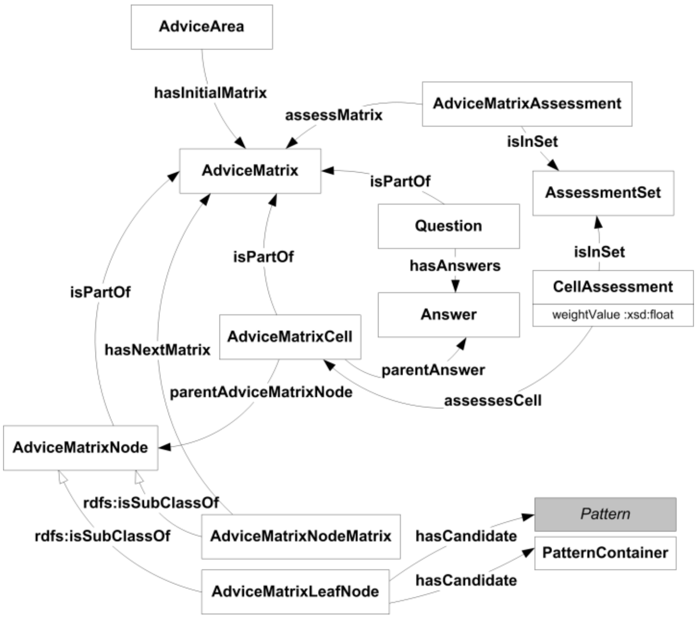


Fig. 3. Concepts in DPAO ontology enabling the selection of relevant design pattern(s)

The second aspect, supported by the new approach, is to verify the relevance or possibility to use a pattern in a known problem situation. In this case the developer has already selected a solution (pattern) to be used and would just like to verify if the provided pattern is relevant or there are more relevant patterns. This aspect is covered by the ontology concepts depicted in Figure 4.

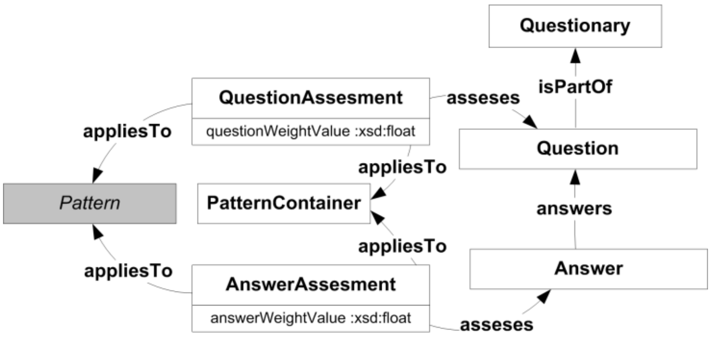


Fig. 4. Concepts in the DPAO ontology enabling relevance verification of a particular design pattern

Obviously, the DPAO ontology enables us to connect some question-answer pairs to a candidate, i.e. pattern or pattern container. Most of the introduced concepts in Figure 4 are known from prior figures. Questionary is attached via Question instances to a particular candidate. The weight of the question is modelled with the questionWeightValue held as a numeric value in QuestionAssessment, and respectively in the answerWeightValue attribute in AnswerAssessment for every possible answer to the question. The details of expert-inserted weights in terms of meaning and usage is described in details in section 4.1. Experts use user-friendly tool for inserting questions, answers and weights, as well possibility to review and alter already existing data in the ontology.

4. The Advisement Procedure

Utilizing knowledge on design patterns as well as a set of question/answers pairs that indicate the applicability of design patterns, we can apply the advisement procedure. In general the question-based advisement procedure consists of five phases:

- **Phase 1: Reducing a set of possible solutions to a subset of pattern containers** Using knowledge in the Ontology-Based Design Pattern Repository, especially advice matrices, and based on user's answers, the overall intention of applying a design pattern should be revealed, e.g. is it related to a specific development phase, technology, domain or any other criteria used to form pattern containers in the repository. The procedure is directed by initial matrices and corresponding question/answer pairs, using algorithms and the usability function presented in chapters 4.1 to 4.3. As a result we get a set of containers that might include a suitable solution.
- **Phase 2: Identifying the most suitable pattern container** An additional narrowing down is necessary in order to identify the atomic container with concrete design pattern candidates. The selection of an appropriate container is also based on advice matrices and algorithms, presented in the sections that follow. As a result, we get a single pattern container. e.g. - after the GoF container is

selected in the first phase, in the second phase we would identify one of the GoF pattern containers, namely behavioural, creational or structural patterns.

- **Phase 3: Selecting the most suitable pattern in a given design pattern container** In this phase, we identify the most appropriate design pattern within the container found in the previous phase. To achieve this we select questions that are common to as many patterns in the container as possible. For this we use groups of questions and answers for relevant patterns. The result is advice based on the most relevant design pattern and eventually related alternative solutions (design patterns).
- **Phase 4: Verification of the proposed design pattern** In this phase, the verification is done to see if any of alternative patterns can be equally or even more relevant. At this stage all questions and trade-offs, connected with a selected pattern, are taken into consideration especially those that can identify negative effects. The same is done for eventual alternative patterns. If the alternative pattern demonstrates less negative consequences it is applied instead of the initially proposed design pattern. The aim of this phase is also to eliminate possible mistakes, caused if a developer misunderstands a question in the advisement procedures.
- **Phase 5: Identifying related design patterns** We want to identify design patterns that might be useful in combination with the already selected pattern. A list of all design patterns that are related to the selected verified design pattern is constructed first. These patterns are then evaluated using the same procedure as applied in phase 4. Thus, composite design patterns can also be identified and applied.

4.1. Calculating the Pattern / Pattern Container Usability

Our aim is to identify and suggest the most appropriate design pattern that could be used by a user in the situation described. For this purpose the usability u(P) of each design pattern container P (in phases 1 and 2) or design pattern P (in phase 3) is calculated as:

$$u(P) = \frac{\sum_{q_i \in Q^A} w_q(P,q_i) \cdot w_a(P,q_i,k)}{\sum_{q_i \in Q^A} w_q(P,q_i)} \qquad \textbf{(1)}$$

where

P is a design pattern or pattern container (category),

$w_q(P,q_i)$ is the weight of a question q_i regarding the given P

$w_a(P,q_i,k)$ is the weight of the obtained answer k for the question q_i regarding the given P

Q^A is the set of already answered questions

As we can see from the (Eq. 1), the same function is used to determine the usability of a design pattern or a pattern container. This is possible because the same mechanism of questions/answers matrices are used throughout the advisement process.

In the following text, we will use the term "design pattern" to explain the algorithm; please note that the same procedure is also used to identify and select the pattern containers.

The usability u(P) of each design pattern P is 0 when no question has been answered. After a user answers a question q_i, the usability u(P) of each pattern P changes regarding the values of the weight for this question ($w_a(P,q_i)$) and weight for the given answer ($w_a(P,q_i,k)$); the weights are pre-determined by experts. A gain for a pattern P, obtained by answering the question q_i with answer k is calculated as:

$$gain(P,q_i,k) = \frac{w_q(P,q_i) \cdot w_a(p,q_i,k)}{\sum\limits_{q_j \in \{Q^A + q_i\}} w_q(P,q_j)} \tag{2}$$

To clarify this concept, let us consider the following example: there are three possible patterns P_1, P_2, and P_3, from which we want to select the most appropriate one. From the advisement matrix the question q_n is chosen as the first question to be asked with the weight value $w_q(P,q_n)=0.3$. There are two possible answers (k=1 or k=2) available for the question q_n with the following weight values:

- When choosing the first answer k=1, the weights are $w_a(P_1,q_n,1)=-0.3$, $w_a(P_2,q_n,1)=0.0$, $w_a(P_3,q_n,1)=0.8$.
- When choosing the second answer k=2, the weights are $w_a(P_1,q_n,2)=0.5$, $w_a(P_2,q_n,2)=0.7$, $w_a(P_3,q_n,2)=0.0$.

Let us now presume that the user chooses the second answer (k=2). Using (Eq. 2) the gains for patterns P_1 through P_3 are the following (note that no questions have been answered yet): $gain(P_1,q_n,2)=0.5$, $gain(P_2,q_n,2)=0.7$, $gain(P_3,q_n,2)=0.0$.

For the chosen question q_n and the given answer k=2, the pattern P_2 gains the most and is temporarily (after answering only one question q_n) considered to be the most appropriate design pattern. Naturally, the design pattern having the highest usability score u(P) (Eq. 1) is ultimately selected after obtaining enough answers to the given questions.

The weight values for questions range from 0 to 1, meaning:

- If the weight value for a question is 0, the answer will have no effect on the selection of design patterns (unimportant question),
- If the weight value for a question is 1, the answer will have the maximum effect on the selection of the design pattern (the most important question), and
- Weights between 0 and 1 determine the importance of the question – the higher the weight value, the more important the question.

The weight values for the answers also range from -1 to 1, meaning:

- If the weight value of an answer is less than zero, choosing this answer will include some penalty for a design pattern,
- If the weight value of the answer is 0, choosing this answer will not change the usability of the design pattern, and if the weight value of the answer is greater than zero, choosing this answer will add something to the usability of the design pattern.

4.2. Dynamic Question Selection

Since the weights for questions and answers are pre-determined by experts, the order of choosing questions and/or assessment matrices is determined by the selected strategy and goals of the advisement process. If we want to identify and suggest an appropriate design pattern as soon as possible (with a minimum number of questions), a question should be chosen that maximizes the difference in usability functions for the most appropriate design pattern Pfirst and the second most appropriate design pattern Psecond. This can be determined by calculating specific gains for each of the patterns and all the remaining (unanswered) questions. As, naturally, we do not know which answer will be given by a user, we have decided to use the min-max algorithm, well known from game playing, for selecting the next question; it has been implemented in the prototype tool, used to obtain the results in the experiment performed (see section 5). The algorithm is presented in Figure 5.

```
findTheMostRelevantQuestion(Matrix m)
create a list L of all unasked questions from a matrix m
if notEmpty(L)
  foreach design pattern Pi
    foreach question q in list L
     calculate gain(Pi,q,k) for each possible ans. k according to (Eq. 2)
     determine minimal gain(Pi,q,k)
    determine maximal of all mininal gains
    select question q where the mininal gain(Pi,q,k) is maximal
else
  find next matrix or solution
```

Fig. 5. OQBA pseudo code algorithm for choosing the most relevant question

4.3. The Basic Design Pattern Selection Algorithm

The proposed ontology and question-based design pattern advisement approach is based on the proposition that the developers can adequately describe their specific situations following an interactive question/answer session. From their answers, enough information should be obtained to automatically identify and suggest appropriate design patterns. Based on the set of already answered questions, a level of usability is calculated that determines the appropriateness of each specific design pattern or pattern container. This set of answered questions is also used in combination with the remaining questions to determine the next most relevant question, until the usability of the most appropriate design pattern is dominant over all the other patterns. The overall algorithm that can be used for both the solicitation of pattern containers (phases 1 and 2) and for the selection of design patterns (phase 3) is presented in

Figure 6. The process of identifying appropriate candidate solutions (pattern containers or design patterns) is presented. The integral part of the algorithm (line 5) is also the procedure for finding the most relevant question described above.

```
QBA algorithm(AdviceArea aa)
1  for a given AdviceArea aa
2    ma = getMatrix(aa)
3    Qa = {}   ; set of already answered questions
4    do
5      q = findMostRelevantQuestion(aa)
6      a = getAnswer(q)
7      Qa = Qa + q
8      foreach pattern P in aa
9        calculate u(P) using (Eq. 1)
10       calculate remaining gain(P) using (Eq. 2)
11       ;the remaining gain is the sum of max possible
12       ;gains of all the remaining questions until
13       ;remaining gain (P) < u(Pfirst)-u(Psecond)
14       ;if the dominant pattern cannot be changed
15       ;with the remaining questions
16   if isContainer(Pfirst)
17     ma=nextMatrix(Pfirst)
18   else
19     displaySolution(Pfirst)
```

Fig. 6. OQBA pseudo code algorithm for choosing the most appropriate pattern container/design pattern

5. Concept Verification: The Experiment

In order to evaluate the efficiency of the proposed OQBA approach, a controlled experiment was designed and conducted. Two surveys were also taken – one before and one after the experiment. This section presents methodology, results and discussion on results.

5.1. Experiment Introduction and Methodology

The experiment was aimed at confirming or rejecting the following propositions:

- (P1) The OQBA approach contributes to more successful design pattern identification in comparison to manual identification.
- (P2) Less experienced developers benefit the most from using the OQBA approach in identifying the appropriate design patterns.

The experiment was carried out in four phases: in the first phase, participants had to answer a few questions concerning their development background and level of expertise (see Table 2). In the second phase, they were given the opportunity to solve a set of design problems manually, without the DPEX tool (but they were allowed to use other instruments such as books and established search facilities). In the third phase, the DPEX tool was provided to help them solve the same design problems. The fourth phase was aimed at gathering participants' opinions and reflections on the tool. For this purpose a post-experiment survey was conducted.

Two groups of developers were involved in the experiment. The basic profile of the participants and differences between groups are summarized in Table 2. The results of this self-assessment on GoF patterns knowledge is given in the last row in Table 2.

Table 2. Group profiles – used in experiment

Group no.	I	II
No. of participants (profile)	10 (software engineers / developers)	11 (students)
Formal training in GoF design patterns	None	All – one year ago
Active in production software development	less than 3 years: 0 3-6 years: 5 7-9 years: 3 10 years or more: 2	less than 3 years:4 3-6 years:6 7-9 years:1
Knowledge of GoF design patterns (self-assessment)	Weak: 1 Good: 5 Very good: 2 Excellent: 2	Weak: 4 Good: 4 Very good: 3 Excellent: none
How often they use design patterns	Never: 2 Rarely: 7 Often: 1	Never: 3 Rarely: 7 Often: 1

Each participant was given the descriptions for nineteen problem situations in which a particular GoF design pattern might represent a suitable solution. The participants were asked to propose the most appropriate design pattern to solve each of the given design problems. Design problems were chosen to represent real-life situations, not academic ones.

When preparing problem descriptions we intentionally selected problem situations with higher complexity and as we will describe later, some were even too complex. We wanted to evaluate the approach for situations that are similar to those that developers are faced with during their daily work.

Using the tool, they needed on average 40 minutes to complete all 19 tasks whereas an average completion time without the tool was 65 minutes (55 minutes for the fastest participant, 80 minutes for the slowest). Undoubtedly this difference is not caused by the tool, but due to the fact that they were familiar with the problems since they had been solving them in the previous stage of the experiment.

5.2. Results of The Experiment

We can observe that using DPEX tool, on average developers in the first group identified the appropriate design pattern in 58% of all cases. The approach and tool helped achieve more than 50 percent more correct solutions to design problems (40% without the tool).

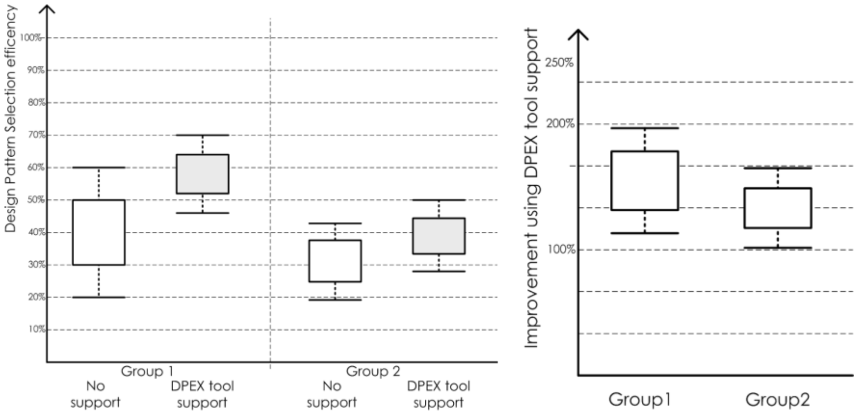


Fig. 7. Efficiency and improvement factors on solving problems with and without the DPEX tool – Group 1 and Group 2

Using the paired t-test we determined that for Group I the difference between the number of correct solutions found with and without the DPEX tool is statistically significant (P = 0.000706). The results of the post-experiment survey have shown that only one of the participants found the use of the tool to be less efficient than searching for an appropriate solution without the tool.

Only in four cases did the developer not accept and agree with the solution advised by the tool. In general the tool increased the efficiency of less experienced developers, and also of experts. The information on standard deviation for successfully solved problems with and without the tool confirms that by using the tool, less experienced developers manage to achieve results that are closer to those demonstrated by experts (Figure 7). After all, this was one of our main goals: to bridge the gap between the pattern expert community and the typical pattern user.

The use of the tool also resulted in improvements in the second experimental student group. The average improvement factor was 1.33. A paired t-test for Group 2 confirmed that there is a significant difference between the results achieved with and without the tool (P=0.001575).

The post-experiment survey showed that three developers believed it was easier to search for a solution using a tool, while four participants were convinced that it was easier to work without the tool. Four developers found the tool easy to use and beneficial to them.

The efficiency of both groups have improved significantly using the DPEX tool (group 1 by factor 1.63 and group 2 by factor 1.33). We can observe that the improvement factor for Group 1 is surprisingly higher. The discussion of the results follows in next section.

5.3. Discussion and Future Work

As shown by the experiment, most developers were able to solve more design problems by using DPEX tool. This shows that our approach assure better achievements regarding design pattern use. Let us discuss the propositions.

- P1: The OQBA approach contributes to a more successful design pattern identification in comparison to manual identification. P1 was confirmed since results confirmed that the use of tool and implemented approach improved the achievements of developers. We can detect the improvements in both groups.
- P2: Less experienced developers benefit the most from using the OQBA approach in identifying the appropriate design patterns. Proposition is not confirmed with the experiment. Group I average improvement factor is 1.63, while average improvement factor for Group II is 1.33.

It is a surprise that improvement factor is higher in Group I (experts). The reason might be in experiences, that professional developers have in terms of understanding questions and giving appropriate answers. However, interesting observation could be identified while comparing standard deviations for both groups. Decreased standard deviation show that the design pattern selection efficiency is becoming more equal using the DPEX tool. Less-experienced developers are obviously becoming more similar to experienced developers.

Additional analysis will be necessary (pre and post-experiment survey, log files) in order to improve the advisement on selecting a suitable pattern. It is also our aim to upgrade the approach with learning capabilities to ask personalized questions.

As future work, we plan to develop a holistic methodology for design pattern selection including automatic question forming based on the analysis of paths (recorded in logs) taken by developers interacting with the system during the question-answer session. We also plan to develop/use ontology based reasoning and techniques in order to automatically create relevant questions and answers based on information, concepts and relationships stored in the ontology. Nevertheless, the proposed Ontology and Question Based Design Patterns Advisement Approach might contribute to our common goal: to bridge the identified gap between pattern experts and the typical pattern user from the software community. For that purpose we also plan to join initiatives such as [23] aimed at resolving this challenge by networking, sharing ideas and joining resources.

6. Conclusion

The potential of using patterns has not yet been fully realized. Many challenges and issues still remain to be solved. Finding a suitable design pattern for a given situation obviously represents a great challenge for a typical developer. Tools assisting in this process have become of the utmost importance [11]. There are some facilities and approaches based on pattern review, browsing and full-text search. They might be helpful for the pattern expert community, but not for less-experienced developers, who are unaware as to which patterns exist in their work domains. The novel ontology and question-based approach presented in this paper was aimed at improving the use of design patterns. The proposed OQBA approach and corresponding DPEX tool assist software developers in choosing design patterns suitable for a given problem. The main contributions of the conducted research were:

- Definition of the Question and Ontology-Based Design Pattern Advisement approach.
- Extension of existing ODOL ontology with concepts needed to capture additional knowledge on patterns.
- Definition of the DPAO ontology that is used to provide knowledge for applying the OQBA approach.

Using the defined ontological concepts, information and knowledge on design patterns could be shared and automatically analyzed. We also provided the semantic based description of GoF patterns not available in the WoP project. The DPEX tool was developed and a controlled experiment was conducted. The main aim of the experiment was to explore if the DPEX tool-enabled and OQBA approach for selecting an appropriate pattern would do better than a manual selection. The results of the controlled experiment show that the developers were essentially able to solve more problem cases when the tool was available. The proposed approach also provided promising results with regard to the use of design patterns in the community of less-experienced developers.

Introducing the concepts and technologies of the semantic web into the field of design-pattern research creates new possibilities for making design patterns more approachable to software engineers. One could envisage a global knowledge database on design patterns, which would be specified using the ODOL ontology and other derived ontologies, such as DPAO ontology. This knowledge base could then be accessible through some form of a service-oriented architecture and available to various software development tools. This would allow the integration of facilities of a global knowledge base into Web-based knowledge platform on design patterns as defined in [4]. At the time of publishing the results we are extending and improving both approach and tools. Because of the platforms ability to insert new knowledge easily we are also introducing it to new fields, which include selecting design patterns in other domains (e.g. Service Oriented Architecture) or as a helper in expert systems for selecting resources (e.g. e-services).

References

1. L. Rising, Understanding the Power of Abstraction in Patterns, IEEE Software, 24 (4) (2007) 46-51.
2. Schmidt, D.C., Using Design Patterns to Develop Reusable Object-Oriented Communication Software, Communications of the ACM, 38 (10) (1995) 65-74.
3. Microsoft Patterns & Practices Group, http://msdn.microsoft.com/practices

4. D.Manolescu, W. Kozaczynski, A. Miller, J. Hogg, The Growing Divide in the Patterns World, IEEE Software, 24 (4) (2007), 61-67.
5. E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley, 1998.
6. L. Rising, The Pattern Almanac 2000, Addison Wesley, 2000.
7. F. J. Budinsky, M. A. Finnie, J. M. Vlissides, P. S. Yu, Automatic Code Generation from Design Patterns, IBM Systems Journal, 36 (2)(1996) 151-171.
8. J. Dietrich, C. Elgar, The Web of Patterns Project, http://www-ist.massey.ac.nz/wop
9. A. H. Eden, A. Yehudai, J. Gil, Precise Specification and Automatic Application of Design Patterns, 12th IEEE International Conference on Automated Software Engineering, IEEE Press, 1997, Proceedings, pp. 143-152.
10. P. Gomes, F.C. Pereira, P. Paiva, N. Seco, P. Carreiro, J. Ferreira, C. Bento, Selection and Reuse of Software Design Patterns Using CBR and WordNet, 15th International Conference on Software Engineering and Knowledge Engineering, SEKE 2003, Proceedings, (SEKE'03), pp. 289-296.
11. A. Birukuo, E. Blanzieri, P. Giorgini, Choosing the Right Design Pattern: The Implicit Culture Approach, Arturo Hinojosa, A Cognitive Model of Design Pattern Selection. Technical Report DIT-06-007, Informatica e Telecomunicazioni, University of Trento, 2006.
12. ODOL- Object-Oriented Design Ontology, http://svn.sourceforge.net/viewvc/webofpatterns/wop-patterndefinitions/20060324/ wop.owl
13. J. M. Rosengard, M. F. Ursu, Ontological Representations of Software Patterns, Lecture Notes in Computer Science (Proceedings of the of KES'04), 3215 (2004), pp. 31-37.
14. M. Fontoura, C. Lucena , Extending UML to Improve the Representation of Design Patterns, Journal of Object-Oriented Programming, 13(11) (2001) 12-19.
15. G. Sunyé, A.L. Guennec, J-M Jézéquel, Design Patterns Application in UML, Lecture Notes in Computer Science (ECOOP'2000 proceedings), 1850 (2000), pp. 44-62.
16. R. B. France, D. K. Kim, S. Ghosh and E. Song. A UML-Based Pattern Specification Technique, IEEE Transactions on Software Engineering, 30(3)(2004), 193-206.
17. T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American, 284 (5) (2001) 28-37.
18. W3C Semantic Web Activity, http://www.w3.org/2001/sw/
19. M. Safiz, P. Adamczyk, R.E. Johnson, Organizing Security Patterns, IEEE Software, 24 (4) 52-60.
20. M. Fowler, Patterns of Enterprise Application Architecture, Addison-Wesley, 2003.
21. C. Alexander et al, A Pattern Language, Oxford University Press, New York, 1977.
22. Kung, D. C., H. Bhambhani, R. Shah, and G. Pancholi. 2003. An expert system for suggesting design patterns: a methodology and a prototype. In Software Engineering With Computational Intelligence, ed. T. M. Khoshgoftaar. Kluwer Int. Lazovik, A., M. Aiello, and M. Papazoglou. 2006.
23. M. Weiss, A. Birukou, P. Giorgini, EuroPLoP 2007 Focus Group on Pattern Repositories, http://www.patternforge.net/wiki/index.php?title=EuroPLoP 2007 Focus Group.

Luka Pavlič received his Ph.D. degree in computer science in 2009 from the University of Maribor, Slovenia. He is currently a senior researcher with the Institute of Informatics, FERI, at the University of Maribor. His main research interests include all aspects of IS development, reuse in software engineering, object orientation, information system architecture, Java, UML and XML-related technologies, semantic technologies, intelligent systems, big data and nosql. He has appeared as an author and co-author in several peer-reviewed scientific journals. He has also presented his work at a number of international conferences. In addition, he has participated in many national and international research projects.

Vili Podgorelec is a professor of computer science at the University of Maribor, Slovenia. His main research interests include intelligent systems, semantic technologies, and medical informatics. He has been participating in many international research projects and is author of several journal papers on computational intelligence, software engineering and medical informatics. Dr. Podgorelec has worked as a visiting professor and researcher at several universities around the world, including University of Osaka, Japan, University of Nantes, France, University of La Laguna, Spain, University of Madeira, Portugal, and University of Applied Sciences Seinäjoko, Finland. He received several international awards and grants for his research activities.

Marjan Heričko is a full professor at the Institute of Informatics. He is the head of the Information systems laboratory and Deputy Head of the Institute of informatics. He received his PhD in Computer Science from University of Maribor in 1998. His main research interests include all aspects of information systems development, software and service engineering, agile methods, process frameworks, software metrics, functional size measurement, SOA, component-based development, object-orientation, software reuse and software patterns. Dr. Heričko has been a project or work co-ordinator in several applied projects, project or work co-ordinator in several international research projects and committee member and chair of several international conferences.

Received: August 24, 2013; Accepted: January 25, 2014

Computer Science and Information Systems 11(2):665–678 DOI: 10.2298/CSIS140212036O

Efficient Data Abstraction Using Weighted IB2 Prototypes

Stefanos Ougiaroglou and Georgios Evangelidis

Department of Applied Informatics, School of Information Sciences, University of Macedonia
156 Egnatia str., 54006, Thessaloniki, Greece
{stoug,gevan}@uom.gr

Abstract. Data reduction techniques improve the efficiency of k-Nearest Neighbour classification on large datasets since they accelerate the classification process and reduce storage requirements for the training data. IB2 is an effective prototype selection data reduction technique. It selects some items from the initial training dataset and uses them as representatives (prototypes). Contrary to many other techniques, IB2 is a very fast, one-pass method that builds its reduced (condensing) set in an incremental manner. New training data can update the condensing set without the need of the "old" removed items. This paper proposes a variation of IB2, that generates new prototypes instead of selecting them. The variation is called AIB2 and attempts to improve the efficiency of IB2 by positioning the prototypes in the center of the data areas they represent. The empirical experimental study conducted in the present work as well as the Wilcoxon signed ranks test show that AIB2 performs better than IB2.

Keywords: k-NN classification, data reduction, abstraction, prototypes.

1. Introduction

The k-Nearest Neighbour (k-NN) algorithm is an effective lazy classifier [12]. For each new, unclassified item x, it searches in the available training database and retrieves the k nearest items (neighbours) to x according to a distance metric (e.g., Euclidean distance). Then, x is assigned to the class where most of the retrieved k nearest neighbours belong to. The k-NN classifier has some properties that make it a popular classifier: (i) it is considered to be an accurate classifier, (ii) it has many applications, (iii) it is very easy to implement, (iv) it is analytically tractable, and, (v) for $k = 1$ and unlimited items the error rate is asymptotically never worse than twice the minimum possible, which is the Bayes rate [11].

On the other hand, the k-NN classifier has two main drawbacks. The first one is the high computational cost involved. Since all distances between an unclassified item and the available training data must be estimated, the computational cost is high, and thus, classification is a time consuming process. In cases of very large datasets, using the k-NN classifier may be prohibitive. The second drawback is the high storage requirements for the training set. Contrary to the eager classifiers that can discard the training data after the construction of the classification model, the k-NN classifier needs all training data available for accessing when a new item has to be classified. These two weaknesses render the k-NN classifier inefficient for large datasets.

Multi-attribute indexes [30] can efficiently accelerate the nearest neighbours search and speed-up the k-NN classifier. However, storage requirements increase, since in addition to the training data, the index must be stored, too. Moreover, indexes can be applied only on datasets with moderate dimensionality (e.g., 2-10). In higher dimensions, the phenomenon of the dimensionality curse makes even sequential scans more effective than indexes [34], thus, it is essential to first apply a dimensionality reduction technique.

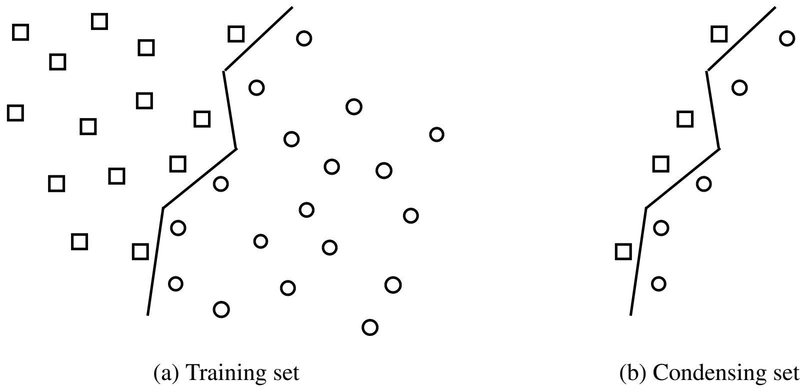
Many Data Reduction Techniques (DRTs) have been proposed to address both weaknesses. DRTs are distinguished into prototype selection [18] and prototype abstraction [32] algorithms. Prototype selection algorithms select items from the training set, whereas, prototype abstraction algorithms generate prototypes by summarizing similar training items. Prototype selection algorithms are also divided into two categories. They can be either condensing or editing algorithms. Prototype abstraction and condensing algorithms have the same motivation. They aim to build a small representative set, usually called condensing set, of the initial training data. Usage of a condensing set has the benefits of low computational cost and storage requirements and at the same time keeps classification accuracy at high levels. On the other hand, editing algorithms aim to improve accuracy rather than achieve high reduction rates. To do this, they improve the quality of the training set by removing outliers, noisy and mislabelled items as well as by smoothing the decision boundaries between classes.

Although editing has a completely different goal than prototype abstraction and condensing, it can be used to improve their efficiency by increasing their reduction rates and/or accuracy levels. More specifically, the reduction rates of many prototype abstraction and condensing algorithms highly depend on the level of noise in the training set. Actually, the higher the level of noise in the training set, the lower the reduction rates achieved. Consequently, effective data reduction implies removal of noise from the data, i.e., application of an editing algorithm beforehand [13, 24]. Note that some condensing algorithms integrate the idea of editing. These algorithms are called Hybrid [18].

A great number of DRTs have been proposed and are available in the literature. Also, many the state-of-the-art surveys have been published since 2000. Prototype abstraction and prototype selection algorithms are recently reviewed in [18, 32]. Moreover, the particular papers present interesting taxonomies and experimental results. Other well-known surveys can be found in [9, 21, 23, 25].

IB2 belongs to the popular family of Instance-Based Learning (IBL) algorithms [3, 4]. IB2 is an one-pass incremental condensing algorithm. Actually, it is based on the earliest and well known condensing algorithm, CNN-rule [22]. Contrary to CNN-rule and many other state-of-the-art DRTs, IB2 can build its condensing set in an incremental manner[1]. It can take into consideration new training items after the construction of the condensing set and without needing the old removed items. Hence, it can be used in dynamic/streaming environments [1] where new training data is gradually available. In addition, the incremental nature of IB2 allows its execution on training sets that cannot fit into main memory.

[1] In the literature [18, 32], DRTs can be incremental or decremental depending on the way they build their condensing set. An incremental technique begins with an empty condensing set and adds items to it, whereas a decremental technique uses all training data as the initial condensing set and then removes items. In this paper, the use of term incremental refers to the ability of the DRT to update an already constructed condensing set.


(a) Training set

(b) Condensing set

Fig. 1. Initial training data and close-class-border data

In this paper, we attempt to improve the performance of IB2 by considering the idea of prototype abstraction. Our contribution is the development and evaluation of a prototype abstraction version of IB2 that we call Abstraction IB2 (AIB2). The proposed prototype abstraction algorithm retains all the properties of IB2, but it is faster and achieves higher reduction rates and improved accuracy.

CNN-rule and IB2 are presented in Section 2. Section 3 describes in detail the proposed AIB2 algorithm. Section 4 presents an experimental study where k-NN is applied on the original training sets and the corresponding condensed sets built by CNN, IB2 and AIB2 on nine datasets. The experimental study is complemented by the statistical comparison of the three DRTs through the Wilcoxon signed ranks test [14]. Section 5 concludes the paper.

2. CNN-rule and IB2

The Condensing Nearest Neighbour (CNN) rule [22] is the earliest and also a reference condensing algorithm. CNN-rule (and many other DRTs) builds its condensing set based on the following simple idea. Items that lie in the "internal" data area of a class (i.e., far from class decision boundaries) are useless during the classification process. Thus, they can be removed without loss of accuracy. By adopting this idea, CNN-rule tries to place into the condensing set only the items that lie in the close-class-border data areas. These are the only essential items for the classification process. Figure 1 depicts this strategy. The idea is that the k-NN classifier will be able to have similar accuracy using either the training set (Figure 1(a)) or the condensing set (Figure 1(b)). However, a scan of the condensing set involves much lower computational cost than that of the training set.

CNN-rule tries to keep the close-class-border items as follows (see Algorithm 1). Initially, a random item from the training set (TS) is moved to the condensing set (CS) (line 2). Then, CNN-rule applies the single nearest neighbour rule and classifies the items of TS by scanning the items of CS (line 6). If an item is misclassified, it is moved from TS to CS (lines 7–11). The algorithm continues until there are no moves from TS to CS during a complete scan of TS (lines 3,4,10,13). This ensures that the content of TS is

Algorithm 1 CNN-rule

```
Input: $TS$
Output: $CS$
 1: $CS \leftarrow \varnothing$
 2: pick a random item of $TS$ and move it to $CS$
 3: repeat
 4:     $stop \leftarrow TRUE$
 5:     for each $x \in TS$ do
 6:         $NN \leftarrow$ Nearest Neighbour of $x$ in $CS$
 7:         if $NN_{class} \neq x_{class}$ then
 8:             $CS \leftarrow CS \cup \{x\}$
 9:             $TS \leftarrow TS - \{x\}$
10:             $stop \leftarrow FALSE$
11:         end if
12:     end for
13: until $stop == TRUE$ {no move during a pass of $TS$}
14: discard $TS$
15: return $CS$
```

correctly classified by the content of CS. The remaining content of TS is discarded (line 14) to save space.

CNN-rule considers that misclassified items are probably close to decision boundaries and so they must be included in the condensing set. An advantage is that it is a non-parametric approach. It determines the number of the prototypes automatically, without user-defined parameters. A drawback is that the resulting condensing set depends on the order in which class labels of items enter the condensing set. Therefore, CNN-rule builds a different condensing set by examining the same training data in a different order.

Furthermore, CNN-rule cannot handle new training data, i.e., is non-incremental. The algorithm involves multiple passes over the data and it cannot use training data available at a later time to update its condensing set. To construct an updated condensing set, CNN-rule needs the complete training set (new and old data), thus, the training items that do not enter the original condensing set should be retained. In addition, CNN-rule requires that all training items are memory resident.

Although IB2 is based on CNN-rule, it is faster and incremental. Therefore, it can handle new training data or datasets that cannot fit into memory. Actually, IB2 is an one-pass version of CNN-rule (see Algorithm 2) and works as follows. When a new item x of the training set (TS) arrives (line 3), it is classified by the single nearest neighbour rule by scanning the contents of the current condensing set (CS) (line 4). If x is wrongly classified, it is placed in CS (lines 5-7). Then, x is discarded since it has been examined (line 8).

Similarly to CNN-rule, IB2 is non-parametric and its condensing set depends on the order of training items. Contrary to CNN-rule, there is no guarantee that the condensing set built by IB2 can correctly classify all examined training data. IB2 is very fast because it is a one pass algorithm. Training data segments can update an existing condensing set in a simple manner and without considering the “old” (removed) data that had been used for the construction of the condensing set. Each new training item can be examined, be

Algorithm 2 IB2

```
Input: $TS$
Output: $CS$
1: $CS \leftarrow \varnothing$
2: pick a random item of $TS$ and move it to $CS$
3: for each $x \in TS$ do
4:     $NN \leftarrow$ Nearest Neighbour of $x$ in $CS$
5:     if $NN_{class} \neq x_{class}$ then
6:         $CS \leftarrow CS \cup \{x\}$
7:     end if
8:     $TS \leftarrow TS - \{x\}$
9: end for
10: return $CS$
```

placed or not in the condensing set, and then, removed. Additionally, IB2 can handle new class labels. All these properties render IB2 an appropriate condensing algorithm for dynamic/streaming environments where new training data may be gradually available or for training databases that cannot fit into the device's memory.

Although many DRTs have been proposed during the past decades, here, we present only CNN-rule and IB2 because they are ancestors of the proposed algorithm. There are many other condensing algorithms that either extend the CNN-rule or are based on the same idea, i.e., the removal of the non-close-border data. Some of the these algorithms are the Reduced Nearest Neighbour (RNN) rule [20], the Selective Nearest Neighbour (SNN) rule [29], the Modified CNN rule [15], the Generalized CNN rule [10], the Fast CNN algorithms [6, 7], Tomek's CNN rule [31], the Patterns with Ordered Projection (POP) algorithm [2, 28], the recently proposed Template Reduction for k-NN (TRkNN) [16], the Prototype Selection by Clustering (PSC) algorithm [26, 27] and of course the IB2 algorithm [3, 4].

3. The proposed AIB2 algorithm

The proposed Abstraction IB2 (AIB2) algorithm is a prototype abstraction version of IB2. Therefore, AIB2 is a non-parametric, fast, one-pass prototype abstraction algorithm. It is also appropriate for dynamic/streaming environments and can handle new class labels. Like IB2, it can be applied on very large datasets that cannot fit into main memory or on devices with limited main memory (e.g., sensor networks), without transferring data to a server over a network for processing. The latter is usually costly and time-consuming. Certainly, like IB2, AIB2 does not take into account the phenomenon of concept drift [33] that may exist in data streams. IBL-DS [8] is based on the idea of IBL family of algorithms that can effectively deal with this phenomenon.

The main difference between AIB2 and IB2 is the following. The items that are correctly classified by the 1-NN rule are not ignored. They contribute to the construction of the condensing set by updating their nearest prototype. The main idea behind AIB2 is that prototypes should be at the center of the data area they represent. To achieve this, AIB2 adopts the concept of prototype weight. Each prototype has a weight value as an extra

Algorithm 3 AIB2

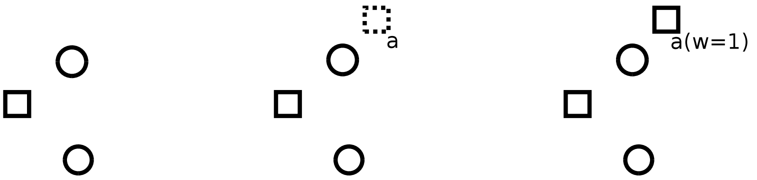
```
Input: $TS$
Output: $CS$
 1: $CS \leftarrow \varnothing$
 2: pick a random item $y$ of $TS$ and move it to $CS$
 3: $y_{weight} \leftarrow 1$
 4: for each $x \in TS$ do
 5:     $NN \leftarrow$ Nearest Neighbour of $x$ in $CS$
 6:     if $NN_{class} \neq x_{class}$ then
 7:         $x_{weight} \leftarrow 1$
 8:         $CS \leftarrow CS \cup \{x\}$
 9:     else
10:         for each attribute $attr(i)$ do
11:             $NN_{attr(i)} \leftarrow \frac{NN_{attr(i)} \times NN_{weight} + x_{attr(i)}}{NN_{weight}+1}$
12:         end for
13:         $NN_{weight} \leftarrow NN_{weight} + 1$
14:     end if
15:     $TS \leftarrow TS - \{x\}$
16: end for
17: return $CS$
```

attribute that denotes the number of training items it represents. The weight values are used for updating the prototype attributes in the multidimensional space.

AIB2 is presented in Algorithm 3. Initially, the condensing set (CS) is populated by a random item of the training set (TS) whose weight is initialized to 1 (lines 2–3). For each item x of TS, AIB2 searches in CS and retrieves its nearest prototype NN (line 5). If x is misclassified, it is placed in CS and its weight is initialized to one (lines 6–8). If x is correctly classified, NN's attributes are updated by taking into consideration its current weight and the attributes of x. Effectively, NN "moves" towards x in the multidimensional space (lines 10–12). Finally, the weight of NN is increased by 1 (line 13) and x is removed (line 15).

Figures 2 and 3 illustrate two-dimensional examples of AIB2 execution. Suppose that the current condensing set includes three prototypes, two of class circle and one of class square (Figure 2(a) and Figure 3(a)). Suppose that a new square item, a, arrives (Figure 2(b)). AIB2 should decide whether a will enter the condensing set or it will be used to update an existing prototype. Since a is closer to a prototype of a different class, a enters the condensing set and its weight value is initialized to one (Figure 2(c)). In contrast, suppose that the new item a is a circle item (Figure 3(b)). Since a is closer to P, which is a prototype of the same class, a updates P. Therefore P moves towards a and its weight is increased by one (Figure 3(c)). In our example, the updated P lies in between the original P and a because the weight of P was 1).

By updating the prototypes, AIB2 ensures that each prototype lies near the center of the data area it represents. Notice that the weight of a prototype practically denotes the population of the original items that it represents. Thus, although each time a new training item arrives and is assigned to an existing prototype, it causes the prototype to "move" towards the new item, the larger the weight of the prototype is, the smaller the "move"

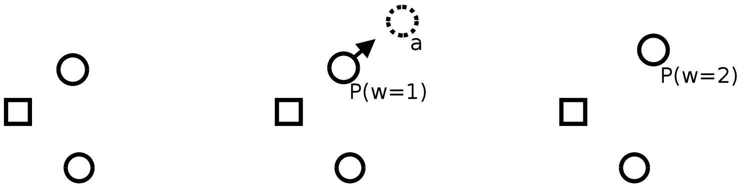


(a) Current condensing set (3 prototypes)

(b) New item arrival

(c) The new item enters the condensing set

Fig. 2. AIB2 example: new prototype enters the condensing set



(a) Current condensing set (3 prototypes)

(b) New item arrival

(c) Updating of nearest prototype

Fig. 3. AIB2 example: repositioning an existing prototype

is towards the new training item. Like IB2, AIB2 prototypes depend on the order of the training items. It is possible that over time and because of a non-uniform arrival of training data items, an AIB2 prototype will gradually move far away from some of the original training items that it represents. Still, AIB2 prototypes with large weights will move very slowly towards the new training data items. This is also a problem with IB2 that contrary to CNN-rule cannot guarantee correct classification of all examined training data.

The idea is that a condensing set built by AIB2 will contain better prototypes compared to the condensing set built by IB2 and will achieve higher classification accuracy. In addition, we expect that updating the prototypes will reduce the number of items that enter the condensing set (lines 6–8), and thus, AIB2 will achieve higher reduction rates and lower preprocessing cost compared to IB2.

4. Performance Evaluation

4.1. Experimental setup

The three presented DRTs were evaluated using eight datasets distributed by the KEEL repository[2] [5] and summarized in Table 1. Seven datasets (all except KDD) were used

[2] http://sci2s.ugr.es/keel/datasets.php

without applying data normalization. Datasets MGT, SI and TXR are distributed sorted on the class label. This affects all examined DRTs because their performance depends on the order of data in the training set. Therefore, we randomized the items of these datasets. The three algorithm implementations were written in C and Euclidean distance was the distance metric used.

Table 1. Datasets description

Dataset	Size	Attributes	Classes
Letter Recognition (LR)	20000	16	26
Magic Gamma Telescope (MGT)	19020	10	2
Pen-Digits (PD)	10992	16	10
Sat Image (SI)	6435	36	6
Shuttle (SH)	58000	9	7
Texture (TXR)	5500	40	11
Phoneme (PH)	5404	5	2
KddCup (KDD)	141481	23	36

The KDD dataset includes 494,020 items and 41 attributes. However, huge amounts of data are duplicates and some attributes are unnecessary (three nominal and two fixed-value attributes). We removed all duplicates and the five unnecessary attributes. Consequently, the transformed KDD dataset includes 141,481 items and 36 attributes. It is worth mentioning that duplicates are useless especially for k-NN classification with $k = 1$. They do not influence the accuracy and negatively affect the computational cost. Although, for $k > 1$, duplicates may influence the accuracy, they should be removed, especially when fast execution of classification is required. In addition, the value ranges of the KDD dataset attributes vary extremely. Therefore, we normalized all attributes to the interval $[0, 1]$ as follows. Assume that the dataset includes n items and an attribute e should be normalized to $[0, 1]$. The normalized attribute value of the i-th item, $i = 1, \ldots, n$ is:

$$norm(e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}}$$

where E_{min} and E_{max} are the minimum and maximum values for attribute e, respectively. Finally, we randomized the dataset.

We did not include more DRTs in our experimentation because our purpose was to focus on incremental and non-parametric DRTs. Of course, CNN-rule is non-incremental, but it is considered a reference DRT and is being used in many papers for comparison purposes. In addition, CNN-rule is the ancestor of IB2 and AIB2 and it makes sense to use it in our experiments.

DRTs were evaluated by estimating three measurements: (i) classification accuracy, (ii) reduction rate, and, (iii) preprocessing cost in terms of distance computations. Accuracy measurements were estimated by executing the k-NN classifier with $k = 1$ over the condensing set built by each DRT. For each dataset and DRT, we report the average values of these measurements obtained via five-fold cross-validation. With the exception of the KDD dataset, we used the five already constructed pairs of training and testing sets distributed by the KEEL repository. For the transformed KDD dataset, we created the appropriate for cross-validation training/testing pairs. Of course, all measurements were estimated after the arrival of all training items.

The original form of the MGT dataset contains high levels of noise. Noisy items are misleading for the three DRTs and negatively affect reduction rates and the corresponding accuracies. Thus, we tried to build a noise-free version of the MGT dataset. To achieve this, we ran an editing algorithm before the execution of the DRTs. In particular, we coded in C the Edited Nearest Neighbour (ENN) rule [36], which is the reference editing algorithm. ENN-rule removes noisy items by using the following rule: a training item is removed if its class label does not agree with the majority of its k nearest neighbours in the initial training set. Consequently, ENN-rule needs to compute all distances among the training items, i.e., $\frac{N*(N-1)}{2}$ distances, where N is the number of the training items. By applying ENN-rule (with $k = 3$, see [19, 25, 35]) on each training portion of each fold, we built an extra dataset that we call MGT-ENN. Of course, the testing portions were not edited by the ENN-rule. We tested the three DRTs on both forms of the MGT dataset.

4.2. Comparisons

The results of our experimental study are presented in Table 2. Each column lists the results related to a classifier. Best measurements are in bold. The preprocessing cost measurements are in million distance computations. For reference, Table 2 presents the accuracy measurements obtained by the conventional 1-NN classifier on the original training set under the label Conv-1-NN.

We should clarify that the reduction rates of MGT-ENN do not take into account the items removed by editing. They concern the size of the condensing set in relation to that of the edited set. Note that ENN-rule considered as noise 20.08% of MGT items on average. Therefore, the training portions of MGT-ENN include 20.08% fewer items than MGT on average.

Although the three DRTs did not reach the accuracy levels of Conv-1-NN, they were very close to them. With the exception of LR, SI and TXR, CNN-rule was more accurate than IB2 and AIB2. However, it required much higher preprocessing cost and it achieved lower reduction rates than IB2 and AIB2.

As we anticipated, the proposed algorithm performed better than IB2 in most datasets and in terms of all comparison criteria. KDD is the only dataset where IB2 seems to be better than AIB2. In the LR, SI and TXR datasets, AIB2 was more accurate than CNN-rule. Although we expected even better performance, we believe that the improvements achieved by AIB2 are noteworthy.

Moreover, as we expected, ENN-rule improved the efficiency of the classification process on MGT data. Accuracies and reduction rates achieved by the three DRTs on MGT-ENN (edited data) are higher than the corresponding measurements for MGT (non edited data).

4.3. Non-parametric statistical test

In this subsection, we provide the results of a Wilcoxon signed ranks test [14]. We ran the test four times: once on the nine measurements of each comparison criterion (classification accuracy (ACC), reduction rate (RR), preprocessing cost (PC)), and once on the nine

Table 2. DRTs Comparison in terms of Accuracy (Acc(%)), Reduction Rate (RR(%)) and Preprocessing Cost (PC(M))

Dataset		Conv-1-NN	CNN-rule	IB2	AIB2
LR	Acc:	95.83	92.84	91.98	**94.12**
	RR:	-	83,54	85.66	**88.12**
	PC:	-	163.03	23.37	**20.10**
MGT	Acc:	78.14	**74.54**	71.97	73.36
	RR:	-	60.08	70.60	**71.90**
	PC:	-	281.49	34.61	**33.05**
MGT-ENN	Acc:	80.44	**79.26**	78.01	78.81
	RR:	-	87.62	90.07	**91.06**
	PC:	-	68.61	8.48	**7.65**
PD	Acc:	99.35	**98.68**	98.04	98.33
	RR:	-	95.36	96.23	**97.19**
	PC:	-	11.75	1.78	**1.38**
SI	Acc:	90.60	88.21	86.87	**89.42**
	RR:	-	80.22	84.62	**86.72**
	PC:	-	17.99	2.22	**1.92**
SH	Acc:	99.82	**99.76**	99.73	99.72
	RR:	-	99.37	99.44	**99.46**
	PC:	-	45.30	8.26	**7.89**
TXR	Acc:	99.02	97.16	96.35	**97.69**
	RR:	-	91.90	93.33	**94.95**
	PC:	-	5.65	0.84	**0.66**
PH	Acc:	90.10	**87.82**	85.57	84.92
	RR:	-	76.04	80.85	**81.75**
	PC:	-	13.45	1.96	**1.84**
KDD	Acc:	99.71	**99.66**	99.48	99.41
	RR:	-	99.12	**99.26**	99.21
	PC:	-	384.90	**55.58**	58.78

measurements of an extra criterion that estimates the overall classification performance of an algorithm. Our criterion adopts the idea presented in the statistical comparissons in [17, 18]. More specifically, the overall classification performance measurements were computed by averaging the normalized (to the range $[0, 1]$) measurements of the three criteria. Note that since the higher the preprocessing cost, the lower the classification performance, we used $1 - normalized(PC)$ as the preprocessing cost. The overall classification performance criterion combines the three criteria by considering all of them as having the same significance.

Table 3 illustrates the results of the Wilcoxon test. The columns with label "W/L" count the number of wins and loses, respectively. The columns with label "Wilcoxon" list the Wilcoxon significance level. When that value is lower than 0.05, one can safely claim that the difference between the compared algorithms is statistically significant. We observe that this is true in most cases. In terms of accuracy, CNN-rule has more wins than AIB2 and AIB2 has more wins than IB2. However, there is no statistical difference between the corresponding algorithms. In addition, AIB2 is statistically better than both CNN-rule and IB2 in terms of reduction rate, and AIB2 is statistically better than CNN-rule in terms of preprocessing cost. On the other hand, although it is clear that AIB2 is faster than IB2, this is not statistically supported (note that we have adopted the strict threshold $Wilc. = 0.05$). Finally, we observe that AIB2 is statistically better than the other algorithms in terms of the overall classification performance.

Concerning CNN-rule and IB2 (last table row), we observe that IB2 is statistically better than CNN-rule in terms of reduction rate and preprocessing cost. In contrast, CNN-rule is statistically better than IB2 in terms of accuracy.

Table 3. Results of Wilcoxon signed ranks test

Methods	**ACC**		**RR**		**PC**		**Overall performance**	
	W/L	**Wilcoxon**	**W/L**	**Wilcoxon**	**W/L**	**Wilcoxon**	**W/L**	**Wilcoxon**
AIB2 vs CNN	3/6	0.767	9/0	**0.008**	9/0	**0.008**	9/0	**0.008**
AIB2 vs IB2	6/3	0.066	8/1	**0.015**	8/1	0.086	7/2	**0.028**
IB2 vs CNN	0/9	**0.008**	9/0	**0.008**	9/0	**0.008**	9/0	**0.008**

5. Conclusions

Data reduction is an important research issue in the context of k-NN classification. This paper proposed the AIB2 algorithm, an abstraction DRT that is based on the well-known condensing IB2 algorithm. The main concept behind AIB2 is that each generated prototype should be near to the center of the data area it represents. Like IB2 and contrary to CNN-rule and many other DRTs, AIB2 is an incremental DRT. This property renders AIB2 appropriate for dynamic domains where new training data is gradually available and for datasets that cannot fit into main memory.

The experimental results illustrate that AIB2 can achieve higher reduction rates and classification accuracy and lower preprocessing cost than IB2. In addition, AIB2 is preferable to CNN-rule when preprocessing cost and/or reduction rates are more critical criteria than accuracy and, of course, when an incremental DRT is required. The improvement offered by AIB2 was statistically confirmed by the Wilcoxon signed ranks test.

Acknowledgments. Stefanos Ougiaroglou is supported by a scholarship from the State Scholarships Foundation (I.K.Y.) of Greece.

References

1. Aggarwal, C.: Data Streams: Models and Algorithms. Advances in Database Systems Series, Springer Science+Business Media, LLC (2007)
2. Aguilar, J.S., Riquelme, J.C., Toro, M.: Data set editing by ordered projection. Intell. Data Anal. 5(5), 405–417 (Oct 2001), http://dl.acm.org/citation.cfm?id=1294007.1294010
3. Aha, D.W.: Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. Int. J. Man-Mach. Stud. 36(2), 267–287 (Feb 1992), http://dx.doi.org/10.1016/0020-7373(92)90018-G

4. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. 6(1), 37–66 (Jan 1991), http://dx.doi.org/10.1023/A:1022689900470
5. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Multiple-Valued Logic and Soft Computing 17(2-3), 255–287 (2011)
6. Angiulli, F.: Fast condensed nearest neighbor rule. In: Proceedings of the 22nd international conference on Machine learning. pp. 25–32. ICML '05, ACM, New York, NY, USA (2005)
7. Angiulli, F.: Fast nearest neighbor condensation for large data sets classification. IEEE Trans. on Knowl. and Data Eng. 19(11), 1450–1464 (Nov 2007), http://dx.doi.org/10.1109/TKDE.2007.190645
8. Beringer, J., Hüllermeier, E.: Efficient instance-based learning on data streams. Intell. Data Anal. 11(6), 627–650 (Dec 2007), http://dl.acm.org/citation.cfm?id=1368018.1368022
9. Brighton, H., Mellish, C.: Advances in instance selection for instance-based learning algorithms. Data Min. Knowl. Discov. 6(2), 153–172 (Apr 2002), http://dx.doi.org/10.1023/A:1014043630878
10. Chou, C.H., Kuo, B.H., Chang, F.: The generalized condensed nearest neighbor rule as a data reduction method. In: Proceedings of the 18th International Conference on Pattern Recognition - Volume 02. pp. 556–559. ICPR '06, IEEE Computer Society, Washington, DC, USA (2006), http://dx.doi.org/10.1109/ICPR.2006.1119
11. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13(1), 21–27 (Sep 2006), http://dx.doi.org/10.1109/TIT.1967.1053964
12. Dasarathy, B.V.: Nearest neighbor (NN) norms : NN pattern classification techniques. IEEE Computer Society Press (1991)
13. Dasarathy, B.V., Snchez, J.S., Townsend, S.: Nearest neighbour editing and condensing toolssynergy exploitation. Pattern Analysis & Applications 3(1), 19–30 (2000), http://dx.doi.org/10.1007/s100440050003
14. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (Dec 2006), http://dl.acm.org/citation.cfm?id=1248547.1248548
15. Devi, V.S., Murty, M.N.: An incremental prototype set building technique. Pattern Recognition 35(2), 505–513 (2002)
16. Fayed, H.A., Atiya, A.F.: A novel template reduction approach for the k-nearest neighbor method. Trans. Neur. Netw. 20(5), 890–896 (May 2009), http://dx.doi.org/10.1109/TNN.2009.2018547
17. García, S., Cano, J.R., Herrera, F.: A memetic algorithm for evolutionary prototype selection: A scaling up approach. Pattern Recogn. 41(8), 2693–2709 (Aug 2008), http://dx.doi.org/10.1016/j.patcog.2008.02.006
18. Garcia, S., Derrac, J., Cano, J., Herrera, F.: Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Trans. Pattern Anal. Mach. Intell. 34(3), 417–435 (Mar 2012), http://dx.doi.org/10.1109/TPAMI.2011.142
19. García-Borroto, M., Villuendas-Rey, Y., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: Using maximum similarity graphs to edit nearest neighbor classifiers. In: Proceedings of the 14th Iberoamerican Conference on Pattern Recognition: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. pp. 489–496. CIARP '09, Springer-Verlag, Berlin, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-10268-4_57
20. Gates, G.W.: The reduced nearest neighbor rule. ieee transactions on information theory. IEEE Transactions on Information Theory 18(3), 431–433 (1972)
21. Grochowski, M., Jankowski, N.: Comparison of instance selection algorithms ii. results and comments. In: Artificial Intelligence and Soft Computing - ICAISC 2004, LNCS, vol. 3070, pp. 580–585. Springer Berlin / Heidelberg (2004)

22. Hart, P.E.: The condensed nearest neighbor rule. IEEE Transactions on Information Theory 14(3), 515–516 (1968)
23. Jankowski, N., Grochowski, M.: Comparison of instances seletion algorithms i. algorithms survey. In: Artificial Intelligence and Soft Computing - ICAISC 2004, LNCS, vol. 3070, pp. 598–603. Springer Berlin / Heidelberg (2004)
24. Lozano, M.: Data Reduction Techniques in Classification processes (Phd Thesis). Universitat Jaume I (2007)
25. Olvera-López, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Kittler, J.: A review of instance selection methods. Artif. Intell. Rev. 34(2), 133–143 (Aug 2010), `http://dx.doi.org/10.1007/s10462-010-9165-y`
26. Olvera-Lopez, J.A., Carrasco-Ochoa, J.A., Trinidad, J.F.M.: A new fast prototype selection method based on clustering. Pattern Anal. Appl. 13(2), 131–141 (2010)
27. Ougiaroglou, S., Evangelidis, G.: Fast and accurate k-nearest neighbor classification using prototype selection by clustering. In: 16th Panhellenic Conference on Informatics (PCI), 2012. pp. 168–173 (2012)
28. Riquelme, J.C., Aguilar-Ruiz, J.S., Toro, M.: Finding representative patterns with ordered projections. Pattern Recognition 36(4), 1009 – 1018 (2003), `http://www.sciencedirect.com/science/article/pii/S003132030200119X`
29. Ritter, G., Woodruff, H., Lowry, S., Isenhour, T.: An algorithm for a selective nearest neighbor decision rule. IEEE Trans. on Inf. Theory 21(6), 665–669 (1975)
30. Samet, H.: Foundations of multidimensional and metric data structures. The Morgan Kaufmann series in computer graphics, Elsevier/Morgan Kaufmann (2006)
31. Tomek, I.: Two modifications of cnn. Systems, Man and Cybernetics, IEEE Transactions on SMC-6(11), 769–772 (Nov 1976)
32. Triguero, I., Derrac, J., Garcia, S., Herrera, F.: A taxonomy and experimental study on prototype generation for nearest neighbor classification. Trans. Sys. Man Cyber Part C 42(1), 86–100 (Jan 2012), `http://dx.doi.org/10.1109/TSMCC.2010.2103939`
33. Tsymbal, A.: The problem of concept drift: definitions and related work. Tech. Rep. TCD-CS-2004-15, The University of Dublin, Trinity College, Department of Computer Science, Dublin, Ireland (2004)
34. Weber, R., Schek, H.J., Blott, S.: A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: Proceedings of the 24rd International Conference on Very Large Data Bases. pp. 194–205. VLDB ’98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998), `http://dl.acm.org/citation.cfm?id=645924.671192`
35. Wilson, D.R., Martinez, T.R.: Reduction techniques for instance-basedlearning algorithms. Mach. Learn. 38(3), 257–286 (Mar 2000), `http://dx.doi.org/10.1023/A:1007626913721`
36. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE trans. on systems, man, and cybernetics 2(3), 408–421 (July 1972)

Stefanos Ougiaroglou holds a B.Sc. in Computer Science (2004) from Alexander TEI of Thessaloniki, Greece, and, a M.Sc. in Computer Science (2006) from Aristotle University of Thessaloniki, Greece. Currently, he is a computer science teacher in secondary education and a Ph.D. candidate of the department of Applied Informatics at the University of Macedonia, Greece. His Ph.D. studies are supported by a scholarship from the Greek State Scholarships Foundation (IKY), and focus on Algorithms and Techniques for Efficient k-Nearest Neighbours Classification. He has been employed as an adjunct instructor at ATEI of Messolonghi, Greece. His research interests include Data Mining, Mobile Computing, and Educational Technology.

Georgios Evangelidis is a Professor at the Department of Applied Informatics, University of Macedonia, Greece. He has a B.Sc. in Mathematics (Aristotle University of Thessaloniki, Greece, 1987) and a M.Sc. and Ph.D. in Computer Science (Northeastern University, Boston, MA, USA, 1990 and 1994, respectively). His interests include database technologies, data mining, citation analysis and software engineering. He has participated in various EU funded projects and has published over 100 scientific papers in refereed journals and conferences. He is a member of the DBTechNet initiative.

Received: February 12, 2014; Accepted: May 28, 2014.

Computer Science and Information Systems 11(2):679–696 DOI:10.2298/CSIS140216037R

Generic and Standard Database Constraint Meta-Models

Sonja Ristić[1], Slavica Aleksić[2], Milan Čeliković[2] and Ivan Luković[2]

[1]University of Novi Sad, Faculty of Technical Sciences,
Department for Industrial Engineering and Management
Trg Dositeja Obradovića 6
21000 Novi Sad, Serbia
sdristic@uns.ac.rs
[2]University of Novi Sad, Faculty of Technical Sciences,
Department of Computing and Control
Trg Dositeja Obradovića 6
21000 Novi Sad, Serbia
{slavica, ivan, milancel}@uns.ac.rs

Abstract. Many software engineering activities entail dealing with legacy information systems. When these systems become too costly to maintain, or when new technologies need to be incorporated, they need to be replaced or somehow reengineered. This can be done with significantly reduced amount of effort and cost if the conceptual models of these systems are available. Reverse engineering is the process of analyzing a subject system to create representations of the system at a higher level of abstraction. Relational databases are a common source of reverse engineering. Starting from a physical database schema, that is recorded into relational database schema data repository, the conceptual database schema or logical database schema could be extracted. The extraction process may be seen as a chain of model-to-model transformations that trace model elements from a model at the lower level of abstraction to a model at the higher level of abstraction, achieved through meta-modeling. In the paper we present generic and standard database constraint meta-models, focusing on multi-relational database constraints captured in a legacy database. These meta-models are aimed at support of model transformations to create conceptual models, as a useful source for the system reengineering process.

Keywords: Model-driven Software Engineering, Meta-modeling, Inclusion Dependency, Database Reengineering.

1. Introduction

The evolution in organization procedures and objectives over the time significantly reduces the effectiveness of an information system implemented to fulfill organizational information requirements. Coupled with the technological development it becomes the major cause for a legacy information system replacement or any form of its reengineering. A new system can be redeveloped from scratch, but in that case the knowledge captured in the legacy system is lost. Legacy system replacement or reengineering can be done with significantly reduced amount of effort and cost if the conceptual models are reconstructed from them. Reverse engineering is the process of analyzing a subject system to create models of the system at a higher level of abstraction. It encompasses a broad set of methods and tools related to understanding and modifying information systems. Relational databases are at the core of most company information systems, hosting critical information for the day to day operation of the company. The knowledge captured in them can serve as an important resource in a legacy information system modernization project and they are a common source of reverse engineering processes. Starting from a physical database schema, that is recorded into the relational database schema data repository, the conceptual database schema or logical database schema may be extracted. All of these database schemas represent models at different levels of abstraction. An extraction process may be seen as a chain of model-to-model (M2M) transformations that trace model elements from a model at the lower level of abstraction to a model at the higher level of abstraction.

Models are widely used in engineering disciplines. In the model-driven approach to software engineering (MDSE) the idea of abstracting implementation details by focusing on models as first class entities is promoted in [30]. Models are used to specify, simulate, test, verify and generate code for the application to be built [9]. Each model is expressed by the concepts of a modeling language that is specified by means of a meta-model. A meta-model defines a set of valid models [5]. An M2M transformation is based on meta-models that are conformed by the source and target models of the transformation. These meta-models are said to be in support of M2M transformation.

In a forward engineering process, designers start with a high-level model, abstracting from all kinds of platform issues. Through a chain of M2M transformations, ending up with a model-to-text (M2T) transformation, the initial platform independent model transforms iteratively to a series of models with less degree of platform independency, introducing more and more platform specific extensions. Conversely, in a reverse engineering process, the abstraction level of models and degree of platform independency are increasing throughout the chain of transformations.

Here we present a part of our research efforts focused on meta-models relating to databases that we call database meta-models. These meta-models are in support of database M2M transformations. In [28] we proposed the classification of database meta-models as follows: i) data model (dm) meta-models; ii) generic database schema meta-models; iii) standard physical database schema meta-models; and iv) vendor-specific physical database schema meta-models.

In [29] we have proposed a meta-model of relational database schema concerning inclusion dependency (IND) constraints. Both meta-model of the relational database schema and meta-model of the Universal Relational Schema (URS) are presented there. In the context of forward engineering, these meta-models enable the platform independent specification of a broad class of INDs and the development of M2M and M2T transformations. The meta-model of INDs is important in the context of database and information system reverse engineering, too. A discovery of inclusion dependencies has attracted a lot of research interests together with methods for discovery of INDs. A formal specification of discovered INDs by means of proposed IND meta-model provides a better support of the automated reengineering and improvement of legacy databases.

Here we shift focus on M2M transformation of a physical database schema to a logical relational database schema. Therefore, we present one generic and one standard physical database schema meta-model. Generic database schema meta-models are based on theoretical foundations of a data model as it is, for example, relational data model. The relational data model is the focus of a continuous standardization process, and therefore we have extracted the standard physical database schema meta-models according to the specific SQL standard. We have developed a meta-model of relational database schema (RDSMM), which can be classified as a generic database schema meta-model. Also, we developed a meta-model of a standard physical database schema (SPMM). We selected these meta-models, because they are in support of database model transformations that capture logical database schemas based on RDSMM from legacy databases based on SPMM. In this way, the extraction and conceptualization of a database schema from a legacy database is provided. Then, it can be analyzed, restructured or improved and it becomes the input of a forward engineering process to get a modernized database schema.

To specify and manage RDSMM and SPMM, we used the Eclipse Modeling Framework (EMF) [16]. Both of these two meta-models are complex, and here we focus on their parts aimed at specifying multi-relational constraints, and particularly INDs. An overview of a complete meta-model of the relational database schema may be found in [28]. It comprises several modeling concepts, like: Attribute Constraint, Relation Scheme, Universal Relational Schema (URS) and Relational Database Schema.

Apart from Introduction and Conclusion the paper has five sections. Section 2 is devoted to the recall of basic notions. A generic database schema meta-model is presented in Section 3. A standard physical database schema meta-model is explained in Section 4. In Section 5 we present a case study of an M2M transformation. Related work is presented in Section 6.

2. Relational Database Schema

In this section, we briefly recall some basic notions of the relational data model used in the text to assist the reader in easier following the rest of the paper. They are borrowed from many sources, as well as [13] and [15], and are slightly adapted to the needs of our research.

Let R be a finite set of **attributes**. For each attribute $A \in R$, the set of all its possible values is called the domain of A. The domain associated with an attribute A is denoted by $Dom(A)$, and the set of possible values of attribute A (A-values) is denoted by $dom(A)$ [25]. A **domain constraint** restricts allowed values within a certain domain. A **tuple** t over $R = \{A_1, ..., A_m\}$ is a sequence of values $(a_1, ..., a_m)$ where: $(\forall i \in \{1, ..., m\})(a_i \in dom(A_i))$. A **relation** over R, denoted with $r(R)$, is a set of tuples over R.

A universal relational schema (URS) is a pair (R, UC), where R is a set that contains all the attributes of the Universe of Discourse (UoD) with associated domain constraints, and UC is a set of URS constraints that comprises a set of functional dependencies and non-trivial inclusion dependencies. R is called **universal attribute set** (UAS). A **universal relation** u(UAS) is a relation over the UAS. A f**unctional dependency** (FD) is a relationship that exists when each X-value uniquely determines a Y-value. Formally, given a set of attributes R, a functional dependency between attribute sets X and Y is represented as $X \rightarrow Y$, which specifies that Y is functionally dependent on X. A **non-trivial inclusion dependency** is a statement of the form $[X] \subseteq [Y]$, where X and Y are non-empty sequences of attributes from UAS. The cardinalities of X and Y have to be equal (unlike the cardinalities of attribute sets X and Y in FD), and the corresponding sequence elements from X and Y have to be domain compatible (again, unlike FD). A universal relation u is said to satisfy the non-trivial inclusion dependency if for each tuple $t \in u$ exists at least one tuple $s \in u$ such that $t[X] = s[Y]$, where $t[X]$ represents the projection of tuple t on X. There are different database design approaches ([15], [25]). One of them is based on the URS assumption. Using the set of FDs, the URS is decomposed into a set of relation schemes, resulting in a relational database schema.

Formally, a **relational database schema** is a pair (S, I), where S is a finite set of relation schemes and I a finite set of multiple relational constraints. A **relation scheme** is a named pair $N(R, C)$ where N is the name of relation scheme, R is a finite set of attributes (from UAS) and C a finite set of relational constraints. C contains **attribute value constraints** alongside with **null constraints**, **tuple check constraints**, **uniqueness constraints** and **key constraints**. A set of **multiple relational constraints** I, contains **extended tuple constraints** (an example can be seen in [28]) and **inclusion dependencies**. Here we give only the definition of inclusion dependency.

Let $N_l(R_l, C_l)$ and $N_r(R_r, C_r)$ be two relation schemes, where N_l and N_r are their names, R_l and $R_{r,}$ their corresponding sets of attributes, and C_l and C_r their corresponding sets of relation scheme constraints. An **inclusion dependency (IND)** is a statement of the form $N_l[LHS] \subseteq N_r[RHS]$, where LHS and RHS are non-empty sequences of attributes from R_l and R_r respectively. Having the inclusion operator ($\subseteq$) orientated from the left to right we say that relation scheme N_l is on the left-hand side of the IND, while the relation scheme N_r is on its right-hand side. We use the indexes l and r, and the names of attribute sequences LHS and RHS, in order to indicate the left and right hand side of the IND, respectively. To define a validation rule of the IND we use the following notation: (i) the relation $r(N_l)$ is a set of tuples $u(R_l)$ (or just u) satisfying all constraints from the constraint set C_l; (ii) X-value is a projection of a tuple u on the set of attributes X; and (iii) according to the aforementioned orientation of the inclusion operator, $r(N_l)$ is called the referencing relation, while $r(N_r)$ is called the referenced relation. Informally, a database satisfies the inclusion dependency if the set of LHS-values in the referencing relation $r(N_l)$ is a subset of the set of RHS-values in the referenced relation $r(N_r)$.

There are two basic kinds of INDs: key-based INDs and non-key-based INDs. An IND is said to be **key-based** if the RHS is a key of the relation scheme N_r. Otherwise, it is a **non-key-based**. More often a key-based IND is called **referential integrity constraint (RIC)**. A non-key-based IND with a LHS that is a key of the relation scheme N_l, where a RIC $N_r[RHS] \subseteq N_l[LHS]$ is specified at the same time, is called **inverse referential integrity constraint (IRIC)**. The detailed explanation of these constraints may be found in [4].

In Fig. 1, a simplified part of a University database is given. The database satisfies inclusion dependencies:

$Ind1$: $Course[DepID] \subseteq Department[DepID]$
$Ind2$: $Department[DepID] \subseteq Course[DepID]$
$Ind3$: $Employed\_At[EID] \subseteq Employee[EID]$
$Ind4$: $Employee[EID] \subseteq Employed\_At[EID]$
$Ind5$: $Employed\_At[DepID] \subseteq Department[DepID]$
$Ind6$: $Taught\_By[DepID + CID] \subseteq Course[DepID + CID]$
$Ind7$: $Employee[SupervisorId] \subseteq Employee[EID]$
$Ind8$:$Taught\_By[EID + DepID] \subseteq \sigma_{Position = \text{'Prof.' or } Position = \text{'Ass.'}}$
$Employed\_At \rhd\lhd Employee[EID + DepID]$.

Ind1, *Ind3*, *Ind5*, *Ind6* and *Ind7* are the RICs since *DepID*, *EID*, *DepID*, *DepID* + *CID* and *EID* on the RHS of INDs are the keys of relation schemes *Department*, *Employee*, *Department*, *Course* and *Employee*, respectively. *Ind1*, *Ind3*, *Ind5* and *Ind7* are unary RICs, since the cardinality of the attribute sequence is 1, while *Ind6* is a binary RIC since the cardinality of the attribute sequence is 2. Generally, if the cardinality of attribute sequence is *n*, an IND is said to be ***n*-ary**. *Ind7* from the relational database schema is the consequence of a non-trivial IND from URS: [*SupervisorId*] $\subseteq$ [*EID*], while other RICs are the consequence of the decomposition of URS. *Ind2* and *Ind4* are the IRICs since: i) there are specified RICs *Ind1* and *Ind3*, respectively; and ii) *DepID* and *EID* on the LHS of *Ind2* and *Ind4* are the keys of relation schemes *Department* and *Employee*, respectively.

Employee

EID	FName	LName	Position	SupervisorID
003	Iva	Ilic	Ass.	007
007	Aca	Jovic	Prof.	009
009	Ina	Ras	Prof.	
010	Mila	Kun	PR	009

Department

DepID	DName
D1	Comp.
D2	Mech.
D3	Art

Employed_At

EID	DepID	Percent
003	D1	100
007	D1	70
007	D3	30
009	D2	100
010	D1	100

Course

DepID	CID	CName
D1	001	Java
D1	002	Databases
D2	003	Robotics
D3	001	Painting

Taught_By

EID	DepID	CID	ClassPerWeek
003	D1	001	3
007	D1	002	2
007	D3	001	2
009	D2	003	4

Fig. 1. A part of University database

IND *Ind8* requires further explanation. This type of IND is called **extended IND** [28] due to the fact that at least on the one side of the IND there is the natural join of two or more relation schemes. In our example, the RHS of the IND contains the natural join of two relation schemes *Employed_At* $\triangleright\triangleleft$ *Employee*. *Ind8* is also an example of **selective (conditional) IND** ([8], [17], [28]). An IND is said to be selective if there is a selection condition at least on the one side of the IND. The semantics of constraint *Ind8* is that a course can be taught only by an employee that is employed at the department that contains the course, and that the employee must be either professor or assistant. Therefore, a database with a relation *Taught_by* that would contain one of the tuples: (003, D3, 001, 1) or (010, D1, 002, 2) would not obey the constraint *Ind8* and would not be formally consistent. The first tuple insertion would fail due to the fact that employee with *EID* = 003 is not employed at department D3, and the second due to the fact that employee with *EID* = 010 has the position of PR, and is neither a professor nor an assistant.

Integrity has always been an important issue for database design and implementation. Its importance grows with increasing demands regarding the quality and reliability of data. Integrity constraint specifications are translated into constraint enforcing mechanisms provided by the Database Management System (DBMS) used to implement a database. Most of the commercial DBMSs offer efficient declarative support for the domain constraints, null value constraints, uniqueness constraints and RICs (by means of foreign key constraints). On the contrary, non-key-based INDs are completely disregarded by actual RDBMSs, obliging the users to manage them via custom procedures or triggers. That is the reason why these kinds of constraints are ignored by database designers in a way that they do not recognize, specify and implement them. In the paper we present two meta-models of relational database constraints: the Generic Constraint Meta-Model (GCMM) and the Standard Constraint Meta-Model (SCMM). In that way two abstract syntaxes of two modeling languages are defined to enable database schema specification. That is a prerequisite for the development of M2M transformations that would enable automated transformation of a physical database schema extracted from a database to a relational database schema based on theoretical foundations of the relational data model. These specifications may be transformed into declarative scripts, procedures or triggers for integrity constraint enforcement supported by a DBMS. These transformations are M2T transformations. In the following section the fundamental GCMM concepts are presented.

3. Fundamental Generic Constraint Meta-Modeling Concepts

There are two approaches to perform relational database design: top-down (design by analysis) and bottom-up (design by synthesis). Top-down design methodology involves conceptual schema design (e.g., using Entity-Relational data model) followed by its mapping into relational database schema that can be improved in the subsequent analysis process. Bottom-up approach presupposes that the set of UoD attributes and functional dependencies among them have been given as a URS. Several algorithms may be used to decompose URS into a relational database schema. Our meta-model comprises modeling concepts to specify both: URS and relational database schema. Our main motive to design a meta-model of URS was to support the database design approaches based on the URS assumption and appliance of a synthesis algorithm to generate relational database schema starting with a URS. In our ongoing research we develop M2M transformations of legacy relational database schema into URS. We use our IIS*Studio development environment (presented in [3] and [22]) aimed at relational database schema generation and integration, to reengineer relational database schema and to further generate application prototype.

Fundamental GCMM concepts are presented in Fig. 2. *Project* concept encompasses *URS* concept and *RelationDBSchema* concept.

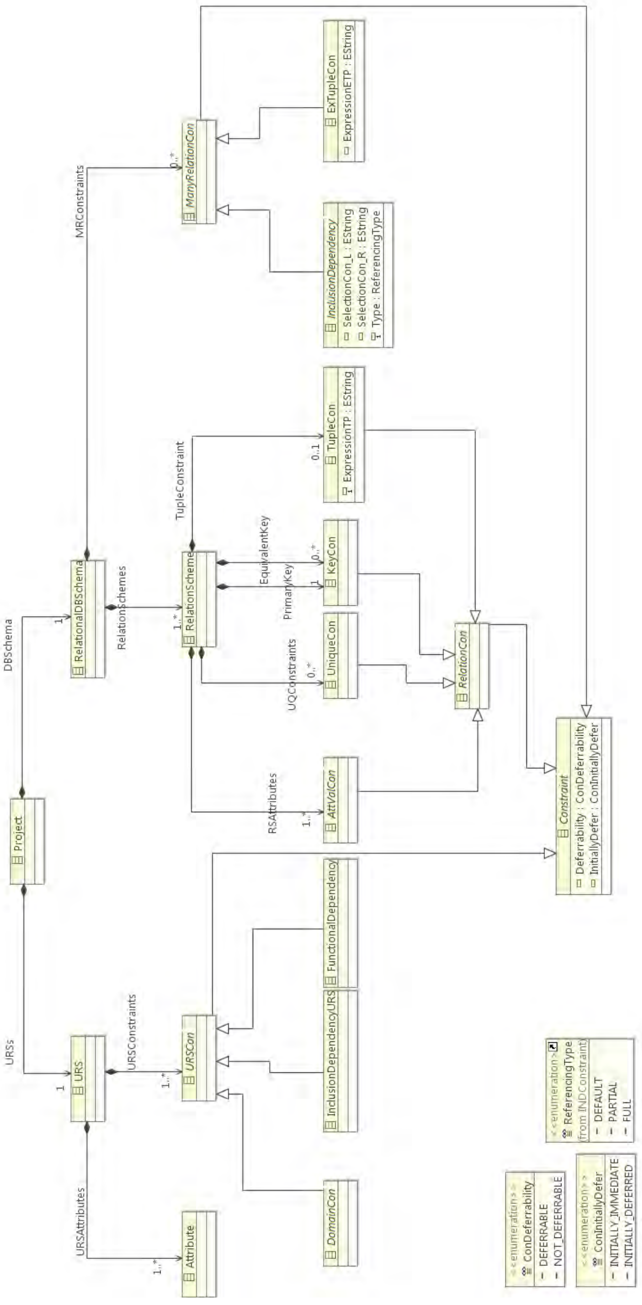


Fig. 2. Fundamental GCMM concepts

The *Constraint* is an abstract concept that has two properties: *Deferrability* and *InitiallyDefer*. *Deferrability* is aimed to specify whether or not constraint checking can be deferred until the end of the transaction. *InitiallyDefer* is used to specify the default checking behavior for constraints that are deferrable. The *Constraint* is specialized as: URS constraint (abstract concept *URSCon*), relation scheme constraint (abstract concept *RelationCon*) or multi relation constraint (abstract concept *ManyRelationCon*). There are three types of URS constraints: domain, functional dependency and non-trivial inclusion dependency. A detailed description of URS constraint meta-model may be found in [29]. Relation scheme constraints are specialized as: attribute value constraints (*AttValCon*), uniqueness constraints (*UniqueCon*), key constraints (*KeyCon*) and check constraints (*CheckCon*). Inclusion dependencies (*InclusionDependency*) and extended tuple constraint concept (*ExTupleCon*) specialize *ManyRelationCon* concept. Hereinafter we give a detailed description of inclusion dependency meta-model.

3.1. Inclusion Dependency Meta-Model

The *InclusionDependency* modeling concept is abstract and it generalizes concepts for modeling several kinds of INDs listed in Section 2. As can be seen in Fig. 3, *InclusionDependency* is first specialized with *ReferentialIntegrityCon* and *NonKeyBased IND*. The first is a concrete modeling concept aimed at modeling RICs. The second concept is abstract and is further specialized with concrete modeling concepts: *InverseReferentialIntegrityCon* and *NonInverseReferentialIntegrityCon* aimed at modeling IRICs and other INDs (that are neither RIC nor IRIC), respectively. Each of these three concrete concepts is further specialized with concrete concepts aimed at modeling: extended RICs, extended IRICs and others extended INDs. The *InclusionDependency* modeling concept has two properties: *SelectionCon_L* and *SelectionCon_R* that are used to specify selection conditions on the left or right side of IND. These properties are optional and if at least one of them is specified that implies that the modeled constraint is selective (conditional) IND. The third property *ReferencingType* is used to specify whether the referencing is default, partial or full, for *n*-ary INDs. For unary INDs there are no differences between these referencing types.

The relation schemes specified in an IND may have two roles: referencing, if it is on the LHS of the IND and referenced, if it is on the RHS of the IND. In Fig. 3, roles are modeled with concrete concepts *RoleReferenced* and *RoleReferencing*. For the referenced role, critical database operations that may violate the IND constraint are deletes and updates, and for the referencing role, critical operations are inserts and updates. The *RoleReferenced* and *RoleReferencing* concepts contain properties aimed at specifying the actions that would take place to preserve database from violation of an IND constraint in case when a critical operation occurs.

For an IND it has to be specified at least one relation scheme in each of the roles. If an IND is an extended RIC, than at least one more relation scheme must be specified as the referencing role of the IND. In the case of extended IRICs at least one more relation scheme must be specified as the referenced role of IND. For other extended INDs at least one more relation schema has to be specified as the referencing or as the referenced role of the IND.

For each IND, the attribute sequences on the LHS and RHS of the IND must be specified. In particular, for a RIC on the RHS is specified a key of a referenced relation scheme and for an IRIC on the LHS is specified a key of a referencing relation scheme, instead of an arbitrary attribute sequence.

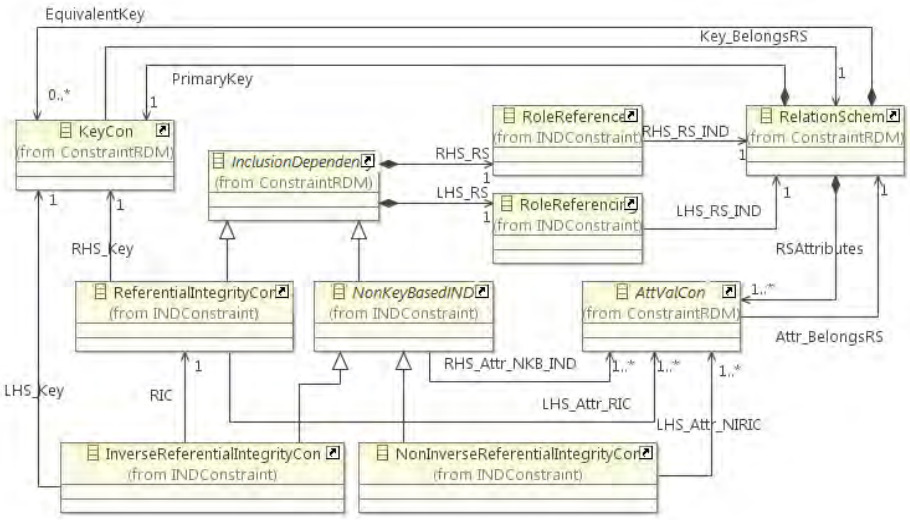


Fig. 3. Meta-model of Inclusion dependency concept

3.2. OCL Invariants to Improve Constraint Meta-Model Semantics

A meta-model defines the modeling language, i.e. the constructs that can be used to make a model and, consequently, defines a set of valid models [5]. Meta-models presented in Fig. 2 and Fig. 3 comprise a lot of knowledge about the real system (relational database schema) that would be modeled by means of it. But, it is still not enough. In order to improve the semantics of our meta-model we use Object Constraint Language (OCL) to specify constraints concerning modeling concepts that can not be expressed solely by means of UML class diagram concepts. Here we present some OCL constraint examples.

In Fig. 4, an OCL invariant is presented that enables checking if the sequence of attributes on the LHS of a RIC belongs to the set of attributes of the relation scheme that is specified as the referencing in the RIC. In addition, it checks if the key specified at the RHS of a RIC is a key of the relation scheme specified as referenced in the RIC.

A similar OCL invariant for the extended RIC can be seen in Fig. 5. The difference is in the fact that attributes for attribute sequence must belong to at least one of the relation schemes specified as referencing in the RIC.

In addition to checking of conformance between attribute sets and specified relation schemes, in Fig. 6 is presented an OCL invariant that checks if the appropriate RIC is specified for an IRIC.

```
invariant CheckLHSAttributeRIC:
LHS_Attr_RIC->forAll(a : ConstraintRDM::AttValCon |
                                          a.Attr_BelongsRS = self.LHS_RS.LHS_RS_IND);
invariant CheckRHSAttributeRIC:
RHS_Key->forAll(a : ConstraintRDM::KeyCon |
                                          a.Key_BelongsRS = self.RHS_RS.RHS_RS_IND);
```

Fig. 4. OCL invariant for RIC attribute checking

```
invariant CheckLHSAttributeExRIC:
LHS_Attr_RIC->forAll(a : ConstraintRDM::AttValCon |
                                     a.Attr_BelongsRS = self.LHS_RS.LHS_RS_IND or
                                     a.Attr_BelongsRS = LHS_RS_Ex_RIC.LHS_RS_IND);
invariant CheckRHSAttributeExRIC:
RHS_Key->forAll(a : ConstraintRDM::KeyCon |
                                          a.Key_BelongsRS = self.RHS_RS.RHS_RS_IND);
```

Fig. 5. OCL invariant for extended RIC attributes checking

```
invariant CheckRIC:
LHS_Key = RIC.RHS_Key and RHS_Attr_NKB_IND = RIC.LHS_Attr_RIC;
invariant CheckRHSAttributeIRIC:
RHS_Attr_NKB_IND->forAll(a : ConstraintRDM::AttValCon |
                                          a.Attr_BelongsRS = self.RHS_RS.RHS_RS_IND);
invariant CheckLHSAttributeIRIC:
LHS_Key->forAll(a : ConstraintRDM::KeyCon |
                                           a.Key_BelongsRS = self.LHS_RS.LHS_RS_IND);
```

Fig. 6. OCL invariant for the RIC – IRIC pair existence checking

4. Fundamental Standard Constraint Meta-Model Concepts

The relational data model is a superior logical data model and the acceptance of RDBMSs is widespread, too. Structured Query Language (SQL) is currently available in most commercial and open-source RDBMSs. It is also the focus of a continuous standardization process, resulting in SQL standards (the current revision is: SQL:2011, ISO/IEC 9075:2011). RDBMS products more or less comply with an SQL standard. Here we propose a standard physical database schema meta-model (Fig. 7). Instead of attribute and relation scheme concepts from GCMM, in SCMM there are column (*Column*) and table (*Table*) concepts, respectively. The *Constraint* abstract concept has the same properties, as the corresponding concept in GCMM. There are four concepts to specialize the *Constraint* concept: check constraint (*CheckCon*), primary key constraint (*PrimaryKeyCon*), uniqueness constraint (*UniqueCon*) and foreign key constraint (*ForeignKey*). The *UniqueCon* and *PrimaryKeyCon* concepts indirectly specialize the *Constraint* concept, via their generalization (*PKeyUnique*). A detailed description of SCMM, alongside with appropriate OCL invariants may be found in [3].

In the next section an illustration of an M2M transformation is presented. The transformation is based on SCMM and GCMM.

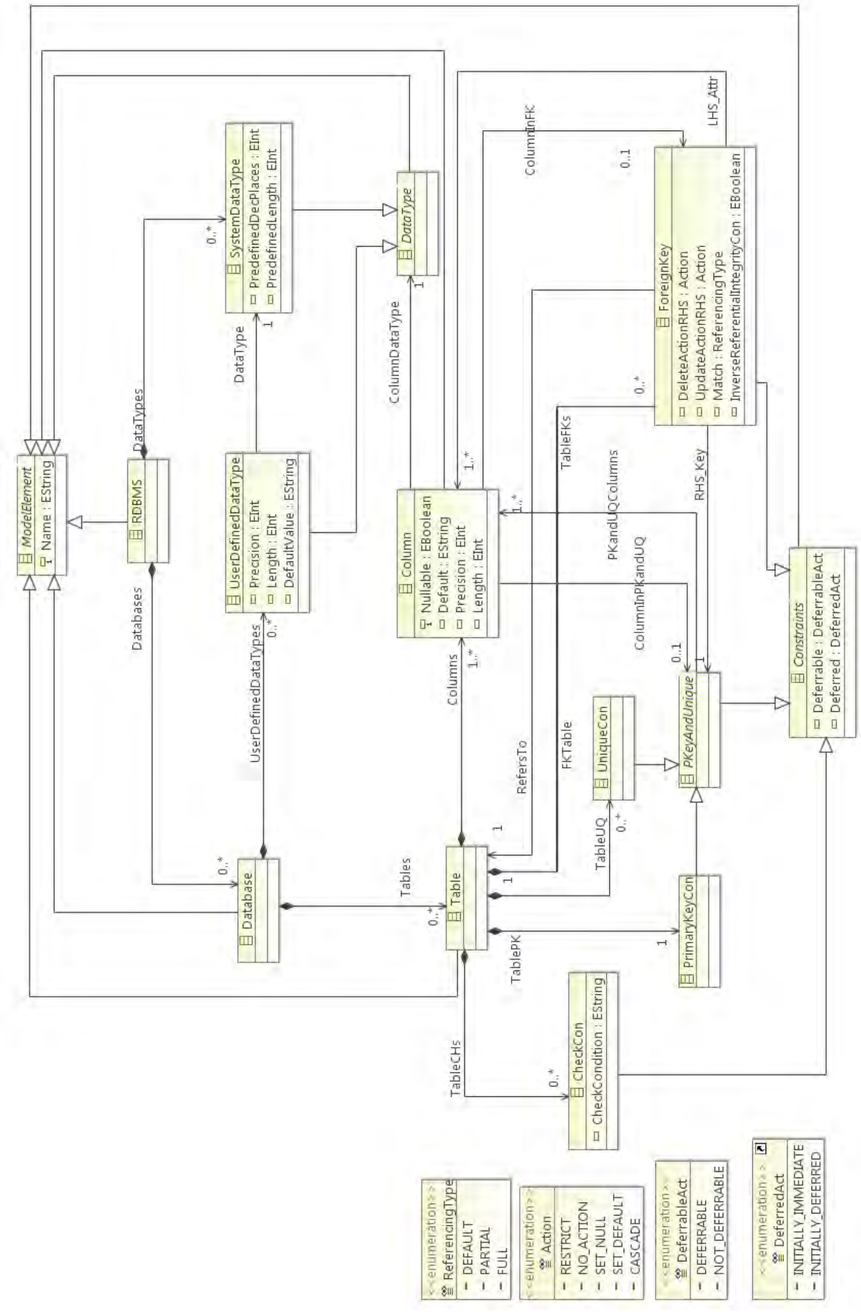


Fig. 7. Fundamental SCMM concepts

5. Case Study – Model-To-Model Transformation

To specify and manage presented meta-models we used the Eclipse Modeling Framework (EMF), Eclipse Juno 4.2.1. and OCL 3.2.1. After specifying the meta-model, a fully functional Eclipse editor can be generated for it. The editor guides a database schema specification process and ensures the conformance of the database schema with the proposed meta-model. By means of such an editor, we have specified a relational database scheme of the University database whose part is presented in Section 2. In Fig. 8, the conceptual database schema of the University database is visually represented by means of a UML class diagram to facilitate better understanding of database constructs and relationships between them. The relational database schema University contains the set of relation schemes: *Employee*, *University*, *Department*, *WorkSite*, *Course*, *EmployedAt* and *Taught_By*, accompanied with the set of multi-relational constraints. The detailed specification of the aforementioned relational database schema may be found in [28].

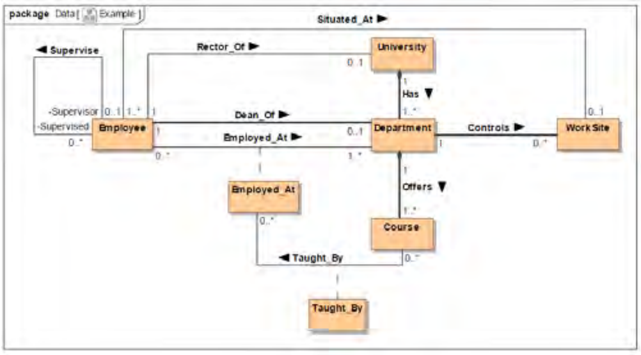
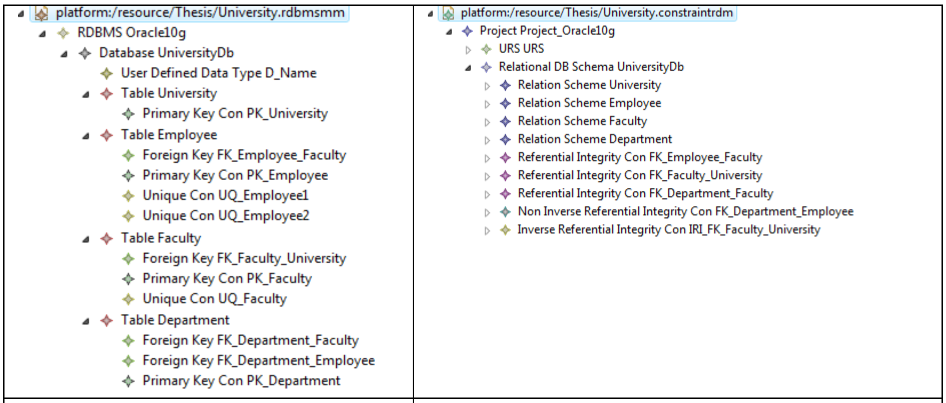


Fig. 8. The conceptual database schema of University database

A part of *University_PDBS* database schema modeled by means of the Eclipse editor generated from presented SCMM is shown in Fig. 9 (a). In this example it is captured from database repository.

In Fig. 9 (b) a part of the *University_SDBS* is presented. It is the output of the *University_PDBS*2*University_SDBS* model-to-model transformation that is based on SCMM and GCMM. The transformation will transform a vendor specific physical database schema that conforms to the SQL standard database schema, into generic, relational database schema. The Atlas Transformation Language (ATL) is used to implement *University_PDBS*2*University_SDBS* transformation.

In Fig. 10 (a), ATL rules for mapping the *ForeignKey* concept onto *ReferentialIntegrity Con* are presented. Rules for mapping the *ForeignKey* concept onto the *NonInvReferentialIntegrityCon* concept are shown in Fig. 10 (b).



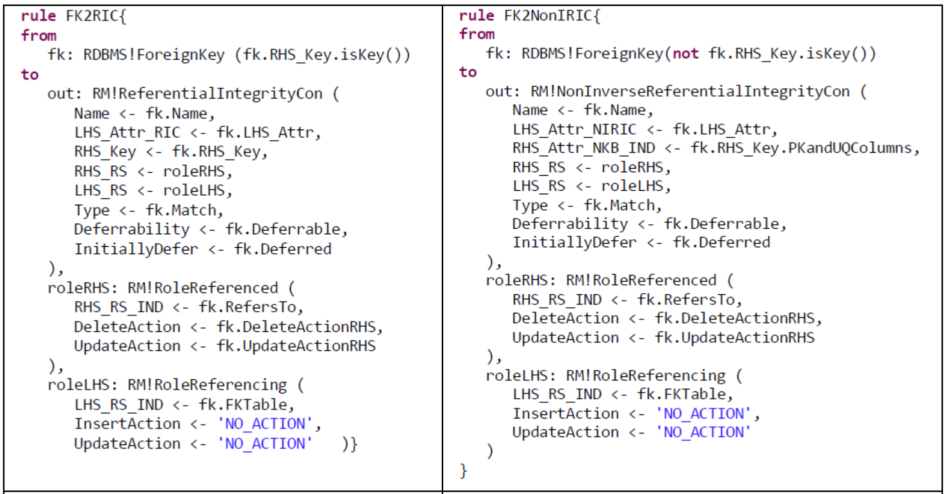
- platform:/resource/Thesis/University.rdbmsmm
  - RDBMS Oracle10g
    - Database UniversityDb
      - User Defined Data Type D_Name
      - Table University
        - Primary Key Con PK_University
      - Table Employee
        - Foreign Key FK_Employee_Faculty
        - Primary Key Con PK_Employee
        - Unique Con UQ_Employee1
        - Unique Con UQ_Employee2
      - Table Faculty
        - Foreign Key FK_Faculty_University
        - Primary Key Con PK_Faculty
        - Unique Con UQ_Faculty
      - Table Department
        - Foreign Key FK_Department_Faculty
        - Foreign Key FK_Department_Employee
        - Primary Key Con PK_Department

a) Model of *University_PDBS* conformant with SCMM

- platform:/resource/Thesis/University.constraintrdm
  - Project Project_Oracle10g
    - URS URS
    - Relational DB Schema UniversityDb
      - Relation Scheme University
      - Relation Scheme Employee
      - Relation Scheme Faculty
      - Relation Scheme Department
      - Referential Integrity Con FK_Employee_Faculty
      - Referential Integrity Con FK_Faculty_University
      - Referential Integrity Con FK_Department_Faculty
      - Non Inverse Referential Integrity Con FK_Department_Employee
      - Inverse Referential Integrity Con IRI_FK_Faculty_University

b) Model of *University_RDBS* conformant with GCMM

Fig. 9. Models of University_PDBS and University_RDBS



```
rule FK2RIC{
from
   fk: RDBMS!ForeignKey (fk.RHS_Key.isKey())
to
   out: RM!ReferentialIntegrityCon (
      Name <- fk.Name,
      LHS_Attr_RIC <- fk.LHS_Attr,
      RHS_Key <- fk.RHS_Key,
      RHS_RS <- roleRHS,
      LHS_RS <- roleLHS,
      Type <- fk.Match,
      Deferrability <- fk.Deferrable,
      InitiallyDefer <- fk.Deferred
   ),
   roleRHS: RM!RoleReferenced (
      RHS_RS_IND <- fk.RefersTo,
      DeleteAction <- fk.DeleteActionRHS,
      UpdateAction <- fk.UpdateActionRHS
   ),
   roleLHS: RM!RoleReferencing (
      LHS_RS_IND <- fk.FKTable,
      InsertAction <- 'NO_ACTION',
      UpdateAction <- 'NO_ACTION'    )}
```

a)ATL rules to map *ForeignKey* concept onto *ReferentialIntegrityCon* concept

```
rule FK2NonIRIC{
from
   fk: RDBMS!ForeignKey(not fk.RHS_Key.isKey())
to
   out: RM!NonInverseReferentialIntegrityCon (
      Name <- fk.Name,
      LHS_Attr_NIRIC <- fk.LHS_Attr,
      RHS_Attr_NKB_IND <- fk.RHS_Key.PKandUQColumns,
      RHS_RS <- roleRHS,
      LHS_RS <- roleLHS,
      Type <- fk.Match,
      Deferrability <- fk.Deferrable,
      InitiallyDefer <- fk.Deferred
   ),
   roleRHS: RM!RoleReferenced (
      RHS_RS_IND <- fk.RefersTo,
      DeleteAction <- fk.DeleteActionRHS,
      UpdateAction <- fk.UpdateActionRHS
   ),
   roleLHS: RM!RoleReferencing (
      LHS_RS_IND <- fk.FKTable,
      InsertAction <- 'NO_ACTION',
      UpdateAction <- 'NO_ACTION'
   )
}
```

b) ATL rules to map *ForeignKey* concept onto *NonInvReferentialIntegrityCon* concept

Fig. 10. ATL transformation rules

6. Related Work

Meta-modeling is widely spread area of research. OMG's Model-Driven Architecture (MDA) [26] currently is the most mature formulation of the MDSE paradigm. It refers to a high-level description of an application as a Platform Independent Model (PIM) and a more concrete implementation-oriented description as a Platform Specific Model (PSM) [26]. The OMG's Meta Object Facility (MOF) defines the metadata architecture that lies at the heart of MDA [24]. MOF is used to define semantics and structure of generic meta-models or domain specific ones. There are numerous references covering MOF based meta-models.

In the paper [14], Eessaar explained why it is advantageous to create meta-model of a data model. He demonstrated that a meta-model could be used in order to find similarities and differences between other data models. Polo, Garcia-Rodriguez and Piattini in [27] propose a very simplified relational and object-oriented meta-model. A similar, simplified RDBMS meta-model is presented in [32], by Wang, Shen and Chen. Vara et al. in [31] presented Oracle 10g meta-model that can be classified as vendor-specific physical database schema meta-model. Lano and Kolahdouz-Rahimi in [20] and [21] propose a rather simplified relational database model not discerning standard and vendor specific constructs. SQL standard meta-models can be found in Calero et al. [11] and del Castillo et al. [12]. The importance of generic models is emphasized by Atzeni, Gianforme and Cappellari in [1] and [2]. Cabot and Teniente in [10] and Cabot et al. in [9] present an OCL meta-model and a case study where they used a simplified UML class meta-model, that can be classified as a generic database schema meta-model. The paper of Gogolla et al. [18] is interesting in another context: it presents the intensional and extensional ER/relational meta-models. The relational database schema meta-models that we presented in this paper are intensional meta-models. Our future research has to consider extensional database meta-models, too.

Most of the presented database meta-models can be classified as standard physical database schema meta-models or vendor-specific physical database schema meta-models. On the contrary, our relational database schema meta-model is generic database schema meta-model and it comprises a broader set of concepts with more properties then the aforementioned database meta-models. Such a meta-model will enable the creation of semantically rich models of relational database schemas. Such models can be further transformed through the chain of M2M and M2T transformations, ending up with automatically generated executable program code for the implementation of all specified database constraints.

There is one more issue that is important in the context of results presented in this paper. Inclusion dependency discovery has attracted a lot of research interests from the communities of database design, machine learning and knowledge discovery. Bauckmann, Leser, and Naumann in [6], Koeller and Rundensteiner in [19] and De Marchi, Lopes and Petit in [23] propose different techniques to discover INDs. In recent years, some studies to extend traditional inclusion dependencies have been made, like Bravo, Fan and Ma in [8] and Fan in [17]. The methods for conditional IND discovery are suggested, too. Bi and Shan in [7] notify new interest in dependencies to extend traditional dependencies, such as conditional functional dependencies and conditional inclusion dependencies. They state that data dependencies play an important role in data repair, too. Once discovered, INDs would be properly specified by means of a meta-model that is semantically rich enough. We have developed our meta-models keeping that in mind.

7. Conclusion

Some kinds of relational database constraints are well-known and can be implemented by the declarative DBMS mechanisms, like the key constraint and the referential integrity constraint. However, some kinds of constraints are not recognized by contemporary DBMSs and have to be implemented through the procedural mechanisms. Very often these kinds of constraints are ignored by database designers in a way that they do not recognize, specify and implement them. The striking examples are some kinds of inclusion dependency (IND) constraints, like: inverse referential integrity constraint, conditional IND and extended IND. In the paper we present a part of our research efforts focused on meta-models relating to databases. We developed a generic meta-model of relational database schema and a standard physical database schema meta-model. Here we deal with their parts concerning multi-relational constraints: inclusion dependencies in the generic meta-model and foreign key constraint in the standard meta-model. We consider all kinds of constraints as important to be specified and implemented. In the context of forward engineering, the proposed generic database meta-model will enable platform independent specification of a broad class of INDs and development of further M2M and M2T transformations ending up with executable program code. On the other side, a meta-model of INDs is important in the context of database and information system reengineering, too. Considering growing interest in IND discovery in legacy databases, we plan to integrate these results with our IIS*Studio development environment. That would enable the reconstruction of a relational database schema and its improvement through the mechanisms implemented in IIS*Studio.

Acknowledgements. Research presented in this paper was supported by Ministry of Science and Technological Development of Republic of Serbia, Grant III-44010, Title: *Intelligent Systems for Software Product Development and Business Support based on Models*.

References

1. Atzeni, P., Cappellari, P., and Gianforme, G.: MIDST: model independent schema and data translation. In Proceedings of the 2007 ACM SIGMOD international Conference on Management of Data (Beijing, China, June 11 - 14, 2007). SIGMOD '07. ACM, New York, NY, 1134-1136. (2007)
2. Atzeni, P., Gianforme, G., Cappellari, P.: A universal meta-model and its dictionary. T. Large-Scale Data and Knowledge-Centered Systems 1, 38–62. (2009)
3. Aleksic, S. Methods of Database Schema Transformations in Support of the Information System Reengineering Process, Ph.D thesis, University of Novi Sad, Faculty of Technical Sciences (2013).
4. Aleksic, S., Ristic, S., Lukovic, I., Celikovic, M.: A Design Specification and a Server Implementation of the Inverse Referential Integrity Constraints. Computer Science and Information Systems (ComSIS), Consortium of Faculties of Serbia and Montenegro, Belgrade, Serbia, ISSN: 1820-0214, Vol. 10, No.1, pp. 283-320. (2013)
5. Assmann, U., Zchaler and S., Wagner, G.: Ontologies, Meta-Models, and the Model-Driven Paradigm. In: Calero, C., Ruiz, F., Piattini, M. (eds.) Ontologies for Software Engineering and Software Technology (2006)
6. Bauckmann, J., Leser, U., and Naumann, F.: Efficiently Computing Inclusion Dependencies for Schema Discovery. Proc. Second Int'l Workshop Database Interoperability. (2006)
7. Bi, Z., and Shan, M.: Review of Data Dependencies in Data Repair. Journal of Information & Computational Science 9: 15 4623–4630. (2012)

8. Bravo, L., Fan, W., and Ma, S. Extending dependencies with conditions. In Proceedings of the 33rd international conference on Very large data bases (VLDB '07). VLDB Endowment 243-254. (2007)
9. Cabot, J., Clarisó, R., Guerra, E., and De Lara, J.: Verification and validation of declarative model-to-model transformations through invariants. Journal of Systems and Software, 83(2), 283-302. (2010)
10. Cabot, J. and Teniente, E.: Incremental integrity checking of uml/ocl conceptual schemas. Journal of Systems and Software, 82(9), 1459–1478. (2009)
11. Calero, C., Ruiz, F., Baroni, A., Brito e Abreu, F. , Piattini, M.: An ontological approach to describe the SQL:2003 object-relational features. Computer Standards & Interfaces, Volume 28, Issue 6, September 2006, Pages 695–713. (2006)
12. del Castillo, R.P., García-Rodríguez, I., and Caballero, I.: PRECISO: a reengineering process and a tool for database modernization through web services. In: Jacobson Jr., M.J., Rijmen, V., Safavi-Naini, R. (eds.) SAC 2009. LNCS, vol. 5867, pp. 2126–2133. Springer, Heidelberg. (2009)
13. Date, C.J. and Darwen, H.: Types and the Relational Model. The Third Manifesto, 3rd ed. Addison Wesley, Reading. (2006)
14. Eessaar, E.: Using Meta-modeling in order to Evaluate Data Models. In Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Corfu Island, Greece, February 16-19. (2007)
15. Elmasri R., Navathe B. S.: Database Systems: Models, Languages, Design and Application Programming, Sixth Edition, Pearson Global Edition, ISBN 978-0-13-214498-8. (2011)
16. Eclipse Modeling Framework, [Online] Available: http://www.eclipse.org/modeling/emf/. (retrieved February, 2014).
17. Fan, W.: Extending dependencies with conditions for data cleaning, Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference on, pp.185-190. (2008)
18. Gogolla, M., Lindow, A., Richters, M. and Ziemann, P.: Meta-model transformation of data models. Position paper. WISME at the UML (2002)
19. Koeller, A., and Rundensteiner, E.A.: Heuristic Strategies for Inclusion Dependency Discovery, On the Move to Meaningful Internet Systems 2004: Proc. Int'l Conf. CoopIS, DOA, and ODBASE, pp. 891-908. (2004)
20. Lano, K., and Kolahdouz-Rahimi, S.: Model-driven development of model transformations. Theory and Practice of Model Transformations, 47-61. (2011)
21. Lano, K., and Kolahdouz-Rahimi, S.: Constraint-based specification of model transformations Journal of Systems and Software, Volume 86, Issue 2, Pages 412–436. (2013)
22. Luković, I., Mogin, P., Pavićević, J. and Ristić, S.: An Approach to Developing Complex Database Schemas Using Form Types, Software: Practice and Experience, John Wiley & Sons Inc, Hoboken, USA, ISSN: 0038-0644, DOI: 10.1002/spe.820 Vol. 37, No. 15, pp. 1621-1656. (2007)
23. De Marchi, F., Lopes, S., and Petit, J.-M.: Unary and N-Ary Inclusion Dependency Discovery in Relational Databases, J. Intelligent Information Systems, vol. 32, no. 1, pp. 53-73. (2009)
24. Meta-Object Facility, [Online] Available: http://www.omg.org/mof/. (retrieved February, 2014)
25. Mogin, P., Luković, I., Govedarica, M.. Database Design Principles, University of Novi Sad, Faculty of Technical Sciences & MP "Stylos", Novi Sad, Serbia. (2004)
26. Mukerji, J. and Miller, J., 2003. MDA Guide Version 1.0.1, document omg/03-06-01 (MDA Guide V1.0.1), http://www.omg.org/, (retrieved February, 2014)
27. Polo, M., Garcia-Rodriguez, I., and Piattini, M.: An MDA-based approach for database reengineering. J. Softw. Maint. Evol. 19, 6 (November 2007), 383-417. (2007)

28. Ristic, S., Aleksic, S., Celikovic, M., and Lukovic, I.: An EMF Ecore based relational dB schema meta-model. In Proceedings of the 6th International Conference on Information Technology ICIT 2013. Amman, Jordan. (2013).
29. Ristic, S., Aleksic, S., Celikovic, M., and Luković, I. Meta-modeling of inclusion dependency constraints. In Proceedings of the 6th Balkan Conference in Informatics (BCI '13). ACM, New York, NY, USA, 114-121. (2013)
30. Stahl T, Völter M, Bettin J, Haase A, Helsen S.: Model Driven Software Development: Technology, Engineering, Management. John Wiley & Sons, Ltd. (2006)
31. Vara, J., Vela, B., Bollati, V.A. and Marcos, E.: Supporting model-driven development of object-relational database schemas: a case study, in: R. Paige (Ed.), Theory and Practice of Model Transformations, Heidelberg, Springer Berlin, pp. 181–196. (2009)
32. Wang, H., Shen, B. and Chen, C.: Model-Driven Reengineering of Database. Software Engineering, 2009. WCSE '09. WRI World Congress on , vol.3, no., pp.113-117. (2009)

Sonja Ristić works as an associate professor at the University of Novi Sad, Faculty of Technical Sciences, Serbia. She received two bachelor degrees with honors from University of Novi Sad (UNS), one in Mathematics, Faculty of Science in 1983, and the other in Economics from Faculty of Economics, in 1989. She received her Mr (2 year) and Ph.D. degrees in Informatics, both from Faculty of Economics (UNS), in 1994 and 2003. From 1984 till 1990 she worked with the Novi Sad Cable Company NOVKABEL–Factory of Electronic Computers. From 1990 till 2006 she was with High School of Business Studies – Novi Sad, and since 2006 she has been with the Faculty of Technical Sciences (UNS). Her research interests are related to Database Systems, Information Systems and Software Engineering.

Slavica Aleksić received her M.Sc. degree from the Faculty of Technical Sciences, at University of Novi Sad in 1998. She received her Mr (2 year) degree in 2006 and Ph.D. degree in 2013, at the University of Novi Sad, Faculty of Technical Sciences. Currently, she works as a assistant professor at the Faculty of Technical Sciences at the University of Novi Sad, where she lectures in several Computer Science and Informatics courses. Her research interests are related to Information Systems, Database Systems and Software Engineering.

Milan Čeliković graduated in 2009 at the Faculty of Technical Sciences, Novi Sad, at the Department of Computing and Control. Since 2009 he has worked as a teaching assistant at the Faculty of Technical Sciences, Novi Sad, at the Chair for Applied Computer Science. In 2010, he started his Ph.D. studies at the Faculty of Technical Sciences, Novi Sad. His main research interests are focused on: Domain specific modeling, Domain specific languages, Databases and Database management systems. At the moment, he is involved in the projects concerning application of DSLs in the field of software engineering.

Ivan Luković received his M.Sc. degree in Informatics from the Faculty of Military and Technical Sciences in Zagreb in 1990. He completed his Mr (2 year) degree at the University of Belgrade, Faculty of Electrical Engineering in 1993, and his Ph.D. at the University of Novi Sad, Faculty of Technical Sciences in 1996. Currently, he works as a Full Professor at the Faculty of Technical Sciences at the University of Novi Sad, where he lectures in several Computer Science and Informatics courses. His research interests are related to Database Systems and Software Engineering. He is the author or co-author of over 100 papers, 4 books, and 30 industry projects and software solutions in the area.

Received: *February 16*, *2014*, *Accepted*: *May 28*, *2014*.

Computer Science and Information Systems 11(2):697–721 DOI: 10.2298/CSIS140301038B

Agent Reasoning on the Web using Web Services

Costin Bădică[1], Nick Bassiliades[2], Sorin Ilie[1], and Kalliopi Kravari[2]

[1] University of Craiova
Bvd. Decebal, 107, RO-200440, Craiova, Romania
{cbadica,silie}@software.ucv.ro
[2] Aristotle University of Thessaloniki, GR-54124, Thessaloniki, Greece
{nbassili,kkravari}@csd.auth.gr

Abstract. In this paper we present an approach for reusing agent-based reasoning capabilities by making them available for invocation as Web services. In this way, we provide the missing link between the highly interoperable Web services and the autonomicity and intelligence of agent-based systems, so that the latter can be seamlessly integrated into the knowledge-rich Semantic Web environment without being compromised by isolated communication platforms and languages or restricted to only one or just few reasoning formalisms. We have achieved this by extending the EMERALD framework for agent based reasoning with a Web service interface. Our approach is exemplified by the development of an online system for intelligent brokering of apartment rentals. The broker intelligence is captured as a defeasible knowledge base, while its problem solving process involves the invocation of third party defeasible reasoning Web services included into the EMERALD framework.

Keywords: Web service, reasoning, software agent, UML.

1. Introduction

The Web was originally envisioned as a platform for information dissemination to human consumers. The next generation of Web systems shifted its interest to make that information available for machine consumption by realizing that Web data can be reused for various problem solving purposes. This idea was put into practice through the development of the Semantic Web [2] that promotes the enrichment of Web content with meaningful semantic metadata.

In parallel with the progress of the Semantic Web, the field of distributed systems has seen a progress by the proposal of a new approach known as service oriented architecture. By promoting the new ideas of interoperability and integration, the adoption of this approach enabled the development of significantly more complex distributed applications by integration of various existing or new heterogeneous components.

Rapidly, this idea was embraced by the Web community resulting in the adoption of the service oriented architecture as a method for developing more complex Web-based applications using the Web services concepts, standards and technologies. Naturally, the areas of the Semantic Web and Web services converged to the new concept of Semantic Web services. They can be broadly understood either as Web services whose description is enhanced with semantic metadata to allow automated discovery and composition, or as Web services capable of providing declarative semantics in the form of reasoning methods, beyond the operational semantics of procedural invocation.

Software agents are defined as computer systems situated in an environment that are able to achieve their objectives by: (i) acting autonomously, i.e. by deciding themselves what to do, and (ii) being sociable, i.e. by interacting with other software or human agents [33]. If we place software agents into the Web environment then we can easily regard them as a potential incarnation of the Semantic Web vision. According to this view, the Semantic Web is regarded as a Web of semantic services that are provided and consumed by software agents over the Web protocols.

We already have service provisioning agents that act similarly with Web services being able to provide reasoning services [18] or negotiation services [9]. However, we have noticed that there are both conceptual and technological "missing links" between software agents and Web services in the context of the Semantic Web. From the conceptual point of view, services lack autonomy and they are rather passive components that behave reactively by waiting to be invoked by client software. They also lack social abilities, as well as the ability to adapt and evolve. On the other hand agents have interesting features regarding autonomy, sociability and adaptivity. From the technological point of view, multi-agent systems are developed using specialized packages and deployed on distributed multi-agent platforms. Those platforms are providing appropriate middleware services required for running distributed multi-agent systems [7]. Moreover, while available for invocation within the agent platform, agent services are not immediately available for invocation over the Web. In this paper we propose a system that integrates and updates existing software tools to enable the invocation of agent reasoning services over the Web. In particular we highlight the problems, as well as our solution details of the integration process of a defeasible reasoning service into a coherent system for supporting a realistic e-commerce scenario in the domain of brokering apartment rental.

Our work contributes to bridging gap between software agents and Web services in the context of the Semantic Web. Using our approach, a Web application can reuse a set of declarative rules that capture valuable knowledge about the various constraints, requirements and particularities of the problem domain. The knowledge can be represented using one of the available rule-based knowledge representation formalisms. Using our extension, Web applications can exploit this knowledge by invoking appropriate reasoning services over the Web protocols on specific sets of facts that capture the particular instance of the problem in hand.

The main contributions reported in this paper can be summarized as follows:

(i) The extension of EMERALD multi-agent knowledge-based framework [17] to enable the exposure of agent-based reasoning services as Web services.

(ii) The development of a prototype system supporting a realistic scenario for intelligent brokering of apartment rentals. The broker intelligence is captured as a defeasible knowledge base, while its problem solving process involves the invocation of third party defeasible reasoning Web services included into the EMERALD framework. The prototype provides a Web-based user interface, incorporates a rule-based knowledge base about user preferences and invokes external Web services for reasoning. This prototype is intended as a proof-of-concept for checking the viability of our proposed approach.

Note that, although in our experimental system we used the particular *DR Reasoner* included into the EMERALD framework [16] and the scenario involves an e-commerce

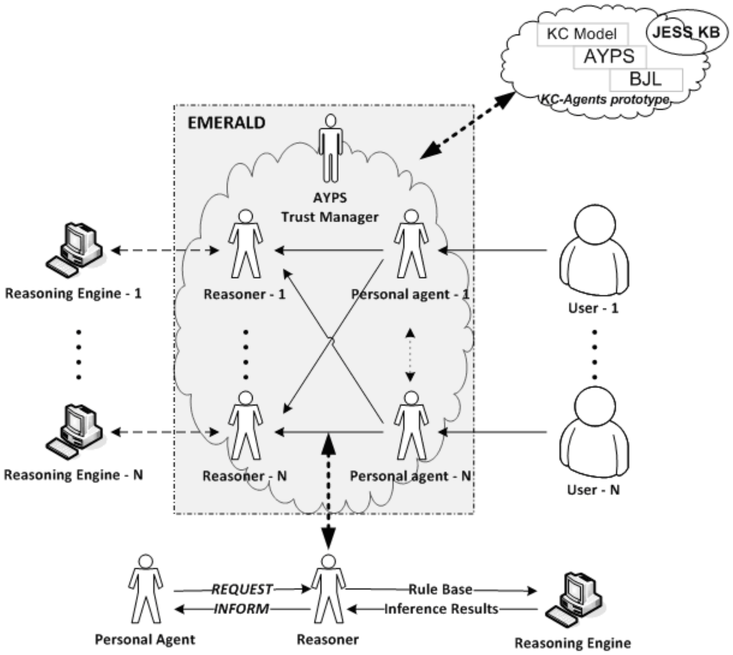


Fig. 1. EMERALD abstract architecture.

brokering activity for apartments rental, the presented approach is general and customizable so it can be also applied (i) to other types of reasoners, as well as (ii) to other application domains.

The paper is structured as follows. In section 2 we give an overview of background results and software tools that we have utilized to develop our prototype system and we also provide a review of related works. Section 3 introduces our method for exposing agent-based reasoners incorporated into EMERALD as Web services. The method was used for the development of the *DR Web Service* for defeasible reasoning. In section 4 we present our use case and we outline the design and implementation of an intelligent broker for the domain of apartment rental. Section 5 presents the details of our experiment carried out with the prototype system. The last section presents our conclusions and identifies a number of paths for continuing this research.

2. Background and Related Works

In this section we provide a brief overview of software platforms and components used in our proposed system, as well as an outlook to the results and related works from the literature.

2.1. Background

EMERALD ([17], see Fig. 1) is a multi-agent knowledge-based framework, based on the Semantic Web and FIPA standards [10], that enables reusability and interoperability of behavior between agents. It is built on Jade [4], a reliable and widely used framework. EMERALD supported so far the implementation of various applications, like brokering and agent negotiations. It provides a generic, reusable agent prototype for knowledge-customizable agents (KC-Agents), consisted of an agent model (KC Model), a directory service (AYPS - Advanced Yellow Pages Service) and external Java methods (Basic Java Library). Hence, EMERALD supports service provisioning by allowing agent service descriptions to be published onto the "yellow pages" service - known as Directory Facilitator (DF) in FIPA and Jade. Whenever a consumer agent decides to outsource a certain function, it can search the AYPS for an appropriate service that meets its specific requirements. AYPS actually provides a service retrieval ability, which groups and sorts the registered (advertised) services according, among others, to their domain and their synonyms, allowing (requesting) agents to make complex queries and receive the best available service. Yet since trust has been recognized as a key issue in Semantic Web Multi-agent Systems, EMERALD adopts a variety of reputation mechanisms, both decentralized and centralized [26].

Another important issue in this framework is reasoning interoperability. Agents do not necessarily share a common rule or logic formalism, thus, it is vital for them to find a way to exchange their position arguments seamlessly. To this end, EMERALD proposes the use of Reasoners [18], which are actually agents that offer reasoning services to the rest of the agent community. This approach does not rely on translation between rule formalisms, but on exchanging the results of the reasoning process of the rule base over the input data. The receiving agent uses an external reasoning service to grasp the semantics of the rule base, namely the set of entailments of the knowledge base. Thus, although Reasoners are built as agents, actually they act more like Web services.

Currently, EMERALD implements a number of Reasoners that offer reasoning services in two major reasoning paradigms: deductive rules and defeasible logic. Deductive reasoning is based on classical logic arguments, where conclusions are proved to be valid, when the premises of the argument (i.e. rule conditions) are true. Defeasible reasoning [23], on the other hand, constitutes a non-monotonic rule-based approach for efficient reasoning with incomplete and inconsistent information, which is useful in many applications, such as security policies, business rules, e-contracting, personalization, brokering and agent negotiations.

Among these Reasoners available in EMERALD, DR-Reasoner based on DR-DEVICE [3] has an exceptional interest. It is based on defeasible logic (DL), a nonmonotonic logic which is capable of modeling the way intelligent agents, like humans, draw reasonable conclusions from inconclusive information. Knowledge in DL is represented in the form of facts and rules. Facts are indisputable statements, represented either in form of states of affairs or actions that have been performed. On the other hand, rules describe the relationship between premises and conclusion. Three types of rules are available strict rules, defeasible rules and defeaters. Strict rules ($A_1, \ldots, A_n \rightarrow B$) are rules in the classical sense: whenever the premises are indisputable then so is the conclusion. Thus, they can be used for definitional clauses. Defeasible rules ($A_1, \ldots, A_n \Rightarrow B$) are rules that can be defeated by contrary evidence. Defeaters ($A \rightsquigarrow p$), finally, are used to prevent conclusions, not to support them which is actually the main concept in DL. This logic does not support contradictory conclusions, instead seeks to resolve conflicts.

Hence, in cases where there is some support for concluding A but there is also support for concluding $\neg A$, namely where there are conflicting (mutually exclusive) literals, no conclusion can be derived unless one of them has priority over the other. This priority is expressed through a superiority relation among rules which defines priorities among them; namely where one rule may override the conclusion of another rule. For example, suppose that there is a conflict set $\{(c(x), c(d)|x \neq d\}$; it is consisted of two competing rules $r1$ $(r1 : a(X) \Rightarrow c(X))$ and $r2$ $(r2 : b(X) \Rightarrow c(X))$, where $r2$ could be superior $(r2 > r1)$. To this end, *DR Reasoner* in order to exploit the DL's capabilities uses, as already mentioned, the DR-DEVICE inference system. DR-DEVICE is capable of reasoning about RDF metadata over multiple Web sources using defeasible logic rules. More specifically, DR-DEVICE accepts as input the address of a defeasible logic rule base, written in DR-RuleML language, an extension of the OORuleML syntax. The rule base contains the rules and one or more RDF documents that contain the facts for the rule program, while the final conclusions are exported as an RDF document. Furthermore, DR-DEVICE supports all defeasible logic features, like rule types, rule superiorities etc., applies two types of negation (strong, negation-as-failure) and conflicting (mutually exclusive) literals. DR-DEVICE is based on a CLIPS-based implementation of deductive rules. The core of the system consists of a translation of defeasible knowledge into a set of deductive rules, including derived and aggregate attributes. The implementation is declarative as it interprets the not operator using Well-Founded Semantics [11].

An important aspect of our approach was the integration of multi-agent technologies with Web services. Authors of [25] provide a good overview of such possibilities with a special focus on interoperability between FIPA (Foundation for Intelligent Physical agents) and Web services. Among them, a useful tool for our own work is the Jade Web service Integration Gateway – WSIG (see [4], chapter 10). It allows the provision of Jade agent services as Web services. In particular, services provided by Reasoner agents available in EMERALD framework can be exposed as Web services. This allows the convenient incorporation of reasoning functions into complex Semantic Web applications.

2.2. Related Works

Concerning interoperability, Rule Responder [5] is quite similar to EMERALD. It builds a service-oriented methodology and a rule-based middleware for interchanging rules in virtual organizations. It demonstrates the interoperation of distributed platform-specific rule execution environments, with Reaction RuleML as a platform-independent rule interchange format. It has a similar view of reasoning service for agents and usage of RuleML but it is not based on FIPA specifications. EMERALD supports multiple rule formalisms via trusted third-party reasoning services rather than a single rule interchange language. In EMERALD, Reasoners instead of interchanging rule bases using a specific rule language, they have a simpler, but more general interface, that requires just the URL or the file path of the rulebase, in the form of Java Strings. This information is exchanged via ACL messages either from a requesting agent to a rule engine or from the rule engine to the requesting agent. Hence, it is up to the requesting agent to provide the appropriate files, by taking each time into consideration the rule engines' specifications. For instance, an agent who wants to use the DR-DEVICE rule engine has to provide valid RuleML files. To this end, whenever a Reasoner receives a new valid request, it launches a new instance of the associated reasoning engine that performs inference. Reasoners actually act like proxies, allowing other agents to contact with the appropriate rule engines with low computational and time cost. Otherwise, each single agent should have had hard-coded rule engines for every rule or logic formalism, which is infeasible. Moreover, EMERALD provides trust mechanisms that ensure the reliability of each agent or Reasoner, in particular, acting in the environment. Hence, it is ensured that Reasoners in EMERALD are trusted parties that have no access or interference in agents' internal information.

A similar to EMERALD architecture for intelligent agents is presented in [32], where reasoning engines are employed as plug-in components, while agents intercommunicate via FIPA-based communication protocols. The framework is build on top of the OPAL agent platform [24] and, similarly to EMERALD, features distinct types of reasoning services that are implemented as reasoner agents. The featured reasoning engines are 3APL [8], JPRS (Java Procedural Reasoning System) and ROC (Rule-driven Object-oriented Knowledge-based System). 3APL agents incorporate BDI logic elements and first-order logic features, providing constructs for implementing agent beliefs, declarative goals, basic capabilities and reasoning rules. JPRS agents perform goal-driven procedural reasoning and each JPRS agent is composed of a world model (agent beliefs), a plan library, a plan executor (reasoning module) and a set of goals. Finally, ROC agents are composed of a working memory, a rule-base (consisting of first-order, forward-chaining production rules) and a conflict set. The primary difference between the two frameworks lies in the variety of reasoning services offered by EMERALD. While the three reasoners featured in [8] are all based on declarative rule languages, EMERALD proposes a variety of reasoning services, including deductive, defeasible and modal defeasible reasoning, thus, comprising a more integrated solution. Furthermore, EMERALD not only features trust and reputation mechanisms but also it is based on the Semantic Web standards, like, for rule and data interchange.

The One Ring project [22] on the other hand is a promising scripting rules engine service. It aims to be used as a Web service (SOA) for multiple applications to gain access to scripted processing of arbitrary parameters. It centralizes processing of common or business rules for multiple applications that need access to the same rules. Rules are defined using a simple language understood by domain experts. The primary difference between One Ring and EMERALD lies in the variety of reasoning services offered by EMERALD. While One Ring supports only a simple but not widely understood language, EMERALD proposes a variety of reasoning services, including deductive, defeasible and modal defeasible reasoning. Yet EMERALD, featuring additionally trust mechanisms, comprises a more integrated solution complying with the Semantic Web standards. However, both approached share a similar view of rule importance and usage.

Our literature review revealed also a number of systems and approaches that combine distributed multi-agent systems, reasoning services and the Semantic Web technologies for the development of application in several areas including: medical rehabilitation, agent auctions, supply-chain management, and service personalization.

Reference [29] proposes an ontology-based medical rehabilitation system called OntoRis that provides rehabilitation-related expertise to patients and therapists. The system uses agent technology for exploring and integrating external knowledge resources into the application. Sharing of users experience is encouraged in OntoRis with the help of an

Web 2.0-enabled discussion forum. To ensure a certain level of correctness of information, the users experience in the forum is filtered by evidence-based medicine, thus only certified information will be integrated into OntoRis. The OntoRis component incorporates an ontology and a reasoning engine. Nevertheless, while OntoRis agents are useful for information searching and knowledge enhancement, they have no role in the automation of reasoning processes, requiring either the manual intervention of the OntoRis administrator, as well as knowledge certification via methods of evidence-based medicine.

Knowledge-based declarative approaches are also useful for the specification of interaction mechanisms between software agents. Reference [21] proposes a declarative framework for auctions mechanisms in multi-agent systems. The key point of the proposal is the factoring of the model into two architectural components, the auction host and the auction participants. Their interaction is governed by a protocol with a declarative specification decomposed into: generic negotiation protocol (GNP), declarative negotiation mechanism (DNM) and custom negotiation mechanism (CNS). The GNP makes possible the interaction between auction host and auction participants. The specific rules of a particular auction are described using the DNM, while the CNS captures the strategic behavior which is private to each auction participant. A prototype implementation of this model is proposed using the Jason agent programming language [6]. This implementation benefitted from several features of Jason including: Belief-Desire-Intention model, Prolog-style rule-based reasoning, meta-programming and extensibility. Nevertheless, a cleaner separation between the CNS and DNM is desirable, especially taking into account that negotiation strategies could require more specialized types of reasonings, for example to account for priorities or conflicts between strategy rules.

Authors of reference [31] propose an ontology-based approach for capturing negotiation knowledge in supply chain management multi-agent systems. Similarly to [21], negotiation knowledge is partitioned into shared negotiation ontology and private negotiation ontology to ensure agent communicative interoperability and privacy of strategic knowledge. The agents' negotiation ability is further enhanced with private inference rules defined on top of private negotiation ontology. Nevertheless, although conceptually quite similar to [21], the implementation approach is different. It is based on Jade multi-agent platform and Jess expert system shell [12], as well as on the Semantic Web technologies OWL and SWRL [30].

Personalization is an important concern in context-aware applications. Authors of [27] propose a method that combines ontological modeling of user profiles, rule-based personalization mechanisms and service-oriented architecture for providing personalized Help-on-Demand services to mobile users in pervasive environments. The flexibility of the method is achieved by using an intelligent personalization service that incorporates a rule-based knowledge and a reasoning engine. While interesting, one limitation of this approach is that it is restricted to a single type of reasoning based on standard forward chaining for determining cause-effect relationships.

Reference [20] proposes the use of agent-based Web services to enhance collaboration within a supply chain of retailers, manufacturers and suppliers. However, rather than looking at the possibilities for integrating agent and Web services technologies to enhance flexibility of reasoning, the paper is focused on defining and experimentally evaluating of two scenarios: (i) independence level, where collaboration is kept minimal by letting agents have individual goals and communicate minimally by peer-to-peer information exchange, and (ii) collaboration level, by letting partner agents develop strategic relationships through intense communication thus allowing to achieve global goals for the entire supply chain.

3. Agent-Based Reasoning Web Services

The aim of this section is to introduce the design and implementation of a Reasoning Web Service. An external Web-based application will be able to invoke this Web service by providing a rulebase and an initial set of facts. The Web service returns the output of the reasoning process to the invoker application as a set of resulted facts.

3.1. Architecture and Design

Our design presents a method for wrapping an EMERALD reasoning service as a Web service using the Jade WSIG. For that purpose we exploit:

- The architecture and interoperability protocols of Reasoner agents inside EMERALD framework.
- The architecture of Jade WSIG.

The Reasoning Web Service takes as input a rule base represented in RuleML and outputs a result file using a simple request-reply interaction protocol. The meaning of the rulesbase, as well as of the initial set of facts is application dependent. It must be appropriately defined and interpreted by the client application in the context of the application domain of the system that is using the reasoning process.

The rulebase will specify a link to an RDF file containing an initial set of facts using the *rdf_import* attribute of RuleML. The output of the reasoning process is determined as an RDF file and saved by the Reasoning Web Service in a temporary location on the service side, while its URL is returned to the client. Consequently, the client will be able to download the result file and further process it according to the application requirements.

The architecture of the Reasoning Web Service is shown in Fig. 2. In particular, this diagram shows the relation between EMERALD, Jade, WSIG and the “outside world”. The Reasoning Web Service is hosted by a Web server and implemented with the help of a WSIG servlet that handles reasoning requests. The servlet is interfaced with Jade via a special agent – the WSIG agent using a special component known as the Jade Gateway. The WSIG agent is able to pass reasoning requests to the appropriate Reasoner agent of the EMERALD framework, as well as to return reasoning results to the WSIG servlet.

For the design of the Reasoning Web Service we have produced an UML activity diagram that shows the activities of the client application and the reasoning platform, shown in Fig. 3, as well as an UML sequence diagram detailing the interactions between them, shown in Fig.4. For the reasoning platform we show two actors, namely the *Reasoning Web Service* which was developed by us, as well as the *Reasoner* which is part of EMERALD. The system functioning can be better understood if we follow both diagrams.

The *Client Application* invokes the *Reasoning Web Service* (interaction 1 in Fig.4) by passing it the *rulebase*. The *Reasoning Web Service* simply extracts the *rulebase* from the request and forwards it to the *Reasoner Agent* (interaction 2 in Fig.4). The result

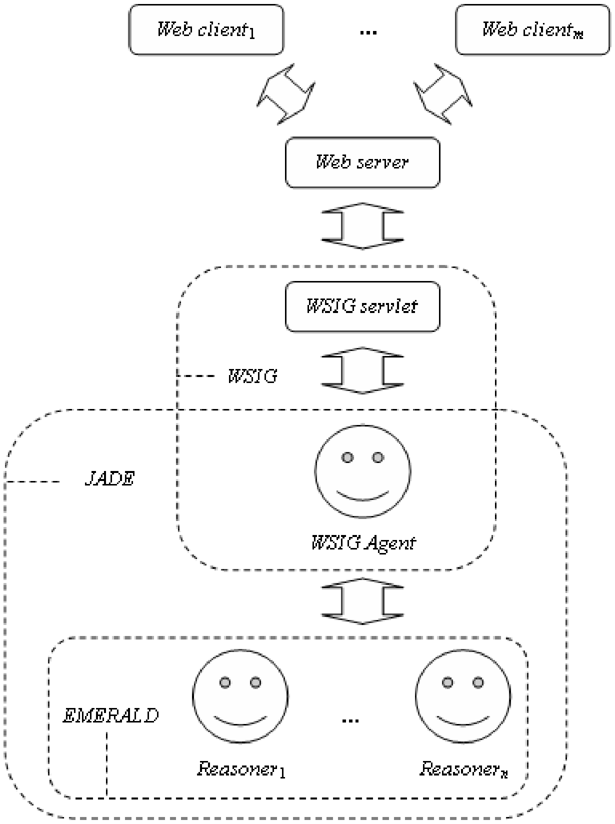


Fig. 2. Block diagram of Reasoning Web Service.

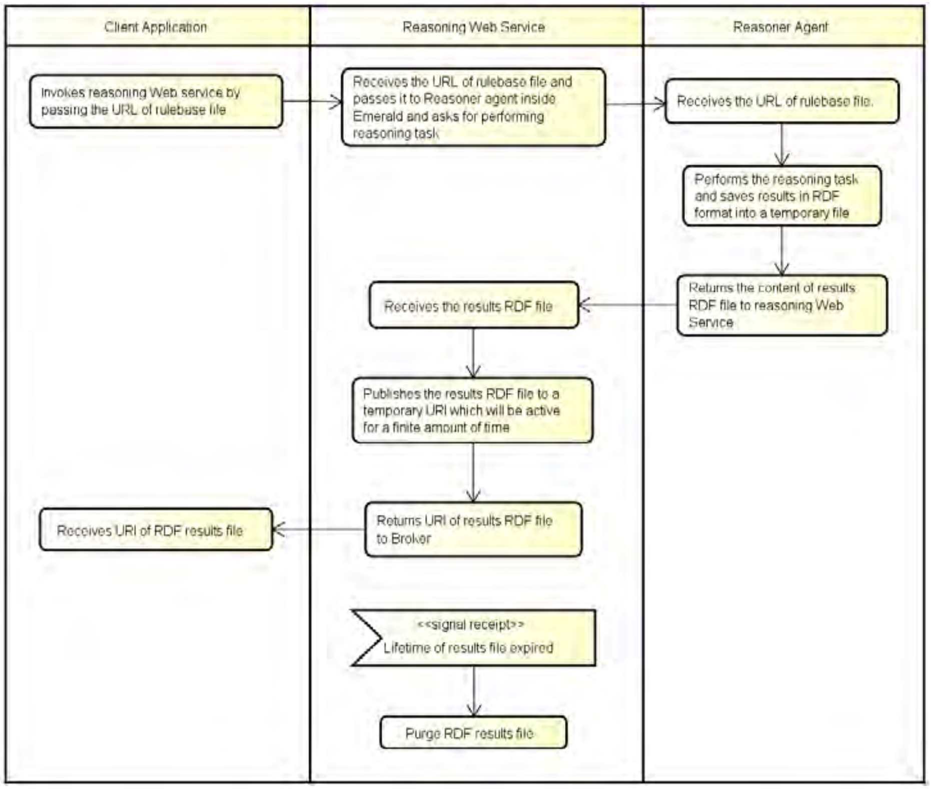

Client Application
Reasoning Web Service
Reasoner Agent
Invokes reasoning Web service by passing the URL of rulebase file
Receives the URL of rulebase file and passes it to Reasoner agent inside Emerald and asks for performing reasoning task
Receives the URL of rulebase file.
Performs the reasoning task and saves results in RDF format into a temporary file
Returns the content of results RDF file to reasoning Web Service
Receives the results RDF file
Publishes the results RDF file to a temporary URI which will be active for a finite amount of time
Returns URI of results RDF file to Broker
Receives URI of RDF results file
<<signal receipt>>
Lifetime of results file expired
Purge RDF results file

Fig. 3. Roles and activities in the Reasoning Web Service.

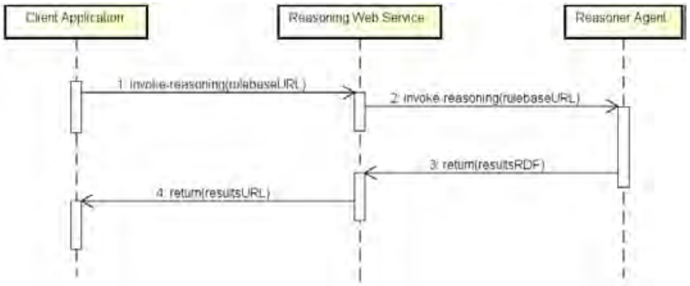


Fig. 4. Roles interaction in the Reasoning Web Service.

```
<wsdl:definitions ... declarations of namespaces>
  <wsdl:types>
    <xsd:schema xmlns:impl="urn:Defeasible_Reasoning_Service"
      ... declarations of namespaces>
      <xsd:annotation/>
      <xsd:element name="DRReasoning">
        <xsd:complexType>
          <xsd:sequence><xsd:element
            name="inputRulemlPath" type="xsd:string"/></xsd:sequence>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="DRReasoningResponse">
        <xsd:complexType>
          <xsd:sequence><xsd:element
            name="DRReasoningReturn" type="xsd:string"/></xsd:sequence>
        </xsd:complexType>
      </xsd:element>
    </xsd:schema>
  </wsdl:types>
  <wsdl:message name="DRReasoningRequest">
    <wsdl:part name="parameters" element="impl:DRReasoning"></wsdl:part>
  </wsdl:message>
  <wsdl:message name="DRReasoningResponse">
    <wsdl:part name="parameters" element="impl:DRReasoningResponse"></wsdl:part>
  </wsdl:message>
  <wsdl:portType name="Defeasible_Reasoning_ServicePort">
    <wsdl:operation name="DRReasoning">
      <wsdl:input message="impl:DRReasoningRequest"></wsdl:input>
      <wsdl:output message="impl:DRReasoningResponse"></wsdl:output>
    </wsdl:operation>
  </wsdl:portType>
  <wsdl:binding
    name="Defeasible_Reasoning_ServiceBinding"
    type="impl:Defeasible_Reasoning_ServicePort">
    <wsdlsoap:binding style="document"
      transport="http://schemas.xmlsoap.org/soap/http"/>
    <wsdl:operation name="DRReasoning">
      <wsdlsoap:operation soapAction="urn:Defeasible_Reasoning_ServiceAction"/>
      <wsdl:input><wsdlsoap:body use="literal"/></wsdl:input>
      <wsdl:output><wsdlsoap:body use="literal"/></wsdl:output>
    </wsdl:operation>
  </wsdl:binding>
  <wsdl:service name="Defeasible_Reasoning_ServiceService">
    <wsdl:port name="Defeasible_Reasoning_ServicePort"
      binding="impl:Defeasible_Reasoning_ServiceBinding">
      <wsdlsoap:address location=
        "http://ids.software.ucv.ro:8080/KSWAN/ws/Defeasible_Reasoning_Service"/>
    </wsdl:port>
  </wsdl:service>
</wsdl:definitions>
```

Fig. 5. WSDL of the DR Web service.

of the reasoning process expressed in RDF (*resultsRDF*) is returned to the *Reasoning Web Service* (interaction 3 in Fig.4). The *Reasoning Web Service* publishes the RDF file to a temporary URL and then returns this URL (*resultsURL*) to the *Client Application* (interaction 4 in Fig.4). The *Client Application* receives the URL and consequently can download the RDF file and further process it according to the application requirements. The temporary RDF file containing the reasoning results will reside on the Web server that hosts the Reasoning Web Service for a preestablished finite amount of time. When this time elapses the results file will be automatically purged, as can be seen in Fig.3.

3.2. Implementation

For demonstration purposes we have only addressed the integration of the *DR Reasoner* agent available in EMERALD into our prototype system. Nevertheless, the principles behind this integration process are the same for any other Reasoner that is part of the EMERALD platform. As a general procedure, a Reasoner (including the *DR Reasoner*): (i) receives a rulebase that contains the knowledge base as a set of rules expressed for interoperability purposes using RuleML, as well as possibly one or more RDF files containing initial facts that are required for the reasoning process, (ii) performs the reasoning and (iii) returns the reasoning result as an RDF file.

For the implementation we have used the Jade WSIG add-on [13] that gives the possibility of invoking agent services exposed as Web services from Web service clients. WSIG provides the standard Web services stack including: WSDL service descriptions, SOAP message transport and UDDI "yellow pages" repository. WSIG runs as part of a Web application powered by a Web server and a servlet container. WSIG incorporates the WSIG servlet and the WSIG agent and acts as a gateway between the Web application and the Jade multi-agent system. The WSIG servlet handles SOAP over HTTP requests, extracts the SOAP content, maps it to an agent action and passes it to the WSIG agent. After the action execution, WSIG servlet converts the result received from the WSIG agent to SOAP, prepares a SOAP over HTTP reply and sends it back to the Web client (see Fig. 2). The implementation follows the architecture diagram shown in Fig. 2.

The WSIG add-on provides a method for exposing Jade agent services whose descriptions are published in the *Directory Facilitator* (DF) agent as Web services. For this purpose, the DF description of an agent service is automatically mapped by WSIG onto an WSDL description. For example, the WSDL for the reasoning service provided by *DR Reasoner* is shown in Fig. 5. Analyzing this example we observe that this WSDL file defines a Web service named *Defeasible_Reasoning_Service* that was derived from the description of the *Defeasible_Reasoning_Service* service provided by *DR Reasoner* agent.

The WSDL is structured in an abstract part, containing data types, messages and operations specification, followed by a concrete part, containing bindings and services specification [1]. The abstract part contains: (i) type definitions specified using XML Schema – the *wsdl:type* section; (ii) message types, specified as *wsdl:message* sections; and (iii) port types that represent service interfaces (expressed using *wsdl:portype* element) made up of sets of operations, represented as *wsdl:operation* elements. For example our *Defeasible_Reasoning_Service* Web service has an interface with a single operation that takes two parameters: an input parameter representing a *DRReasoningRequest* message and an output parameter representing a *DRReasoningResponse* message. Going deeper, for example the *DRReasoningRequest* message transports an object of type *DRReasoning* that encapsulates a value named *inputRulemlPath* representing the URL of the rulebase that is passed to the service. The concrete part contains bindings for operations and messages (element *wsdl:binding*) and service endpoints specification (element *wsdl:port*). For example, *Defeasible_Reasoning_Service* Web service exchanges messages that are specified in "document" style and are formatted using SOAP and transported over HTTP. This binding is combined with the URL where this service is available: *http://ids.software.ucv.ro:8080/KSWAN/ws/Defeasible_Reasoning_Service*.

4. Web-Enabled Intelligent Broker

This section contains the details of the use case, as well as the design and implementation of our prototype Web-enabled Intelligent Broker that is using third party reasoning services provided by our Reasoning Web Service.

4.1. Use Case

We consider the intelligent broker use case introduced in [3] and we adapt it for the Web environment. The actors of this use case are:

(i) The *User* that represents a buyer interested to purchase or rent an apartment.
(ii) The *Broker* that matches buyer requested features with seller capabilities and recommends one or more possible transactions that meet the requirements of both trading parties.
(iii) The EMERALD framework that provides on-demand third party reasoning services for supporting the "intelligence" of the *Broker*.

Placing the actors in the Web environment, the situation presents as follows. The *User* is using a Web browser to connect to the *Broker* Web site and to search for apartment offers. The business of the *Broker* is powered by a Web application that provides at least the following two functionalities: (i) to assist the human users looking for renting apartments to search for suitable offers according to their requirements; (ii) to allow landlords to describe and publish their offers. Apartment offers will be registered and saved in a database that is available to the Web application of the *Broker*. In this work we assume the existence of this database and we shall only focus on the *User* side that is looking for renting or buying a suitable apartment.

The intelligence of the *Broker* is driven by a knowledge base. However, taking into account the large variety of knowledge representation formalisms, in particular based on rules, the *Broker* decided to outsource its reasoning functions to a third party reasoning platform (EMERALD in this case) that provides trusted reasoning capabilities that can be invoked over the Web as Web services. Separating the representation of the *Broker* intelligence from the reasoning services has the advantage that the *Broker* can update the knowledge-base according to the interests of the *User*. This update can either involve the update of the rules, as well as the adoption of a new rule representation formalism that better fits the *User* needs.

Briefly, the brokering use case proceeds as follows. In the first step the *User* connects to the *Broker* Web site and receives a form where he or she can fill-in his or her request. When the input process is finished, the *Broker* generates a parameterized rulebase represented using RuleML and an RDF file that contains parameter values specified by the *User*. Parameters can specify for example: minimum number of bedrooms, maximum floor, minim size of the garden (in m^2), if pets are allowed or not, maximum price for a central or suburb location, extra price per m^2 (above the minimum preferred size), the minimum preferred size, the available budget.

A more difficult task is however to provide the *User* with the facility of describing his or her general preferences, like for example: main preference is the cheapest option, second preference is the presence of a garden, and third preference is additional space, or preferences that cannot be easily parameterized, like for example: "if the flat is on the 3^{rd} floor or above the presence of an elevator is necessary". Therefore we have simplified the scenario by restricting our attention to a "prototypical user" for whom we defined a rulebase characterized by a given fixed set of parameters. The rulebase captures the "intelligence" of the *Broker* in understanding the preferences of a typical *User*, while the specific values of the parameters are input by the *User* at run-time using the form provided by the *Broker*. Eventually, an "experienced user" can still be provided with the possibility to visualize and update the rulebase. Moreover, if the user does not provide values for some of the preference parameters, then the corresponding rules will not be added to the generated rulebase.

In the second step the *Broker* will consult its internal database to retrieve a set of apartments which are relevant for the *User* request. In the third step, the *User* preferences (captured as the rulebase accompanied by a set of parameter-value pairs), together with the set of relevant apartments are submitted by the *Broker* to the reasoning service. Finally, the reasoning service returns to the *Broker* a set of available options. The *Broker* converts them into a meaningful presentation (eg. HTML) and sends it to the *User* for visualization in the browser.

4.2. Design

For the system design we have produced an UML activity diagram that shows the activities of the *User*, *Broker* and the reasoning service, shown in Fig. 6, as well as an UML sequence diagram detailing the interactions between them, shown in Fig.7. The reasoning service is named *DR Web service* and it is based on the *DR Reasoner* which is part of EMERALD reasoning platform. The system functioning can be better understood if we follow both diagrams.

The *User* connects to the Web site of the *Broker* by submitting the URL. The *Broker* delivers back to the *User* a Web page containing a form for preference input as a set of parameter-value pairs describing his options concerning the rental of an apartment. The *User* fills-in this form and submits it to the *Broker* Web site. Behind this form the *Broker* preprocesses the preferences and produces a rulebase represented in RuleML notation, following the example from [3]. The rulebase contains a set of defeasible rules that are used by the *Broker* to capture the generic preferences of a typical *User*. Differently from [3], the rulebase is kept separate from the set of parameter-value pairs (called *preferences* in Fig.7, interaction 3). This approach simplifies the operation of instantiating a particular scenario with the specific values input by the *User*. Intuitively, the rulebase represents the "intelligence" of the *Broker* in understanding and formalizing the requirements and preference of a prototypical user regarding the rental of apartments.

The *Broker* receives the *preferences*, converts them to RDF and publishes them to a temporary RDF file inside the Web server of the *Broker* Web site at a given URL. Next, the *Broker* determines an initial set of relevant apartments for the *User* request, maps this list to RDF using the method already shown in [3], and publishes it to a temporary RDF file inside the Web server of the *Broker* Web site at a given URL. These two RDF files are included into the rulebase using the *rdf_import* attribute of RuleML, thus creating an updated rulebase that contains the generic rules as well as the two RDF files that specify the initial set of facts for the reasoning process.

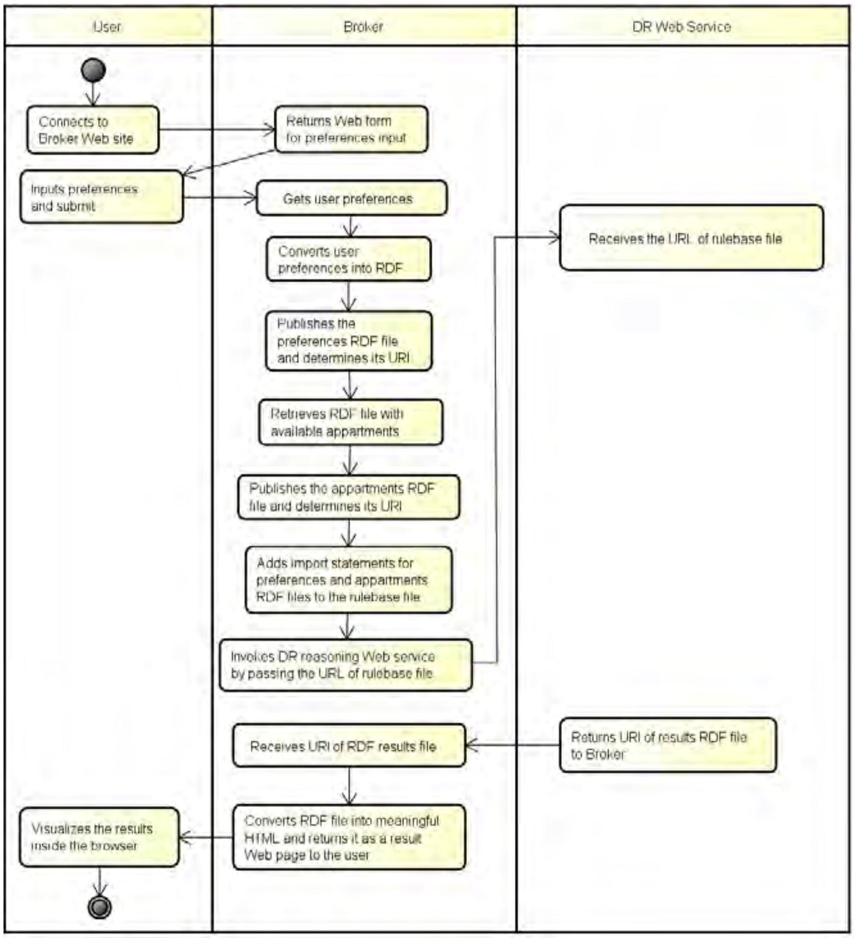

User
Broker
DR Web Service
Connects to Broker Web site
Returns Web form for preferences input
Inputs preferences and submit
Gets user preferences
Receives the URL of rulebase file
Converts user preferences into RDF
Publishes the preferences RDF file and determines its URI
Retrieves RDF file with available appartments
Publishes the appartments RDF file and determines its URI
Adds import statements for preferences and appartments RDF files to the rulebase file
Invokes DR reasoning Web service by passing the URL of rulebase file
Receives URI of RDF results file
Returns URI of results RDF file to Broker
Visualizes the results inside the browser
Converts RDF file into meaningful HTML and returns it as a result Web page to the user

Fig. 6. Roles and activities in the brokering scenario.

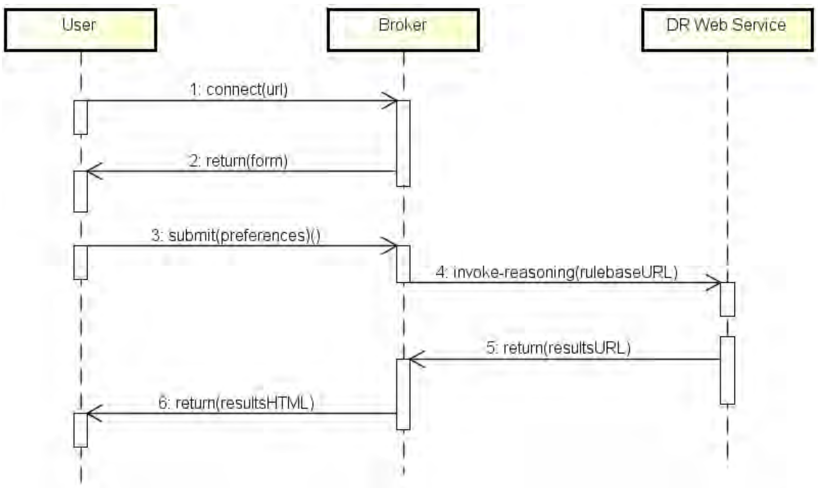


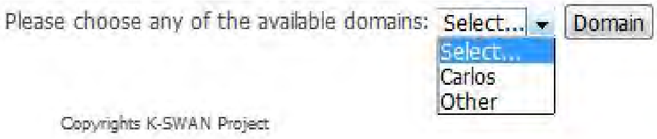
Fig. 7. Roles interaction in the brokering scenario.

Then the *Broker* invokes the *DR Web service* (interaction 4 in Fig.7) by passing it the *rulebase*. The *DR Web service* performs the reasoning task by using the EMERALD reasoning infrastructure, determines the output RDF file and returns its URL (*resultsURL*) to the *Broker* (interaction 5 in Fig.7). The *Broker* receives the URL, downloads the RDF file, converts it to meaningful HTML and returns the HTML (*resultsHTML*) to the *User* for visualization in the browser (interaction 6 in Fig.7).

4.3. Implementation

Following the above described design principles, a user-friendly Web site that hosts the *Broker* functionality was developed. Here, it is called *Broker Web site* since it is currently focused on this specific domain, however it can host any kind and any number of domains. The key issue behind this possibility is the fact that the site is hosted in an independent server. Having two separated servers, one for hosting the services such as the *DR Web Service* and one for hosting the registered parties such as the *Broker*, enables flexibility in the setting. In other words, it is important to preserve intact and unaffected the core framework which provides the services and at the same time provide an easy and efficient way for agent parties to communicate with human users. Adding new domains is in principle easily achievable, since it requires parsing the ontology of the input data and then using this ontology to author the appropriate rulebase for the end-user preferences. There exist graphical tools to achieve these, such as [14] and [15]; however, in this paper we do not cover this aspect.

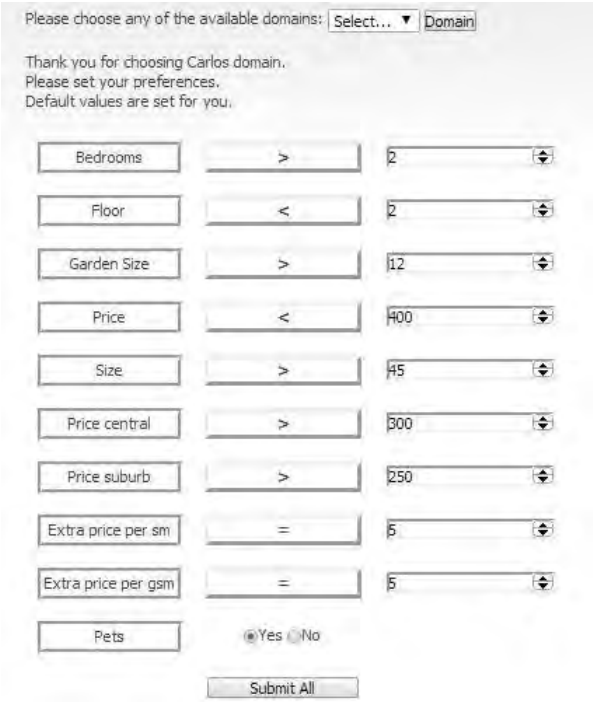
In this study we used a new separated server for the newly developed GUI. This interface is implemented mainly in PHP, a server side language that allows all calculations and functions to be performed on the server leaving for the end user only what it is designed for him/her. Hence, when a user visits the Web site he/she has to choose from a dropdown



Please choose any of the available domains:

Copyrights K-SWAN Project

Fig. 8. Choosing domain in the GUI.



Please choose any of the available domains:

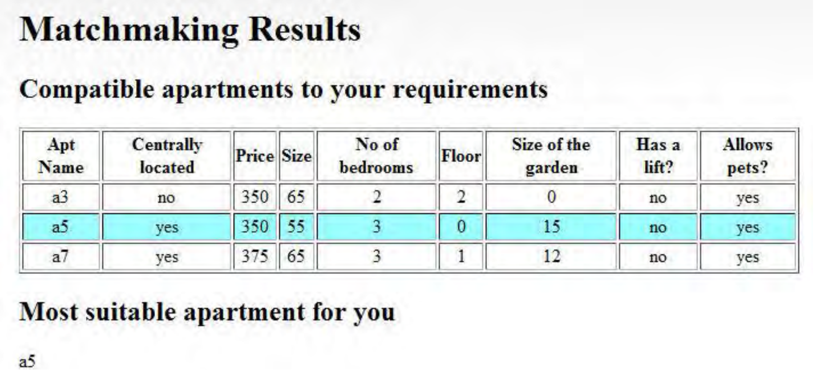
Thank you for choosing Carlos domain.
Please set your preferences.
Default values are set for you.

Bedrooms	>	2
Floor	<	2
Garden Size	>	12
Price	<	400
Size	>	45
Price central	>	300
Price suburb	>	250
Extra price per sm	=	5
Extra price per gsm	=	5
Pets	● Yes ○ No	

Fig. 9. Setting preferences.

menu the domain in which he/she is interested in; here it is Broker domain called Carlo in the implementation (Fig.8).

After the domain is chosen, the GUI is dynamically updated providing a list with all the available attributes for this domain (Fig.9). Hence, the user will be able to indicate his/her personal preferences, e.g. the number of bedrooms. For user convenience we provide some default values in the interface. When the user finishes, “Submit all” button is clicked, the data are collected and stored in the RDF data file that will be used in the *DR Web Service*. This file is stored in the server and each time users provide their preferred values appropriate PHP functions update the stored attribute values. More specifically, we have stored an RDF template data file, hence each time a new attribute values is set, the template is parsed and the value is stored under the appropriate instance.



Matchmaking Results

Compatible apartments to your requirements

Apt Name	Centrally located	Price	Size	No of bedrooms	Floor	Size of the garden	Has a lift?	Allows pets?
a3	no	350	65	2	2	0	no	yes
a5	yes	350	55	3	0	15	no	yes
a7	yes	375	65	3	1	12	no	yes

Most suitable apartment for you

a5

Fig. 10. GUI output.

Next, the updated RDF file, filled with the user’s requirements and preferences, is used by the Broker in its request to the DR Web Service. As soon as the *DR Reasoner* processes the request, it sends back a reply containing the results in RDF syntax, here the compatible apartments to user’s requirements, as well as the most suitable for him/her. The Broker on its side parses the RDF results and transforms then into HTML, in order to represent them in a user-friendly and easily understandable output to the user (Fig.10).

5. Proof-of-concept Scenario

In this section we present a proof-of-concept scenario of using the *Defeasible_Reasoning_Service* Web service for implementing the intelligent broker introduced in section 4. This scenario was adapted following [3], while it actually originates from [2].

We assume that a generic user with name Carlo has the following requirements for an apartment rental:

a) Apartment requirements
- Minimum size: $45m^2$.

- Minimum number of bedrooms: 2.
- Maximum floor: 2.
- Minimum garden size: 12.
- Pets allowed: yes.

b) Price Requirements

- Maximum budget: 400.
- Maximum price for a central apartment: 300.
- Maximum price for a suburb apartment: 250.
- Extra price per additional apartment m^2: 5.
- Extra price per additional garden m^2: 2.

c) Preferences

- The cheapest apartment is mostly preferred.
- The second option is the garden availability.
- The third option is availability of additional space.

Now, differently from [3], in order to fit this scenario into our prototype system, we extracted from this set of requirements the generic rulebase representing the requirements of a typical user. Moreover, we captured the specific values of the requirements' parameters for Carlo as a separate RDF file shown in Fig. 12.

The rulebase that captures the Broker's intelligence as a set of defeasible rules is expressed in RuleML[3]. A snapshot of the rulebase is exemplified in Fig .11. The set of apartments available for rent that are registered with the the *Broker* is captured as an RDF file[4]. An example is shown in , shown in Fig. 12. This file, as well as the RDF file with parameter values from Fig. 12 are included into the rulebase using the *rdf_import* attribute of the *RuleML* root element of the rulebase file expressed in RuleML [3].

The *Broker* invokes the reasoning task by sending a request message to the *Defeasible_Reasoning_Service* Web service. This message is packaged using SOAP as shown in Fig. 13. The request specifies the URL of the rulebase file that is located on the *Broker*'s Web server, using the *inputRulemlPath* element.

The *Defeasible_Reasoning_Service* Web service replies with a response packaged as a SOAP file that contains a link to the RDF file with the results, published on the *Defeasible_Reasoning_Service* Web server. A part of the results file in this case is shown in Fig. 14. It contains the list of "acceptable" apartments – *carlo:acceptable* element with the *truthStatus* set to *defeasibly_proven_positive*, as well as the apartment which is chosen by Carlo – *carlo:rent* element. You can easily note that among the acceptable apartments, Carlo chose to rent apartment *a5*.

[3] The file is available at `http://lpis.csd.auth.gr/systems/dr-device/carlo/carlo-rbase-flex-0.91.ruleml`.

[4] The file is available at `http://lpis.csd.auth.gr/systems/dr-device/carlo/carlo-flex_ex.rdf`.

```
<RuleML
  rdf_export="export.rdf"
  rdf_export_classes="acceptable rent"
  rdf_import="http://lpis.csd.auth.gr/systems/dr-device/carlo/carlo_ex.rdf"
  namespace declarations>
  ...
   <Assert>
      <Implies ruletype="defeasiblerule">
         <oid><Ind uri="r1">r1</Ind></oid>
         <head>
            <Atom><op><Rel>acceptable</Rel></op>
               <slot><Ind>apartment</Ind>
                  <Var>x</Var>
               </slot></Atom></head>
         <body>
            <Atom><op><Rel uri="carlo:apartment"/></op>
               <slot><Ind uri="carlo:name"/>
                  <Var>x</Var>
               </slot></Atom></body>
      </Implies>
      <Implies ruletype="defeasiblerule">
         <oid><Ind uri="r2">r2</Ind></oid>
         <head>
            <Neg><Atom><op><Rel>acceptable</Rel></op>
                  <slot><Ind>apartment</Ind>
                     <Var>x</Var>
                  </slot></Atom></Neg></head>
         <body><And>
               <Atom><op><Rel uri="carlo:apartment"/></op>
                  <slot><Ind uri="carlo:name"/>
                     <Var>x</Var></slot>
                  <slot><Ind uri="carlo:bedrooms"/>
                     <Var>y</Var></slot></Atom>
               <Atom><op><Rel uri="carlo:requirement"/></op>
                  <slot><Ind uri="carlo:min-bedrooms"/>
                     <Var>mb</Var></slot></Atom>
               <Test><Expr>
                     <Fun in="yes">&lt;</Fun>
                     <Var>y</Var>
                     <Var>mb</Var></Expr></Test></And></body>
         <superior><Ind uri="r1"/></superior>
      </Implies>
      ...
      <Implies ruletype="defeasiblerule">
         <oid><Ind uri="r11">r11</Ind></oid>
         <head>
            <Atom><op><Rel>rent</Rel></op>
               <slot><Ind>apartment</Ind>
                  <Var>x</Var></slot></Atom></head>
         <body> <And>
               <Atom><op><Rel>cheapest</Rel></op>
                  <slot><Ind>apartment</Ind>
                     <Var>x</Var></slot></Atom>
               <Atom><op><Rel>largestGarden</Rel></op>
                  <slot><Ind>apartment</Ind>
                     <Var>x</Var></slot></Atom></And></body>
         <superior><Ind uri="r10"/></superior>
      </Implies>
      ...
   </Assert>
</RuleML>
```

Fig. 11. Rule base fragment.

```
<rdf:RDF ... declarations of namespaces>
...
<carlo:apartment rdf:about="&carlo_ex;a5">
   <carlo:bedrooms
     rdf:datatype="&xsd;integer">3</carlo:bedrooms>
   <carlo:central>yes</carlo:central>
   <carlo:floor
     rdf:datatype="&xsd;integer">0</carlo:floor>
   <carlo:gardenSize
     rdf:datatype="&xsd;integer">15</carlo:gardenSize>
   <carlo:lift>no</carlo:lift>
   <carlo:name>a5</carlo:name>
   <carlo:pets>yes</carlo:pets>
   <carlo:price
     rdf:datatype="&xsd;integer">350</carlo:price>
   <carlo:size
     rdf:datatype="&xsd;integer">55</carlo:size>
   <rdfs:label>a5</rdfs:label>
</carlo:apartment>
...
<carlo:requirement rdf:about="&carlo_ex;req1">
   <carlo:min-bedrooms
     rdf:datatype="&xsd;integer">2</carlo:min-bedrooms>
   <carlo:max-floor
     rdf:datatype="&xsd;integer">2</carlo:max-floor>
   <carlo:min-gardenSize
     rdf:datatype="&xsd;integer">12</carlo:min-gardenSize>
   <carlo:pets-req>yes</carlo:pets-req>
   <carlo:max-price
     rdf:datatype="&xsd;integer">400</carlo:max-price>
   <carlo:min-size
     rdf:datatype="&xsd;integer">45</carlo:min-size>
   <carlo:min-price-central
     rdf:datatype="&xsd;integer">300</carlo:min-price-central>
   <carlo:min-price-suburb
     rdf:datatype="&xsd;integer">250</carlo:min-price-suburb>
   <carlo:extra-price-per-sm
     rdf:datatype="&xsd;integer">5</carlo:extra-price-per-sm>
   <carlo:extra-price-per-gsm
     rdf:datatype="&xsd;integer">5</carlo:extra-price-per-gsm>
</carlo:requirement>
</rdf:RDF>
```

Fig. 12. RDF document for available apartments and parameter values.

```
<soapenv:Envelope ... declarations of namespaces>
  <soapenv:Header/>
  <soapenv:Body>
    <urn:DRReasoning
      soapenv:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
      <inputRulemlPath xsi:type="xsd:string">
       http://lpis.csd.auth.gr/systems/dr-device/carlo/carlo-rbase-flex-0.91.ruleml
      </inputRulemlPath>
    </urn:DRReasoning>
  </soapenv:Body>
</soapenv:Envelope>
```

Fig. 13. SOAP message used by the *Broker* to request a reasoning task to the *DR Web service*.

```
<rdf:RDF ... declarations of namespaces>
   ...
   <export:acceptable rdf:about="&export;acceptable4">
      <export:apartment>a4</export:apartment>
      <defeasible:truthStatus>defeasibly_proven_negative</defeasible:truthStatus>
   </export:acceptable>
   <export:acceptable rdf:about="&export;acceptable2">
      <export:apartment>a5</export:apartment>
      <defeasible:truthStatus>defeasibly_proven_positive</defeasible:truthStatus>
   </export:acceptable>
   ...
   <export:rent rdf:about="&export;rent1">
      <export:apartment>a5</export:apartment>
      <defeasible:truthStatus>defeasibly_proven_positive</defeasible:truthStatus>
   </export:rent>
</rdf:RDF>
```

Fig. 14. RDF file with the reasoning results.

6. Conclusions

In this paper we proposed an approach for reusing agent-based reasoning capabilities by making them available for invocation as Web services. Our approach is supported by a prototype system that provides an online brokering function in the domain of apartments rental. Firstly, we proposed an extension of the EMERALD framework for agent based reasoning services with a Web service interface. Then we considered an intelligent brokering proof-of-concept scenario that involves the defeasible reasoner available in the EMERALD framework.

Our approach is general enough so it can be applied to other types of reasoners, as well as to other problem domains. As future work we plan to (i) extend the system by providing other types of reasoning services over the Web such as the deductive reasoner of EMERALD [17]; (ii) experiment with more complex scenarios that require multiple and possibly different reasoning tasks; and (iii) extend the Web services to serve multiple operations regarding the reasoning tasks, such as proof explanation ([19]); (iv) enhance the description of reasoning Web services with metadata (including for example attributes like trust, performance, quality, reasoning type, and representation formalism) for allowing clients to search and select the most appropriate reasoner for their task in hand.

Finally, a longer term goal is to provide an open, flexible and customizable Web-agent platform for hosting various agent communities and activities over the Web, such as negotiation and auction platforms for e-business applications. They can benefit from the reasoning services provided by our Web-agent-based extension of EMERALD, for example by declarative representation of private agent strategies, as well as of public negotiation mechanisms for agent-based negotiation activities [28,9,21].

Acknowledgments. This work was supported by the *K-SWAN: An Interoperable Knowledge-based Framework for Negotiating Semantic Web Agents* Greek-Romanian bilateral research project carried out between 2012-2013 and partly funded by Romanian UEFISCDI and by the Greek R&D General Secretariat.

References

1. Alonso, G., Casati, F., Kuno, H., Machiraju, V.: Web Services: Concepts, Architectures and Applications. Springer (2004)
2. Antoniou, G., van Harmelen, F.: A Semantic Web Primer – Second Edition. MIT Press (2008)
3. Bassiliades, N., Antoniou, G., Vlahavas, I.P.: A defeasible logic reasoner for the semantic web. International Journal on Semantic Web and Information Systems 3(1), 1–41 (2006)
4. Bellifemine, F.L., Caire, G., Greenwood, D.: Developing Multi-Agent Systems with JADE. John Wiley & Sons Ltd (2007)
5. Boley, H., Paschke, A.: Rule responder agents framework and instantiations. In: Elçi, A., Koné, M.T., Orgun, M.A. (eds.) Semantic Agent Systems, Studies in Computational Intelligence, vol. 344, pp. 3–23. Springer (2011)
6. Bordini, R.H., Hübner, J.F., Wooldridge, M.: Programming Multi-Agent Systems in AgentSpeak using Jason. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England (2007)
7. Bădică, C., Budimac, Z., Burkhard, H.D., Ivanović, M.: Software agents: Languages, tools, platforms. Computer Science and Information Systems 8(2), 255–298 (2011)
8. Dastani, M., Birna Riemsdijk, M., Meyer, J.J.C.: Programming multi-agent systems in 3apl. In: Bordini, R., Dastani, M., Dix, J., Fallah Seghrouchni, A. (eds.) Multi-agent programming: Languages platforms and applications, Multiagent Systems, Artificial Societies, and Simulated Organizations, vol. 15, pp. 39–67. Springer, US (2005)
9. Dobriceanu, A., Biscu, L., Bădică, A., Bădică, C.: The design and implementation of an agent-based auction service. International Journal of Agent-Oriented Software Engineering 3(2/3), 116–134 (2009)
10. FIPA: The foundation for intelligent physical agents (fipa): Fipa communicative act library specification (2002), retrieved November 29, 2012 from `http://www.fipa.org/specs/fipa00037/`
11. Gelder, A.V., Ross, K.A., Schlipf, J.S.: The well-founded semantics for general logic programs. J. ACM 38(3), 620–650 (1991)
12. Giarratano, J.C., Riley, G.D.: Expert Systems: Principles and Programming, Fourth Edition. Course Technology (2004)
13. JADE: Jade web services integration gateway (wsig) guide (2012), retrieved March 23, 2013 from `http://jade.tilab.com/doc/tutorials/WSIG_Guide.pdf`
14. Kontopoulos, E., Bassiliades, N., Antoniou, G., Seridou, A.: Visual modeling of defeasible logic rules with dr-vismo. International Journal of Artificial Intelligence Tools (IJAIT) 17(5), 903–924 (2008)
15. Kontopoulos, E., Zetta, T., Bassiliades, N.: Semantically-enhanced authoring of defeasible logic rule bases in the semantic web. In: Proc. 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS'12). Craiova, Romania, June 13-15, 2012. pp. 489–492, Article 56. ACM (2012)
16. Kravari, K., Kontopoulos, E., Bassiliades, N.: A trusted defeasible reasoning service for brokering agents in the semantic web. In: Papadopoulos, G.A., Bădică, C. (eds.) Intelligent Distributed Computing III, Proc. 3rd International Symposium on Intelligent Distributed Computing, IDC 2009. Studies in Computational Intelligence, vol. 237, pp. 243–248. Springer (2009)
17. Kravari, K., Kontopoulos, E., Bassiliades, N.: Emerald: A multi-agent system for knowledge-based reasoning interoperability in the semantic web. In: Konstantopoulos, S., Perantonis, S.J., Karkaletsis, V., Spyropoulos, C.D., Vouros, G.A. (eds.) Artificial Intelligence: Theories, Models and Applications, Proc. 6th Hellenic Conference on AI, SETN 2010. Lecture Notes in Computer Science, vol. 6040, pp. 173–182. Springer (2010)
18. Kravari, K., Kontopoulos, E., Bassiliades, N.: Trusted reasoning services for semantic web agents. Informatica (Slovenia) 34(4), 429–440 (2010)

19. Kravari, K., Papatheodorou, K., Antoniou, G., Bassiliades, N.: Extending a multi-agent reasoning interoperability framework with services for the semantic web logic and proof layers. In: Bassiliades, N., Governatori, G., Paschke, A. (eds.) Rule-Based Reasoning, Programming, and Applications, Lecture Notes in Computer Science, vol. 6826, pp. 29–43. Springer Berlin Heidelberg (2011)
20. Kwon, O., Im, G., Lee, K.: An agent-based web service approach for supply chain collaboration. Scientia Iranica 18(6), 1545–1552 (2011)
21. Muscar, A., Bădică, C.: Towards a declarative framework for the specification of agent-driven auctions. Engineering Intelligent Systems 21(2-3), 642–651 (2013)
22. nerdErg: The one ring project, nerderg pty ltd 2012 (2013), retrieved June 10, 2013 from `http://nerderg.com/One+Ring`
23. Nute, D.: Defeasible reasoning. In: Proc. 20th International Conference on Systems Science. pp. 470–477. IEEE Press (1987)
24. Purvis, M., Cranefield, S., Nowostawski, M., Carter, D.: Opal: A multi-level infrastructure for agent-oriented software development. In: Information Science Discussion Paper Series No. 2002/01. University of Otago, New Zealand (2002)
25. Ryzko, D., Radziszewska, W.: Integration between web services and multi-agent systems with applications for multi-commodity markets. In: Kaleta, M., Traczyk, T. (eds.) Modeling Multi-commodity Trade: Information Exchange Methods, Advances in Intelligent and Soft Computing, vol. 121, pp. 65–77. Springer Berlin Heidelberg (2012)
26. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review 24(1), 33–60 (2005)
27. Skillen, K.L., Chen, L., Nugent, C.D., Donnelly, M.P., Burns, W., Solheim, I.: Ontological user modelling and semantic rule-based reasoning for personalisation of help-on-demand services in pervasive environments. Future Generation Computer Systems 34, 97–109 (2014)
28. Skylogiannis, T., Antoniou, G., Bassiliades, N., Governatori, G., Bikakis, A.: Dr-negotiate - a system for automated agent negotiation with defeasible logic-based strategies. Data & Knowledge Engineering 63(2), 362–380 (2007)
29. Su, C.J., Peng, C.W.: Multi-agent ontology-based web 2.0 platform for medical rehabilitation. Expert Systems with Applications 39(12), 10311–10323 (2012)
30. Walton, C.D.: Agency and the Semantic Web. Oxford University Press (2007)
31. Wang, G., Wong, T., Wang, X.: An ontology based approach to organize multi-agent assisted supply chain negotiations. Computers & Industrial Engineering 65(1), 2–15 (2013)
32. Wang, M., Purvis, M., Nowostawski, M.: An internal agent architecture incorporating standard reasoning components and standards-based agent communication. In: IEEE/WIC/ACM international Conference on intelligent Agent Technology (IAT'05), Washington, DC. pp. 58–64 (2005)
33. Wooldridge, M.: An Introduction to MultiAgent Systems – Second Edition. John Wiley & Sons (2009)

Costin Bădică is Professor at the Department of Computers and Information Technology, University of Craiova from 2006. In 2001 and 2002 he was Postdoctoral Researcher at the Department of Computer Science, King's College London, UK. His research interests are located at the intersection of artificial intelligence, distributed systems and software engineering. Costin Bădică was involved in many national and international research projects. He is coordinating the Intelligent Distributed Systems research group (`http://ids.software.ucv.ro`). Costin Bădică co-authored more than 120 research publications as: journal papers; conference proceedings papers; co-edited books;

book chapters; editorials of journal special issues; books. His published research has received over 500 citations (h-index 13). Costin Bădică is member of the editorial board of several international journals and served as member of the program or organizing committee of many international conferences. He has co-initiated the Intelligent Distributed Computing – IDC series of international conferences that is being held annually. He has been Program co-Chair of IDC-2007 and WIMS-2012. He has been Conference co-Chair of ICCCI-2013.

Nick Bassiliades (`http://tinyurl.com/nbassili`) received his MSc in Applied Artificial Intelligence from the Computing Science Department of Aberdeen University, in 1992, and his PhD degree in parallel knowledge base systems from the Department of Informatics at the Aristotle University of Thessaloniki, Greece, in 1998, where he is currently an Associate Professor. He currently also serves as an academic coordinator at the School of Science & Technology of the International Hellenic University. His research interests include knowledge-based and rule systems, multiagent systems, ontologies, linked data and the Semantic Web. He has published more than 150 papers in journals, conferences, and books, and has coauthored two books. His published research has received over 1000 citations (h-index 17). He was on the Program Committee of more than 70 and on the Organizational Committee of 6 conferences / workshops. He has been the Program co-Chair of RuleML-2008, RuleML-2011@IJCAI and recently of WIMS-2014. He has been involved in 30 R&D projects leading 9 of them. He has been the general secretary of the Board of the Greek Artificial Intelligence Society; he is a director of RuleML, Inc., and also a member of the Greek Computer Society, the IEEE, and the ACM.

Sorin Ilie is research assistant at the Department of Computers and Information Technology, University of Craiova from 2009. He defended his PhD thesis in December 2012 in the area of Distributed Ant Colony Optimization. His research interests are Multi-Agent Systems, Distributed Systems and Software Engineering. Sorin Ilie was involved in four national and international research projects. He is a member of the Intelligent Distributed Systems research group (`http://ids.software.ucv.ro`). Sorin Ilie co-authored more than 30 research publications as: journal papers, conference proceedings papers and book chapters. Sorin Ilie served as member of the program or organizing committee of two international conferences.

Kalliopi Kravari is currently a PhD student at the Department of Informatics of Aristotle University of Thessaloniki (AUTH), Greece, and she is member of Intelligent Systems and Knowledge Processing (ISKP) Group. She holds a BSc in Informatics and an MSc in Information Systems from AUTH. Her research interests are mainly focused on Semantic Web and Intelligent Agents, including among others Trust Management, Knowledge Representation and Reasoning, Logic and Rule-based Programming, Ontologies and Rules. She has already published 3 journal papers, 1 book chapter and 12 conference papers at the above issues. She has received a best paper award at RuleML-2010 and one of her papers was selected among the best 2 papers of RuleML-2011@IJCAI to be presented at IJCAI-2011.

Received: March 01, 2014; Accepted: June 10, 2014.

Computer Science and Information Systems 11(2):723–743 DOI: 10.2298/CSIS121212041

A Novel Educational Game for teaching Emotion Identification Skills to Preschoolers with Autism Diagnosis

Eirini Christinaki[1], Nikolas Vidakis[1], and Georgios Triantafyllidis[2]

[1] Technological Educational Institute of Crete, Dep. Informatics Engineering, Heraklion, Crete
echrist@ics.forth.gr, nv@ie.teicrete.gr
[2]Aalborg University, Medialogy Section, AD:MT, Copenhagen, A.C.Meyers Vænge 15
gt@create.aau.dk

Abstract. Emotion recognition is essential in human communication and social interaction. Children with autism have been reported to exhibit deficits in understanding and expressing emotions. Those deficits seem to be rather permanent so intervention tools for improving those impairments are desirable. Educational interventions for teaching emotion recognition should occur as early as possible. It is argued that Serious Games can be very effective in the areas of therapy and education for children with autism. However, those computer interventions require considerable skills for interaction. Before the age of 6, most children with autism do not have such basic motor skills in order to manipulate a mouse or a keyboard. Our approach takes account of the specific characteristics of preschoolers with autism and their physical inabilities. By creating an educational computer game, which provides physical interaction with natural user interface (NUI), we aim to support early intervention and to enhance emotion recognition skills.

Keywords: autism, facial emotion recognition, gesture-based interaction, Kinect, natural user interface.

1. Introduction

Autism, according to the Diagnostic and Statistical Manual (DSM-IV-TR) of Mental Disorders [1], is a Pervasive Developmental Disorder (PDD) characterized by impairments in social interaction, in communication and by restricted, repetitive and stereotyped patterns of behavior, interests and activities. Children with the same diagnosis of autism may exhibit different symptoms and may demonstrate markedly different behaviors and skills. The spectrum of symptoms can range from mild to severe and efficient diagnosis is difficult. Autism is also described as an Autistic Spectrum Disorder (ASD) due to the variability with which the disorder is manifested. In this article autism and ASD will be referred interchangeably.

Social interaction impairments, a core feature of ASD, involve difficulties in understanding and expressing emotions [2]. Social interaction is the mutual influence between individuals, during which people exchange verbal and nonverbal messages, in order to provide and receive information for themselves and for the others. It is a process of production and exchange of signs that usually affects the behavior of people involved in it. Marked impairments in the use of nonverbal behaviors such as facial expressions, according to DSM-IV-TR, can be responsible for qualitative impaired social interaction. Nonverbal communication is the expression of emotions, moods, attitudes and the general inner world expressed through the body. It constitutes the largest part of a communication process and facilitates clarification between communicators to define the possible role that each of them has at each moment during this transaction. Source of nonverbal signals is the human body and particularly the face, the eyes, the postures, the body movements and the gestures.

Facial expressions give important clues about emotions and provide a key mechanism for understanding, identifying and conveying them. Facial expressions underline the feelings of emotions. Children with autism often fail to recognize the qualitative differences and associations between various expressions of emotions [3]. Due to limited social and emotional understanding they do not know how to adequate interact with other people which sometimes leads to inappropriate behaviors. Studies have reported that young children with autism experience difficulties recognizing expressions and adults with autism are not as good as typically developing adults at emotion recognition [4]. It has also been recorded that individuals with autism are significantly impaired and exhibit heterogeneity in their emotion processing [5], have difficulties in recognizing emotions from the upper part of the face [6] and exhibit a deficit in overall emotion recognition from facial expressions [7]. Toddlers with autism focus their attention on a single facial feature and treat people as objects [8]. Visual scanning of faces in autism revealed an increased visual fixation time on mouth region and a significant less time on eyes region that was associated with impairment in daily social interaction [9]. There is also evidence that individuals with autism and typically developing individuals decode facial expressions through different mechanisms [10]. A recent study was conducted to evaluate verbal and perceptual skills implicated in the recognition of facial emotions. Individuals with ASD, 5-20 years old, were tested for their perception and identification of facial emotions. Results indicated unimpaired ability to label basic facial emotions and impaired ability to generalize them [11]. Other studies support that individuals with ASDs do not show impaired emotion recognition [12]and these results indicate the wide variation in the manner that individuals are affected.

2. Related Work

Treatment approaches aim to improve social interaction, conquest communication and control inappropriate behavior. Children with autism are more likely to initiate positive interaction after treatment [13]. Education is considered as the main solution for the socio-emotional deficits and training is claimed to improve face processing abilities and strategies in autism [14]. A variety of educational interventions have been proposed for children with autism and many proponents have supported developmental improvement and other benefits [15].

2.1. Traditional Educational Interventions

Traditional therapy techniques and tools for teaching children with autism about emotions are Social Stories™, Developmental Individualized and Relationship based (DIR) / Floortime, Social Communication Emotional Regulation, Transactional Support (SCERTS), Applied Behavioral Analysis (ABA), Early Intensive Behavioral Intervention (EIBI), Picture Exchange Communication System (PECS) and Treatment and Education of Autistic and Related Communication Handicapped CHildren (TEACCH).

Social Stories™, developed by Carol Gray [16] are an approach for teaching individuals with ASD social behaviors. They are short stories that describe and give information about social situations or interaction which children may find difficult or confusing. Their main goal is to provide accurate social information for improving understanding of events that may lead to more effective responses and not for effecting changes. DIR / Floortime parent intervention [17] can help children with autism to connect emotionally and to build social and intellectual skills. This approach follows the child's emotional interests and provides one to one, intensive, play - based intervention. SCERTS is a framework that incorporates practices from other approaches and evidence-based practices [18]. It is a supporting multidisciplinary model for children with autism to enhance socio-emotional and communication abilities. The ABA therapy is often used as a treatment for children with autism. It is a widespread and recognized method for increasing learning, communication and appropriate social behavior [19]. Positive reinforcement, reward for desired behaviors and ignorance for inappropriate behaviors are the main characteristics. According to the ABA intervention method, teaching children with autism about emotions can be achieved by providing examples of appropriate emotional behavior and then rewarding when a child gives the correct emotional response. Young children with ASDs usually receive a home-based program, the EIBI, which is based on the principles and technologies of ABA [20]. Early intervention appears to lessen the effects of autism and children with ASD appear most able to benefit when intervention begins very early (2-4 years old) [21], [22]. Toddlers, who attended an early intervention program, appear to benefit from this, since 31% of those children were functioning in the typical developing range after the intervention [23].

Children with ASD are reported to have strong visual processing capability and a good performance on difficult visual search tasks [24], [25]. Augmentative and Alternative Communication (AAC), such as PECS, is frequently used to increase functional communication in children with ASD. PECS is an effective visually-based system for communication via icons [26]. Visual supports are also suggested for social skill development enhancement in young children with ASD [27]. Picture cards with cartoon faces or illustrations of people faces or photos showing different facial emotional expressions reinforce and support learning. Programs that emphasize visual support strategies are commonly used in education classrooms. A well-known and widely applied model, the TEACCH, is a teaching strategy which emphasizes in structured and predictable learning environment. It utilizes visual cues to increase independence and to teach new skills, such as facial emotions, to children with autism. It involves daily schedules, visual materials, individualized treatment and parental support [28].

2.2. Computer Interventions

Computers are rich, stable, predictable, consistent teaching tools and they provide a proper and more motivational learning environment for individuals with autism [29]. They are great educational tools since children with autism often experience discomfort with unpredictable social environment and they prefer a controlled learning environment [30]. Autistic individuals can enjoy learning and improve their skills with computer-based intervention. Nowadays computers are the most adaptable assistive technology devices available for children with autism and various computer games have been developed to help them to manipulate their impairments. Some of those games focus on their social interaction training and especially learning about emotions [31]. Although effects on social and emotional skills are mixed, computer intervention is a promising practice. In a recent review, Wainer and Ingersoll [30] examined a number of innovation computer programs as educational interventions for people with ASD. They focused on studies describing programs to teach language, emotions or social skills. Their analysis showed that those tools are promising strategies for delivering direct intervention to individuals with ASD.

Currently, ample researches in Serious Games for children with autism have been done. Serious Games are designed for a primary purpose other than pure entertainment, fun or enjoyment [32] and in relations to autism they cover matters related to education, therapy for communication, psychomotor treatment and social behavior enhancement [33]. Serious Games for ASD education are designed to help teachers or student during the teaching and/or learning process. Studies show that Serious Games are very effective in the areas of therapy and education for such children.

The Emotion Trainer [34] is a multimedia computer program for improving school age students' ability to recognize and predict emotions in others. It has positive effects on users' understanding of emotion, particularly with repeated use. Mind Reading [35] is an interactive systematic emotion guide for teaching individuals with Asperger syndrome. Experiments show that Mind Reading is effective in teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions. Another computer training program for teenagers with autism is the "What to choose" game [36]. It is software with human and cartoon facial expressions, 3D images, text and audio for training individuals with autism in understanding dialogues that contain pragmatic subtleties. Children with autism have failed to take into account the causal link between facial expressions and the outcome of the dialogue. cMotion [37] is a computer game that uses virtual characters to teach emotion recognition and the programming concept to children with autism. The game is designed to teach the users how to recognize facial expressions and manipulate an interactive virtual character by using a visual drag-and-drop programming interface. A computer-based intervention for face training is the Let's Face It! (LFI!) program [38]. This program is comprised of seven interactive computer games that target the specific face impairments associated with autism. Those games are organized into a theoretical hierarchy of face processing domains. The Let's Face It! Program shows promise as an effective intervention tool and treatment alternative. Emotion Mirror is a project that integrates Computer Expression Recognition Toolbox (CERT) with the program Let's Face it! [39]. It is a computer assisted intervention system to enhance facial expression perception and production in children with ASD. This game could be beneficial for children who already have learned labeling and understanding of emotions. The LIFEisGAME is a facial emotion recognition learning system based on the interaction between humans and 3D avatars [40]. It is a computer-based approach that uses real time facial synthesis of 3D characters in order to teach autistic people to recognize emotions from facial expressions.

Bernardini et al. [41] proposed ECHOES a Serious Game for children with ASD to practice social communication skills. This project presents an intelligent virtual character that acts both as a peer and as a tutor on a number of different learning activities. These activities can be selected manually by a human operator (practitioner, parent or other carer) through a graphical interface. The experimental results showed encouraging tendencies by relating the effectiveness of the children's interaction with the virtual character acting as a social partner to them. Porayska-Pomsta et al. [42] suggest SHARE-IT, an intelligent and authorable environment to assist children with ASD in gaining social interaction skills. Their tool contains an intelligent agent and a play environment that allows teachers and parents to become co-creators and tailor the game according to the needs of the individual children in their care. Although the design and creation of personalized games is crucial for children with ASD, as reported by the authors, limitations in the agent's intelligence (agent inability to deal with inappropriate or unexpected behavior from the user) contradicts the structured, stable and predictable learning environment that is also crucial.

Table 1. Computer and other interventions

Intervention	Refers to	Age	Test/Evaluation	Mouse/Keyboard	2D/3D	Virtual Characters	Images/Photos	Text/Audio/Video
Emotion Trainer [34]	*Autism / Asperger Syndrome*	*school age*	√	√	*2D*	-	√	T
Mind Reading [35]	*Asperger Syndrome / High Functioning Autism*	*from 8 years old*	√	√	*2D*	-	√	T, A, V
What to choose [36]	*High Functioning Autism*	*teenagers*	√	√	*3D*	√	√	T, A
cMotion [37]	*Autism*	-	-	√	*3D*	-	-	-
Let's Face it! [38]	*Autistic Disorder / Asperger Syndrome / PDD-NOS*	*school age*	√	√	*2D*	-	√	-
Emotion Mirror [39]	*ASD*	-	-	√	-	-	√	V
LIFEisGAME [40]	*ASD*	*from 6 years old*	√	√	*3D*	√	-	A
ECHOES [41]	*Autism Spectrum Conditions*	*young children*	√	*other*	*2D/3D*	√	-	T, A
SHARE-IT [42]	*Autism*	-	-	*other*	*3D*	√	-	-
The Transporters [46]	*Autism / Asperger Syndrome*	*4-8 years old*	√	*other*	*3D*	√	-	A, V
Pix Talk [50]	*Autism*	*2-8 years old*	-	*other*	*2D*	-	√	T

2.3. Other Interventions

Nowadays, there is a research and development trend in using motion-based touchless games for children with disabilities. An ongoing EU funded project called M4ALL[1] that aims to develop a "professional" version of these games and a company called Kinems[2]

[1] http://www.m4allproject.eu
[2] http://www.kinems.com

that develops learning games with natural interaction for children with multiple learning disabilities, are examples of this tendency.

Current studies have also gone considerably beyond the simple use of computers. Diverse technology-based interventions have been employed for empowerment and skill acquisition. Recent reviews [43], [44] have shown that there is a growing number of interventions and report a variety of technologies such as multimedia presentations and virtual environments. Other kinds of interventions include interactive DVDs, wearable devices and mobile applications.

MIT Media Lab proposes a wearable device [45] that perceives and reports on social-emotional information in real-time human interaction. The system records and analyzes the facial expressions and the head movements of the person with whom the wearer is interacting. This wearable platform is suggested as an exploratory and monitoring tool to assist individuals with ASD in perceiving communication in a natural environment.

The Transporters [46], a DVD for teaching emotion recognition, is an animation series for children with ASD (preschoolers or with significant learning difficulties). It involves toy vehicles and real life faces of actors that show emotional expressions in social context. The aim of this program was to teach not just some basic emotions but also some more complex ones. Evaluations of this project showed improvements in emotion comprehension and recognition skills [47], [48] but limited efficiency in teaching basic emotion recognition skills to young children with autism with lower range of cognitive ability [49].

PixTalk [50] is a software application for Windows Mobile Smart-phones which can be used as part of ongoing therapy. Teachers and caregivers can access a web site and select from an online library the images to be downloaded on to the Smart-phone. Children also can browse and select images to express their intentions, desires, and emotions using PixTalk.

3. Our Approach

Difficulties in identifying and describing feelings are assumed to be an integral part of autism. Educational interventions can be used as a tool to help individuals with autism to cope with those deficits. As the number of children diagnosed with autism increased [51], [52], new methods for educating this population become necessary. In several studies, computer-based programs have been widely used with success to teach people with ASD to recognize emotions. However, those computer interventions require considerable skills for interaction as the users have to control a mouse or a keyboard. Such abilities are beyond very young children with autism as they have major restriction in their efficiency to interact with computers. Our aim is to design and develop a gesture controlled Serious Game, as an early intervention, for preschoolers (2-6 years old) with autism to teach them basic facial emotions. Our approach takes into account the specific characteristics of preschoolers with autism and their physical abilities. By creating an educational computer game which proposes physical interaction, we hope to make early intervention more appealing and to foster learning of emotions. This software intends to support individuals with autism, their families and their trainers and also to assist the preschool teaching and learning process.

Educating children with ASD and particularly teaching them to recognize and identify emotions from facial expressions is a significant challenge and a complex task. The use of educational software for teaching social and emotional skills could help students with ASD to improve those abilities [53]. Learning tasks developed in digital environments using information technology can motivate the desire to learn in ASD students. As we have already mentioned in the Related Work, computer interventions appear to be particularly appropriate for people with ASD for several reasons. Traditional interventions previously mentioned can be transferred into a computerized version. In our game, we will provide one to one intensive play-based intervention, visual support, positive reinforcement and reward, structural and predictable learning environment and we will incorporate practices from all traditional approaches already mentioned in order to make intervention more efficient.

Many approaches to technology-enhanced intervention rely on educational methods shown to result in good outcomes. Computer interventions such as Serious Games could be a good approach as they also have positive effects on children with autism. Serious Games have the potential to support the process of learning and to motivate individuals to learn more. They usually run on a personal computer and they provide a different style of learning. We will attempt to apply a Serious Game approach to teach children with autism facial emotion recognition.

Previews studies that used Serious Games with children with autism have taken for granted the required skills and in particular the ability to use the mouse or the keyboard. People with ASD demonstrate delay in fine motor skills which causes difficulties in grasping and manipulating objects, such as a mouse [54]. The difficulties they encounter include moving the computer mouse to designated area, corresponding to the location of the cursor to the mouse movement, pressing down and lifting up fingers for the clicking motion, and clicking the correct button on the mouse to make selection [55]. Those difficulties are also found in typically developing individuals. Very young children (autistic and non-autistic) are unable to functionally operate the computer mouse. Experiments showed that four and five year old non-autistic children make more and less accurate mouse sub-movements on approach to targets than young adults [56]. With the increase of age, children in this study demonstrated significantly faster and more accurate use of the mouse which gives evidence of a strong link between speed and accuracy skills and developmental maturity. The same problems occur when players are required to use other game controllers such as joysticks and gamepads. Various research studies have been conducted to evaluate game controllers with the use of ISO 9241-9 [57] based on Fitts' law tasks which is an effective and widely used predictor of performance and comfort. For point and select task, evaluation showed that participants prefer using a mouse and results indicated that remote pointing devices perform poorly in terms of throughput, speed and error rate when compared to a mouse [58]. Touchpads are consider to be a good solution for young children but evaluations have also shown inferior performance, increased movement time, lower throughput and higher error rates when compared with a joystick [59], a trackball [60] and a mouse [61]. Thus, we propose an interaction based on gestures and a controller-free interface as a different interaction technique. Our game will allow user to employ gestures for navigation and interaction.

4. Design and Implementation

In this paper we present an educational computer-based single player game specially designed for Greek preschoolers with autism. The primary objective of this game is to assist young children with ASD to identify different emotions from facial expressions. The main goal of our game is to provide recognition and understanding of facial emotions through early intervention.

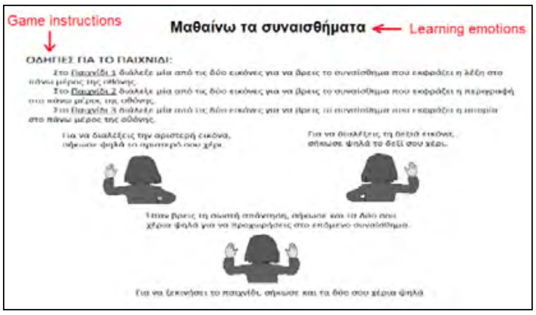
As a Serious Game, our development was based on the following principles: (a) it should have an impact on the player in a real life context [62], (b) it is explicitly designed to reach a specific purpose beyond the game itself and (c) it aims to teach preschoolers with autism, facial emotion recognition in order to enhance their social interaction.

Taking into account specific characteristics of autism, we designed the game in order to meet the needs of students with ASD. A recent study conducted to analyze user needs for Serious Games for teaching children with ASD emotions, revealed the characteristics of the children's game play behaviors [63]. The observation showed repetition, matching instead of learning the features, lack of holistic face processing and deliberately incorrect selection. According to those findings, our game was intentionally designed to avoid such behaviors. For repetition, we decided not to give them the opportunity to choose the same emotion again and again. For matching, instead of learning the features, we describe the features of facial expressions in a separate level. For lack of holistic face processing, we describe all the face features that reveal the emotion and ask them to look at each feature separately. For deliberate incorrect selection, nothing special happens when they give a wrong answer.

Our design also incorporates a theory-driven game design framework supported by learning and developmental theories [64]. The framework is based on the integration of Kolb's experiential learning model and Piaget's cognitive model. From this systematic approach were extracted six essential elements for designing games to teach children with ASD emotions. Those elements are: matching, recognition, observation, understanding, generalizing and mimicking. We took those elements into consideration during the design process.

4.1. Game Environment

The game environment is simple and less detailed in order to avoid children's distraction. Individuals with autism are reported to have enhanced perception of details [65] which may causes distraction. For those reasons we select black context presented on a white background and grayscale stimuli. Black and white contrast may also help to increase and retain child's attention and keep them focused on the screen.



Μαθαίνω τα συναισθήματα

ΟΔΗΓΙΕΣ ΓΙΑ ΤΟ ΠΑΙΧΝΙΔΙ:

Στο Παιχνίδι 1 διάλεξε μία από τις δύο εικόνες για να βρεις το συναίσθημα που εκφράζει η λέξη στο πάνω μέρος της οθόνης.

Στο Παιχνίδι 2 διάλεξε μία από τις δύο εικόνες για να βρεις το συναίσθημα που εκφράζει η περιγραφή στο πάνω μέρος της οθόνης.

Στο Παιχνίδι 3 διάλεξε μία από τις δύο εικόνες για να βρεις το συναίσθημα που εκφράζει η ιστορία στο πάνω μέρος της οθόνης.

Για να διαλέξεις την αριστερή εικόνα, σήκωσε ψηλά το αριστερό σου χέρι.

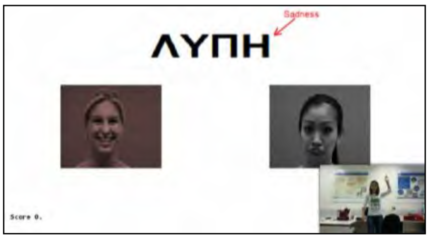
Για να διαλέξεις τη δεξιά εικόνα, σήκωσε ψηλά το δεξί σου χέρι.

Όταν βρεις τη σωστή απάντηση, σήκωσε και τα δύο σου χέρια ψηλά για να προχωρήσεις στο επόμενο συναίσθημα.

Για να ξεκινήσει το παιχνίδι, σήκωσε και τα δύο σου χέρια ψηλά.

Fig. 1. Example of game instruction page

Game begins with an instruction page where the child is informed what is going to happen, what she/he has to do and how she/he can do it. Apart from the text on the screen, audio instructions are also provided. Audio cues are important as the information presented is clear and age-appropriate. When the child feels ready, she/he can choose to start the game. The game provides a structure learning environment which consists of 3 different levels with increasing difficulty. Breaking the teaching intervention into small learning steps makes the task easier to perform. In the first level children should learn labeling emotions by correlating emotion terms with images. In the second level they should learn to recognize emotions from their description and their association with facial features. In the third level they should learn to identify the causes of various feelings in different situations, obtained through the use of social stories. Those three levels provide recognition, matching, observation, understanding and generalization of facial emotions.



ΛΥΠΗ

Fig. 2. First level – labeling emotions (incorrect selection) [66]

Individuals with autism are usually visual learners, which mean that they understand written words, photos and visual information better than spoken language. Information is good to be presented through their strongest processing area. When teaching individuals with autism about emotions, it is important to keep explanations as simple and as concrete as possible. It is also recommended to describe each feeling pictorially by using pictures with clear outline, minimal details and color [67]. For young children it is advisable to keep to the basic emotions. In our approach, the basic emotions selected include happy, sad, angry, scared and surprised. Those emotions were chosen because typically developing children can recognize and understand them between 2 and 7 years of age. The face stimuli we used are 15 grayscale photographs of male and female faces, taken from the CAlifornia Facial Expressions (CAFE) dataset [66]. This dataset was selected as the most appropriate with respect to the emotion recognition task since all images meet FACS criteria [68] and all faces have been certified as "FACS-correct" [69]. The stimuli are presented on each trial with different pairs of photos and the goal is to choose the correct image.

4.2. System Development

Our game is implemented with the use of Microsoft Kinect, which is a motion sensing input device able to track movement and voice, and even identify faces, without the need for any additional devices. Kinect is a low cost and simple way for motion capturing. In more detail, it offers simple, reliable skeleton tracking and a Software Development Kit (SDK). The Microsoft Kinect SDK is used with C# as backend. Among the advantages using C# is that we can integrate XNA Game Studio to develop our game, as well as to utilize the XNA libraries (provided by Microsoft) for graphics creation.

4.3. Interaction with the System

Interaction may be one of the areas that need to be developed with extreme care, depending on the activities and skills to be worked on each of the games and target group. The computer-based interventions that use a keyboard or a mouse for interaction might cause problem with the younger children which may not be able to use a computer. Our gesture-based interaction approach moves the control of computer from a mouse and keyboard, to the motions of the body via new input devices.

Our game is designed to use non-touch based NUI and to be controlled by hand gestures. The gestures are translated into control commands. The player has three possible actions in all game states, to choose left or right image and move to the next play area. These basic actions are implemented with efficient and easy to use gestures. Moving to the next play area requires a two-hand gesture which is performed by moving both hands above the head. Selecting the left image requires a one-hand gesture which is performed by moving the left hand above head. Respectively, selecting the right image requires a one-hand gesture which is performed by moving the right hand above head.

During the game, if the player selects the correct or incorrect stimuli, the system will inform player that he/she gave the correct or incorrect answer. Each answer provides an audio and a visual feedback such as operation-related sounds and changing the images' color. A voice telling "Bravo" rewards player for the correct answer and a voice telling "Try again" encourages the player to try again when the user provides an incorrect answer. There are no other sound effects because individuals with ASD may suffer from auditory sensitivity [70], may demonstrate oversensitivity to certain sounds, even at low volume and may feel discomfort when exposed to certain sounds [71]. Visual feedback is also provided by changing the image's color into light green for the correct answer and into light red for incorrect answer. Light colors were selected because in ASD occur a reduced chromatic discrimination that is due to general reduction in sensitivity [72]. Children with autism are less accurate at detecting the differences between colors and have less accurate color perception [73]. Individuals with ASD also experience hypersensitivity which includes increased light sensitivity and harshness of colors. Acute sensitivity to color is presented by their preference or avoidance of a particular color [74].

5. Research Focus, Methodology and Findings

The rationale for the present work summarizes in the cumulative effect that the above-mentioned practices have on teaching children with ASD. In this context, new teaching techniques and interaction methods have been exploited primarily as gaming tool for improved child - learning relationship.

5.1. Context

Our game exhibits several novel characteristics, which differentiate it from other forms of educational computer games and platforms. It introduces novel interaction techniques which allow the “player” to focus on the learning process without distraction from the use of “complex” interaction devices that need skills. Consequently, its primary focus is to enable preschoolers, with the use of NUI devices to perform learning tasks and provide an effective and engaging learning experience. To achieve this, we build on technologies such as, game engines and advanced human-computer interaction. To illustrate some of the concepts described so far and to provide insight into the features of our game, we present an indicative scenario emphasizing on structured and NUI based game execution for educating preschoolers with ASD. Our game begins with an instruction page where the child is informed on what is going to happen, what it has to be done and how it is done. A two-hand gesture which is performed by moving both hands above the head is required to start the game. In the first level, children should learn labeling emotions by correlating emotion terms with images. The stimuli are presented on each trial with different pair of photos and the goal is to choose the correct image among the two. Selecting the left image requires a one-hand gesture which is performed by moving the left hand above the head. Image's color changes into light red for incorrect answer. Selecting the right image requires a one-hand gesture which is performed by moving the right hand above the head. Image's color changes into light green for the correct answer. Moving to the next play area requires a two-hand gesture which is performed by moving both hands above the head. In the second level they should learn to recognize emotions from their description and their association with facial features. In the third level they should learn to identify the causes of various feelings in different situations, obtained through the use of social stories. At the end of the game, there is a congratulation message.

5.2. Research question and methods

Having outlined relevant general issues, we will attempt to briefly elaborate on some key research questions. These can be broadly grouped into three constituents, namely:

- NUI devises related questions such as what are the advantages of NUI devices when preschoolers are involved; which is the impact of them to the player involved, and how they are enhancing the learning process.
- Emotional state questions as to how player's emotional state is affecting its learning abilities and how such games help to overcome these obstacles.
- Play area settings related questions aiming to unfold player's cooperation matters in conjunction with surroundings while in learning process with the use of such games.

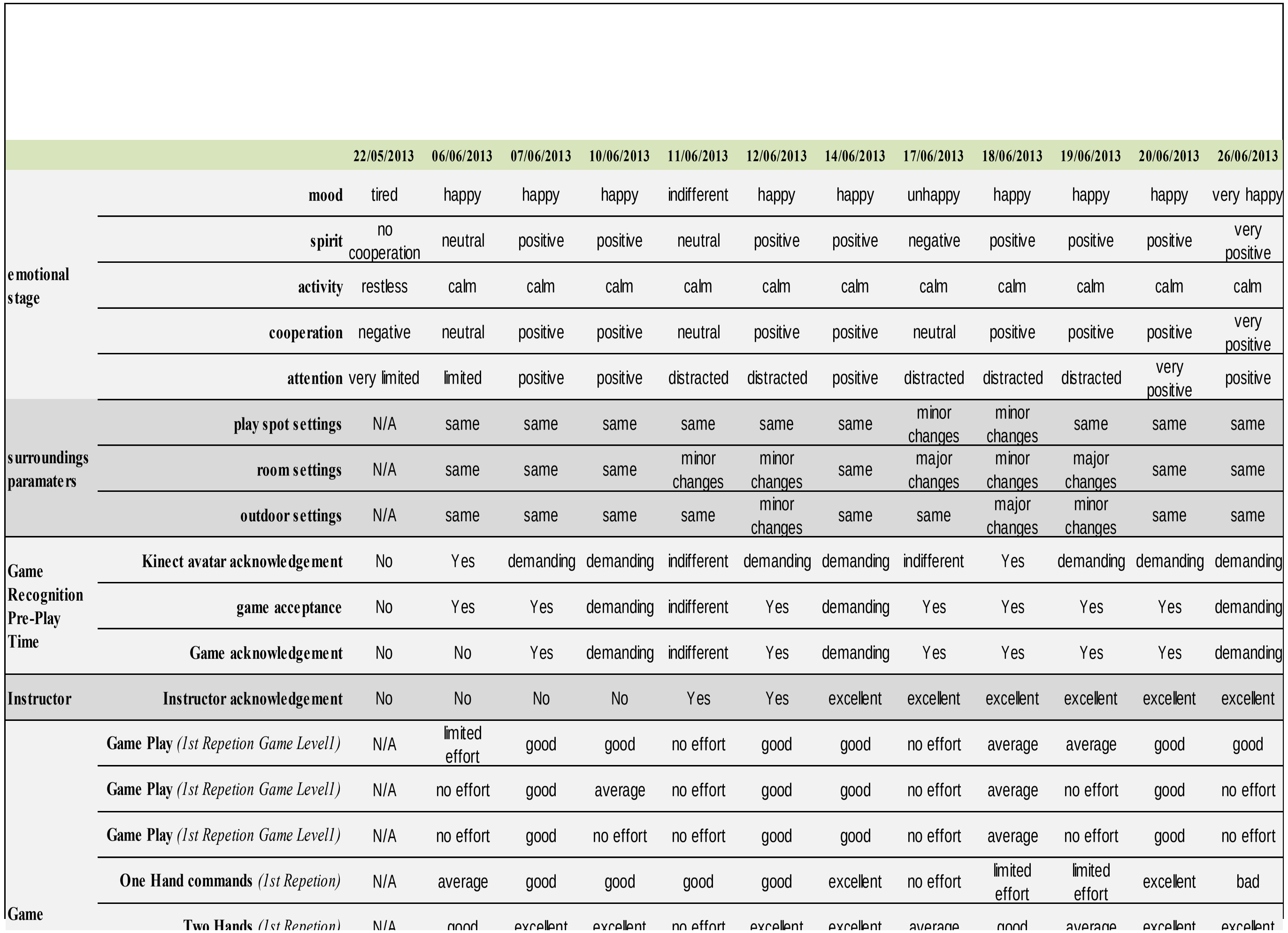
To address the above questions, research methods were used to collect data and envision new capabilities for improved learning processes. Specifically, an experimental descriptive survey was conducted during game execution sessions, which included, videos and text notes in order to provide the insights required.

As our intention was to unfold hidden or implicit elements of educational games with the use of NUI devices for preschoolers with ASD diagnosis, game sessions were tailored so as to feed envisioning of new (improved) learning practices. The focus of the findings is put on player's emotional state, usage of NUI devices and play area settings.

5.3. Findings

Our survey was directed to preschoolers with ASD and was complemented by documented materials. The findings can be analyzed into emotional state versus game performance, emotional state versus surroundings, concentration and game performance, and NUI device and game acknowledgement.

Figure 3 summarize data collected during the learning sessions performed. Specifically, the survey identified 5 areas of importance namely the emotional state, the surroundings, the pre session start game acknowledgement, the instructor pre session start recognition and game interaction during playing. Areas 1, 2, and 5 were expected to be identified. Areas 3 and 4 were unexpected and strengthened our assumptions for NUI devices.



		22/05/2013	06/06/2013	07/06/2013	10/06/2013	11/06/2013	12/06/2013	14/06/2013	17/06/2013	18/06/2013	19/06/2013	20/06/2013	26/06/2013
emotional stage	**mood**	tired	happy	happy	happy	indifferent	happy	happy	unhappy	happy	happy	happy	very happy
	spirit	no cooperation	neutral	positive	positive	neutral	positive	positive	negative	positive	positive	positive	very positive
	activity	restless	calm	calm	calm	calm	calm	calm	calm	calm	calm	calm	calm
	cooperation	negative	neutral	positive	positive	neutral	positive	positive	neutral	positive	positive	positive	very positive
	attention	very limited	limited	positive	positive	distracted	distracted	positive	distracted	distracted	distracted	very positive	positive
surroundings paramaters	**play spot settings**	N/A	same	same	same	same	same	same	minor changes	minor changes	same	same	same
	room settings	N/A	same	same	same	minor changes	minor changes	same	major changes	minor changes	major changes	same	same
	outdoor settings	N/A	same	same	same	same	minor changes	same	same	major changes	minor changes	same	same
Game Recognition Pre-Play Time	**Kinect avatar acknowledgement**	No	Yes	demanding	demanding	indifferent	demanding	demanding	indifferent	Yes	demanding	demanding	demanding
	game acceptance	No	Yes	Yes	demanding	indifferent	Yes	demanding	Yes	Yes	Yes	Yes	demanding
	Game acknowledgement	No	No	Yes	demanding	indifferent	Yes	demanding	Yes	Yes	Yes	Yes	demanding
Instructor	**Instructor acknowledgement**	No	No	No	No	Yes	Yes	excellent	excellent	excellent	excellent	excellent	excellent
Game	**Game Play** *(1st Repetion Game Level1)*	N/A	limited effort	good	good	no effort	good	good	no effort	average	average	good	good
	Game Play *(1st Repetion Game Level1)*	N/A	no effort	good	average	no effort	good	good	no effort	average	no effort	good	no effort
	Game Play *(1st Repetion Game Level1)*	N/A	no effort	good	no effort	no effort	good	good	no effort	average	no effort	good	no effort
	One Hand commands *(1st Repetion)*	N/A	average	good	good	good	good	excellent	no effort	limited effort	limited effort	excellent	bad
	Two Hands *(1st Repetion)*	N/A	good	excellent	excellent	no effort	excellent	excellent	average	good	average	excellent	excellent

Fig. 3. Raw experimental data

Figure 4 refers to our first group of research questions namely: (a) what are the advantages of NUI devices when preschoolers are involved, (b) which is the impact of them to the player involved and (c) how they are enhancing the learning process. As shown, regardless of the emotional state of the participant (happy, unhappy etc.) and regardless of minor changes of the surroundings, participant were acknowledging the use of NUI devices and especially the device avatar which was demanded even before the session start. Thus we can safely deduct that the use of NUI devices enhance game acceptance, game recognition and player involvement & participation, i.e. the learning process.

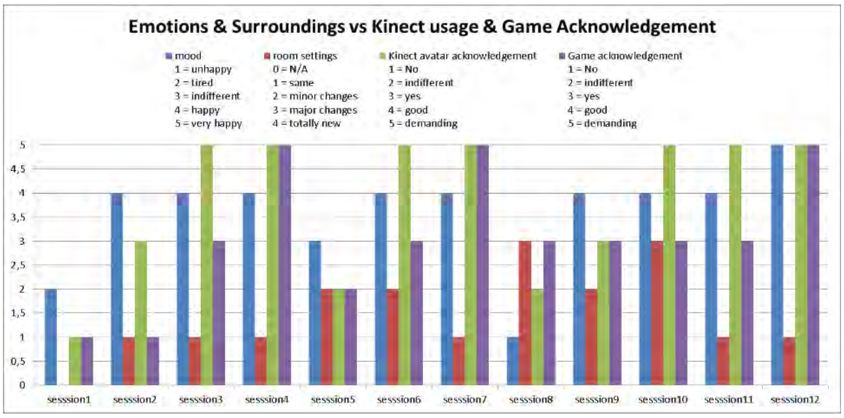


Fig. 4. NUI devices and Game Acknowledgement

Figure 5 refers to the second group of research questions i.e. (a) how player's emotional state is affecting its learning abilities and (b) how such games help to overcome these obstacles. We can see that player's emotional state is affecting its learning abilities in such a way that sometimes it makes the learning process impossible. Even so we can see that game was accepted and an effort was made by the player, not always with success.

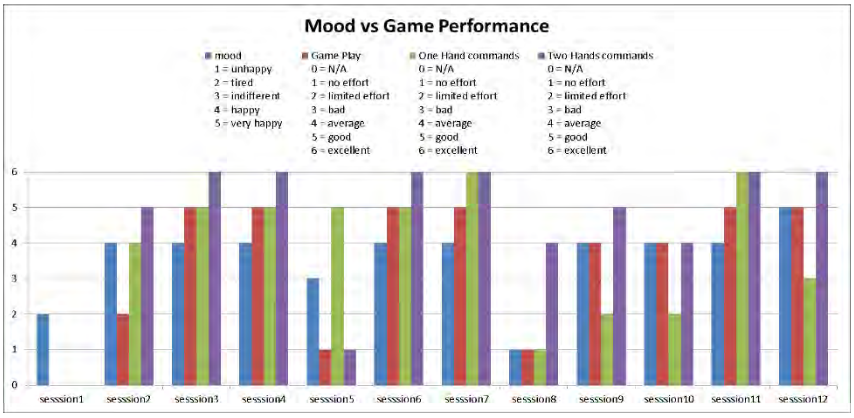


Fig. 5. Emotional state versus Game performance

Figure 6 refers to the third group of research questions, i.e. Play area settings related questions aiming to unfold player's cooperation matters in conjunction with surroundings while in learning process with the use of such games. As shown even minor changes in the surroundings (play spot, room, outdoor etc.) affect dramatically the game acknowledgement, game acceptance, and game interaction but have a small or no effect on the NUI device avatar acknowledgement. Thus we can also here safely deduct that the use of NUI devices enhance the game acceptance i.e. the learning process.

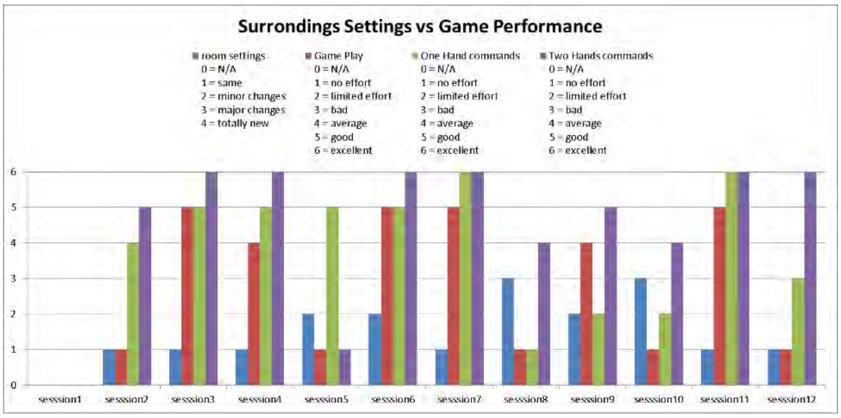


Fig. 6. Play Area surroundings versus Game performance

Summarizing our findings analysis we are lead to the following conclusions: (a) NUI devices enhance the game acceptance i.e. the learning process, (b) Player's emotional state is affecting its learning abilities and NUI devices can help to overcome, up to a degree these obstacles, and (c) Play area settings affect the player's attention which in turn affects the learning process. The use of NUI devices and especially recognizable avatars help to minimize the phenomenon.

6. Discussion and Future Work

Increased interest in the potential of technology for users with autism is motivated by the rapidly growing needs for providing intervention. Research shows that early intervention can greatly improve the lives of children with autism. Computer-based tools have been widely used with success to teach individuals with autism. However, early intervention cannot be achieved with computers due to lack of skills. Gesture-based interaction aims to contribute to overcoming this restriction [75].

NUI devices have been used in learning environments successfully [76]. Therefore, Serious games with NUI interaction could be a promising intervention strategy because they are appealing, motivating for young children to use and convenient to access. When control is achieved through natural gestures, the user does not have to learn how to perform actions or how to operate games. Gesture-based interventions in order to help children with autism to improve their skills must be carefully designed in accordance with their abilities and needs.

Kinect offers an unlimited number of opportunities for new applications such as gesture-based interventions for individuals with autism. As Kinect can support gesture and speech recognition and offer motion-sensing and interaction, children with autism can benefit to improve a broader range of skills.

Our experimental results are twofold, (a) technology interventions provide a positive reaction to alternative interaction techniques (i.e. NUI I/O) to individuals with ASD and (b) technology acceptance is highly connected with individual's emotional state and game surroundings settings.

Future work could include further involvement of multimodal NUI devices so that the roles, between player and machine, are reversed and the player performs gestures, sounds, grimaces etc. and the machine responds. Such a design, of Serious Games that will enable children to perform and the system to respond will support a "learn by doing" methodology. Furthermore, special gestures could be designed to improve interaction between children and machine

References

1. American Psychiatric Association, "Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition – Text Revision (DSM-IV-TR)," 2000.
2. S. W. White, K. Keonig and L. Scahill, "Social Skills Development in Children with Autism Spectrum Disorders: A review of the Intervention Research," Journal of autism and developmental disorders, vol. 37, no. 10, pp. 1858-1868, 2007.
3. P. R. Hobson, "The autistic child's appraisal of expressions of emotion," Journal of Child Psychology and Psychiatry, vol. 27, no. 3, pp. 321-342, 1986.
4. K. M. Rump, et al., "The Development of Emotion Recognition in Individuals With Autism," Child development, vol. 80, no. 5, pp. 1434-1447, 2009.
5. M. D. Lerner, J. C. McPartland and J. P. Morris, "Multimodal emotion processing in autism spectrum disorders: An event-related potential study," Developmental cognitive neuroscience, vol. 3, pp. 11-21, 2013.
6. S. Kuusikko, et al., "Emotion recognition in children and adolescents with autism spectrum disorders," Journal of autism and developmental disorders, vol. 39, no. 6, pp. 938-945, 2009.
7. M. Uljarevic and A. Hamilton, "Recognition of emotions in autism: a formal meta-analysis," Journal of autism and developmental disorders, vol. 43, no. 7, pp. 1517-1526, 2013.
8. K. A. Pelphrey, N. J. Sasson and J. S. Rezni, "Visual scanning of faces in autism," Journal of autism and developmental disorders, vol. 32, no. 4, pp. 249-261, 2002.
9. W. Jones, K. Carr and A. Klin, "Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder," Arch Gen Psychiatry, vol. 65, no. 8, pp. 946-954, 2008.
10. M. B. Harms, A. Martin and G. L. Wallace, "Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies," Neuropsychology review, vol. 20, no. 3, pp. 290-322, 2010.
11. J. W. Tanaka, et al., "The perception and identification of facial emotions in individuals with autism spectrum disorders using the Let's Face It! Emotion Skills Battery," Journal of Child Psychology and Psychiatry, vol. 53, no. 12, pp. 1259-1267, 2012.

12. J. L. Tracy, et al., "Is emotion recognition impaired in individuals with autism spectrum disorders?," Journal of autism and developmental disorders, vol. 41, no. 1, pp. 102-109, 2011.
13. N. Bauminger, "The facilitation of social-emotional understanding and social interaction in high-functioning children with autism: Intervention outcomes.," Journal of autism and developmental disorders, vol. 32, no. 4, pp. 283-298, 2002.
14. S. Faja, et al., "Becoming a face expert: a computerized face-training program for high-functioning individuals with autism spectrum disorders," Developmental neuropsychology, vol. 33, no. 1, pp. 1-24, 2007.
15. S. Eikeseth, "Outcome of comprehensive psycho-educational interventions for young children with autism," Research in developmental disabilities, vol. 30, no. 1, pp. 158-178, 2009.
16. C. A. Gray and J. D. Garand, "Social stories: Improving responses of students with autism with accurate social information," Focus on Autism and Other Developmental Disabilities, vol. 8, no. 1, pp. 1-10, 1993.
17. K. Pajareya and K. Nopmaneejumruslers, "A One-Year Prospective Follow-Up Study of a DIR/Floortime™ Parent Training Intervention for Preschool Children with Autistic Spectrum Disorders," Journal of the Medical Association of Thailand, vol. 95, no. 9, pp. 1184-1193, 2012.
18. B. M. Prizant, et al., "The SCERTS Model: A transactional, family-centered approach to enhancing communication and socioemotional abilities of children with autism spectrum disorder," Infants & Young Children, vol. 16, no. 4, pp. 296-316, 2003.
19. S. Axelrod, K. K. McElrath and B. Wine, "Applied behavior analysis: autism and beyond," Behavioral Interventions, vol. 27, no. 1, pp. 1-15, 2012.
20. B. Reichow, "Overview of meta-analyses on early intensive behavioral intervention for young children with autism spectrum disorders," Journal of autism and developmental disorders, vol. 42, no. 4, pp. 512-520, 2012.
21. P. Tzanakaki, et al., "How and why do parents choose Early Intensive Behavioral Intervention for their young child with Autism?," Education and training in autism and developmental disabilities, vol. 47, no. 1, pp. 58-71, 2012.
22. N. Peters-Scheffer, et al., "A meta-analytic study on the effectiveness of comprehensive ABA-based early intervention programs for children with Autism Spectrum Disorders," Research in Autism Spectrum Disorders, vol. 5, no. 1, pp. 60-69, 2011.
23. A. C. Stahmer, N. Akshoomoff and A. B. Cunningham, "Inclusion for toddlers with autism spectrum disorders The first ten years of a community program," autism, vol. 15, no. 5, pp. 625-641, 2011.
24. Z. Kaldy, et al., "Toddlers with autism spectrum disorder are more successful at visual search than typically developing toddlers," Developmental science, vol. 14, no. 5, pp. 980-988, 2011.
25. C. Gonzalez , et al., "Practice makes improvement: How adults with autism out-perform others in a naturalistic visual search task," Journal of autism and developmental disorders, vol. 43, no. 10, pp. 2259-2268, 2013.
26. J. B. Ganz, R. L. Simpson and E. M. Lund, "The picture exchange communication system (PECS): A promising method for improving communication skills of learners with autism spectrum disorders," Education and Training in Autism and Developmental Disabilities , vol. 47, no. 2, pp. 176-186, 2012.
27. A. K. Moody,, "Family Connections: Visual Supports for Promoting Social Skills in Young Children: A Family Perspective," Childhood Education, vol. 88, no. 3, pp. 191-194, 2012.

28. G. B. Mesibov, V. Shea and E. Schopler, The TEACCH approach to autism spectrum disorders, New York, NY, US: Springer Science + Business Media, 2005.
29. T. R. Goldsmith and L. A. LeBlanc, "Use of technology in interventions for children with autism," Journal of Early and Intensive Behavior Intervention, vol.1, no.2, pp.166-178, 2004.
30. A. L. Wainer and B. R. Ingersoll, "The use of innovative computer technology for teaching social communication to individuals with autism spectrum disorders," Research in Autism Spectrum Disorders, vol. 5, no. 1, pp. 96-107, 2011.
31. B. O. Ploog, et al., "Use of Computer-Assisted Technologies (CAT) to Enhance Social, Communicative, and Language Development in Children with Autism Spectrum Disorders," Journal of Autism and Developmental Disorders, pp. 1-22, 2012.
32. D. Djaouti, et al., "Origins of Serious Games," Serious Games and Edutainment Applications, pp. 25-43, 2011.
33. M. Noor, F. Shahbodin and C. Pee, "Serious Game for Autism Children: Review of Literature," World Academy of Science, Engineering and Technology, vol.64, no.124, pp 647-652.
34. M. Silver and P. Oakes, "Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others," Autism, vol. 5, no. 3, pp. 299-316, 2001.
35. O. Golan and S. Baron-Cohen, "Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia," Development and Psychopathology, vol. 18, no. 2, pp. 591-617, 2006.
36. O. Grynszpan, J. C. Martin and J. Nadel, "Multimedia interfaces for users with high functioning autism: An empirical investigation," International Journal of Human-Computer Studies, vol. 66, no. 8, pp. 628-639, 2008.
37. S. L. Finkelstein, et al., "cMotion: A new game design to teach emotion recognition and programming logic to children using virtual humans," IEEE Virtual Reality Conference, pp. 249-250, 2009.
38. J. W. Tanaka, et al., "Using computerized games to teach face recognition skills to children with autism spectrum disorder: The Let's Face It! program," Journal of Child Psychology and Psychiatry, vol. 51, no. 8, pp. 944-952, 2010.
39. D. Deriso, et al., "Emotion Mirror: A Novel Intervention for Autism Based on Real-Time Expression Recognition," vol. 7585, pp. 671-674, 2012.
40. J. C. Miranda, et al., "Interactive Technology: Teaching People with Autism to Recognize Facial Emotions," Autism Spectrum Disorders - From Genes to Environment, 2011.
41. S. Bernardini, et al., "ECHOES: An intelligent serious game for fostering social communication in children with autism," Information Sciences, in press 2013.
42. K. Porayska-Pomsta, et al., "Building an Intelligent, Authorable Serious Game for Autistic Children and Their Carers," Advances in Computer Entertainment, Springer International Publishing, pp. 456-475, 2013.
43. S. Ramdoss, et al., "Use of computer-based interventions to teach communication skills to children with autism spectrum disorders: A systematic review," Journal of Behavioral Education, vol.20, no.1, pp.55-76, 2011.
44. S. Parsons and S. Cobb, "State-of-the-art of virtual reality technologies for children on the autism spectrum," European Journal of Special Needs Education, vol. 26, no. 3, pp. 355-366, 2011.
45. R. el Kaliouby, A. Teeters and R. W. Picard, "An exploratory social-emotional

prosthetic for autism spectrum disorders," International Workshop on Wearable and Implantable Body Sensor Networks, pp. 2-4, 2006.

46. S. Baron-Cohen, et al., "Transported to a world of emotion," The Psychologist, vol. 20, no. 2, pp. 76-77, 2007.
47. S. Baron-Cohen, O. Golan and E. Ashwin, "Can emotion recognition be taught to children with autism spectrum conditions?," Philosophical Transactions of the Royal Society B-Biological Sciences, vol. 364, no. 1535, pp. 3567-3574, 2009.
48. O. Golan, et al., "Enhancing emotion recognition in children with autism spectrum conditions: an intervention using animated vehicles with real emotional faces," Journal of Autism and Developmental Disorders, vol. 40, no. 3, pp. 269-279, 2010.
49. B. T. Williams, K. M. Gray and B. J. Tonge, "Teaching emotion recognition skills to young children with autism: a randomised controlled trial of an emotion training programme," Journal of Child Psychology and Psychiatry, vol. 53, no. 12, pp. 1268-1276, 2012.
50. G. De Leo, et al., "A smart-phone application and a companion website for the improvement of the communication skills of children with autism: clinical rationale, technical development and preliminary results," Journal of Medical Systems, vol. 35, no. 4, pp. 703-711, 2011.
51. J. Baio, "Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008," MMWR Surveillance Summaries, vol. 61, no. 3, pp. 1-19, 30 March 2012.
52. M. Elsabbagh et al., "Global Prevalence of Autism and Other Pervasive Developmental Disorders," Autism Research, vol. 5, no. 3, pp. 160-179, 11 April 2012.
53. J. Lozano-Martínez, et al., "Software for Teaching Emotions to Students with Autism Spectrum Disorder," Revista Comunicar, vol. 18, no. 36, pp. 139-148, 2011.
54. E. Jasmin, et al., "Sensori-motor and daily living skills of preschool children with autism spectrum disorders," Journal of Autism and Developmental Disorders, vol.39, no.2, pp.231-241, 2009.
55. K. Sitdhisanguan, et al., "Using tangible user interfaces in computer-based training systems for low-functioning autistic children," Personal and Ubiquitous Computing, vol. 16, no. 2, pp. 143-155, 2012.
56. A. E. Lane and J. M. Ziviani, "Factors influencing skilled use of the computer mouse by school-aged children," Computers & Education, vol. 55, no. 3, pp. 1112-1122, 2010.
57. ISO, ISO 9241-9 Ergonomic requirements for office work with visual display terminals (VDTs) - Part 9: Requirements for non-keyboard input devices, International Organization for Standardization, 2000.
58. D. Natapov, S. J. Castellucci and I. S. MacKenzie, "ISO 9241-9 evaluation of video game controllers," Proceedings of Graphics Interface 2009, pp. 223-230, May 2009.
59. S. A. Douglas, A. E. Kirkpatrick and I. S. MacKenzie, "Testing pointing device performance and user assessment with the ISO 9241, Part 9 standard," Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 215-222, May 1999.
60. I. S. MacKenzie, T. Kauppinen and M. Silfverberg, "Accuracy measures for evaluating computer pointing devices," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 9-16, 2001.
61. M. Akamatsu and, I. S. MacKenzie, "Changes in applied force to a touchpad during pointing tasks," International Journal of Industrial Ergonomics, vol. 29, no. 3, pp. 171-182, 2002.

62. K. Mitgutsch and N. Alvarado, "Purposeful by design?: a serious game design assessment framework," Proceedings of the International Conference on the Foundations of Digital Games, pp. 121-128, 2012.
63. B. Abirached, Y. Zhang and J. H. Park, "Understanding User Needs for Serious Games for Teaching Children with Autism Spectrum Disorders Emotions," Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 1054-1063, 2012.
64. J. H. Park, B. Abirached and Y. Zhang, "A framework for designing assistive technologies for teaching children with ASDs emotions," Extended Abstracts on Human Factors in Computing Systems, pp. 2423-2428, 2012.
65. E. Ashwin, et al., "Eagle-eyed visual acuity: an experimental investigation of enhanced perception in autism," Biological Psychiatry, vol. 65, no. 1, pp. 17-21, 2009.
66. M. N. Dailey, G. W. Cottrell and J. Reilly, "CAlifornia Facial Expressions (CAFE)," 2001. [Online]. Available: http://www.cs.ucsd.edu/users/gary/CAFE/.
67. S. Dodd, Understanding Autism, Sydney: Elsevier, 2004.
68. P. Ekman and W. V. Friesen, Facial Action Coding System, Palo Alto: Consulting Psychologist Press, 1978.
69. M. L. Smith, et al., "Transmitting and Decoding Facial Expressions," Psychological Science, vol. 16, no. 3, pp. 184-189, 2005.
70. E. Gomes, F. S. Pedroso and M. B. Wagner, "Auditory hypersensitivity in the autistic spectrum disorder (original title: Hipersensibilidade auditiva no transtorno do espectro autístico)," Pró-Fono Revista de Atualização Científica, vol. 20, no. 4, pp. 279-284, 2008.
71. Y. H. Tan, et al., "Auditory abnormalities in children with autism," Open Journal of Psychiatry, vol. 2, no. 1, pp. 33-37, 2012.
72. A. Franklin, et al., "Reduced chromatic discrimination in children with autism spectrum disorders," Developmental Science, vol. 13, no. 1, pp. 188-200, 2010.
73. A. Franklin, et al., "Color perception in children with autism," Journal of Autism and Developmental Disorders, vol. 38, no. 10, pp. 1837-1847, 2008.
74. R. A. Coulter, "Understanding the Visual Symptoms of Individuals with Autism Spectrum Disorder (ASD)," Optometry & Vision Development, vol. 40, no. 3, pp. 164-175, 2009.
75. E. Christinaki, N. Vidakis and G. Triantafyllidis, "Facial expression recognition teaching to preschoolers with autism: a natural user interface approach," Proceedings of the 6th Balkan Conference in Informatics. ACM, pp. 141-148, 2013.
76. N. Maldonado, "Technology in the Classroom: Wii: An Innovative Learning Tool in the Classroom." Childhood Education, vol, 86, no 4, pp. 284-285, 2010.

Eirini Christinaki was born in Crete, Greece. She received her BSc degree in Applied Informatics and Multimedia from the Technological Educational Institute of Crete in 2013. She is currently a postgraduate student in Informatics and Multimedia at the same Institute. In March 2013, she joined the Computational Medicine Laboratory (CML) of Institute of Computer Science at Foundation for Research and Technology - Hellas (FORTH). Since November 2013, she participates in research project SEMEOTICONS: SEMEiotic Oriented Technology for Individual's CardiOmetabolic risk self-assessmeNt and Self-monitoring. Her current research interests are mainly in the areas of Serious Games, Affective Computing, Biomedical Engineering and Health Informatics.

Nikolas Vidakis was born in Crete, Greece. He received a BSc(Hons) in Computing in Business from the University of Northumbria at Newcastle in 1993. Dr Vidakis obtained his Ph.D degree from the Technical University of Vienna in 1997. In the past, he has worked for the Technical University of Vienna (Austria), Minoan Lines S.A. (Greece) and the Technological Education Institution of Crete (Greece). During his work at Minoan Lines and the Technological Education Institution of Crete he has been involved in a few Research and Technological Development projects, funded by national and European organizations. Since October 2005, Dr Vidakis is an Assistant Professor at the Department of Informatics Engineering, Faculty of Applied Technologies at the Technological Education Institution of Crete and since 2006 a member of the search laboratory iSTLab. His current research interests are in the areas of HCI, SE, Serious Games, VCoP. More information about publication, research interests, as well as a full CV can be found at the following address (http://www.istl.teiher.gr)

Georgios A. Triantafyllidis was born in Thessaloniki, Greece, in 1975. He received the Diploma Degree in electrical engineering and the Ph.D. degree in image/video processing from Aristotle University of Thessaloniki, Greece, in 1997 and 2002, respectively. From 1997 till 2000 he worked as a research assistant at the Information Processing Lab of the Aristotle University of Thessaloniki. From 2000 to 2008 he joined the Informatics and Telematics Institute of CERTH in Thessaloniki as an associate and since 2003 as a senior researcher. From 2005 to 2008 he was also with the Electrical and Computer Eng Dept of the Aristotle University of Thessaloniki as a visiting lecturer. Then, he joined the Applied Informatics and Multimedia Dept of the Technological Educational Institute of Crete, Greece, as an assistant professor. Since August 2012, he has been appointed assistant professor in the Department of Medialogy at Aalborg University Copenhagen (Denmark). Dr Triantafyllidis has been involved in 16 national and international research competitive projects, where he conducted research in various fields of image/video processing/analysis, intelligent systems and 2D/3D computer vision. He was member of the 3DTV Network of Excellence and the general chair of the 3DTV CON 2007 and CGIM 2012 conferences. He is an editorial board member in 3D Research Journal of Springer Publishing. He edited three special issues on 3DTV and published more than 65 referred research papers with more than 300 citations.

Received: February 15, 2014; Accepted: May 28, 2014

Computer Science and Information Systems 11(2):745–763 DOI: 10.2298/CSIS140215040K

A Distributed Near-Optimal LSH-based Framework for Privacy-Preserving Record Linkage

Dimitrios Karapiperis and Vassilios S. Verykios

School of Science and Technology
Hellenic Open University
{dkarapiperis,verykios}@eap.gr

Abstract. In this paper, we present a framework which relies on the Map/Reduce paradigm in order to distribute computations among underutilized commodity hardware resources uniformly, without imposing an extra overhead on the existing infrastructure. The volume of the distance computations, required for records comparison, is largely reduced by utilizing the so-called Locality-Sensitive Hashing technique, which is optimally tuned in order to avoid highly redundant computations. Experimental results illustrate the effectiveness of our distributed framework in finding the matched record pairs in voluminous data sets.

Keywords: Locality-Sensitive Hashing, Bloom filter, Map/Reduce.

1. Introduction

Recently, a series of bank and insurance company failures triggered a financial crisis of unprecedented severity. Institutions had to engage in negotiations in order to get back on their feet [1]. An important parameter of these negotiations are the customer bases that need to be cleaned up and possibly merged. The process of merging the customer bases and finding out records that refer to the same real entity is known as the Record Linkage, Entity Resolution or Data Matching problem [11]. In our case, we also assume that records belong to different owners, who are bound to protect the privacy of the data they hold by the legislation framework. If privacy should be preserved during the linkage of the records, special techniques should be developed that leverage similarity and simultaneously respect the privacy of sensitive data. This type of linkage is known as Privacy-Preserving Record Linkage and is picking up a lot of steam lately.

More specifically, the process of linking records from various data sets consists basically of two steps. The first step is the *searching* of potentially matching pairs and the second is the actual *matching* of these pairs. The *searching* step refers to the smooth scaling of computational cost in terms of time and consumed resources with respect to the data volume explosion. So long as the data is increasing, we should be able to link records efficiently, in an anticipated manner and offer linkage solutions of low resource consumption. The first step relies on techniques such as the traditional blocking [6], the Mapping-based Indexing [24], the Sorted Neighborhood approach [21], the Q-Gram-based indexing [6] and the Canopy Clustering [15]. The second step, known as the *matching* step, is implemented either in an exact or in an approximate manner. Exact matching of two records can be regarded as a binary decision problem of exact agreement or disagreement. Approximate matching entails the calculation of value similarity, falling in the range of $[0..1]$.

Several methods have been developed for approximate matching of values such as Edit distance [33], Jaro-Winkler distance [46] and SoftTF-IDF [14]. Variations in the data values, mainly due to typographical errors, and privacy concerns, which need to be taken into consideration, lead us to the choice of secure approximate matching [37] solutions for our problem, where value variations are preserved in the transformations that occur to protect their privacy.

By going back to our scenario, let us suppose that two banks started negotiations for a merger and that their corresponding customer bases are voluminous. Since no bank is willing to disclose any confidential data to the other bank, in order to protect their privacy, we should employ high-dimensional data structures to embed that data in, which in combination with the huge collection of records at hand, introduce high computational overhead in the linkage process. Hence, we need an efficient method to identify similar data, given the complex representational structure and the huge size of data sets.

The secure searching solutions, which have been developed to solve the PPRL problem, rely mostly on traditional blocking, where all records that have the same value in a specific field(s) are blocked together for comparison. However, the proposed solutions exhibit a considerable overhead in terms of performance, when applied to voluminous data and especially to high-dimensional data. In [22], as an example, blocking relies on the categorization of records into generalized hierarchies based on the semantics of values of selected fields, which may lead to load imbalance problems, if most values semantically belong to certain categories. Karakasidis and Verykios in [25] present a blocking technique which relies on a sliding window that creates blocks of records. Its performance is considerably degraded, when the size of that window is increased in order to produce more accurate results, as shown experimentally in Sect. 6. Authors in [26], [28] and [17] use redundant probabilistic methods, where each record is blocked into several independent blocking groups, in order to amplify the probability of bringing together similar records for comparison. These techniques though utilize an arbitrary number of blocking groups as well as an arbitrary number of hash functions to shape up the blocking keys for each group. This process has as a result either unnecessary and expensive comparisons or a large number of missed similar record pairs.

In this paper we propose the optimal configuration, denoted by oBfJ, of a naive methodology for PPRL, as introduced in [17]. As shown experimentally, oBfJ can reduce the number of record pairs that are brought together for comparison up to 95% of the total comparison space while it maintains high levels of recall, constantly above 93%. Moreover, by exploiting a number of computational resources of commodity hardware using a distributed framework based on oBfJ, as illustrated in Sect. 5, we will manage these computations with respect to large-scale data volumes. Experimental results in Sect. 6 indicate that our proposed framework outperforms the Multi-Dimensional Privacy-Preserving Blocking (MPPB) method [25], in terms of running time and accuracy in the results.

The structure of the paper is organized as follows. Related work is described in Sect. 2. In Sect. 3 we present an outline of the various building blocks we utilize in our framework. In Sect. 5 we illustrate our proposed framework, which is evaluated in Sect. 6. Conclusions and ideas for future extensions are presented in Sect. 7.

2. Related Work

Several solutions have been presented in the literature in the field of efficient searching for similar records [15, 17, 22–26, 28, 38]. However, these solutions exhibit poor performance when applied to massive data sets. In [15] for example, a cheap distance metric is used for creating clusters of records and then a more expensive, accurate distance metric is used to evaluate the record pairs that are tagged for further evaluation. Nevertheless, the number of record pairs, that should be compared eventually, can still be excessively large. The tree-based indexing methods used in [23, 24, 38] in order to reduce the number of candidate record pairs, as reported and proved in [20, 45] and [4], exhibit quadratic complexity, because they need to scan the whole index structure repeatedly, when these structures are used for representing records even with moderate dimensionality (≥ 10). A detailed survey of blocking techniques for Record Linkage can be found in [12]. An overview of privacy-preserving blocking techniques is provided in [44].

The Bloom filter-based encoding method, as was presented in [40] and is used by our oBfJ methodology, is easily implemented and preserves the similarity of the original records. However under certain circumstances, this method is susceptible to constraint satisfaction cryptanalysis [30]. Scannapieco et al. in [38] present an encoding method that embeds string values in the Euclidean space by using reference sets of random string values. This embedding method imposes certain strict requirements, like the use of random strings of length approximately equal to the length of the values to be embedded, which cannot be applicable to data sets exhibiting quite large variation in the length of their values. Pang et al. in [35] use public reference tables in order to compute the distance of field values from those in the reference tables and to assign them into clusters. However, accurate results are attained only when the reference tables are a superset of the values in the data sets. The encoding method in [13] relies on the extracted q-grams from string values which are sent in an encrypted format to a third-party for comparison. Q-grams are susceptible to frequency attacks and add high communication cost due to the large number of encrypted data that should be sent to the third-party.

Our methodology utilizes a trusted third-party in order to conduct the linkage of the encoded data sets. Two-party techniques, like the ones in [42] and [43], may reduce the risk of privacy breach, like colluding parties, but they are complex and they add high communication cost. Authors in [22, 31], instead of encoding data and submitting it to a third-party, suggest Secure Multi-Party Computation (SMC) protocols [34] to the matching step. These protocols are effective and reliable but they add high computational overhead leading to prohibitive running times.

3. Background and Problem Formulation

For illustration purposes and without loss of generality, let us consider that we have to link two data sets A, which belongs to Alice, and B, which belongs to Bob, as shown in Table 1 and 2 correspondingly. Alice and Bob are allowed to make use of the services offered by an independent party, whom we call Charlie. We assume that Charlie can help in the process by exhibiting what is known as a semi-honest or honest-but-curious behavior [44], which means that even if Charlie is not trying to collude with Alice or Bob, and follows the protocol prespecified by the two custodians, he is still curious to find out as much information, from what he is presented with, as possible. The cardinality of A and B is N_A and N_B respectively.

Table 1. Data set A belonging to Alice

Id	*FirstName*	*LastName*
1	George	Peters
2	John	Smith

Table 2. Data set B belonging to Bob

Id	*FirstName*	*LastName*
1	William	Grace
2	John	Smyth

Alice and Bob will transform their data sets into secure structures so that linkage will be conducted in a private manner. Also, Alice and Bob need the result of the linkage within a reasonable amount of time, although the cardinality of the corresponding data sets might be millions of records. In the next subsections we present the basic components of a naive methodology for PPRL, as introduced in [17], denoted by BfJ, which relies on the Bloom filter-based encoding method [40] and on the Min-Hash Locality-Sensitive Hashing approach [9]. By applying BfJ and setting certain configuration parameters appropriately, as shown in 4, the truly matched record pairs can be found efficiently and accurately.

3.1. Bloom filters for Private Representation of Data

A string value is anonymized by encoding it as a Bloom filter. A Bloom filter is a data structure used to represent the elements of a set, in order to support membership queries for these elements efficiently, in terms of time and space required [7]. It has been shown in [40] that Bloom filters are able to preserve the distance between the pair of unencoded string values, and for this reason they can be used to compare string values in a private manner. More specifically, a bitmap array of size L, initialized with zeros, is created by hashing all consecutive bigrams of a string (sequences of pairs of adjacent characters), by using F independent composite cryptographic hash functions G_is (see Fig. 1). For example, *MD5* and *SHA1* [39] may play the role of G_is along with other more advanced keyed hash message authentication code (*HMAC*) functions like *HMAC-MD5* and *HMAC-SHA1* [29], which utilize a secret key, that should be shared by the data custodians, for increased security.

Let us suppose that Alice and Bob encode the *FirstName* and the *LastName* attributes of their sets, into field-level Bloom filters with length equal to 6 bits by using one cryptographic hash function for each bigram, as shown in Tables 3 and 4 respectively. By concatenating those field-level Bloom filters for each record, we construct the encoded representations of the original records. The *Id* attribute indicates both the initial *Id* and

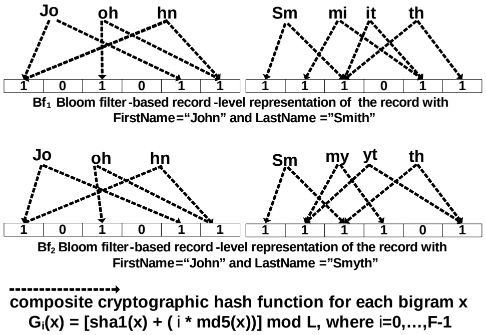

Jo
oh
hn
Sm
mi
it
th
1 0 1 0 1 1 1 1 1 0 1 1
Bf$_1$ Bloom filter-based record-level representation of the record with FirstName="John" and LastName ="Smith"
Jo
oh
hn
Sm
my
yt
th
1 0 1 0 1 1 1 1 1 1 0 1
Bf$_2$ Bloom filter-based record-level representation of the record with FirstName="John" and LastName ="Smyth"
composite cryptographic hash function for each bigram x
G$_i$(x) = [sha1(x) + (i * md5(x))] mod L, where i=0,…,F-1

Fig. 1. Concatenation of field-level Bloom filters into an encoded record-level structure

the source data set. Durham in [17, 18] develops a Bloom filter-based record-level encoding format by sampling random bits from field-level Bloom filters either of dynamic or static size she experimentally shows that accuracy is maintained in similarity calculations. The encoded data sets are denoted as A' and B' respectively. The distance of two Bloom filters Bf_1 and Bf_2 can be measured by the Jaccard metric as $d_J(Bf_1, Bf_2) = |BP(Bf_1) \cap BP(Bf_2)| / |BP(Bf_1) \cup BP(Bf_2)|$, where $BP(\cdot)$ returns all bit positions of a Bloom filter set to one.

Table 3. Transforming data set A into A'

Id_A	*Bf*(*FirstName*)	*Bf*(*LastName*)
A1	<1,0,1,1,0,0 >	<1,0,1,1,0,0>
A2	<0,1,1,1,1,0 >	<1,0,1,1,0,0>

Table 4. Transforming data set B into B'

Id_B	*Bf*(*FirstName*)	*Bf*(*LastName*)
B1	<1,1,0,0,1,1 >	<0,0,0,1,0,0 >
B2	<0,1,1,1,1,0 >	<1,0,1,1,0,1>

3.2. Min-Hash Locality-Sensitive Hashing for Efficient Grouping of Similar Bloom filters

An efficient well-known method that is used in finding similar items among huge data sets, is the Locality-Sensitive Hashing (LSH) [19] technique. The LSH technique performs probabilistic dimensionality reduction of high-dimensional data. Locality-sensitive hash functions that are sensitive to the Jaccard metric, comprise the Min-Hash family, which is denoted by $\mathcal{H}$ [8,10,36]. The locality-sensitive property assumes that the probability of generating the same result if we hash two Bloom filters by the same function of $\mathcal{H}$, is dependent upon their Jaccard distance. In essence, the application of the Min-Hash LSH method to the Bloom filters, can be considered as a Λ-independent and redundant blocking technique. In order to generate these Λ independent blocking groups, we should make use of Λ composite hash functions H^j, where $j = 1, \ldots, \Lambda$. Each H^j consists of a fixed number, say K of base hash functions h^j_k, where $k = 1, \ldots, K$ chosen randomly and uniformly from $\mathcal{H}$. The result of each base hash function $h^j_k(\mathit{Bf}) = \mathit{min}\{\pi^j_k(\mathit{Bf})\}$ applied to a Bloom filter *Bf*, is the bit position of the minimum non-zero element after permuting its elements according to the π^j_k, which is the k-th permutation of the j-th blocking group.

For example, say we have to permute $\mathit{Bf} = <1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0>$ according to the permutation $\pi^1_1 = <3, 5, 0, 2, 11, 8, 9, 10, 4, 6, 7, 1>$ which transforms *Bf* into $<1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0>$, hence $min\{\pi^1_1(A'_1)\} = 3$. Each H^j is used to hashing a Bloom filter to one of the separate blocks of each independent group. By applying to the Bloom filter *Bf* each of the K base hash functions of H^j and by concatenating the results, a Key^j is built, which hashes *Bf* to the j-th blocking group, namely $\mathit{Key}^j = H^j(\mathit{Bf})$, where $H^j = \mathit{Concat}(\mathit{min}\{\pi^j_k(\mathit{Bf})\})$. Each of the blocking groups can be implemented as a simple hash table consisting of key-bucket pairs, where a bucket hosts a linked list where we place the *Id*s of the Bloom filters that have been hashed to this bucket. Another way of implementing blocking groups, especially suited for massive data, is shown in Sects. 5.1 and 5.2.

3.3. Secure Matching of Bloom filter Pairs

Charlie, by using common permutations, hashes the Bloom filters of A and B once to each blocking group. In the blocks of those groups, where Bloom filters of both data sets are located, Charlie compares them in a pairwise manner. To be more precise, the stored *Id*s are used to retrieving the corresponding Bloom filters. A simple decision model classifies those Bloom filter pairs either to matched or to non-matched pairs according to their distance d_J compared to a threshold ϑ, which is defined by the data custodians. Given two Bloom filters Bf_1 and Bf_2, we compute the intermediate variables a, b and c, which will be used in the distance calculation of Bf_1 and Bf_2, as follows:

$$a = \sum_{l \in [0,L)} [(\mathit{Bf}_1[l] == 1) \wedge (\mathit{Bf}_2[l] == 1)] \tag{1}$$

$$b = \sum_{l \in [0,L)} [(\mathit{Bf}_1[l] == 1) \wedge (\mathit{Bf}_2[l] == 0)] \tag{2}$$

$$c = \sum_{l \in [0,L)} [(\mathit{Bf}_1[l] == 0) \wedge (\mathit{Bf}_2[l] == 1)] \tag{3}$$

The Jaccard distance can be calculated as: $d_J(\mathit{Bf}_1, \mathit{Bf}_2) = 1 - [a/(a + b + c)]$. We refer the reader to [44] where various classification techniques are discussed. Also, as shown in Fig. 1, all underlying fields of records in A and B, participate equally in the composition of the encoded representation of the record. This means that each field is weighted equally in both blocking mechanism and distance calculation.

3.4. Map/Reduce Framework for Scaling Up Computations

Map/Reduce [16] is a computational paradigm, where an application is divided into fragments of computation, each of which may be executed on any node of a cluster, so that parallelism can speed-up the whole process. The Map/Reduce system runs on top of a distributed file system where data is fragmented into equally-sized subsets and stored on the nodes of a cluster, providing reliability and integrity by managing redundancy of data and scalability by easily adding compute nodes on demand. When a job is submitted to the Map/Reduce framework, map and reduce tasks are allocated to the compute nodes where data, that the job has specified, exists. Each map task, denoted by M_x, processes a subset of data and outputs intermediate <*key, value*> pairs. Each unique *key*, as generated by the map tasks, is distributed by the partitioner tasks to a single reduce task R_x, along with its associated *value*. This mechanism results to the allocation of each unique *key* with its associated vector of *values* to a reduce task which processes this vector of *values* according to the application logic.

4. The oBfJ Methodology

From the LSH theory [19], given two Bloom filters Bf_1 and Bf_2, if it holds that $d_J(Bf_1, Bf_2) \leq \vartheta$ then $\Pr[h_k^j(Bf_1) = h_k^j(Bf_2)] \geq p_\vartheta$, where $p_\vartheta = 1 - \vartheta$ [36] and ϑ is the maximum distance that two Bloom filters should exhibit in order to be considered as similar. Since each H^j consists of K base hash functions, the probability of Bf_1 and Bf_2 colliding in the same block of the j-th group is:

$$\Pr[H^j(Bf_1) = H^j(Bf_2)] \geq p_\vartheta^K. \tag{4}$$

In the naive BfJ (Bloom filters-Jaccard) methodology, the probability of collision is amplified by blocking the Bloom filters independently to Λ blocking groups. While, though, the probability of collision is amplified, the redundant blocking groups increase the running time and the utilized space. Therefore, the naive BfJ should be optimized in order to produce the smallest set of Bloom filter pairs that include as many as possible from the matched pairs and simultaneously to consume the least possible computational resources. For a given value of K, by setting $\Lambda_{opt} = \lceil \ln(\delta)/\ln(1 - p_\vartheta^K) \rceil$ [5, 19], each Bloom filter pair is returned by this scheme with probability at least $(1 - \delta)$, where δ is an input parameter indicating the probability of failing to return a matched pair. For example, for a distance threshold $\vartheta = 0.40$, by setting $K = 5$ and $\delta = 0.1$, we generate $\Lambda_{opt} = 29$ blocking groups. By utilizing this structure, theoretical guarantees are provided that each matched Bloom filter pair is returned with probability at least 0.9. An arbitrary value for Λ ($\Lambda > 28$), would return the matched Bloom filter pairs but would result to useless additional running time and space. Since the Bloom filter-based method is highly accurate and distance-preserving, by finding the matched Bloom filter pairs, we also find the original matched record pairs, which are a subset of those Bloom filter pairs.

The choice of the optimal value for K, that is K_{opt}, is more complex since we should choose a value that minimizes the running time, which actually depends on the number of distance computations. We estimate the expected running time $E[RT]$ by using several different values of K, and their corresponding Λ_{opt} values, and we choose the one that minimizes the $E[RT]$ value. In the same way as authors do in [41], we decompose $E[RT]$

into two basic components: $E[\mathit{RT}] = \mathit{RT}_T + E[\mathit{UC}]$. The first component RT_T, which is equal to

$$\mathit{RT}_T = \Lambda_{\mathit{opt}}\ K(N_{A'} + N_{B'}), \tag{5}$$

is the time required to compute the *Key*^{j}s for each record of A' and B'. The second component $E[\mathit{UC}]$ is the expected number of unique collisions (unique pairs) in the blocks. Let $Q(\tau)$ denote the probability of a collision for a Bloom filter pair, exhibiting distance τ. Then, $Q(\tau)$ is equal to $1-(1-(1-\tau)^K)^{\Lambda_{\mathit{opt}}}$ [5]. A tight estimation of $E[\mathit{UC}]$ is given by $E[\mathit{UC}] = \sum_{\iota=1}^{N_{A'}} \sum_{\kappa=1}^{N_{B'}} Q(d_J(\mathit{Bf}_\iota, \mathit{Bf}_\kappa))$. However, it is not practical to compute $E[\mathit{UC}]$ in this way since it includes the calculation of distances of the whole comparison space $A' \times B'$. By exploiting statistical information about the distribution of the Bloom filter pairs in distance intervals and by using the Monte Carlo method [32], we can make inferences about the number of Bloom filter pairs falling to those intervals. More specifically, we sample the minimum required number of pairs for each interval from the comparison space and by computing their distances, we make global inferences about the Bloom filter pairs distribution to those distance intervals. For example, for our data sets, the probability that a Bloom filter pair falls within the distance interval $[0.80, 0.85)$ is 0.23. This probability distribution follows the Poisson distribution with $\lambda = 16$ (the 16th interval $[0.80, 0.85)$ starting from $[0.0, 0.05)$). Given such empirical data, let $\widehat{p}_\alpha$ be the estimated proportion of pairs that fall to the α-th interval. Let also $\widehat{N}_\alpha$ be the estimated number of those pairs which equals to $\widehat{p}_\alpha\ |A' \times B'|$. Then, the estimated number of unique collisions, denoted by $\widehat{\mathit{UC}}$, for a given K and Λ_{opt} is

$$\widehat{\mathit{UC}} = \sum_{\alpha=1}^{N_{\mathit{int}}} \widehat{N}_\alpha\ Q(\widetilde{X}_\alpha), \tag{6}$$

where N_{int} is the number of the intervals and $\widetilde{X}_\alpha$ is the midpoint of each interval. Thus, by setting several values for K, we compute for each of them, Λ_{opt} and $E[\mathit{RT}]$. The value of K that exhibits the smallest $E[\mathit{RT}]$, is chosen as K_{opt}. For example, by setting $K = 2, \ldots, 12$, we observe from the curve in Fig. 2 that as K approaches 5 from the left, $E[\mathit{RT}]$ decreases, it is minimized when $K = 5$ and for $K > 5$, $E[\mathit{RT}]$ grows substantially. Therefore, the value of 5 is chosen as K_{opt}.

We can also set K empirically since the correctness of the scheme is guaranteed by setting appropriately Λ_{opt} , but by doing so we may not be optimal in terms of running time.

5. A Distributed LSH-based Framework

In this section we present a distributed framework that relies on the Min-Hash LSH technique for searching similar records efficiently and on the Map/Reduce programming paradigm in order to enable scalability, by providing compute nodes on the fly. Alice and Bob submit their data formatted as shown in Tables 3 and 4 to the trusted third party, Charlie, who utilizes a Map/Reduce system on top of a distributed file system to run the computations. Input data sets are horizontally partitioned into independent subsets D^s of equal predefined size, and then each subset is replicated ρ times on different compute

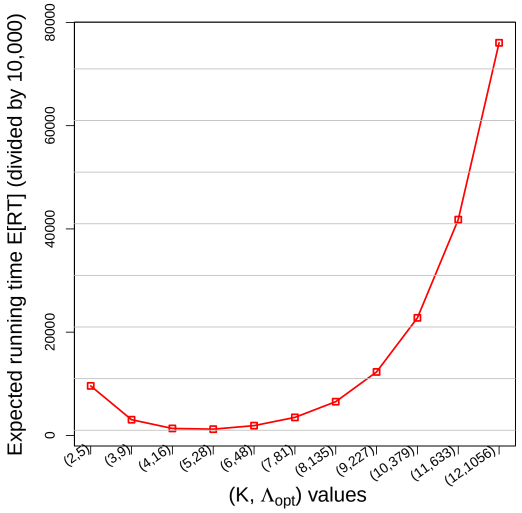


Fig. 2. Expected running time $E[RT]$ for specific (K, Λ_{opt}) values

nodes, for fault tolerance purposes. Map tasks receive the subset D^s stored on the node they run. To illustrate the framework, Charlie's cluster consists of two nodes C_1 and C_2 that hold data and a master node C_0 which coordinates the functionality of the cluster. The master node splits each data set as submitted by Alice and Bob in two subsets of equal size. Each subset in our example has one row. By setting the replication factor ρ to 1, each subset of A' and B' is distributed to different nodes. For example, Bf_{A2}, Bf_{B1} are stored on C_1 and Bf_{A1} and Bf_{B2} are stored on C_2. If we set $\rho = 2$, then each subset should be replicated twice, which would result to the total omnipresence of subsets to nodes. We present two large-scale linkage strategies that use the described cluster of nodes. The first strategy, denoted by St_1, consists of one Map/Reduce job, while the second strategy, that is St_2, utilizes two chain Map/Reduce jobs. Both strategies have two distinct logical steps, (a) the generation of the Key^js and (b) the matching of the formulated Bloom filter pairs.

5.1. Generation of the Key^js

When Charlie submits the Map/Reduce job of linkage, Algorithm 1 runs on each node by the map tasks M_1 and M_2. For each Bloom filter contained in a subset D^s, the Key^js are constructed by applying the corresponding permutations, as agreed by Alice and Bob and shown in Algorithm 1 in line 7. For illustration purposes, by letting $K = 2$, Λ_{opt} is set to 2 (to be precise Λ_{opt} should be set to 6, given $\delta = 0.1$ and $\vartheta = 0.40$). On the compute node C_1, which holds Bf_{A2} and Bf_{B1}, M_1 builds the following Key^js for Bf_{A2}: <1B,3,1> and <2B,0,5>. The map task M_2, running on C_2, builds also for Bf_{B2} the same Key^js. The labels 1B and 2B denote that the corresponding Key^js refer to the first and the second blocking group respectively. The Key^js, along with the corresponding Ids, are emitted to the partitioner tasks (Algorithm 1, line 10). This phase is common for both linkage strategies.

Algorithm 1 Transformation of Bloom filters into $<Key^j, Id>$ pairs in the map phases of both St_1 and St_2

```
Input: $D^s$, $\pi_k^j$s permutations, $K$
Output: list ($<Key^j, Id>$)
 1: $\Lambda_{opt}$ = $computeOptimal(K)$
 2: for i=1 to $|D^s|$ do
 3:     $Id \leftarrow D_i^s.Id$
 4:     $Bf \leftarrow D_i^s.Bf$
 5:     for j=1 to $\Lambda_{opt}$ do
 6:         for k=1 to $K$ do
 7:             $Key^j = Concat(min\{\pi_k^j(Bf)\})$
 8:         end for
 9:     end for
10:     emit($Key^j$, $Id$)
11: end for
```

5.2. Matching of the Formulated Bloom Filter Pairs

The *Id*s of the Bloom filters corresponding to a unique Key^j are routed to the same reduce task R_x, by the partitioner tasks. Then in strategy St_1, each R_x calculates the Jaccard distance of each Bloom filter pair, as illustrated in Sect. 3.3 and shown in Algorithm 2, lines 2 and 4. If the distance of a pair is below or equal to a predefined threshold (ϑ) then it is classified as a matched pair, otherwise as a non-matched pair (Algorithm 2 lines 7, 9). In our running example, the Key^1=<1B,3,1>, which refers to the first blocking group and it is generated by both Bf_{A2} and Bf_{B2}, has as a result their grouping to the same block and their classification as a matched pair. It is also noted in Fig. 3, where the strategy St_1 is outlined, that these Bloom filters also exhibit the same keys in the second blocking group, which results to a redundant distance computation.

The notation $[Id]$ in Algorithm 2, in the input statement, denotes that we get a list of *Id*s for each Key^j. In lines 3 and 5 function $get(\cdot)$ retrieves a Bloom filter from a data store. The same function uses a naive cache mechanism in order to cache locally a limited amount of the retrieved Bloom filters, for efficient use in subsequent computations that these Bloom filters participate. In our implementation the data store is a relational database. Linking two data sets of $200,000$ Bloom filters each, the Min-Hash scheme produces around $130,000,000$ pairs ($K = 5, \Lambda_{opt} = 29$), including duplicates which may be up to 12% of the total number of pairs.

In order to avoid the calculation of distance of duplicate pairs across the R_xs, in strategy St_2, we employ two chain Map/Reduce jobs. The map phase of the first job includes the generation of the Key^js, as illustrated in Algorithm 1, while during the reduce phase, the Bloom filter pairs are formulated without computing their distance (Algorithm 3).

Algorithm 2 Matching of the grouped Bloom filters in the reduce phase in St_1

```
Input: list(<$Key^j$, [$Id$] >)
Output: list ({"M","U"}, $Id_A$, $Id_B$)
 1: for each $Key^j$ do
 2:     for each $Id_A \in A'$ do
 3:         $Bf_A \leftarrow get(Id_A)$
 4:         for each $Id_B \in B'$ do
 5:             $Bf_B \leftarrow get(Id_B)$
 6:             if $d_J(Bf_A, Bf_B) \leq \vartheta$ then
 7:                 output("M", $Id_A$, $Id_B$);
 8:             else
 9:                 output("U", $Id_A$, $Id_B$);
10:             end if
11:         end for
12:     end for
13: end for
```

Algorithm 3 Formulation of Bloom filter pairs in the first reduce phase in St_2

```
Input: list(<$Key^j$, [$Id$] >)
Output: list ($Id_A$, $Id_B$, $\cdot$)
for each $Key^j$ do
    for each $Id_A \in A'$ do
        for each $Id_B \in B'$ do
            output($Id_A$, $Id_B$, $\cdot$);
        end for
    end for
end for
```

The second map phase is an identity function that simply passes its input data to the second reduce phase, where unique record pairs, as represented by their *Id*s, are distributed to the reduce tasks which perform the distance computations. This generating-unique-pairs strategy limits the comparisons only to the unique Bloom filter pairs, as specified by oBfJ. However, the use of a relational database to provide the Bloom filters results to long running time, because each pair requires the retrieval of two Bloom filters since the probability of having at least one of them cached is very low. On the contrary to St_1, as illustrated in line 3 of Algorithm 2, the Bloom filter of A' is retrieved once from the database, it is then cached and compared to all the Bloom filters of B' for the specific Key^j (line 5). The two strategies are outlined in Figs. 3 and 4 respectively.

6. Evaluation

In the experiments we evaluate oBfJ in terms of (a) the accuracy in finding the encoded matched record pairs, (b) the accuracy in finding the truly matched record pairs, (c) the efficiency in reducing the number of candidate record pairs and (d) the scalability in voluminous data sets. For (a),(b) and (c) we use semi-synthetic data sets of 50,000 records each, extracted from the NCVR [2] list. Each record includes 4 fields, namely *Id*, *LastName*,

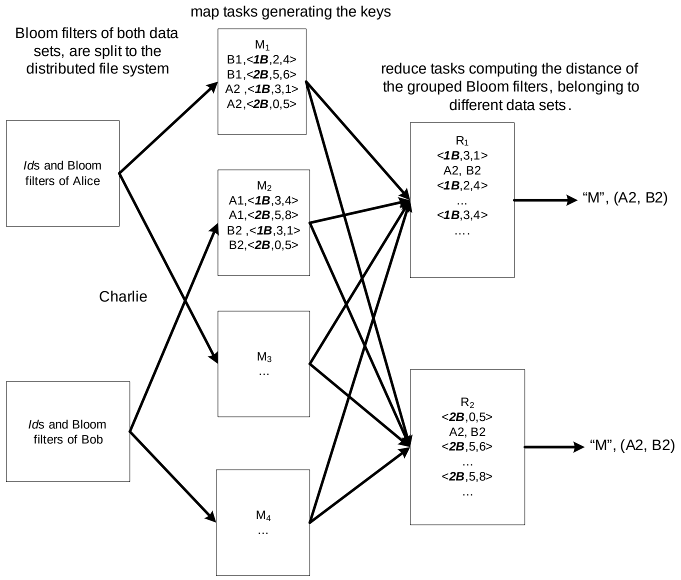

map tasks generating the keys
Bloom filters of both data sets, are split to the distributed file system
Ids and Bloom filters of Alice
Ids and Bloom filters of Bob
Charlie
M1
B1,<1B,2,4>
B1,<2B,5,6>
A2 ,<1B,3,1>
A2,<2B,0,5>
M2
A1,<1B,3,4>
A1,<2B,5,8>
B2 ,<1B,3,1>
B2,<2B,0,5>
M3
...
M4
...
reduce tasks computing the distance of the grouped Bloom filters, belonging to different data sets.
R1
<1B,3,1>
A2, B2
<1B,2,4>
...
<1B,3,4>
....
R2
<2B,0,5>
A2, B2
<2B,5,6>
...
<2B,5,8>
...
"M", (A2, B2)
"M", (A2, B2)

Fig. 3. Outline of strategy St_1

Algorithm 4 Matching of the grouped Bloom filters in the second reduce phase in St_2

```
Input: list($Id_A$, $Id_B$, [·])
Output: list ({"M","U"}, $Id_A$, $Id_B$)
$Bf_A \leftarrow get(Id_A)$
$Bf_B \leftarrow get(Id_B)$
if $d_J(Bf_A, Bf_B) \leq \vartheta$ then
    output("M", $Id_A$, $Id_B$);
else
    output("U", $Id_A$, $Id_B$);
end if
```

FirstName and *Address*. We develop a software prototype that extracts data sets, A and B of user-defined size and perturbation frequency, from the NCVR list [2]. Records are chosen randomly for perturbation, according to the specified frequency. For each perturbed record, two fields are chosen randomly and undergo (a) a single perturbation scheme (Pt_1) or (b) a double perturbation scheme (Pt_2). Insert, edit, delete and transpose operations are used to perturb the values of those chosen fields. The perturbation frequency of the records is set to 0.3. Field-level Bloom filters are created with size L equal to 500 bits, by using 15 cryptographic hash functions for each bigram, as proposed in [40]. All experiments, except for the large-scale ones (see Sect. 6.3), are executed in a dual-core Pentium PC of 8 GB RAM.

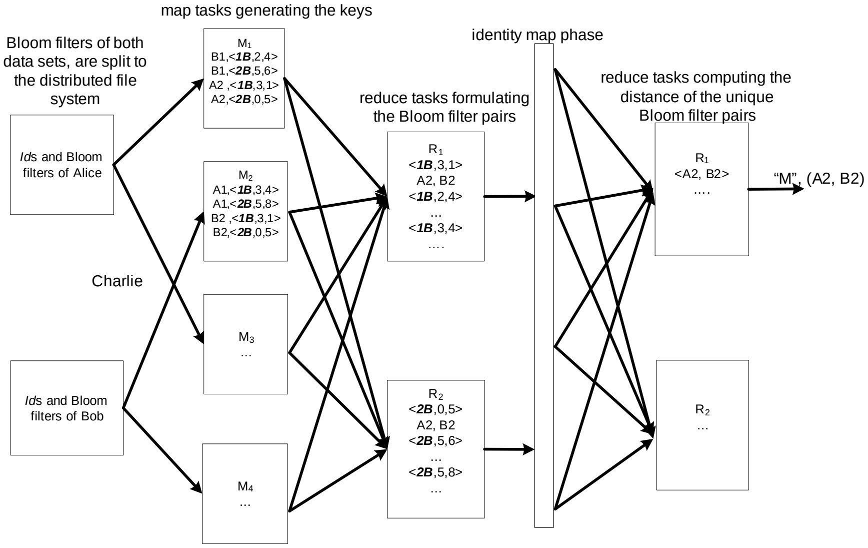

map tasks generating the keys
Bloom filters of both data sets, are split to the distributed file system
Ids and Bloom filters of Alice
Ids and Bloom filters of Bob
Charlie
M1
B1,<1B,2,4>
B1,<2B,5,6>
A2 ,<1B,3,1>
A2,<2B,0,5>
M2
A1,<1B,3,4>
A1,<2B,5,8>
B2 ,<1B,3,1>
B2,<2B,0,5>
M3
...
M4
...
reduce tasks formulating the Bloom filter pairs
R1
<1B,3,1>
A2, B2
<1B,2,4>
...
<1B,3,4>
....
R2
<2B,0,5>
A2, B2
<2B,5,6>
...
<2B,5,8>
...
identity map phase
reduce tasks computing the distance of the unique Bloom filter pairs
R1
<A2, B2>
....
R2
...
"M", (A2, B2)

Fig. 4. Outline of strategy St_2

6.1. Selected Measures

The Pairs Completeness (*PC*), the Pairs Quality (*PQ*) and the Reduction Ratio (*RR*) metrics [12] are employed to evaluate the efficiency of our methodology in finding the truly matched record pairs with respect to accuracy. The set of the identified encoded matched record pairs is denoted by $\mathcal{M}$ and the set of the original matched record pairs by M. Accuracy is measured by the PC_M metric, which is equal to $|\mathcal{M} \cap M|/|M|$. The efficiency in generating mostly matched pairs with respect to the number of the candidate pairs is indicated by the *PQ* metric which is equal to $|\mathcal{M} \cap M|/|CR|$, where *CR* is the set of the candidate record pairs. The Reduction Ratio illustrates the percentage in the reduction of the comparison space and it is equal to $1.0 - |CR|/|A' \times B'|$.

6.2. Comparative Results

The efficiency of the Min-Hash LSH scheme and the accuracy of the Bloom filter-based encoding jointly determine the overall performance of our methodology. In Fig. 5 we demonstrate the results by performing experiments for various values for K ($\vartheta = 0.40$ and $\delta = 0.1$) along with their corresponding Λ_{opt} values, namely $(5, 29)$,$(6, 49)$,$(7, 82)$ and $(8, 136)$. Furthermore, for a specific value of K, we perform experiments by setting Λ to several arbitrary values. For example, for $K = 5$, we set $\Lambda = <49, 82, 136>$ or for $K = 8$, we set $\Lambda = <29, 49, 82>$. We observe from Fig. 5, for a given value of K, as Λ approaches from the left to Λ_{opt}, the PC_M rate increases but after Λ_{opt}, the PC_M rate remains almost stable. This implies that the largest part of the matched Bloom filter pairs is identified (each pair is returned with probability at least 0.9) and consequently the same happens for the largest part of the original matched record pairs. An amount of

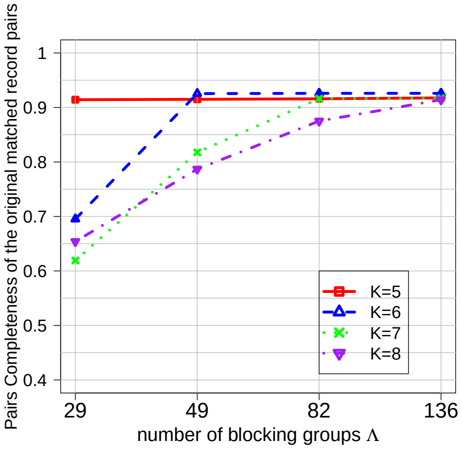
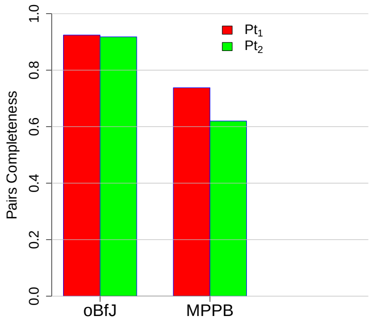


Fig. 5. Trying several values of K along with the corresponding Λ_{opt} values

the original matched record pairs is missed, approximately 8%, due to improper Bloom filter-based encoding, where the initial distances are not preserved.

We compare our methodology to the Multidimensional Privacy-Preserving Blocking (MPPB) method as proposed in [25]. MPPB is based on the K-Medoids algorithm [27] for creating clusters from elements of public reference sets. Next, records are classified to those clusters and they are encoded into Bloom filters, where the Sorted Neighborhood algorithm [21] is used for the selection of the Bloom filter pairs for comparison within each cluster. The Dice coefficient [11] is employed to measure the similarity between those pairs. Thresholds in the comparisons are set to 0.85 and 0.80 for the two perturbation schemes respectively.



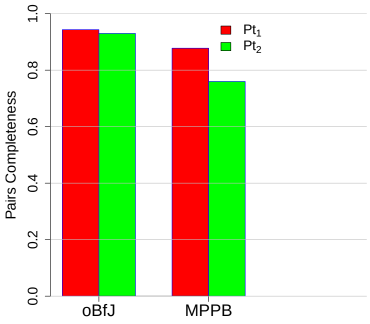


Fig. 6. The Pairs Completeness rates of the original matched record pairs

Fig. 7. The Pairs Completeness rates of the matched Bloom filter pairs

We set $K_{opt} = 5$ and $\delta = 0.1$ as the basic parameters of oBfJ which yield $\Lambda_{opt} = 29$ blocking groups for the *Pt*$_1$ scheme with $\vartheta = 0.40$ and $\Lambda_{opt} = 45$ groups for the *Pt*$_2$ scheme with $\vartheta = 0.45$. Duplicate distance computations are prevented by utilizing an efficient $\mathcal{O}(1)$ structure (like a HashMap in Java programming language) which stores the *Id*s of each compared pair only once, therefore discarding any computation that has already been conducted. Highly accurate results are generated by oBfJ, as shown in Fig. 6, with PC_M rate above 92%. In order to increase the PC_M rate for the MPPB method, we had to increase the window size to a large value (> 70), which had dramatic impact on the performance and a slight increase in the PC_M rate, clearly outperformed by oBfJ, as shown in Figs. 6 and 7.

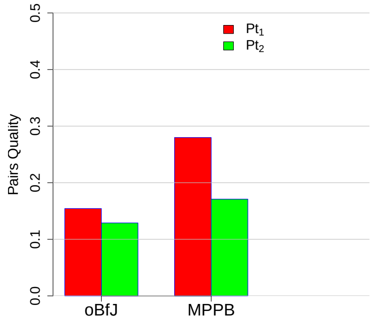


Fig. 8. The Pairs Quality *PQ* rates

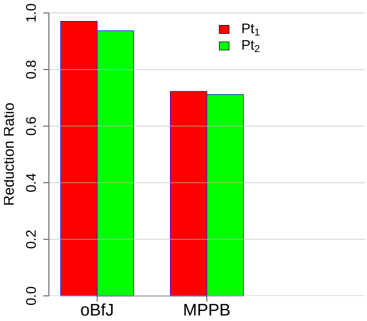


Fig. 9. The Reduction Ratio *RR*

Although, oBfJ displays high *RR*, constantly above 94% (Fig. 9), the *PQ* rates remain rather low (Fig. 8). Therefore, although oBfJ is efficient in reducing the initial $|A' \times B'|$ number of pairs, there are still candidate non-matched Bloom filter pairs with distance d_J close to ϑ but slightly above it. The MPPB method displays better *PQ* rate but lower *RR*.

6.3. Scalability

Scalability is measured in terms of execution time, utilizing a Map/Reduce system. More specifically, Apache Hadoop 1.0.4 is used, as the Map/Reduce implementation, running on a cluster of four compute nodes, which are virtual machines of *Okeanos* [3], the cloud service of the Greek Research and Academic Community. Two data sets are linked, which consist of $300,000$ records. By adding more physical compute nodes in the reduce phase, execution time is decreased as shown in Fig. 10. Permutations are shared across the map tasks by utilizing the distributed file system, since each task runs on its own execution environment. It is clearly shown that by utilizing a relational database, strategy *St*$_2$ becomes inefficient and its main feature, which is the unique distance computation of duplicate pairs, does not contribute a lot to the overall system's performance.

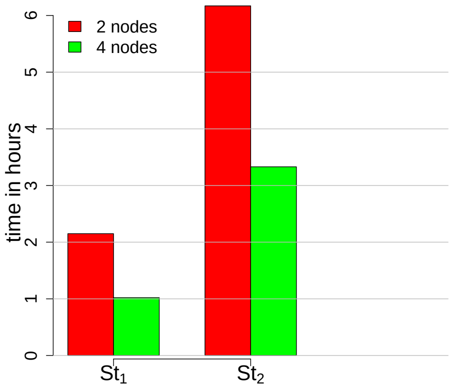


Fig. 10. Execution time exhibited, utilizing a Map/Reduce system using two concrete strategies

7. Conclusions

Linking huge collections of anonymized records is an intriguing problem in the core of the domain of Privacy-Preserving Record Linkage. In this paper, we introduce an LSH-based distributed method on top of a Map/Reduce computational paradigm. Records are encoded into Bloom filters, in order to protect privacy of the underlying data, and then they are submitted to a trusted third party, which splits and distributes them to a file system with replication capabilities. The huge collection of records demands the utilization of a large number of compute nodes that leverage scalability. The Min-Hash LSH technique is applied by producing from each Bloom filter some Key^js which correspond to some blocking groups. The number of those blocking groups (Λ_{opt}) and the number of the hash functions for each Key^j (K_{opt}) are optimized with respect to accuracy and performance. Bloom filters are retrieved from a relational database and they are cached locally to each node for subsequent computations. Those Bloom filters that exhibit identical Key^js are grouped together and their Jaccard distance is calculated on a pairwise manner. We believe that the utilization of a more sophisticated and robust cache mechanism in combination with the use of a distributed storage system, that facilitates efficient queries to massive data sets, is an interesting research direction that may boost the performance of strategy St_2.

References

1. Financial Times US bank mergers (2010), `http://www.ft.com/cms/s/0/5e969dc0-c01f-11df-b77d-00144feab49a.html`
2. North Carolina voter registration database (2013), `ftp://www.app.sboe.state.nc.us/data`

3. Okeanos, GRNET's cloud service for the Greek research and academic community (2013), `https://okeanos.grnet.gr/home/`
4. Aggarwal, C., Yu, P.: The igrid index: Reversing the dimensionality curse for similarity indexing in high dimensional space. In: Proc. 6th ACM international conference on Knowledge discovery and data mining (SIGKDD). pp. 119 – 129 (2000)
5. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Communications of the ACM 51(1), 117 – 122 (2008)
6. Baxter, R., Christen, P., Churches, T.: A comparison of fast blocking methods for record linkage. In: Proc. ACM Workshop on Data Cleaning Record Linkage and Object Consolidation (SIGKDD). pp. 25 – 27 (2003)
7. Bloom, B.: Space/time trade-offs in hash coding with allowable errors. Communications of ACM 13(7), 422 – 426 (1970)
8. Broder, A.Z.: On the resemblance and containment of documents. In: Proc. Compression and Complexity of Sequences. pp. 21 – 29 (1997)
9. Broder, A.Z., Charikar, M., Frieze, A., Mitzenmacher, M.: Minwise independent permutations. In: Proc. 30th ACM symposium on Theory of computing (STOC). pp. 327–336 (1998)
10. Broder, A.Z., Glassman, S., Manasse, M., Zweig, G.: Syntactic clustering of the web. In: Selected papers from the 6th international conference on World Wide Web archive. pp. 1157 – 1166 (1997)
11. Christen, P.: Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Data-Centric Systems and Applications (2012)
12. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE Transactions on Knowledge and Data Engineering (TKDE) 12(9) (2012)
13. Churches, T., Christen, P.: Some methods for blindfolded record linkage. BioMed Central Medical Informatics and Decision Making 4(9) (2004)
14. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. pp. 73 – 78 (2003)
15. Cohen, W., Richman, J.: Learning to match and cluster large high-dimensional datasets for data integration. In: Proc. 8th ACM international conference on Knowledge discovery and data mining (SIGKDD). pp. 475 – 480 (2002)
16. Dean, J., Ghemawat, S.: Mapreduce: Simplifed data processing on large clusters. Communications of the ACM 51(1), 107 – 113 (2008)
17. Durham, E.: A Framework For Accurate Efficient Private Record Linkage. Ph.D. thesis, Vanderbilt University, US (2012)
18. Durham, E., Kantarcioglu, M., Xue, Y., Toth, C., Kuzu, M., Malin, B.: Composite Bloom filters for secure record linkage. IEEE Transactions on Knowledge and Data Engineering 99(PrePrints) (2013)
19. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: Proc. 25th International Conference on Very Large Databases (VLDB). pp. 518–529 (1999)
20. Goodman, J., O'Rourke, J., Indyk, P.: Handbook of Discrete and Computational Geometry. CRC (2004)
21. Hernandez, M., Stolfo, S.: Real world data is dirty: Data cleansing and the merge/purge problem. Data Mining And Knowledge Discovery 2(1), 9 – 37 (1988)
22. Inan, A., Kantarcioglou, M., Bertino, E., Scannapieco, M.: A hybrid approach to private record linkage. pp. 496 – 505. Proc. IEEE 24th International Conference on Data Engineering (ICDE) (2008)
23. Inan, A., Kantarcioglu, M., Ghinita, G., Bertino, E.: Private record matching using differential privacy. In: Proc. 13th International Conference on Extending Database Technology (EDBT). pp. 123 – 134 (2010)
24. Jin, L., Li, C., Mehrotra, S.: Efficient record linkage in large datasets. pp. 137 – 146. 'Proc. 8th International Conference on Database Systems for Advanced Applications (DASFAA) (2003)

25. Karakasidis, A., Verykios, V.: A sorted neighborhood approach to multidimensional privacy preserving blocking. In: IEEE International Conference on Data Mining Workshops. pp. 937 – 944 (2012)
26. Karapiperis, D., Verykios, V.: A distributed framework for scaling up LSH-based computations in privacy preserving record linkage. In: Proc. 6th Balkan Conference in Informatics (BCI). pp. 102 – 109. ACM (2013)
27. Kaufman, L., Rousseeuw, P.: Clustering by means of medoids. Reports of the Faculty of Mathematics and Informatics (1987)
28. Kim, H., Lee, D.: Fast iterative hashed record linkage for large-scale data collections. In: Proc. 13th International Conference on Extending Database Technology (EDBT). pp. 525 – 536 (2010)
29. Krawczyk, H., Bellare, M., Canetti, R.: HMAC: keyed-hashing for message authentication, internet rfc 2104 (1997), http://tools.ietf.org/html/rfc2104
30. Kuzu, M., Kantarcioglou, M., Durham, E., Malin, B.: A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. In: Proc. 11th international conference on Privacy enhancing technologies (PETS). pp. 226 – 245 (2011)
31. Kuzu, M., Kantarcioglu, M., Inan, A., Bertino, E., Durham, E., Malin, B.: Efficient privacy-aware record integration. In: Proc. 16th International Conference on Extending Database Technology (EDBT). pp. 167 – 178 (2013)
32. Motwani, R., Raghavan, P.: Randomized Algorithms. Cambridge University Press (1995)
33. Navarro, G.: A guided tour to approximate string matching. ACM Computing Surveys 33(1), 31 – 88 (2001)
34. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Eurocrypt. pp. 223 – 238 (1999)
35. Pang, C., Gu, L., Hansen, D., Maeder, A.: Privacy-preserving fuzzy matching using a public reference table. Intelligent Patient Management 189, 71 – 89 (2009)
36. Rajaraman, A., Ullman, J.: Mining of Massive Datasets, chap. Finding Similar Items. Cambridge University Press (2010)
37. S., T.: Privacy–preserving string comparisons in record linkage systems: a review. Information Security Journal: A Global Perspective 17(5), 253 – 266 (2008)
38. Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A.: Privacy preserving schema and data matching. In: Proc. ACM international conference on Management of data (SIGMOD). pp. 653 – 664 (2007)
39. Schneier, B.: Applied cryptography: protocols, algorithms, and source code in C. Wiley (1996)
40. Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using Bloom filters. BMC Medical Informatics and Decision Making 9 (2009)
41. Shakhnarovich, G., Darrell, T., Indyk, P.: Locality-sensitive hashing using stable distributions. In: Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing). MIT Press (2004)
42. Vatsalan, D., Christen, P., Verykios, V.: An efficient two-party protocol for approximate matching in private record linkage. pp. 125 – 136 (2011)
43. Vatsalan, D., Christen, P., Verykios, V.: Efficient two-party private blocking based on sorted nearest neighborhood clustering. pp. 1949 – 1958 (2013)
44. Vatsalan, D., Christen, P., Verykios, V.: A taxonomy of privacy-preserving record linkage techniques. Information Systems (Elsevier) 38(6), 946 – 969 (2013)
45. Weber, R., Schek, H., Blott, S.: A quantitative analysis and performance study for similarity search methods in high dimensional spaces. In: Proc. 24th International Conference on Very Large Data Bases (VLDB). pp. 194 – 205 (1998)
46. Winkler, W.: String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. American Statistical Association pp. 778 – 783 (1990)

Dimitrios Karapiperis is a PhD student with the Hellenic Open University. He also holds an MSc degree from the University of York (UK) and a BSc degree from the Technological Institute of Thessaloniki (Greece.). His research interests lie primarily in the area of Privacy-Preserving Record Linkage where he focuses on developing similarity algorithms, data structures, approximation schemes and scalable solutions that rely on various randomization schemes. More specifically, he deals with Locality-Sensitive Hashing-based efficient searching techniques, secure matching on Bloom filter-based encoded data and distributed/scalable solutions using the Map/Reduce programming paradigm.

Vassilios S. Verykios received the Diploma degree in Computer Engineering from the University of Patras in Greece in 1992, and the MS and the PhD degrees from Purdue University in 1997 and 1999, respectively. From 1999 to 2001 he was with the Faculty of Information Systems in the College of Information Science and Technology at Drexel University, Pennsylvania, as a tenure track Assistant Professor. From 2001 to 2005 he held various research positions at Intracom SA, SingularLogic SA and CTI and visiting Assistant Professor positions in the University of Thessaly, Hellenic Open University and the University of Patras. From 2005 to 2011 he was an Assistant Professor in the Department of Computer and Communication Engineering at the University of Thessaly in Volos, Greece. Since January of 2011, he is an Associate Professor in the School of Science and Technology, at the Hellenic Open University in Patras, Greece.

Received: February 15, 2014; Accepted: June 10, 2014.

Computer Science and Information Systems 11(2):765–778 DOI: 10.2298/CSIS121212041R

An Asymmetric Index to Compare Trapezoidal Fuzzy Numbers

Julio Rojas-Mora[1] and Jaime Gil-Lafuente[2]

[1] Institute of Statistics
Faculty of Economics and Administrative Sciences
Universidad Austral de Chile
Valdivia, Chile
julio.rojas@uach.cl
[2] Dpt. of Business Economics and Organization
Faculty of Economics and Administrative Sciencies
Universitat de Barcelona
Barcelona, Spain
j.gil@ub.edu

Abstract. In this paper, we present a tool to help reduce the uncertainty presented in the resource selection problem when information is subjective in nature. The candidates and the “ideal” resource required by evaluators are modeled by fuzzy subsets whose elements are trapezoidal fuzzy numbers (TrFN). By modeling with TrFN the subjective variables used to determine the best among a set of resources, one should take into account in the decision-making process not only their expected value, but also the uncertainty that they express. A mean quadratic distance (MQD) function is defined to measure the separation between two TrFN. It allows us to consider the case when a TrFN is wholly or partially contained in another. Then, for each candidate a weighted mean asymmetric index (WMAI) evaluates the mean distance between the TrFNs for each of the variables and the corresponding TrFNs of the “ideal” candidate, allowing the decision-maker to choose among the candidates. We apply this index to the case of the selection of the product that is best suited for a “pilot test” to be carried out in some market segment.

Keywords: Fuzzy sets, distance, resource selection, pilot test, subjective information, marketing.

1. Introduction

Currently, companies that commercialize a broad range of products, sometimes with high product rotation, tend to have departments specializing in consistently providing technically viable and economically feasible ideas.

The “bank of ideas” that such organizations have produces a multitude of reference points for the delivery of new goods or services. Virtually, any previously screened idea may be of commercial interest, but many of the failures in its commercialization may arise from presenting it to the wrong kind of client. The design of a product must meet customer requirements, making it essential to carry out a post-design verification test with a sample of the intended market segment, before the product’s full-scale commercialization [18].

We have to remember that R&D departments typically have a set of different products, from which they must choose the one best suited to the targeted segment of the market.

The suitability of these products is evaluated with subjective variables, which makes the application of techniques based on fuzzy subsets a straightforward matter.

By means of the theory of fuzzy subsets [22], we can select the best among a group of candidates when information is subjective in nature or comes from expert processed statistical data. As an example of this line of research, we can observe the work of Chen and Wang [6] and its application to search for the perfect home [7], the application to databases developed by Yang et al. [21], the process carried out by the International Olympic Committee for the selection of the venue of the 1st Summer Youth Olympic Games [11], and the evaluation of traffic police centers performance by Sadi-Nezhad and Damghani [17]. This process is based on the comparison between fuzzy numbers, a line of research that have been a cornerstone of the fuzzy sets theory, and of which we can cite the work, for example, of Tran and Duckstein [19], of Zeng and Guo [23], of Zhang, Zhang and Mei [24], of Lee, Pedrycz and Sohn [14], and of Guha and Chakraborty [10].

In this same line of research, we offer a mean quadratic distance (MQD) function between trapezoidal fuzzy numbers (TrFN) that allows us to consider the case when a TrFN is wholly or partially contained in another. In traditional distance functions between TrFN, when a fuzzy number is totally contained in another there is a distance between them, and this distance is symmetric. Nevertheless, in our index, the separation from contained to container is zero, as uncertainty will not allow us to distinguish between them. On the other hand, there is a distinction between container and contained, as some part of the former is beyond the limits of the latter. From this point of view, our MQD calculates the distance needed to "project" one fuzzy number into the another.

Then, the MQD is used for multi-criteria decision making analysis through a weighted mean asymmetric index (WMAI). We apply this index to model subjective information on candidate products for a "pilot test" that will be carried out in a market segment. Among them, there is the need to select the one that fits better to this market and that will help collect valuable information in order to introduce the best possible product. Both, the ideal product and the candidate products, are modeled using TrFN.

The remaining of this paper is organized as follows. Section 2 briefly describes some of the newest research made on ranking of fuzzy numbers. The theoretical framework of fuzzy sets and fuzzy numbers is laid out in Section 3. Our asymmetric index is described in Section 4. Section 5 contains the application to an example. Finally, we present some conclusions in Section 6.

2. Recent work

The traditional application of indexes to evaluate the difference between fuzzy numbers has been in the process of ranking. This line of research was initiated by Jain [12], and followed by many researchers who developed algorithms with different levels of complexity. Recent works that can be mentioned are those of Asady [2,3] who worked on revisions of [20] and [4] to overcome problems that only later arose. In [8], Chou et al. worked on improvements over the widely used method developed by Chen [5], based on the utility value of a fuzzy number. Allahviranloo et al. [1] developed a method based on weighted interval-value approximations defined as crisp-set approximations of fuzzy numbers. Rezvani [16] defined a similarity measure based on the perimeter of generalized fuzzy numbers. Finally, the work more closely related to our approach is that of Nejad and

Mashinchi [15], based on the evaluation of areas between fuzzy numbers in order to rank them.

3. Fuzzy subsets and fuzzy numbers

In cases when information is subjective, models based on the theory of fuzzy subsets can help the decision maker in the evaluation of the alternatives. In this section, we will present the basic definitions of both, fuzzy sets and fuzzy numbers, that we will use throughout this paper.

Definition 1. *A fuzzy subset $\tilde{A}$ can be represented by a set of pairs composed of the elements x of the universal set X, and a grade of membership $\mu_{\tilde{A}}(x)$:*

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) \mid x \in X \, , \, \mu_{\tilde{A}}(x) \in [0,1]\} \, . \tag{1}$$

Definition 2. *An α-cut of a fuzzy subset $\tilde{A}$ is defined by:*

$$A_\alpha = \{x \in X : \mu_{\tilde{A}} \geq \alpha\}, \tag{2}$$

i.e., the subset of all elements that belong to $\tilde{A}$ at least in a degree α.

Definition 3. *A fuzzy subset $\tilde{A}$ is convex, iff:*

$$\lambda x_1 + (1 - \lambda x_2) \in A_\alpha \, \forall x_1, x_2 \in A_\alpha, \, \alpha, \lambda \in [0,1] \, , \tag{3}$$

i.e., all the points in $[x_1, x_2]$ must belong to A_α, for any α.

Definition 4. *A fuzzy subset $\tilde{A}$ is normal, iff:*

$$\max\left(\mu_{\tilde{A}}(x)\right) = 1, \, \forall x \in X. \tag{4}$$

Definition 5. *The core of a fuzzy subset $\tilde{A}$ is:*

$$N_{\tilde{A}} = \{x : \mu_{\tilde{A}}(x) = 1\} \, . \tag{5}$$

Definition 6. *A fuzzy number $\tilde{A}$ is a normal, convex fuzzy subset with domain in $\mathbb{R}$ for which:*

1. *$\bar{x} := N_{\tilde{A}}$, card $(\bar{x}) = 1$, and*
2. *$\mu_{\tilde{A}}(x)$ is at least piecewise continuous.*

The mean value [25] $\bar{x}$, also called maximum of presumption [13], identifies a fuzzy number in such a way that the proposition “about 9” can be modeled with a fuzzy number whose maximum of presumption is $x = 9$. As Zimmermann [25] explains, for computational simplicity there is a tendency to call “fuzzy number” any normal, convex fuzzy subset whose membership function is, at least, piecewise continuous, without taking into consideration the uniqueness of the maximum of presumption. Thus, this definition will include “fuzzy intervals”, fuzzy numbers in which $\bar{x}$ covers an interval. As a matter of fact, Dubois and Prade [9] called them “flat fuzzy numbers”.

Definition 7. *A TrFN is defined by the membership function:*

$$\mu_{\tilde{A}}(x) = \begin{cases} \frac{x-x_1}{x_2-x_1}, & \textit{if } x_1 \leq x < x_2 \\ 1, & \textit{if } x_2 \leq x \leq x_3 \\ \frac{x_4-x}{x_4-x_3}, & \textit{if } x_3 < x \leq x_4 \\ 0 & \textit{otherwise.} \end{cases} \tag{6}$$

A TrFN is represented by a 4-tuple whose first and fourth elements correspond to the extremes from where the membership function begins to grow, and whose second and third components define the interval that limits the maximum of presumption, i.e., $\tilde{A} = (x_1, x_2, x_3, x_4)$.

4. The asymmetric index

In the traditional distance functions between TrFN (based on Manhattan, Euclidean, and, in general, Minkowski's distance functions), the distance between two TrFNs $\tilde{A} = (a_1, a_2, a_3, a_4)$ and $\tilde{B} = (b_1, b_2, b_3, b_4)$, when the former is totally contained in the later, is different from zero.

Nevertheless, one should study the assumption that due to uncertainty the distance from contained to container should be zero, as it would be impossible to distinguish between them. On the other hand, there is some value of distance between container and contained, as some portion of the former is outside the limits of the later.

The objective is, therefore, to calculate a mean index $D(\tilde{A}, \tilde{B})$ that shows the distance needed to make $\tilde{A} \subseteq \tilde{B}$. This function implies a sort of "projection" of $\tilde{A}$ into $\tilde{B}$ in the four regions of the set $Z = \{L_1, L_2, R_1, R_2\}$ shown in Figure 1.

Definition 8. *Given the set $\{\overline{x_1x_2}, \overline{x_2x_3}, \ldots, \overline{x_{n-1}x_n}, \overline{x_nx_1}\}$, where $\overline{x_ix_{\hat{i}}}$ is a segment defined by the points P_i and $P_{\hat{i}}$, whereas x_i and $x_{\hat{i}}$ are the abscissas of these points, we will define the region Λ as the area inscribed in the convex polygon composed by the elements of this set.*

From this definition, we can say that the regions from Z are described as (see Figures 2, 3, 4 and 5):

$$L_1 = \begin{cases} \left\{\overline{a_1b_1}, \overline{b_1x'}, \overline{x'a_1}\right\}, & \text{if } a_1 < b_1 \text{ and } a_2 > b_2, \\ \left\{\overline{a_1b_1}, \overline{b_1b_2}, \overline{b_2a_2}, \overline{a_2a_1}\right\} & \text{if } a_1 \leq b_1 \text{ and } a_2 \leq b_2, \\ \left\{\overline{a_2b_2}, \overline{b_2x'}, \overline{x'a_2}\right\}, & \text{if } a_1 > b_1 \text{ and } a_2 < b_2, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{7}$$

$$R_1 = \begin{cases} \left\{\overline{b_4a_4}, \overline{a_4a_3}, \overline{a_3b_3}, \overline{b_3b_4}\right\}, & \text{if } b_3 \leq a_3 \text{ and } b_4 \leq a_4, \\ \left\{\overline{b_4a_4}, \overline{a_4x''}, \overline{x''b_4}\right\}, & \text{if } b_3 > a_3 \text{ and } b_4 < a_4, \\ \left\{\overline{b_3a_3}, \overline{a_3x''}, \overline{x''b_3}\right\}, & \text{if } b_3 < a_3 \text{ and } b_4 > a_4, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{8}$$

$$R_2 = \begin{cases} \{\overline{b_4a_1}, \overline{a_1a_2}, \overline{a_2b_3}, \overline{b_3b_4}\}, & \text{if } a_1 \geq b_4, \\ \{\overline{b_3a_2}, \overline{a_2x'''}, \overline{x'''b_3}\}, & \text{if } a_1 < b_4 \text{ and } a_2 > b_3, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{9}$$

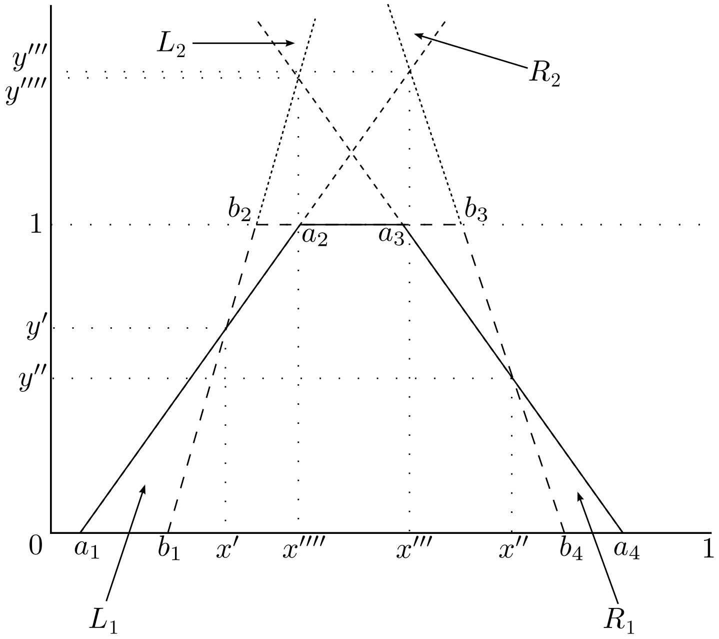


Fig. 1. Regions where the index is calculated.

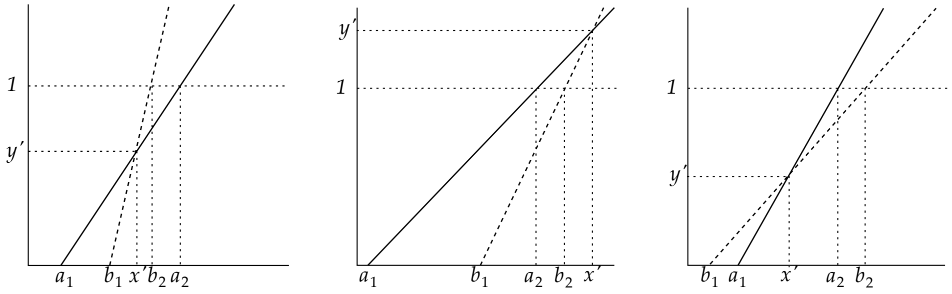

1
y'
a_1 b_1 x' b_2 a_2
y'
1
a_1 b_1 a_2 b_2 x'
1
y'
b_1 a_1 x' a_2 b_2

Fig. 2. Variants of the L_1 region.

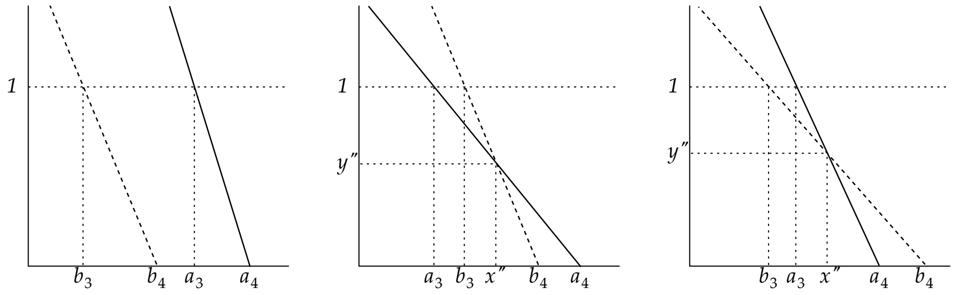


Fig. 3. Variants of the R_1 region.

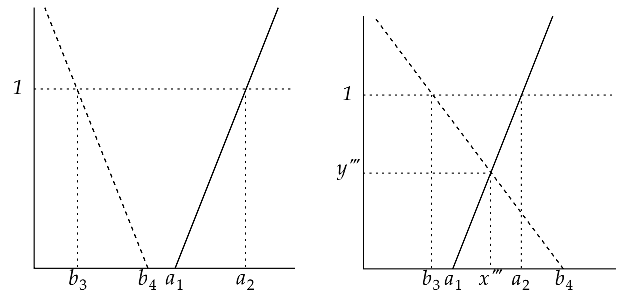


Fig. 4. Variants of the R_2 region.

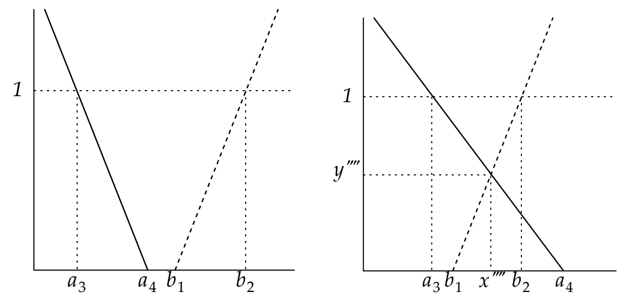


Fig. 5. Variants of the L_2 region.

$$L_2 = \begin{cases} \{\overline{a_4b_1}, \overline{b_1b_2}, \overline{b_2a_3}, \overline{a_3a_4}\}, & \text{if } a_4 \leq b_1, \\ \{\overline{a_3b_2}, \overline{b_2x''''}, \overline{x''''a_3}\}, & \text{if } a_3 < b_2 \text{ and } a_4 > b_1, \\ \emptyset, & \text{otherwise.} \end{cases} \quad (10)$$

Proposition 1. *Two TrFN $\tilde{A}$ and $\tilde{B}$ intersect at maximum four points (x', y'), (x'', y''), (x''', y'''), and (x'''', y'''') such that:*

$$x' = \frac{a_1b_2 - b_1a_2}{a_1 - a_2 - b_1 + b_2} \; ; y' = \frac{a_1 - b_1}{a_1 - a_2 - b_1 + b_2}, \quad (11)$$

$$x'' = \frac{a_3b_4 - b_3a_4}{a_3 - a_4 - b_3 + a_4} \; ; y'' = \frac{b_4 - a_4}{a_3 - a_4 - b_3 + a_4}, \quad (12)$$

$$x''' = \frac{a_2b_4 - a_1b_3}{a_2 - a_1 + b_4 - b_3} \; ; y''' = \frac{b_4 - a_1}{a_2 - a_1 + b_4 - b_3}, \quad (13)$$

$$x'''' = \frac{a_4b_2 - a_3b_1}{a_4 - a_3 + b_2 - b_1} \; ; y'''' = \frac{a_4 - b_1}{a_4 - a_3 + b_2 - b_1}. \quad (14)$$

Proof. Given the equation of the straight line:

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1} \left(x - x_1\right), \quad (15)$$

if segments $\overline{a_1 a_2}$ and $\overline{b_1 b_2}$ intersect at (x', y'), then:

$$y' - 0 - \frac{1-0}{a_2 - a_1}(x' - a_1) = y' - 0 - \frac{1-0}{b_2 - b_1}(x' - b_1) \tag{16}$$

$$\frac{x' - a_1}{a_2 - a_1} = \frac{x' - b_1}{b_2 - b_1} \tag{17}$$

$$(b_2 - b_1 - a_2 + a_1)\, x' = -a_2 b_1 + a_1 b_1 + a_1 b_2 - a_1 b_1 \tag{18}$$

$$x' = \frac{a_1 b_2 - a_2 b_1}{a_1 - a_2 - b_1 + b_2}. \tag{19}$$

Substituting (19) on the equation of the segment $\overline{a_1 a_2}$, we solve for y':

$$y' = \frac{\frac{a_1 b_2 - a_2 b_1}{a_1 - a_2 - b_1 + b_2} - a_1}{(a_2 - a_1)} \tag{20}$$

$$= \frac{\frac{-a_2 b_1 - a_1^2 + a_1 a_2 + a_1 b_1}{a_1 - a_2 - b_1 + b_2}}{(a_2 - a_1)} \tag{21}$$

$$= \frac{(a_1 - b_1)(a_2 - a_1)}{(a_1 - a_2 - b_1 + b_2)(a_2 - a_1)} \tag{22}$$

$$= \frac{a_1 - b_1}{a_1 - a_2 - b_1 + b_2} \tag{23}$$

It is obvious that when segments $\overline{a_1 a_2}$ and $\overline{b_1 b_2}$ are parallel, they will not intersect and, thus, the point (x', y') does not exists. This demonstration is equivalent for the other intersection points.

The first step in the calculation of our index is based on a mean quadratic distance (MQD) function that measures the separation of both TrFNs in the defined regions.

Definition 9. *The MQD function between two TrFN $\tilde{A}$ and $\tilde{B}$ for each region $\zeta \in Z$ is obtained through:*

$$D_\zeta = \frac{\int_{\alpha_\zeta}^{\beta_\zeta} (b_\zeta - a_\zeta)^2 \, dy}{\beta_\zeta - \alpha_\zeta}, \tag{24}$$

where a_ζ is the equation of the line that limits ζ on the left side, b_ζ is the equation of the line that limits this region on the right side, both expressed in terms of y, and $\{\alpha_\zeta, \beta_\zeta\} \in [0,1]$, $\alpha_\zeta \leq \beta_\zeta$, are the integration limits in y that we find through Figures 2, 3, 4 and 5, and Proposition 1.

It seems evident that, from Definition 9, the area of $\tilde{A}$ contained in $\tilde{B}$ generates an MQD equal to zero, like, for example, that from segment $\overline{y' a_2}$ to segment $\overline{y' b_2}$ in Figure 1.

The closed form expressions of (24) for all regions are:

$$D_{L_1} = \begin{cases} \frac{a_1^2 + a_1 a_2 + a_2^2 - 2\, a_1 b_1 + b_1^2 - a_1 b_2 + b_2^2 - a_2 b_1 + b_1 b_2 - 2\, a_2 b_2}{3}, & \text{if } a_1 \leq b_1 \text{ and } a_2 \leq b_2, \\ \frac{(a_1 - b_1)^2}{3}, & \text{if } a_1 < b_1 \text{ and } a_2 > b_2, \\ \frac{(a_2 - b_2)^2}{3}, & \text{if } a_1 > b_1 \text{ and } a_2 < b_2, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

$$D_{R_1} = \begin{cases} \frac{a_3^2+a_3a_4+a_4^2-2\,a_3b_3+b_3^2-a_3b_4+b_4^2-a_4b_3-2\,a_4b_4+b_3b_4}{3}, & \text{if } b_3 \leq a_3 \text{ and } b_4 \leq a_4, \\ \frac{(a_4-b_4)^2}{3}, & \text{if } b_3 > a_3 \text{ and } b_4 < a_4, \\ \frac{(a_3-b_3)^2}{3}, & \text{if } b_3 < a_3 \text{ and } b_4 > a_4, \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

$$D_{R_2} = \begin{cases} \frac{a_1^2+a_1a_2+a_2^2-a_1b_3+b_3^2-2\,a_1b_4+b_4^2-2\,a_2b_3-a_2b_4+b_3b_4}{3}, & \text{if } a_1 \geq b_4, \\ \frac{(a_2-b_3)^2}{3}, & \text{if } a_1 < b_4 \text{ and } a_2 > b_3, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

$$D_{L_2} = \begin{cases} \frac{a_3^2+a_3a_4+a_4^2-a_3b_1+b_1^2-2\,a_3b_2+b_2^2-a_4b_2-2\,a_4b_1+b_1b_2}{3}, & \text{if } a_4 \leq b_1, \\ \frac{(a_3-b_2)^2}{3}, & \text{if } a_3 < b_2 \text{ and } a_4 > b_1, \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Definition 10. *The mean asymmetric index between two TrFN obtained from (25), (26), (27) and (28) is:*

$$D\left(\tilde{A},\tilde{B}\right) = \begin{cases} \sqrt{\frac{SD}{N}}, & \textit{if } N > 0, \\ 0, & \textit{otherwise.} \end{cases} \tag{29}$$

where:

$$SD = D_{L_1} + D_{L_2} + D_{R_1} + D_{R_2},$$
$$N = \mathbb{1}_{\{D_{L_1}>0\}} + \mathbb{1}_{\{D_{L_2}>0\}} + \mathbb{1}_{\{D_{R_1}>0\}} + \mathbb{1}_{\{D_{R_2}>0\}}.$$

Remark 1. It is straightforward to observe that $N = 0 \iff \tilde{A} \subset \tilde{B}$. Thus $D(\tilde{A},\tilde{B}) = 0 \iff \tilde{A} \subset \tilde{B}$.

Remark 2. Because we would like to know the mean distance for the regions of $\tilde{A}$ not covered by $\tilde{B}$, the resulting asymmetric index sometimes behaves as a hemimetric, which implies:

1. $\exists \tilde{A} \neq \tilde{B} : D(\tilde{A},\tilde{B}) = 0$. Again, we would be saying that $\tilde{A} \subset \tilde{B} \iff D(\tilde{A},\tilde{B}) = 0$.
2. $\exists \tilde{A} \neq \tilde{B} : D(\tilde{A},\tilde{B}) \neq D(\tilde{B},\tilde{A})$. As a matter of fact, there are two cases where $D(\tilde{A},\tilde{B}) = D(\tilde{B},\tilde{A})$. Firstly, if for a given point:

$$b_1 \in \mathbb{R}, B = \left(b_1, b_1 + a_4 - a_3, b_1 + a_4 - a_2, b_1 + a_4 - a_1\right), \tag{30}$$

i.e., when $\tilde{A}$ has the inverse shape of $\tilde{B}$. Secondly, if $\tilde{A} = \tilde{B} + c$, with $c \in \mathbb{R}$, i.e., when $\tilde{A}$ and $\tilde{B}$ have the same shape.

The mean asymmetric index based on the MQD function models the separation between the assessment $\tilde{P}_i$ given to some candidate resource P, in any given characteristic i, and the required value $\tilde{I}_i$ asked from an "ideal" candidate I in that same characteristic. We will now define the weighted mean asymmetric index (WMAI) between all the characteristics evaluated in a candidate and the required levels of those characteristics.

Definition 11. *The weighted mean asymmetric index (WMAI) between $\tilde{P} = \{\tilde{P}_1, \tilde{P}_2, \ldots, \tilde{P}_n\}$ and $\tilde{I} = \{\tilde{I}_1, \tilde{I}_2, \ldots, \tilde{I}_n\}$ will be:*

$$\delta(P, I) = \sum_{i=1}^{n} \omega_i \cdot D\left(\tilde{P}_i, \tilde{I}_i\right), \tag{31}$$

where $\omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is a vector of weights such that $\sum_{i=1}^{n} \omega_i = 1$ and $\omega_i \neq 0$.

5. Application of the WMAI to the selection of a product for a pilot test

In order to illustrate the application of the methodology, we will use an example based on the introduction of a dairy product into the market.

5.1. Description of the target market segment

The first thing that needs to be defined is a mathematical descriptor that numerically and accurately reflects the market segment that the Marketing Department is interested in reaching. A hypothetical dairy company wants to introduce a new product for the market segment defined by the following characteristics:

1. Health-conscious consumer.
2. With an age above 50 years.
3. Willing to pay a high price for healthier diary product.
4. Interested in new technology driven products, but not a consumer totally devoted to technology. \item Looking for a product that can be carried around and consumed at any time.

Given this information, the marketing department models the target segment with the fuzzy set $\tilde{S} = \tilde{s}_i$, as we can see in Table 1.

Table 1. Target segment modeled as a fuzzy set.

	c_1	c_2	c_3	c_4	c_5
$\tilde{S}$=	(0.7,1)	(0.6,1)	(0.5,0.8,1)	(0.7,0.8,0.9)	(0.8,1)

Each fuzzy number $\tilde{s}_i$ represents an assessment of the ideal level that the target segment has on each one of the characteristics in the set $C = c_i$. Both, the characteristics and the meaning of the assessments given to this particular segment, are defined as:

1. c_1- Health-consciousness: A consumer is regarded as health-conscious if at least 70% of his food purchases are done taking health into consideration.
2. c_2- Maturity: A consumer is considered mature if he has already lived more than 60% of his life expectancy.
3. c_3- Price level: In this case, the consumer prefers to buy products in the top 50% of the price scale, with a maximum preference for products below 80%.
4. c_4 - Novelty of organoleptic or technological characteristics: the consumer prefers to buy a new product between 70 and 90% of the times, with a maximum of 80%, if new flavors or properties are also involved.
5. c_5 - Easiness of transportation and consumption: when a consumer buys a product he prefers the top 80% in easiness of transportation and consumption, meaning he wants a product that can be carried around with no worries of spillage or spoilage, and that needs almost no additional procedures beyond opening its package for its consumption.

The characteristics will be weighted according to Table 2.

Table 2. Weights for the characteristics.

	c_1	c_2	c_3	c_4	c_5
ω=	0.1	0.35	0.25	0.25	0.05

5.2. Description of the candidates for the pilot test

The R&D department of the company has decided that the pilot test should be run with one of the following five products from the set $\mathfrak{P} = \{\tilde{P}^{(j)}\}, j = 1, \ldots, 5$:

1. $\tilde{P}^{(1)}$- Fructose sweetened soy yogurt with fruits: It has been assessed as a product with healthy properties above the average, without a defined age group, an average price, no real novelty in flavor or technology, and with the same limitations in transportability and consumption as most dairy products.
2. $\tilde{P}^{(2)}$- Inulin sweetened Greek yogurt enriched with calcium: Being a Greek yogurt means this is a fatty product, although this is compensated with the substitution of complex sugars with inulin, so this product is considered to be middle-of-the-pack in healthiness, although with some uncertainty. Nonetheless, by enriching it with calcium, the targeted age group is certainly mature. Price, as well as its technological appeal due to the novelty of inulin, ranges in the middle upper echelon. Of course, a Greek yogurt is a product to keep refrigerated and eaten with a spoon, which limits its transportability.
3. $\tilde{P}^{(3)}$- Aspartame sweetened, fat free chocolate pudding enriched with Omega-3: By removing fat while adding Omega-3 and aspartame, this product can be considered remarkably healthy. Age groups from young adulthood to elderly will be equally attracted to consume it. The price is above average, but not the most expensive. However, this is not a technologically driven product as all its characteristics are found in

many products. This product has the same problems of transportability described for the Greek yogurt.

4. $\tilde{P}^{(4)}$- Digestion helping, cucumber yogurt soup with *Lactobacillus casei*: A yogurt soup that is designed to aid digestion is regarded as an almost perfect element of a healthy diet. Age groups for this product go from middle-aged people to consumers entering maturity, as this flavor is not a favorite of younger groups and the introduction of external bacteria might cause problems to older groups. Price is at the top of the line, as this is considered a gourmet food. Even if, technologically speaking, there is nothing new in this product, its flavor and concept are novel enough to put this product above average in preferences of people looking for new products with new flavors. Finally, easy consumption is not quite feasible with a soup.
5. $\tilde{P}^{(5)}$- Energy boosting, tropical fruits flavored smoothie, enriched with amino acids and taurine: This product might only be considered healthy in the group of people that have an active night life, as well as for those that practice sports. This makes it more suitable for age groups ranging from young adulthood to early maturity. It is in the most expensive level, has a high impact on people looking for technology driven foods and is the easiest product to use, although it is recommended to consume it cold.

Thus, each $\tilde{P}^{(j)} = \{\tilde{\mu}_{i,j}\}$ is a fuzzy set with the same number of elements as $\tilde{S}$, modeled according to the information gathered as shown in Table 3.

Table 3. Assessments made by the company experts on the new products.

	c_1	c_2	c_3	c_4	c_5
$\tilde{P}^{(1)}$=	(0.5,0.6,0.6,0.7)	(0.1,0.1,1,1)	(0.5,0.5,0.5,0.5)	(0,0,0.1,0.1)	(0.5,0.5,0.5,0.5)
$\tilde{P}^{(2)}$=	(0.4,0.5,0.5,0.6)	(0.7,0.7,1,1)	(0.7,0.7,0.8,0.8)	(0.7,0.7,0.9,0.9)	(0.4,0.4,0.4,0.4)
$\tilde{P}^{(3)}$=	(0.8,0.8,1,1)	(0.3,0.3,1,1)	(0.8,0.8,0.8,0.8)	(0.7,0.7,0.7,0.7)	(0.4,0.4,0.4,0.4)
$\tilde{P}^{(4)}$=	(0.8,0.8,0.9,0.9)	(0.4,0.4,0.7,0.7)	(1,1,1,1)	(0.5,0.5,0.7,0.7)	(0.1,0.1,0.1,0.1)
$\tilde{P}^{(5)}$=	(0.2,0.2,0.4,0.4)	(0.3,0.3,0.7,0.7)	(1,1,1,1)	(1,1,1,1)	(0.9,0.9,0.9,0.9)

5.3. Results

We now proceed to calculate the distance between each one of the new products proposed for a test run and the target segment. The product closest to the target segment will be the one selected for this test.

$$\delta\left(\tilde{P}^{(j)}, \tilde{S}\right) = \sum_{i=1}^{5} \omega_i \, D\left(\tilde{\mu}_{i,j}, \tilde{s}_i\right)$$

$$\begin{aligned} \delta\left(\tilde{P}^{(1)},\tilde{S}\right) &= 0.1\cdot 0.12+0.35\cdot 0.5+0.25\cdot 0.17+0.25\cdot 0.7+0.05\cdot 0.3 \\ &= 0.42. \end{aligned}$$

$$\delta\left(\tilde{P}^{(2)},\tilde{S}\right) = 0.1\cdot 0.21+0.35\cdot 0+0.25\cdot 0.06+0.25\cdot 0.06+0.05\cdot 0.4 = 0.07.$$

$$\delta\left(\tilde{P}^{(3)},\tilde{S}\right) = 0.1\cdot 0+0.35\cdot 0.3+0.25\cdot 0+0.25\cdot 0.06+0.05\cdot 0.4 = 0.14.$$

$$\delta\left(\tilde{P}^{(4)},\tilde{S}\right) = 0.1\cdot 0+0.35\cdot 0.2+0.25\cdot 0.12+0.25\cdot 0.18+0.05\cdot 0.7 = 0.18.$$

$$\delta\left(\tilde{P}^{(5)},\tilde{S}\right) = 0.1\cdot 0.41+0.35\cdot 0.3+0.25\cdot 0.12+0.25\cdot 0.15+0.05\cdot 0 = 0.21.$$

As we can see, the product closest to the market segment targeted with our pilot test is the Greek yogurt, a product that meets most of the requirements even if transportability is not a distinguished feature. The chocolate pudding is the closest competitor to the Greek yogurt, but seems that it would need a strong marketing campaign to introduce it in the age group of the target segment. The worst suited is the soy yogurt, maybe because it is a generic product that can be used as a baseline, to check how well targeted are other products.

We can present the results in terms of preference using the precedence operator:

$$\tilde{P}^{(1)} \prec \tilde{P}^{(5)} \prec \tilde{P}^{(4)} \prec \tilde{P}^{(3)} \prec \tilde{P}^{(2)}.$$

This means that the Greek yogurt is the product best suited for a test run, then the chocolate pudding, the cucumber yogurt soup, the energy boosting smoothie, and the soy yogurt, respectively.

6. Conclusions

In this work, we presented a mean quadratic distance (MQD) function that calculates the distance needed to have one TrFN contained by another. This MQD function generates a distance of magnitude zero for the areas of the first TrFN overlapped by the second, while considering distances bigger than zero for those non-overlapped.

By calculating the MQD over these non-overlapped regions, we obtain a weighted mean asymmetric index (WMAI) of separation between two vectors of TrFN. The WMAI, calculated from assessments given over a set of characteristic of a group of candidates to the "ideal" requirements on those characteristics, is then obtained and used as the decision variable.

This methodology is applied to the evaluation of new products that are candidates for a test run in a particular target segment. It allows the comparison of products with different features, some close to the target segment and some others far from it, to take a decision accordingly. A product test run might be an expensive affair, and as such, the decision on which product use for it has to be as fully supported as possibly.

A good feature of this methodology is that statistical information, adequately transformed or processed by experts, can be used together with subjective information, usually disregarded by statistical methods, to get a more comprehensive view of the situation and give better help to the decision maker.

References

1. Allahviranloo, T., Abbasbandy, S., Saneifard, R.: An approximation approach for ranking fuzzy numbers based on weighted interval-value. Mathematical and Computational Applications 16(3), 588 (2011)
2. Asady, B.: The revised method of ranking lr fuzzy number based on deviation degree. Expert Systems with Applications 37(7), 5056–5060 (2010)
3. Asady, B.: Revision of distance minimization method for ranking of fuzzy numbers. Applied Mathematical Modelling 35(3), 1306–1313 (2011)
4. Asady, B., Zendehnam, A.: Ranking fuzzy numbers by distance minimization. Applied Mathematical Modelling 31(11), 2589–2598 (2007)
5. Chen, S.H.: Ranking fuzzy numbers with maximizing set and minimizing set. Fuzzy sets and Systems 17(2), 113–129 (1985)
6. Chen, S.H., Wang, C.C.: Representation, ranking, distance, and similarity of fuzzy numberswith step form membership function using k-preference integration method. IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th 2, 801–806 (2001)
7. Chen, S.H., Wang, C.C.: House selection using fuzzy distance of trapezoidal fuzzy numbers. In: Proceedings of the Sixth International Conference on Machine Learning and Cybernetics. Hong Kong (2007)
8. Chou, S.Y., Dat, L.Q., Yu, V.F.: A revised method for ranking fuzzy numbers using maximizing set and minimizing set. Computers & Industrial Engineering 61(4), 1342–1348 (2011)
9. Dubois, D., Prade, H.: Fuzzy real algebra: some results. Fuzzy Sets and Systems 2, 327–348 (1979)
10. Guha, D., Chakraborty, D.: A new approach to fuzzy distance measure and similarity measure between two generalized fuzzy numbers. Applied Soft Computing 10(1), 90–99 (2010)
11. IOC Panel of Experts: 1st Summer Youth Olympic Games in 2010. Tech. rep., IOC (2007)
12. Jain, R.: Decision making in the presence of fuzzy variables. IEEE Transactions on Systems, Man and Cybernetics 6, 698–702 (1976)
13. Kaufmann, A., Gupta, M.M.: Introduction to Fuzzy Arithmetic. Van Nostrand Reinhold, New York (1985)
14. Lee, S., Pedrycz, W., Sohn, G.: Design of similarity and dissimilarity measures for fuzzy sets on the basis of distance measure. International Journal of Fuzzy Systems 11(2), 67–72 (2009)
15. Nejad, A.M., Mashinchi, M.: Ranking fuzzy numbers based on the areas on the left and the right sides of fuzzy number. Computers & Mathematics with Applications 61(2), 431–442 (2011)
16. Rezvani, S.: A new method for ranking in perimeters of two generalized trapezoidal fuzzy numbers. International Journal of Applied 2(3), 85–92 (2012)
17. Sadi-Nezhad, S., Damghani, K.K.: Application of a fuzzy topsis method base on modified preference ratio and fuzzy distance measurement in assessment of traffic police centers performance. Applied Soft Computing 10(4), 1028–1039 (2010)
18. Tennant, G.: Design for Six Sigma: Launching New Products and Services Without Failure. Gower, Aldershot (2002)

19. Tran, L., Duckstein, L.: Comparison of fuzzy numbers using a fuzzy distance measure. Fuzzy Sets and Systems 130, 331–341 (2002)
20. Wang, Z.X., Liu, Y.J., Fan, Z.P., Feng, B.: Ranking< i> l</i>–< i> r</i> fuzzy number based on deviation degree. Information Sciences 179(13), 2070–2077 (2009)
21. Yang, M.S., Hung, W.L., Chang-Chien, S.J.: On a Similarity Measure between LR-Type Fuzzy Numbers and Its Application to Database Acquisition. International Journal of Intelligent Systems 20(10), 1001–1016 (2005)
22. Zadeh, L.A.: Fuzzy sets. Information and Control 8(3), 338–353 (1965)
23. Zeng, W., Guo, P.: Normalized distance, similarity measure, inclusion measure and entropy of interval-valued fuzzy sets and their relationship. Information Sciences 178, 1334–1342 (2008)
24. Zhang, H., Zhang, W., Mei, C.: Entropy of interval-valued fuzzy sets based on distance and its relationship with similarity measure. Knowledge-Based Systems 22(6), 449–454 (2009)
25. Zimmermann, H.: Fuzzy Sets: Theory and its Applications. Springer, 4th edn. (2005)

Julio Rojas-Mora has a systems engineering degree from the University of Los Andes (Venezuela). His academic career includes an Advanced Studies Diploma (D.E.A.) in Statistics and Operations Research from the Universidade de Santiago de Compostela (Spain) in 2006, and a Ph.D. from the Business Economics and Organization Department of the Universitat de Barcelona (Spain) in 2011. After finishing a postdoctoral fellowship at the UMR Espace CNRS of the Université d'Avignon, he joined the Institute of Statistics of the Universidad Austral de Chile. His research focuses on the comparison of subjectively evaluated individuals using the fuzzy sets theory.

Jaime Gil-Lafuente is a Professor at the Universitat de Barcelona where he obtained his PhD in Business Administration. He has published more than 250 articles in journals, books and conference proceedings, including journals such as Advances in Consumer Research or Annals of Operations Research. He has participated in a wide range of scientific committees and as a reviewer in international journals. He is an academician of the Royal Academy of Doctors of Spain, the Academy Delphinale of France and the Illustrious Iberoamerican Academy of Doctors of Mexico.

Received: December 12, 2012; Accepted: November 15, 2013.

Computer Science and Information Systems 11(2):779–796 DOI:10.2298/CSIS130125042G

Approximation to the Theory of Affinities to Manage the Problems of the Grouping Process

Anna M. Gil-Lafuente[1]and Anna Klimova[1]

[1] Department of Business Administration, University of Barcelona
Av. Diagonal 690, 08034 Barcelona, Spain
{amgil, aklimova}@ub.edu

Abstract. New economic and enterprise needs have increased the interest and utility of the methods of the grouping process based on the theory of uncertainty. A fuzzy grouping (clustering) process is a key phase of knowledge acquisition and reduction complexity regarding different groups of objects. Here, we considered some elements of the theory of affinities and uncertain pretopology that form a significant support tool for a fuzzy clustering process. A Galois lattice is introduced in order to provide a clearer vision of the results. We made an homogeneous grouping process of the economic regions of Russian Federation and Ukraine. The obtained results gave us a large panorama of a regional economic situation of two countries as well as the key guidelines for the decision-making. The mathematical method is very sensible to any changes the regional economy can have. We gave an alternative method of the grouping process under uncertainty.

Keywords: fuzzy logic, clustering process, affinities, pretopology, Galois Lattices.

1. Introduction

The intelligibility of the universe for each person depends on his aptitude to group or classify different objects. Identification of object types is one of the first phases of knowledge acquisition [1]. “Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning” [2]. “A fundamental operation in data mining is the partitioning of a set of objects represented by data into homogeneous groups or clusters” [3].

The analysis of the process of groups making (clustering analysis) is fundamental nowadays for understanding the complex processes of various substantive areas such as medicine [4-9], chemical industry [10-12], engineering [13-16], image processing [17-19], business [20-35] and others.

At the beginning Boolean logic has appeared as one of the most powerful mathematical tools able to make different groups giving the adequate solutions [36]. It has been commonly used in the classical clustering analysis methods that are the most important techniques in the information procedure [1]. Clustering has been extensively studied in machine learning, databases, and statistic from various perspectives. There are two classical hard clustering methods. Statistical clustering methods [37-39] partition objects according to some proximity (similarity or dissimilarity) measures, whereas conceptual clustering methods [40,41] cluster objects according to the concepts the objects have. For automatic clustering, a method of determining proximity between feature vectors as well as a method for determining representatives of clusters is required. Hard clustering method is more adequate for clustering condition with clear boundary and preciseness in databases [1].

Although a basis of classical clustering techniques remains the same, some changes have been recently introduced into their structure. With increasing complexity of the technological and economic processes, an important aspect of cluster analysis comes to the fore which is the rigidity of the partition that is necessary to find. Many authors have proposed a fuzzy setting as the adequate approach to deal with this problem. Fuzzy set theory, introduced by Zadeh [42], gives an idea of uncertainty of belonging, which is described by a membership function. Fuzzy set theory represents a new model of grouping process that allows classifying the elements under uncertainty. During the last decades, extensive research has been carried out with respect to the investigation of fuzzy clustering techniques for classification. See Refs. 43-54.

In the early 90s a new approach to fuzzy clustering process was proposed by some authors [33].The concept of affinity between the elements has been considered as an "angular stone" of segmentation process under uncertainty. The theory of the affinities represents a generalization of the relations of similarity extending the field of their performance to the field of the rectangular matrices.

The notion of similarity and affinity represent different ways of expressing the concept of neighborhood. Similarity indicates a specific resemblance, either partial or total, between two physical or imaginary objects [51-52]. Affinity refers to a collective behavior of objects with respect to certain specific criteria, even when these criteria are not particularly well specified. But both of these notions indicate whether two or more objects are related under adequate conditions with respect to explicit criteria for affinity [45].

In this paper we have proposed an extension of the algorithm based on the theory of affinities for the formation of homogeneous groups of economic regions of Russian Federation and Ukraine. With this in mind, we have pretended to offer an alternative solution to the problems faced by every person in charge of the distribution (investment) of company financial resources. It was essential, for us, to make a good economic and financial assessment of the regions of Russia Federation and Ukraine, providing a wide and faithful basis for making-decision process under uncertainty.

In this study we have also considered some elements of uncertain "pretopology" that have served us as a mathematical support to implement the theory of affinities to the practical case. See Refs. 23,26,55-58. Some concepts of combinatorial analysis [59] have been introduced too. As a culminating element of our paper we have used a Galois lattice. We considered it an effective tool to structure the obtained results.

2. Recent Backgrounds

In a previous part we have already told about general tendencies of the development of the mathematical methods and algorithms related to the grouping/ clustering process. Now let us continue briefly with the recent approaches that are closely related to the present paper and represent special interest for the current topic.

Gil Aluja in his paper titled “Clans, affinities and Moore’s fuzzy pretopology” [60] proposed some new elements that renovate the structures of the economic thought based on the geometric idea. The elements are based on the concept of pretopology. The author considered that it was impossible to conceive, today, formal structures able to represent the Darwininan conception of the economic behavior without appealing to this part of combinatory mathematics. The author proposed the suitable schemes for the treatment of the economic and management phenomenon using the theory of clans, affinities and the Moore’s closures.

In one of his next papers Gil Aluja emphasized that very often in situations of uncertainty in portfolio management it is difficult to apply the numerical methods based on the linearity principle. In this case, nonnumeric techniques were used to assess the situations with a nonlinear attitude. The authors proposed to use a concept of grouping providing, by this manner, good solutions to the problems of the homogeneous groupings. The Pichat and affinities algorithms were proposed to reduce the number of elements of the power sets of the titles listed in the Stock Exchange and to group them correctly [61].

Among other recent works represented by the same authors dedicated to the main topic of this paper we can mention the following “Decision-making techniques with similarity measures and OWA operators” [62]. In the last paper several aggregation techniques such as the Hamming distance, the adequacy coefficient and the index of maximum and minimum level that were very useful for decision-making were considered. These were the useful mathematical tools able to determine similarities between different elements of the system helping in decision making process especially in business and economic areas (in production management in this concrete case).

The article titled “Approaches to managing sustainability among enterprises” [63] analyzed the important changes a business environment where the companies perform experienced during the last decades. It was characterized as uncertain and unstable. Basing on the idea of the importance and complexity the sustainability management represents for the companies the authors proposed the use of flexible tools based on fuzzy logic such as the Theory of Affinities, the Clans Theory, Hungarian Algorithm, etc., in order to assist the enterprises in making decisions and help them to improve management with stakeholders contributing to the treatment of the problems in the future.

Another field of the social sciences where the algorithms and technics based on the mathematics of uncertainty were used to find out the coherent solutions for decision making was a human resource management. The purpose of the article titled “A personal selection model using Galois group theory” was to propose a personal selection model based on the comparison between the qualifications of prospective candidates. The model based on the Galois group theory was used to group different candidates depending on the common characteristics that meet a certain level. See Ref. 64.

In particular, another author proposed to use the Pichat algorithm, among other methods, in the sports area. He demonstrated that this tool was a great help in decision making related to the phenomenon of grouping/ clustering. This algorithm showed whether the members of a human group/a sport team were sufficiently and properly interrelated or not and, as a consequence, whether its performance was optimal or not. See Ref. 65.

Because of the limitations of the article text extension we'll finish our explications here. We hope that this brief analysis showed the importance and usefulness of the grouping methods based on the mathematics of uncertainty in different social sciences nowadays especially in economy and management.

3. The Theory of the Affinities

The theory of the affinities represents a generalization of the relations of similarity extending the field of their performance to the field of the rectangular matrices. This theory is applied in a multitude company's management problems.

The attempt to generalize a notion of similarity was initiated by Jaime Gil Aluja and Arnold Kaufmann in the eighties. The theory of the affinities turned out to be a successful result of the work presented by these scientists at the IX European Congress of the Operative Investigation.

"We determine the affinities as those homogeneous orderly structured groups limited to the established levels which join all elements of two sets of different nature. These two sets are related between themselves by the own essence of the phenomenon which represent them" [33].

There are three main aspects which form the concept of affinity. The first one refers to the fact that the homogeneity of each group is related to the established level. According to the requirement of each characteristic (the element of one of the sets) the high level of a determiner is assigned and that is a threshold from which the homogeneity starts to exist. The second aspect shows a necessity of the fact that the elements of each set are to be connected between themselves by the certain rules of the nature or human will. The third aspect requires creating a structure which has a special order that is capable to be involved into the process of the decision making.

Over the last decades many approaches to the theory of affinities have been proposed and several useful techniques based on this theory have been provided. See Refs. 23-32,45,51,52,58,66-68.

4. Axioms for the Uncertain Pretopology

There are two elements such as a finite referential E and a "power set" $P(E)$ that have a different nature. Let us define uncertain pretopology. See Refs. 26,33,58.

Definition 1. A mapping Γ of *P*(*E*) in *P*(*E*) is uncertain pretopology of *E* if, and only if, the following axioms are given:

1) $\Gamma\varnothing = \varnothing$

2) $\forall \underset{\sim}{A}_k \in P(E): \underset{\sim}{A}_k \subset \Gamma\underset{\sim}{A}_k$

(where $\Gamma E = E$ and $\underset{\sim}{A}_k$ is a fuzzy subset)

Using the same terms as in determinism, a mapping Γ for all elements of *P*(*E*) will be called as an "adherent mapping" or "adherence". It is also possible to associate Γ with an "inner mapping" or "inner" *δ*, defined by:

$$\forall \underset{\sim}{A}_k \in P(E):\ \delta\underset{\sim}{A}_k = \overline{\Gamma\overline{\underset{\sim}{A}_k}} \qquad \textbf{(1)}$$

(where $\overline{\underset{\sim}{A}_k}$ is a complement of $\underset{\sim}{A}_k$).

Definition 2. $\underset{\sim}{A}_k$ is closed set for an uncertain pretopology if $\underset{\sim}{A}_k = \Gamma\underset{\sim}{A}_k$.

Definition 3. $\underset{\sim}{A}_k$ is open set for an uncertain pretopology if $\underset{\sim}{A}_k = \delta\underset{\sim}{A}_k$.

We considered the widest notion of uncertain pretopology. If new axioms are added then new uncertain pretopological spaces appear. They have a special interest for the treatment of the problems of different economic systems.

Definition 4. A mapping Γ of *P*(*E*) in *P*(*E*) is uncertain isotone pretopology of *E* if, and only if, the following axioms are carried out:

1) $\Gamma\varnothing = \varnothing$
2) $\forall \underset{\sim}{A}_k \in P(E): \underset{\sim}{A}_k \subset \Gamma\underset{\sim}{A}_k$, extensivity
3) $\forall \underset{\sim}{A}_k, \underset{\sim}{A}_l \in P(E): \left(\underset{\sim}{A}_k \subset \underset{\sim}{A}_l\right) \Rightarrow \left(\Gamma\underset{\sim}{A}_k \subset \Gamma\underset{\sim}{A}_l\right)$, isotony

Let us pay a special attention to one of the uncertain isotone pretopologies that satisfies the property, which is idempotency, and can be represented as the fourth axiom.

4) $\Gamma\left(\Gamma\underset{\sim}{A}_k\right) = \Gamma\underset{\sim}{A}_k$, idempotency

With these four axioms, we get Moore's closure which, moreover, has the property $\Gamma\varnothing = \varnothing$, not required in the axiomatic of this closure.

Moore's closure is of a special importance in the development of the algorithms capable of dealing with segmentation problems. So it is necessary to establish ways for obtaining Moore's closures on the basis of the concepts that could easily be found in the information supported by different institutions.

5. Obtaining a Moore's Closure

On the basis of previous practice, we have found two concepts that seem to adapt well to the exposed necessities. These are the notion of a Moore's family and a graph or fuzzy relation. See Refs. 33,51,52.

- Axiomatic of a Moore's family.

Definition 5. Let one family be determined through $F(E) \subset P(E)$. If there are two axioms:

1) $E \in F(E)$
2) $(\underset{\sim}{A}_k, \underset{\sim}{A}_l \in F(E)) \Rightarrow (\underset{\sim}{A}_k \cap \underset{\sim}{A}_l \in F(E))$,

then this is a Moore's family.

Moore's closure is obtained by the following way:

1) Once the family $F(E)$ is available it is necessary for each element $\underset{\sim}{A}_k \in P(E)$ to find a subset of elements of $F(E)$ that contain $\underset{\sim}{A}_k$ and will be referred to as $F_{\underset{\sim}{A}_k}(E)$.
2) For each $F_{\underset{\sim}{A}_k}(E)$ its intersection $\bigcap\limits_{F \in F_{\underset{\sim}{A}_k}(E)} F$ is obtained.
3) A mapping $M\underset{\sim}{A}_k$ is determined by the mode that $\bigcap\limits_{F \in F_{\underset{\sim}{A}_k}(E)} F$ will be corresponded to each element $\underset{\sim}{A}_k$, that is to say: $M\underset{\sim}{A}_k = \bigcap\limits_{F \in F_{\underset{\sim}{A}_k}(E)} F$.

So, Moore's closure was found using one family. There is only one closure for each family.

- The concept of relation is at the basis of a majority of algorithms of non-numerical mathematics of uncertainty, considered by Zadeh [54] and Kaufmann [51]. To formalize this relation there are two graphs that can be used represented either in matrix form or arrowlines.

  Let us consider a fuzzy relation $[\underset{\sim}{R}]$ between the elements of a referential $E = \{y_1, y_2, ..., y_n\}$, such as $[\underset{\sim}{R}] \subset E \times E$.

  1) With this information a Boolean graph is determined, taking a threshold α as follows:

  $$[R_\alpha] = \{(y_i, y_j) \subset E \times E / \mu_{\underset{\sim}{R}}(y_i, y_j) \geq \alpha\}, \alpha \in [0,1]. \qquad \textbf{(2)}$$

  2) Predecessors and successors of y_i limited to the level α are determined as:

$$\left.\begin{array}{l} y_j \text{ is a sucessor of} y_i \\ y_i \text{ is a predecessor of} y_j \end{array}\right\} \text{ if } (y_i, y_j) \in [R_\alpha]. \qquad \textbf{(3)}$$

These concepts allow us to find the connection to the right $R_\alpha^+ \underset{\sim}{A}_k$ and to the left $R_\alpha^- \underset{\sim}{A}_k$.

Definition 6. A connection to the right $R_\alpha^+ \underset{\sim}{A}_k$ or, a mapping R_α^+ of $P(E)$ in $P(E)$ such that for all $\underset{\sim}{A}_k \in P(E)$, R_α^+ is a subset of elements of E that are successors, limited to the level α, of all element that belong to $\underset{\sim}{A}_k$, expressed as:

$$R_\alpha^+ \underset{\sim}{A}_k = \{y_j \in E / (y_i, y_j) \in [R_\alpha], \forall\, y_i \in \underset{\sim}{A}_\alpha\}, \qquad \textbf{(4)}$$
$$\text{with } R_\alpha^+ \varnothing = E.$$

Definition 7. A left connection $R_\alpha^- \underset{\sim}{A}_k$ or, a mapping R_α^- of $P(E)$ in $P(E)$ such that for all $\underset{\sim}{A}_k \in P(E)$, R_α^- is a subset of the elements of E that are predecessors, limited to the level α, of all elements that belong to $\underset{\sim}{A}_k$. It is expressed as:

$$R_\alpha^- \underset{\sim}{A}_k = \{y_i \in E / (y_i, y_j) \in [R_\alpha], \forall\, y_j \in \underset{\sim}{A}_\alpha\}, \qquad \textbf{(5)}$$
$$\text{with } R_\alpha^- \varnothing = E.$$

3) The right connection $R_\alpha^+ \underset{\sim}{A}_k$ and the left one $R_\alpha^- \underset{\sim}{A}_k$ can be found directly by simple reading of the relation $[R_\alpha]$, in the following way:

$$\forall\, y_i \subset \underset{\sim}{A}_k \in P(E): R_\alpha^+ \underset{\sim}{A}_k = \bigcap_{y_i \subset \underset{\sim}{A}_k} R_\alpha^+ \{y_i\}, \qquad \textbf{(6)}$$

$$R_\alpha^- \underset{\sim}{A}_k = \bigcap_{y_i \subset \underset{\sim}{A}_k} R_\alpha^- \{y_i\}. \qquad \textbf{(7)}$$

4) Two Moore's closures corresponding to the fuzzy relation $[R_\alpha]$ or to the graph, limited to level α, are obtained by the maxmin convolution R_α^- with R_α^+ and R_α^+ with R_α^- . It is written:

$$M_\alpha^{(1)} = R_\alpha^- \bullet R_\alpha^+ \text{ and } M_\alpha^{(2)} = R_\alpha^+ \bullet R_\alpha^-, \qquad \textbf{(8,9)}$$

where $M_\alpha^{(1)}$ and $M_\alpha^{(2)}$ are the two Moore's closures.

Two concepts, such as a family and a graph, can be used to obtain Moore's closures which are considered to be important for a wide range of economic approaches.

6. Moore's Closed Sets as Maximum Groups

As for any pretopology, a set of Moore's closed sets is formed by the elements that comply with the following condition:

$$\forall \underset{\sim}{A}_k \in P(E):\ \underset{\sim}{A}_k = M\underset{\sim}{A}_k. \qquad \textbf{(10)}$$

Let us determine a subset of closed sets of *P*(*E*) corresponding to Moore's closure $M_\alpha^{(2)}$ as $C(E, M_\alpha^{(2)})$. As $R_\alpha^- \underset{\sim}{A}_k$ is a closed of *P*(*E*) for $M_\alpha^{(2)}$ which can be written as follows:

$$C(E, M_\alpha^{(2)}) = \bigcup_{\underset{\sim}{A}_k \in P(E)} R_\alpha^- \underset{\sim}{A}_k. \qquad \textbf{(11)}$$

and, $R_\alpha^+ \underset{\sim}{A}_k$ is a closed for $M_\alpha^{(1)}$ and we determine the subset of closed sets of *P*(*E*) corresponding to Moore's closure $M_\alpha^{(1)}$ as $C(E, M_\alpha^{(1)})$, we get:

$$C(E, M_\alpha^{(1)}) = \bigcup_{\underset{\sim}{A}_k \in P(E)} R_\alpha^+ \underset{\sim}{A}_k. \qquad \textbf{(12)}$$

Both families $C(E, M_\alpha^{(1)})$ and $C(E, M_\alpha^{(2)})$ are isomorphic and dual for each other. They are also antitone. Therefore, we get:

$$\underset{\sim}{A}_k \in C(E, M_\alpha^{(1)}) \Rightarrow (\underset{\sim}{A}_l = R_\alpha^- \underset{\sim}{A}_k \in C(E, M_\alpha^{(2)}) \text{ and } R_\alpha^+ \underset{\sim}{A}_l = \underset{\sim}{A}_k). \qquad \textbf{(13)}$$

$$\underset{\sim}{A}_l \in C(E, M_\alpha^{(2)}) \Rightarrow (\underset{\sim}{A}_k = R_\alpha^+ \underset{\sim}{A}_l \in C(E, M_\alpha^{(1)}) \text{ and } R_\alpha^- \underset{\sim}{A}_k = \underset{\sim}{A}_l). \qquad \textbf{(14)}$$

These families of closed sets can be associated to each other and each of the families forms a finite lattice [51]. As a result of the maxmin convolution, both families of closed sets provide the groupings with the greatest possible number of elements of the referential *E*. In the formation of groups, the groupings of the elements of one family of the closed sets are accompanied by the other groupings of the other family of closed sets related to them.

7. Galois Lattice and its Importance in the Process of Segmentation

There is a fuzzy rectangular relation $[\underset{\sim}{R}] \subset E_1 \times E_2$. In this case the necessary properties obtained for only one referential are kept. See Refs. 29-33. Only one lattice is obtained after two lattices were joined, where each vertex represents the relation of the groupings of elements of set E_1 with the groupings of elements of set E_2 taking into consideration a previously established level α. If this happens then there is affinity between the groups of the elements limited to the correspondent level. This lattice determines a structure for these relations. See Refs. 25,27,28,32,33,56,57.

If, when the upper end of the lattice is different to (E_2, *ø*) and the lower end is different to (*ø*, E_1), we add these vertices, then we have a Galois lattice that shows up two sets of Moore's closed sets that have the previous relations (E_2,*ø*) and (*ø*, E_1) as the upper and lower ends.

Let us consider the relation of the following order:

$$\forall X, X' \in E^{(1)} \quad \forall Y, Y' \in E^{(2)}.$$

$$(X, Y \leq (X', Y')) \leftrightarrow (X \subset X', Y \supset Y'). \qquad \textbf{(15)}$$

and, ∇ will be taken as an upper extreme of the lattice with the order of the Eq. (15).

By the same manner, the relation of the order:

$$\forall X, X' \in E^{(1)} \quad \forall Y, Y' \in E^{(2)} .$$

$$(X, Y \geq (X', Y')) \leftrightarrow (X \supset X', Y \subset Y'). \tag{16}$$

will be associated with Δ which is a lower extreme of a Galois lattice (we deal with the complementary order).

If the following condition carries out:

$$(U, V) = (X, Y) \nabla (X', Y'). \tag{17}$$

$$\text{then } (U \supset X \bigcup X' \text{ and } Y \subset Y \bigcap Y'). \tag{18),(19}$$

The same situation happens if :

$$(Z, T) = (X, Y) \Delta (X', Y'). \tag{20}$$

$$\text{then } (Z \subset X \bigcap X' \text{ and } T \supset Y \bigcup Y'). \tag{21),(22}$$

Taking into account the Eqs. (3) to (8) we observe that the maximum subrelations have the configuration of a Galois lattice and the maximum subrelations of the similarity also have this configuration (they are not ordered between themselves in the partial order of the lattice, but they are ordered in the relation with the extremes ∇ and Δ).

8. Illustrative Example

We have participated in the process of formation of economic homogeneous groups of the regions of Russian Federation and Ukraine.

We will start our example with Russian Federation. We selected 7 regions and their 6 more significant social and economic attributes. There are two sets of the regions and their indicators:

$$E_1 = \{a, b, c, d, e, f, g\} \text{ and } E_2 = \{A, B, C, D, E, F\}.$$

So we have the following relation:

Table 1. Relations between two subsets.

	A	*B*	*C*	*D*	*E*	*F*
a	75.328	14.620	10.895	8	58.301	8.436

b	244.352	38.362	58.407	34	100.218	10.553
c	254.965	81.399	61.400	64	319.501	9.630
d	125.971	58.760	31.771	32	121.827	7.252
e	746.141	37.033	86.969	117	401.294	14.243
f	148.447	42.607	35.523	38	213.552	10.317
g	11.344	13.835	22.202	29	41.872	9.389

The data of this matrix is normalized. The average values for each column of the indicators are calculated. These values are considered as the presumption levels of homogeneity α from which the homogenous groups are determined. The level for A is 0,308, for B - 0,503, for C - 0,505, for D - 0,393, for E - 0,447 and for F - 0,700. So, the fuzzy relation is turned to a Boolean matrix, see below:

	A	*B*	*C*	*D*	*E*	*F*
a	0	0	0	0	0	0
b	1	0	1	0	0	1
c	1	1	1	1	1	0
d	0	1	0	0	0	0
e	1	0	1	1	1	1
f	0	1	0	0	1	1
g	0	0	0	0	0	0

Fig. 1. Boolean matrix

The right connection B^{+} and the left connection B^{-} are established.

$$B^{+}\varnothing = E_2,\ B^{+}\{a\} = \varnothing,\ B^{+}\{b\} = \{A,C,F\},\ B^{+}\{c\} = \{A,B,C,D,E\},..,$$
$$B^{+}\{a,b,c,d,f\} = \varnothing,..,\ B^{+}\{a,b,c,d,e,f,g\} = \varnothing.$$
$$B^{-}\varnothing = E_1,\ B^{-}\{A\} = \{b,c,e\},\ B^{-}\{B\} = \{c,d,f\},\ B^{-}\{C\} = \{b,c,e\},..,$$
$$B^{-}\{A,B,C,D,E\} = \{c\},..,B^{-}\{A,B,C,D,E,F\} = \varnothing.$$

Then, Moore's closures such as $M^{(1)} = B^{-} \circ B^{+}$ and $M^{(2)} = B^{+} \circ B^{-}$ are obtained.

In the next step the families of closed sets relative to the Moore's closures $M^{(1)}$ and $M^{(2)}$ are formed:

$$f(E,M^{(1)}) = \{\varnothing,\{B\},\{E\},\{F\},\{A,C\},\{B,E\},\{E,F\},\{A,C,F\},\{B,E,F\},\{A,C,D,E\},$$
$$\{A,B,C,D,E\},\{A,C,D,E,F\},E_2\}.$$

$$f(E,M^{(2)}) = \{\varnothing,\{c\},\{e\},\{f\},\{b,e\},\{c,e\},\{c,f\},\{e,f\},\{b,c,e\},\{b,e,f\},$$
$$\{c,d,f\},\{c,e,f\},E_1\}.$$

The families of closed sets which represent the isomorphic lattices are associated to each other. See Fig.2. A Galois lattice is represented at the Fig.3.

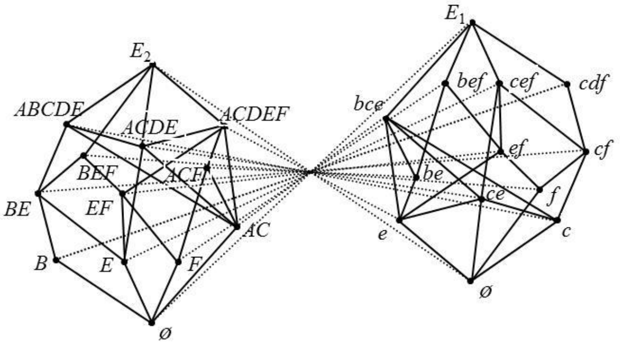


Fig. 2. Isomorphic lattices

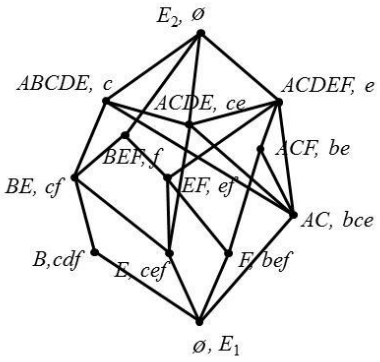


Fig. 3. Galois lattice

The following similar groupings of elements of two sets which are the set of regions and the set of attributes have been obtained.

$$\varnothing \leftrightarrow E_2,\ \{c\} \leftrightarrow \{A,B,C,D,E\},\ \{e\} \leftrightarrow \{A,C,D,E,F\},\ \{f\} \leftrightarrow \{B,E,F\},$$
$$\{b,e\} \leftrightarrow \{A,C,F\},\ \{c,e\} \leftrightarrow \{A,C,D,E\},...,\{b,c,e\} \leftrightarrow \{A,C\},\ \{b,e,f\} \leftrightarrow \{F\},$$
$$\{c,d,f\} \leftrightarrow \{B\},\ \{c,e,f\} \leftrightarrow \{E\},\ E_1 \leftrightarrow \varnothing.$$

Now we will do the same process of calculus for the economic regions of Ukraine. In this case we selected 6 regions and their 5 more significant economic attributes. There are two sets of the regions and their indicators:

$$E_1 = \{a,b,c,d,e,f\} \text{ and } E_2 = \{A,B,C,D,E\}.$$

So we have the following relation:

Table 2. Relations between two subsets.

	A	*B*	*D*	*E*	*F*
a	9.147	30.057	2.148	168	1.72
b	8.040	29.613	1.967	333	1.62
c	3.461	13.539	1.828	336	1.43
d	6.310	22.650	2.908	75	1.52
e	2.068	7.969	779	109	1.18
g	1.149	2.850	652	28	945

The data of this matrix is normalized, see below:

	A	*B*	*C*	*D*	*E*
a	1	1	0,7	0,4	1
b	0,8	0,9	0,6	0,9	0,9
c	0,3	0,4	0,6	1	0,8
d	0,6	0,7	1	0,2	0,8
e	0,2	0,2	0,2	0,3	0,6
f	0,1	0,0	0,2	0,0	0,5

Fig. 4. Normalized matrix

The average values for each column of the indicators are calculated. These values are considered as the presumption levels of homogeneity α from which the homogenous groups are determined. The level for A is 0,550, for B - 0,592, for C - 0,590, for D – 0,520 and for E - 0,814. So, the fuzzy relation is turned to a Boolean matrix, see below:

	A	*B*	*C*	*D*	*E*
a	1	1	1	0	1
b	1	1	1	1	1
c	0	0	1	1	1
d	1	1	1	0	1
e	0	0	0	0	0
f	0	0	0	0	0

Fig. 5. Boolean matrix

The right connection B^{+} and the left connection B^{-} are established.

$$B^{+}\varnothing = E_2,\ B^{+}\{a\} = \{A,B,C,E\},\ B^{+}\{b\} = E_2, \ldots, B^{+}\{a,b\} = \{A,B,C,E\},$$

$$B^{+}\{a,c\} = \{C,E\}, \ldots, B^{+}\{c,d,e\} = \varnothing,\ B^{+}\{c,d,f\} = \varnothing,\ B^{+}\{c,e,f\} = \varnothing, ..,$$

$$B^{+}\{a,b,c,d,e\} = \varnothing,\ B^{+}\{a,b,c,d,f\} = \varnothing, \ldots,\ B^{+}\{a,b,c,d,e,f\} = \varnothing.$$

$$B^{-}\varnothing = E_1,\ B^{-}\{A\} = \{a,b,d\},\ B^{-}\{B\} = \{a,b,d\}, \ldots, B^{-}\{C,D\} = \{b,c\},$$

$$B^{-}\{C,E\} = \{a,b,c,d\}, \ldots, B^{-}\{B,C,D\} = \{b\},\ B^{-}\{B,C,E\} = \{a,b,d\}, \ldots$$

$$B^{-}\{B,C,D,E\} = \{b\},\ B^{-}\{A,B,C,D,E\} = \{b\}.$$

Then, Moore's closures such as $M^{(1)} = B^{-} \circ B^{+}$ and $M^{(2)} = B^{+} \circ B^{-}$ are obtained.

In the next step the families of closed sets relative to the Moore's closures $M^{(1)}$ and $M^{(2)}$ are formed:

$$f(E,M^{(1)}) = \{\varnothing,\{C,E\},\{C,D,E\},\{A,B,C,E\},E_2\}.$$

$$f(E,M^{(2)}) = \{\{b\},\{b,c\},\{a,b,d\},\{a,b,c,d\},E_1\}.$$

The families of closed sets which represent the isomorphic lattices are associated to each other. See Fig. 6. A Galois lattice is represented at the Fig. 7.

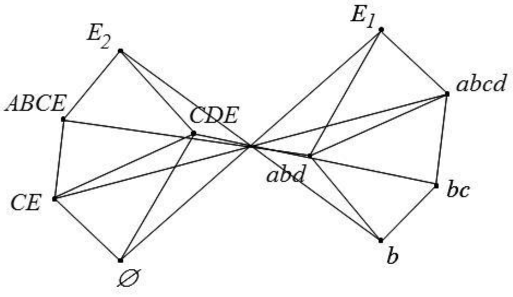


Fig. 6. Isomorphic lattices

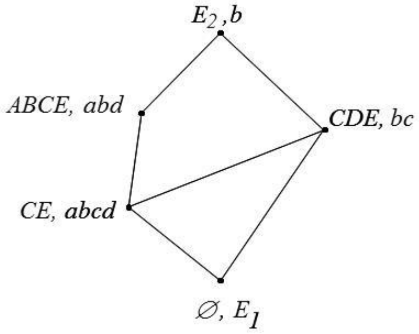


Fig. 7. Galois lattice

The following similar groupings of elements of two sets which are the set of regions and the set of attributes have been obtained.

$\{b\} \leftrightarrow E_2$, $\{b,c\} \leftrightarrow \{C,D,E\}$, $\{a,b,d\} \leftrightarrow \{A,B,C,E\}$, $\{a,b,c,d\} \leftrightarrow \{C,E\}$, $E_2 \leftrightarrow \varnothing$.

Having done the calculus for two different cases we have obtained the following conclusions.

The number of the homogeneous groups of the regions of Russian Federation represented by Moore's closures is much bigger than we have it for Ukraine. The first reason is that the data basis of Russia includes the input elements (social and economic indicators) that are more similar to each other while the economic indicators for the second country have significant dissimilarity between themselves. The second reason is that the method we have used to form the homogeneous groups is very sensible to any minimum changes that data basis have. We have observed this fact setting different levels (degrees) of the membership function for the groups of attributes.

We have obtained very detailed panorama of the economic situation for both countries. The bigger number of the homogeneous groups we have the easier and more complete economic analysis of the regions we can do, and, as a consequence, the more accurate information we contribute to the decision-making process at any required period of time.

Using this type of mathematical tools based on the theory of affinities we have obtained not only the exact number of the homogeneous groups of the economic regions but also the indicators (attributes) that the groups have in common. This information gives us an additional data basis to estimate the level of economic development of each region.

Galois lattices give us a complete structural ordering of the results obtained for two different cases providing by this manner a better visualization of an actual economic situation of all regions.

9. Conclusion

In this article, we proposed an alternative method for the grouping process based on the theory of uncertainty. With this study we pretended to demonstrate the importance and efficiency of the mathematical tools we used for implementing the process of groups (clusters) formation in the practice. The considered technique takes into account the increasing level of impreciseness of the economic information a company has to deal with in its everyday activity. It helps to reduce the complexity of the obtained economic data and set solid basis for the next step of the information procedure and also for the decision making process. This technique is very useful when the company has to deal with a great amount of economic data.

In the first part of the paper we gave brief explanations regarding the concepts such as grouping (clustering) process, similarity and affinity. We also gave the references to the authors who made a significant contribution to the subject of clustering process and fuzzy grouping process, in particular. We gave the theoretical explications of the following questions: 1) generalization of the notion of similarity by using the theory of affinities; 2) transition from one axiomatic to the other in pretopologies; 3) obtaining the strictest uncertain pretopology determined as uncertain pretopology of Moore; 4) process for obtaining maximum groupings in the cases of fuzzy rectangular relations by using a Moore's closure and Moore's closed set.

To demonstrate the method's efficiency we illustrated an empirical example to partition economic regions of Russian Federation and Ukraine into groups due to the different levels of homogeneity. The obtained results allow making a proper selection of one group of the regions or another according to the required levels of economic indicators. As a conclusive phase of our empirical example we introduced Galois lattices in order to form a clearer vision of a regional economic situation in the countries and to give an additional support tool to the decision-making process.

References

1. Brouwer, R.: Clustering feature vectors with mixed numerical and categorical attributes. International Journal of Computational Intelligence Systems, Vol. 1, No. 4, 285-298. (2008)
2. Jain, A. K.; Dubes, R. C.: Algorithms for Clustering Data. Upper Saddle River, NJ, Prentice Hall. (1988)
3. Klosgen, W.; Zytkow, J. M.: Knowledge discovery in databases terminology. In Advances in Knowledge Discovery and Data Mining. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. The MIT Press, 573-592. (1996)
4. Bunyak, F.; et al.: Histopathology tissue segmentation by combining fuzzy clustering with multiphase vector level sets. Advances in Experimental Medicine and Biology, Vol. 696, 413-424. (2011)
5. Kyung, M. W.; et al. Breast tumor classification using fuzzy clustering for breast elastography. Ultrasound in Medicine and Biology, Vol. 37, No. 5, 700-708. (2011)
6. Folkesson, J.; et al.: Local bone enhancement fuzzy clustering for segmentation of MR trabecular bone images. Medical Physics, Vol. 37, No. 1, 295-302. (2010)
7. He, R.; et al.: Generalized fuzzy clustering for segmentation of multi-spectral magnetic resonance images. Computerized Medical Imaging and Graphics, Vol. 32, No. 5, 353-366. (2008)
8. Sakka, E.; et al.: Classification algorithms for microcalcifications in mammogram. Oncology Reports, Vol. 15, 1049-1055. (2006)
9. Hirano, S.; et al.: Comparison of clustering methods for clinical databases. Information Sciences, Vol. 159, 155-165. (2004)
10. Kang, X.; Xu, M.: Forecast about the disaster of Zoumaling tunnel discharge with fuzzy clustering analysis. In Proc. of International Conference on Electric Technology and Civil Engineering (ICETCE), 380-383. (2011)
11. Chang, T.; et al.: Dissolved gas analysis in transformer oil using dynamic tunneling fuzzy c-means algorithm. High Voltage Engineering, Vol. 35, No. 9, 2181-85. (2009)
12. Tari, L.; et al.: Fuzzy c-means clustering with prior biological knowledge. Journal of Biomedical Informatics, Vol. 42, No. 1, 74-81. (2009)
13. Zio, E.; Baraldi, P.: Evolutionary fuzzy clustering for the classification of transients in nuclear components. Progress of Nuclear Energy, Vol. 46, No. 3, 282–296. (2005)
14. Zio, E.; Gola, G.: Neuro-fuzzy patterns classification for fault diagnosis in nuclear components. Annals of Nuclear Energy, Vol. 33, No. 5, 415–426. (2006)
15. Wong, F.; et al.: Civil engineering including earthquake engineering. In Zimmermann, H.J. (ed.) Practical Applications of Fuzzy Technologies. Kluwer Academia Press, Boston, London, Dordrecht, 207-236. (1999)
16. Eom, K.: Fuzzy clustering approach in supervised sea-ice classification. Neurocomputing, Vol. 25, 149–166. (1999)
17. Zhuge, Y.; et al.: Parallel fuzzy connected image segmentation on GPU. Medical Physics, Vol. 38, No. 7, 4365-71. (2011)
18. Zhang, Y.; et al.: Image segmentation using PSO and PCM with Mahalanobis distance. Expert Systems with Applications, Vol. 38, No. 7, 9036-40. (2011)

19. Sowmya, B.; Sheela, R.: Colour image segmentation using fuzzy clustering techniques and competitive neural network. Applied Soft Computing, Vol. 11, No, 3, 3170-78. (2011)
20. Jablonowski, M.: Implications of fuzziness for the practical management of high-stakes risks. International Journal of Computational Intelligence Systems, Vol. 3, No. 1, 1-7. (2011)
21. Gil Aluja, J.; et al.: Using homogeneous groupings in portfolio management, Expert Systems with Applications, Vol. 38, No. 9, 10950-58. (2011)
22. Wang, Y.; Lee, H.: A clustering method to identify representative financial ratios, Information Sciences, Vol. 178, 1087–97.
23. Gil Aluja, J.; Gil Lafuente, A. M.: Algorithms for the treatment of complex economic phenomena (In Spanish). CEURA, Madrid, Spain. (2007)
24. Gil Aluja, J.: Introduction to the theory of uncertainty in business management (In Spanish). Milladoiro, Vigo. (2002)
25. Gil Lafuente, J.; Gil Aluja, J.: Algorithms for excellence. Keys for being successful in sport management (In Spanish). Milladoiro, Vigo. (2002).
26. Gil Aluja, J.: Pretopology in management of uncertainty. Inaugural speech (In Spanish). Public University, Leon. (2001)
27. Gil Lafuente, A. M.: New strategies for financial analysis of a business (In Spanish). Ariel, Barcelona. (2001)
28. Gil Lafuente, J.: Model for the homogeneous groping of the sales force. Proceedings of Congress M.S. China, pp. 332-335. (2001)
29. Gil Aluja, J.: Elements of a theory of decision in uncertainty. Kluwer Academic Publishers, Netherlands. (1999)
30. Gil Aluja, J.: Interactive management of human resources in uncertainty. Kluwer Academic Publishers, London, Dordrecht. (1998)
31. Gil Lafuente, J.: Marketing for the new millennium (In Spanish). Pirámide, Madrid. (1997)
32. Gil Aluja, J.: The multicriteria selection of investments by means of the Galois lattice (In Spanish). Civita, Madrid, cap. 7, pp 139-157. (1995)
33. Kaufmann, A.; Gil Aluja, J.: Selection of affinities by means of fuzzy relation and Galois Lattices. Acts of the XI European Congress O.R., Aachen. (1991)
34. Klir, J.; Folger, A: Fuzzy sets, uncertainty and information. Prentice-Hall International, USA. (1988)
35. Kaufmann, A.: Méthodes et modeles de la recherche opérationnelle. Dunod, París, Vol 1, 65-72. (1970)
36. Kaufmann, A.; et al.: Cours de calcul booléien appliqué. Albin Michel, Francia. (1963)
37. Anderberg, M. R.: Cluster Analysis for Applications. Academic Press. (1973)
38. Everitt, B.: Cluster analysis. Heinemann Educational Books, 117-129. (1996)
39. Kaufman, L.; Rousseeuw, P.: Finding groups in data: an introduction to cluster analysis. John Wiley and Sons, New York. (1990)
40. Michalski, R. S.; Stepp, R. E.: Automated construction of classifications: conceptual clustering versus numerical taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5, 396-410. (1983)
41. Fisher, D. H.: Knowledge acquisition via incremental conceptual clustering. Machine Learning, Vol. 2, No. 2, 139-172. (1987)
42. Zadeh, L.: Fuzzy algorithms. Information and Control, Vol. 12, 94-102. (1968)
43. Li, G.; et al.: Clustering with instance and attribute level side information. International Journal of Computational Intelligence Systems, Vol. 3, No. 6, 770-785. (2010)
44. Kumar, P.: Fuzzy classifier design using modified genetic algorithm. International Journal of Computational Intelligence Systems, Vol. 3, No. 3, 334-342. (2010)

45. Butenko, S.; Gil-Lafuente, J.; Pardalos, P.: Optimal strategies in sports economics and management. Springer-Verlag Berlin Heidelberg, pp. 1-15. (2010)
46. Brouwer, R.: Fuzzy clustering of feature vectors with some ordinal valued attributes using gradient descent for learning. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 6, 195-218. (2008)
47. Honda, K.; Ichihashi, H.: Regularized linear fuzzy clustering and models. IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, 508–516. (2005)
48. Yang, M.; et al.: Fuzzy clustering algorithms for mixed feature variables, Fuzzy Sets and Systems, Vol. 141, 301-317. (2004)
49. Miyamoto, S.: Information clustering based on fuzzy multisets. Information Processing and Management, Vol. 39, 195-213. (2003)
50. Hoppner, F.; et al.: Fuzzy Cluster Analysis. John Wiley and Sons, New York. (1999)
51. Kaufmann, A.: Introduction to the theory of fuzzy sets for use by engineers, Vol. 4, Additions and new applications (In French). Masson, Paris. (1977)
52. Zadeh, L.: Similarity relations and fuzzy orderings. Information Sciences, Vol. 3, No. 2, 177-200. (1971)
53. Ruspini, E.: Numerical methods for fuzzy clustering. Information Sciences, Vol. 2, 319-350. (1970)
54. Zadeh, L.: Fuzzy algorithms. Information and Control, Vol. 12, 94-102. (1968)
55. Dolecki, S.; et al.: Weak regularity and consecutive topologizations and regularizations of pretopologies. Topology and its Applications, Vol. 156, No. 7, 1306-14. (2009)
56. Höhle, U.: Many valued topology and its applications. Kluwer Academic, Dordrecht. (2001)
57. Höhle, U.; et al.: A general theory of fuzzy topological spaces. Fuzzy Sets and Systems, Vol. 73, 131-149. (1995)
58. Kaufmann, A.: Ordinary pretopology and fuzzy pretopology (In French). Working Note 115. (1983)
59. Kaufmann, A.: Introducción a la combinatoria y sus aplicaciones. Campania Editorial Continental, Barcelona. In Spanish. (1971)
60. Gil Aluja, J.: Clan, affinities and Moore’s fuzzy pretopology. Fuzzy Economic Review, Vol. VIII, No. 2, 3-24. (2003)
61. Gil Aluja, J., Gil Lafuente, A. M., Mérigo, J. M.: Using homogeneous groupings in portfolio management. International Journal of Expert Systems with Applications, Vol. 38, No. 9, 10950-58. (2011)
62. Mérigo, J. M., Gil Lafuente, A. M.: Decision-making techniques with similarity measures and OWA operators. Statistics & Operations Research Transactions, Vol. 36, No. 1, 81-102. (2012)
63. Gil Lafuente, A. M., Barcellos Paula, L.: Approaches to managing sustainability among enterprises. Decision making systems in business administration. Vol. 8, pp. 27 -36. (2012)
64. Keropyan, A., Gil Lafuente, A. M.: A personal selection model using Galois group theory, Kybernetes, Vol. 42, No. 5, 711-719. (2013)
65. Gil Lafuente, J.: Are there clans in a wardrobe? “Fuzzy-Pichat” will find out it! Proceedings of the National Congress. Vol. 1, pp. 90. (2008)
66. Gil Lafuente, A. M.; Gil Lafuente, J.: Models and algorithms for the treatment of creativity in business management (In Spanish). Milladoiro, Vigo. (2007)
67. Gil-Aluja, J.; Gil-Lafuente, A. M.; Klimova, A.: M-attributes algorithm for the selection of a company to be affected by a public offering. International Journal of Uncertainty, Fuzziness Knowledge-Based Systems, Vol. 17, No. 3, 333-343. (2009)
68. Sapena, A.: A contribution to the study of fuzzy metric spaces. Applied General Topology, Vol. 2, No. 1, 63-76. (2001)

Anna M. Gil-Lafuente is an Associate Professor in the Department of Business Administration at the University of Barcelona, Spain. She received a MSc and a PhD degree in Business Administration from the University of Barcelona. She has published more than 150 papers in journals, books and conference proceedings including journals such as *Information Sciences*, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* and *Expert Systems with Applications*. She is co-editor-in-chief of 8 journals published by the Association for Modeling and Simulation in Enterprises (AMSE). She is currently interested in Decision Making, Uncertainty and Finance.

Anna Klimova is a Professor in the Department of Business Administration at the University of Barcelona, Spain. She received a PhD degree in Business Administration from the University of Barcelona. She has published more than 20 papers in journals and conference proceedings. She is currently interested in Decision Making, Uncertainty and Finance.

Received: January 25, 2013; Accepted: December 24, 2014

Computer Science and Information Systems 11(2):797–808 DOI:10.2298/CSIS130113043X

An application of Structural Equation Modeling for Continuous Improvement

Yehun Xu[1, 2], Carlos Llopis-Albert[3], and J. González[4]

[1]State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, 210098, PR China
[2]Business School, Hohai University
Jiangning, Nanjing, 211100, China; xuyejohn@163.com
[3] Universitat Politècnica de València, CITV, Campus Vera s/n
Valencia, 46022, Spain; cllopisa@gmail.com
* Corresponding Author
[4] Department of Business Administration, Universitat Politècnica de València, Campus Vera s/n, Valencia, 46022, Spain; jgonzaga@omp.upv.es

Abstract. More and more firms are adopting Web 2.0 technologies for different purposes. Despite some studies about the impact on different aspects of the organization of these technologies, few studies have empirically tested their effects. This study confirms that the there is a positive relationship between Web 2.0 adoption and Innovativeness and between Innovativeness and Business Performance. We can conclude that Innovativeness is a mediating variable between Web 2.0 and Firm Performance. The relationships were tested in a sample of 244 hospitality firms, using a structural modeling approach. The results are presented and discussed. In the last section of the article, the conclusions are discussed and finally, some future lines of research are suggested.

Keywords: Web 2.0, Innovativeness, Organizational Performance, Hospitality firms.

1. Introduction

Innovation has been regarded as a source of sustainable competitive advantage and positively related to firm performance by several authors [1-3]. Innovativeness or the firm´s capacity to innovate is more important in the current dynamic and turbulent markets, in which firms need to constantly develop or adapt their products to the changing demands of their consumers. This challenge brings along the need to improve the firm´s capacity to collect and process external information and to share it within the firm.

On this realm, Web 2.0 technologies are revolutionizing different aspects in the organization, and are fostering collaboration, communication and participation; which has caught the attention of managers and researchers.

Some studies have begun to report the uses of these technologies in the organization in different fields such as: knowledge management, communication and collaboration, customer relationship management, innovation and training[1]. In addition, due to the characteristics of Web 2.0 technologies, even universities and faculty members are using them to support different aspects of higher education [2-3].

Since Web 2.0 technologies are easy to use, more and more people have started to use them for different purposes, both in their personal life and at work. Some authors have analyzed the impacts of Web 2.0 adoption in specific aspects of the organization. However, there a few empirical studies, and to our knowledge only one that analyzes empirically the relationship between Web 2.0 adoption and entrepreneurial orientation.

The purpose of this study is, to analyze the relationships among Web 2.0 adoption, innovativeness and performance. Based on the results reported so far in the literature, we believe that the use of Web 2.0 technologies might positively influence business performance through innovativeness. Later in the paper we will theoretically explain our assumption.

In order to fulfill our purpose we need to achieve the following objectives:

- To briefly review the existing literature about, Web 2.0, innovativeness and performance.
- To theoretically establish the relationships among the variables
- To empirically test the relationships
- To discuss the results and their implications

The document is organized in various sections. The first section, reviews the literature about Web 2.0 technologies, innovativeness and business performance. Once each variable is reviewed, we will theoretically support the relationships. The second section, describes the data and the method employed to test the relationships. The third section, presents the results and the final section, discusses the conclusions and suggests future lines of research.

2. Literature review

2.1. Web 2.0

O´Reilly [2] refers to Web 2.0 as a new version of the internet and technologies that are able to promote collaboration and communication. In this version of the internet most of the content is generated by users. There are various definitions of the term Web 2.0. One of them argues that Web 2.0 is ‘a new philosophy that emphasizes collective intelligence, active participation and collaboration[2].

Web 2.0 can be seen from four perspectives and it has three main characteristics [3]. The four perspectives are: technological, sociological, economical and legal. From a technological perspective, Web 2.0 tools allow information to be used in new ways. From a sociological perspective Web 2.0 increases social interaction. From an economical perspective, Web 2.0 enables and requires new business models. Finally, from a legal perspective, Web 2.0 brings new legal issues to the Web. The three main characteristics of Web 2.0 are: collaboration, participation and openness.

Among the most used technologies of Web 2.0 we have blogs, wikis, podcasts, RSS and social networks. Blogs are a type of Web publication, in which its content is written in a chronological order by one or more authors. Wikis are more a structured Web sites, were users participate adding, editing or erasing information [3]. One of the most famous Wiki is the Wikipedia.

Podcasts are audio files that can be reproduced in a computer, Ipod or any other audio player. The most used format is MP3. RSS or really simple syndication, allows user to receive updates about the content of web pages, such as a news paper web site.

A social network is a Web portal where the users, previously registered, can create a personal profile and contact other users who want to share digital contents. Facebook is one of the most known social networks. Online social networks have become a relevant phenomenon in order to create sound relationships, in a personal and professional way.

It is also necessary to highlight that in all the interaction process between the customer and the software application, the customer should receive all the necessary information in order to perceive a high quality service. In this sense, CEM processes (Customer Experience Management) have been developed, based on the experience of a customer using a product or service.

Various studies about the effects, possible implications and uses have been published in the last years. We will briefly discuss some of the most important findings about Web 2.0 adoption by firms.

Some studies point out that Web 2.0 technologies have a strong bottom-up element and engage a broad base of workers, which in turn, is expected to improve participation and collaboration within the firm[3]. Among the implications of Web 2.0 adoption in the marketing field we have: Built of innovative brands, viral marketing campaigns, consumer behavior and direct marketing [4-5].

Others papers report the implications of Web 2.0 on knowledge management [10], organizational learning [4], organizational innovation4 and entrepreneurial orientation [5].

Some web 2.0 applications that have failed due to they have used many channels for a client, but they have not used the channels in an integrated way. It is not possible to generate synergies if they do not transmit a brand image during all the stages. Secondly, it is necessary to develop a long-term relationship with the client. Firm should not only use a transactional approach, but it should develop processes that let the client interact continuously with the firm. There are some applications such as CRM (*Customer Relationship Management) solutions*, that include the processes related to customer management, and they help to build confidence with clients in the long-term.

2.2. Innovativeness

Innovation has been defined as the generation, acceptance and implementation of new ideas, processes, products and services [5]. Another definition refers to innovation, as the generation (development) or the adoption (use) of new ideas, objects and practices [6]. The Oslo [7] manual defines innovation as the introduction of a new or an improved product, process, commercialization or organizational method. Furthermore, the conceptualization of innovation can vary according to the perspective of the academic field, such as economy, technology or sociology [5].

In the literature, different types of innovations are distinguished such as: product, process, technology and organizational and market innovation [3, 21, 19].

There are various determinants of innovation. According to some authors these antecedents can be grouped in: external determinants, context variables, organizational, individual and group determinants [22].

Finally, various authors report that innovation positively effect, business processes, competencies, financial performance and competitiveness [23].

Innovativeness on the other hand, is somewhat different from innovation because it is a characteristic of an organization [24], some authors refers to it, as the capability of a firm to innovate [25], which implies the generation of new ideas, products, services or processes [26].

Other authors have defined innovativeness as the openness to new ideas [27]; and yet others as the rate of innovation adoption or the organization´s will to change [28].

2.1 Business Performance

Business performance is the variable of the most interest for researchers. Its measurement is essential for both, managers and researchers to know where the firm stands versus its rivals and how the firm is performing. In general, business performance has been consistently considered as a dependent variable. For instance, a review of 439 studies in a period of three years, the variable acted as a dependent variable [29].

According to some authors [30], business performance is composed of three specific areas: financial performance, market performance and return for stockholders. In addition, it is argued that business performance is a multidimensional theoretical construct; therefore, it should be measured in multiple dimensions [31].

There are two different groups of measures to evaluate business performance, objective and subjective. An objective measure is a real and current number about the firm, for instance sales growth. On the other hand, the term subjective is used to mean that the company's performance score is derived using a scale with anchors such as "much lower" to "much higher" compared to competitors [32]. Finally, it is pointed out that subjective evaluations allow more flexibility and consistency that cannot be obtained by objective measures [33].

2.2 Theoretical relations

In order to theoretically establish the relationship among the variables of interest, we will review previous studies that link information technologies with organizational learning —an important antecedent of innovation—, knowledge management processes and innovation, web 2.0 and organizational learning, web 2.0 and entrepreneurial orientation and finally, innovativeness and business performance.

On one hand, earlier studies about information technologies and organizational learning, affirm that the use of the right information technologies under the right environmental and organizational conditions, can considerable benefit organizational learning [34]. On the other hand, the relationship between organizational learning and innovation has been previously studied in the past [35, 26, 31]. For example, it has been pointed out that organizational learning has a positive effect on innovation, and that innovation mediates the relationship between organizational learning with performance [37]. Others have indicated that learning orientation functions as an antecedent to an innovation orientation [30], and similarly, other results have supported that learning orientation is critical for innovation and performance [31].

With regards to the relationship between knowledge management and innovation, among the various studies, one of them concluded that knowledge management components do correlate with innovation, that is, a firm with a capability in knowledge management can also be more innovative [2]. Another study indicated that employees´ willingness to both donate and collect knowledge is related to firm innovation capability [38].

Studies about Web 2.0 and its adoption, affirm that they have a positive effect on organizational learning, knowledge sharing and innovation. First of all, it has been argued that the development of systems, combining social (e.g. Web 2.0 applications) and technical systems could provide an effective solution to the problem of knowledge sharing [39].

Regarding organizational learning, a recent study [40] affirm that Web 2.0 "has the potential not only to improve individual and team learning, but also to promote organizational learning resulting in higher levels of performance". These technologies have the potential to support both, generative and adaptive learning in the organization.

At last, it has been reported that Web 2.0 adoption might has an impact on some aspects of innovation such as: organize innovation, improve research and development success, increase the number of innovation initiatives [41] and productize innovations more effectively [4]. A recent study tested the relationship between Web 2.0 adoption and entrepreneurial orientation [17]. In the mentioned study, the authors argued that entrepreneurial orientation is composed of five dimensions: Innovativeness, Proactiveness, Risk taking, Autonomy and Competitive aggressiveness. Based on the results of the study the authors affirmed that there is positive relationship between Web 2.0 adoption and entrepreneurial orientation and more specifically they pointed out that Web 2.0 adopters showed a stronger mindset for innovation.

As presented in the discussion above, previous studies have established and tested the relationship between organizational learning with innovativeness and the link between knowledge management and innovativeness. Since Web 2.0 adoption is related to both, organizational learning and knowledge sharing it is plausible to infer that Web 2.0 adoption might be related to innovativeness. Finally, in a recent study the liaison between Web 2.0 adoption and entrepreneurial orientation was tested. The entrepreneurial orientation operationalization included innovativeness as a dimension. As a consequence we formulated the first hypothesis as:

H1: *There is a positive relationship between the adoption of Web 2.0 technologies and innovativeness.*

The relationship between innovation and performance and more specifically innovativeness with performance has been previously studied by several authors. Some empirical findings confirmed that innovativeness is a key determinant of business performance [39]. Others authors also found a positive relation between firm innovativeness with firm performance [31]. Finally, technical and administrative innovations have a positive and direct impact on performance [42].

Our second hypothesis links Innovativeness with firm Performance:

H2: *Innovativeness is positively related with business performance.*

3. Measurement Data and Sample

3.1. Measurement

Web 2.0 adoption measurement

There are few validated scales in the literature about Web 2.0 adoption. However some authors^{27} have developed and validated a scale. This scale is composed of 8 items that are measured on a five-point Likert scale, where 1 = never, 3 = sometimes, and 5 = always. The items are:

- WA1: Blogs are used to issue firm release or to spread ideas.
- WA2: Firm uses collaborative software to communicate with the rest of employees.
- WA3: Firm uses an intranet for knowledge management.
- WA4: The site of the firm allows users to introduce contents and express their necessities.
- WA5: Employees know suggestions that customers formulate.
- WA6: The site of the firm has, apart from text, multimedia files to enable the interaction with the user.
- WA7: Firm develops practices so that employees share their knowledge.
- WA8: Employees keep the know-how of the processes in an electronic way.

Innovativeness measurement

There are several scales to measure innovativeness. One of the first scales was developed in 1977 [5]. The scale that was chosen in this study was developed in 2002 by Calantone [28]. The scale is based on previous studies and six items composed it. The items are measured on a seven-point Likert scale where 1 = strongly disagree and 7 = strongly agree.

The items are:

- I1: Our Company frequently tries out new ideas.
- I2: Our Company seeks out new ways to do things.
- I3: Our Company is creative in its methods of operation.
- I4: Our Company is often the first to market with new products and services.
- I5: Innovation in our company is perceived as too risky and is resisted.
- I6: Our new product introduction has increased over the last 5 years.

Business performance measurement

There is a lot of debate about the measurement of organizational performance, namely about the advantages or disadvantages of objective measures versus subjective measures [43]. Based on the recommendations given by some authors and based on a previous study [44] we chose a scale based on subjective measures. The scale was composed of seven items, measured on a five-point Likert scale, where 1=much lower compare to our competitors; 3=equal to our competitors and 5=much higher compare to our competitors

- BP1= Product or service quality
- BP2= New product or service success
- BP3= Client retention rate
- BP4= Sales level
- BP5= Return on equity
- BP6= Gross profit margin
- BP7= Return on investment

3.2. Data and Sample

The questionnaire was sent to Spanish four and five starts hotels that provided their e-mail on the Spanish tourism web page. Between January and July of 2010, we received 255 questionnaires; however 11 were eliminated due to different faults, leaving a total sample of 244. The response rate gave us a sample error of 6 percent for a confidence level of 95 percent.

According to the two step approach for structural equation modeling, first we estimated the measurement modeling using confirmatory factor analysis (CFA) and then we estimated the structural model. We used EQS 5.7 to conduct the tests.

4. Results

According to the method employed first, we will present the confirmatory factor analysis on the entire set of measurement items.

This step resulted in the elimination of 1 item of the innovativeness scale. The factor loading magnitudes are presented in Table I.

Table I. Factor loadings magnitudes and measurement errors

Web 2.0 adoption measurement scale		
Items	λ	Error
W1	0.659	0.569
W2	0.715	0.489
W3	0.884	0.216
W4	0.862	0.256
W5	0.712	0.492
W6	0.683	0.533
W7	0.831	0.309
W8	0.736	0.458
Innovativeness measurement scale		
Items	λ	Error
I1	0.798	0.363
I2	0.861	0.258
I3	0.832	0.306
I4	0.941	0.118

I6	0.788	0.379
Performance measurement scale		
Items	λ	Error
BP1	0.902	0.185
BP2	0.862	0.256
BP3	0.652	0.572
BP4	0.776	0.397
BP5	0.725	0.472
BP6	0.885	0.216
BP7	0.802	0.353

The estimated parameters are statistically significant at a 95% confidence level (t>1.96). The factor loading magnitudes are between 0.652 and 0.941, which are high and above the minimum required of 0.4. The construct´s scales presented a high compound reliability. For the Web 2.0 adoption scale a compound reliability of 0.88, for the Innovativeness scale a compound reliability of 0.89 and for the Business performance scale a compound reliability of 0.89.

4.1. Empirical testing of the hypotheses

H1: The results confirm and adequate global fit, so we can consider the model an adequate representation of the causal relationship between the studied latent variables.

The absolute goodness fits are excellent. The Chi^2 Satorra-Bentler is significant 13.74 for 13 degrees of freedom and the RMSR is 0.019. The incremental goodness fit indicators are according to satisfactory levels, where BB NNFI is close to 1 The GFI index is 0.997, and the parsimonious goodness fit indicator NC is 1.11.

Table II. Coefficient and reliability index for the structural model

Equation Coefficient γ	Reliability of the structural equation
0.762	0.911

The estimated parameter in Table II is statistically significant at a 95% confidence level (t=29.124≥1.96). The equation coefficient and the reliability of the structural model show an adequate fit; therefore, we could confirm a positive and statistically significant relationship between Web 2.0 adoption and innovativeness.

H2: The Chi^2 Satorra-Bentler is significant with a value of 31.61 for 32 degrees of freedom. The incremental goodness fit indicators reached satisfactory levels, where BB NNFI and IFI are close to 1. The parsimonious goodness fit indicator (NC) does not present a good adjustment but it is near one (0.951)

The R^2 of the model for business performance is high, 0.28. Therefore, we could confirm a positive relationship between innovativeness and business performance.

5. Conclusions and Future Research

The purpose of this study was to analyze the relationships among Web 2.0 adoption, innovativeness and performance. To do so, we first conducted a literature review about the variables of interest. Afterwards we theoretically established the relationships among the variables. Based on, previous studies, we argued that Web 2.0 adoption is related to innovativeness, since it could improve knowledge and cognitive processes of the organization and since these aspects are related to innovation, we assumed that Web 2.0 adoption could be related with innovativeness. Then we theoretically sustained the link between innovativeness and performance, a relationship previously studied by several authors.

Based on the empirical analysis, using a structural modeling approach on 244 Spanish firms, we tested the relationships. The results supported the relationships between Web 2.0 and innovativeness and between innovativeness and business performance.

The findings of the present study have several implications for both, researchers and managers. For researchers the results contribute to the understanding of the effects that Web 2.0 adoption has on a very important organizational aspect such as innovativeness. In addition, some of the previous reported impacts of Web 2.0 on innovation, to certain point, now are confirmed, which open additional questions about the effects of Web 2.0 on other organizational variables.

Managers should encourage the adoption of Web 2.0. To do so, managers must decide first which aspect of the organization they want to improve and then they should decide about the Web 2.0 technology that has a positive impact on the chosen aspect. The results indicate that in general these technologies could improve knowledge sharing, organizational learning and innovativeness.

On the other hand, once more it is proven that innovativeness is very important to sustain competitive advantage and that it has a direct and positive effect on performance. Therefore, managers are advised to constantly increase the firm´s capacity to innovative, if they want to achieve a better performance.

We should recognize some limitations in this study. The first, the cross sectional nature of the study impeded the assessment of other effects of Web 2.0 adoption; we believe that a longitudinal analysis could reveal other effects. The scale used to the measure Web 2.0 adoption is composed of 8 items; however not all the technologies are included, it could be interesting to analyze the effect of other technologies. Another limitation is related to the single industry used in the study; therefore, generalizations about the study are not advisable until the relationships in other industries are tested.

Finally some lines of research are suggested. We propose that future studies should analyze the relationship between the adoption of Web 2.0 technologies with knowledge management process. Due to the characteristics of Web 2.0 it is possible to infer that they could have a possible impact on making tacit knowledge explicit and facilitating knowledge sharing. Another interesting study could be the analysis of the effects of these technologies on organizational learning; for example, promoting a culture of learning within the organization.

References

1. W.E. Baker and J.M. Sinkula: The synergistic effect of market orientation and learning orientation on organizational performance. Journal of Academy Market Science, Vol. 27, 411-427. (1999).
2. J. Darroch: Knowledge management, innovation and firm performance. Journal of Knowledge Management, Vol. 9, 101-115. (2005).
3. J. M. Utterback; W. J. Abernathy: A Dynamic Model of Process and Product Innovation. Omega, Vol. 3, No. 6, 639-656. (1975).
4. S. Andriole: Business Impact of Web 2.0 Technologies. Communications ACM, Vol. 53, 67-79. (2010).
5. Y. Koçak; S. Güzin: Adoption of Web 2.0 tools in distance education. International Journal of Human Science, Vol. 6, 75-89 (2009).
6. D. Churchill: A teacher's reflections on educational applications of blogs with a postgraduate class. International Journal of Continuing Engineering Education and Life-Long Learning, Vol. 19, Nos. 2/3, 112–125. (2009).
7. T. O´Reilly: What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software (2005).
8. S. Lim; D. Palacios-Marques: Culture and purpose of Web 2.0 service adoption: a study in the USA, Korea and Spain, Services Industries Journal, Vol. 31, No. 1, 123-131. (2011).
9. S. Lee; D. DeWester; S. Park: Web 2.0 and opportunities for small businesses, Service Business, Vol. 2, 335-345. (2008).
10. M. Levy: Web 2.0 implications on knowledge management, Journal of Knowledge Management, Vol. 13, 120-134. (2009).
11. M. Chui; A. Miller; R. Roberts: Six ways to make Web 2.0 work, McKinsey Quarterly, (2009).
12. E. Constantinides; S. J. Fountain: Web 2.0: Conceptual foundations and marketing issues, Journal Direct Data Digital Marketing Practice, Vol. 9, No. 3, 231–244. (2008).
13. G. Masurek: Web 2.0 Implications on Marketing, Management Organization Systems Research, Vol. 51, 69-82. (2009).
14. R. Boateng; A. Malik; V. Mbarika: Web 2.0 and Organizational Learning: Conceptualizing the Link, AMCIS Proceedings, 546. (2009).
15. N. Choi; L. Horowitz; K. Huang: Web 2.0 Use and Organizational Innovation: A Knowledge Transfer Enabling Perspective, AMCIS Proceedings, (2010).
16. S. Lim; S. Trimi; H. Lee: Web 2.0 service adoption and entrepreneurial orientation, Service Business, Vol. 4, 197-207. (2010).
17. V. A. Thompson: Bureaucracy and Innovation, Administrative Science Quarterly, Vol. 10, No. 1, 1-20. (1965).
18. OECD-EUROSTAT: The Measurement of Scientific and Technological Activities. Oslo Manual. Guidelines for Collecting and Interpreting Innovation Data, Third Edition (Paris: Organisation for Economic Co-operation and Development Eurostat, Paris, (2005).
19. R. M. Walker; F. Damanpour and C.A. Devece: Management Innovation and Organizational Performance: The Mediating Effect of Performance Management, Journal of Public Administration Research Theory, Vol. 21, 367-386. (2011).
20. S. Gopalakrishnan; F. Damanpour: A review of innovation research in economics, sociology and technology management, Omega, Vol. 25, 15-28. (1997).
21. J. R. Kimberly; M. J. Evanisko: Organizational Innovation: The Influence of Individual, Organizational, and Contextual Factors on Hospital Adoption of Technological and Administrative Innovations, Academy Management Journal, Vol. 24, 689-713. (1981).
22. M. Crossan; M. Apaydin: A multi-dimensional framework of organizational innovation: A systematic review of the literature, Journal of Management Studies, Vol. 47, 1154-1191. (2010).

23. T. H. Aas; P. E. Pedersen: The Firm-Level Effects of Service Innovation: A Literature Review, International Journal Innovation Management, Vol. 14, 759-794. (2010).
24. A. J. Johnson; C. C. Dibrell; E. Hansen: Market Orientation, Innovativeness, and Performance of Food Companies, Journal of Agribusiness, Vol. 27, 85-106. (2009).
25. A. Subramanian, Innovativeness: Redefining the concept, Journal Engineering Technology Management, Vol. 13, 223-243. (1996).
26. S. H. Hsu: Human Capital, Organizational Learning, Network Resources and Organizational Innovativeness. Total Quality Management and Business Excellence, Vol. 18, 983-998. (2007).
27. R. F. Hurley; T. M. Hult: Innovation, Market Orientation, and Organizational Learning: An Integration and Empirical Examination, Journal of Marketing, Vol. 3, 42-54. (1998).
28. R. J Calantone; S. T Cavusgil; Y. Zhao: Learning orientation, firm innovation capability, and firm performance, Industrial Marketing Management, Vol. 31, No. 6, 515–524. (2002).
29. J. G. March and R. I. Sutton, Organizational Performance as a Dependent Variable, Organization Science, Vol. 8, 698-706. (1997).
30. P. J. Richard; T. M. Devinney; G. S. Yip; G.Johnson: Measuring Organizational Performance: Towards Methodological Best Practice, Journal of Management, Vol. 35, 718-804. (2009).
31. J. Eskildsen; A. H. Westlund; K. Kristensen: The predictive power of intangibles: Measuring Business Excellence, Vol. 7, 46-54. (2003).
32. J. Dawes: The relationship between subjective and objective company performance measures in market orientation research: further empirical evidence. Marketing Bulleting-Department of Marketing Massey University, Vol. 10, 65–75. (1999).
33. G. G. Dess; R. B. Robinson: Measuring Organizational Performance in the Absence of Objective Measures: The Case of the Privately-held Firm and Conglomerate Business Unit, Strategic Management Journal, Vol. 5, 265-273. (1984).
34. G. C Kane; M. Alavi: Information technology and organizational learning: An investigation of exploration and exploitation processes, Organization Science, Vol. 18, 796–812. (2007).
35. W. E. Baker; J. M. Sinkula: Learning Orientation, Market Orientation, and Innovation: Integrating and Extending Models of Organizational Performance, Journal Market Management, Vol. 4, 295-308. (1999).
36. G. Hult, R.F Hurley and G.A Knight: Innovativeness: Its antecedents and impact on business performance, Industrial Market Management, Vol. 33, 429–438. (2004).
37. D. Jiménez-Jiménez; R. Sanz-Valle: Innovation, organizational learning, and performance, Journal of Business Research, Vol. 64, 408-417. (2011).
38. H. F Lin: Knowledge sharing and firm innovation capability: An empirical study, International Journal of Manpower, Vol. 28, 315–332. (2007).
39. K. Patrick; F. Dotsika: Knowledge sharing: developing from within, Learning Organization, Vol. 14, 395–406. (2007).
40. M. London; M. J. Hall: Web 2.0 support for individual, group and organizational learning, Human Resource Development, Vol. 14, 103-113. (2011)
41. J. K. Han; N. Kim; R. K. Srivastava: Market Orientation and Organizational Performance: Is Innovation a Missing Link?, Journal of Marketing, Vol. 62, 30-45. (1998).
42. G. Prause; D. Palacios; F. Garrigós: What happens if a hotel introduces Web 2.0?, (presented at the International Conference of the Network of Business and Management Journals Proceedings, Valencia, Spain, (2010).
43. H. T. Hurt; K. Joseph; C. D. Cook: Scales for the measurement of innovativeness, Human Communications Research, Vol. 4, 58-65. (1977).

44. C. Nakata; Z. Zhu; M. L. Kraimer: The Complex Contribution of Information Technology Capability to Business Performance, Journal of Management Issues, Vol. 20, 485-506. (2008).

Yejun Xu was born in 1979. He received an MS degree in 2005 and Ph.D degree in 2009 both in management science and engineering from SouthEast University, Nanjing, China. He is currently an associate professor with the Business School of Hohai University, Nanjing, China. He has contributed over 50 journal articles to professional journals such as *Fuzzy Sets and Systems, Information Sciences, International Journal of Approximate Reasoning, Knowledge-Based Systems, Soft Computing, Applied Soft Computing, Applied Mathematical Modelling*, *International Journal of Computational Intelligence Systems, Expert Systems with Applications, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *Journal of Systems Engineering and Electronics*, etc. His current research interests include multicriteria decision making, computing with words, and information sciences.

C. Llopis-Albert received the MSc in industrial engineering and the PhD degree (with Extraordinary Award) in Hydraulic and Environmental Engineering at Universitat Politècnica de València, Spain. He works as an adjunct faculty in the Department of Mechanical Engineering at Universitat Politècnica de València and as senior researcher at Geological Survey of Spain. He has published tens of papers on different research areas in top journals, books and conference proceedings. He is on the editorial board of several journals and scientific committees and serves as a reviewer in a wide range of journals. He has been involved in several multidisciplinary national and international research projects. His research focuses on mathematical modeling and numerical analysis in different fields such as mechanics, water resources, environmental science and economics. Special attention is paid to the integration of those fields.

Dr. Joaquín González-Garcia is currently a Tenured Associate Professor of the Department of Business Organization at the Universidad Politécnica de Valencia (U.P.V.) Spain. He is Bachelor with Honours in Electronic Engineering (BEng (Hons)) from Middlesex University of London U.K. . He is a Chartered Engineer (CEng), Member of the U.K. professional body Institution of Engineering and Technology (I.E.T.)(formerly The Institution af Electrical and Electronic Engineers) (M.I.E.T.).He worked for I.B.M. as a Quality Engineer (Q.E.) in the Engineering Department at Valencia´s Manufacturing Plant. He is a PhD in Industrial Engineering from Universidad Politécnica de Valencia (U.P.V.) Spain. Graduate in Law from The University of Valencia (U.V.) and a Member of the Lawers Professional Body (I.C.A.V.).

Received: January 13, 2014; Accepted: February 5, 2014

Computer Science and Information Systems 11(2):809–823 DOI: 10.2298/CSIS121207044E

A Business Intelligence Software Made in Romania, A Solution for Romanian Companies During the Economic Crisis

Eduard Edelhauser[1], and Andreea Ionică[2]

[1]Head of the Management Department, University of Petrosani,
Universitatii Street, No. 20, 332006, Petrosani, Hunedoara, Romania
edi1ro2001@yahoo.com
[2]Management Department, University of Petrosani,
Universitatii Street, No. 20, 332006, Petrosani, Hunedoara, Romania
andreeaionica2000@yahoo.com

Abstract. The paper focuses on a modern prospective of IT&C sector. The methodology used is both quantitative and qualitative, and the results were obtained with the use of a questionnaire. Our purpose was to demonstrate some hypothesis concerning the size of the organization, the management method and the IT&C based decision. The study is limited to the SIVECO companies portfolio and has a high level of originality, such a study has been never conducted before for computer based advanced management methods implementation. The present interest of the approach consists in the powerful impact IT&C technologies have upon the development of nowadays organizations as well as upon the daily life of each individual; nevertheless, one should not ignore the special condition Romanian economy and especially Romanian society as a whole is confronted with during the current period.

Keywords: Decision Support Systems, Business Intelligence, Romanian Companies, SIVECO Business Analyzer

1. Introduction

In the present context of economic crisis, considering our opinion and from a managerial point of view, in Eastern Europe regarding the particular case of Romania, one of the solutions for economic revival is constituted by the usage of some advanced management methods, based on IT&C tools and modeling and simulation of the economic and management processes methods that can offer the foundation for a high-performance management.

Using the IT&C systems and their applications in management decision, a company can develop a competitive advantage for the organization. In the 2013 economic crises, the use of the IT&C advanced management methods can be a solution for overcoming this crises.

We are in the middle of an economic-financial crisis and I would say that Romania is affected more due to the defective management of some leaders than to the crisis. In a country where we ask ourselves if we have more pensioners than employees or if the wages of the public employees are higher than those of the employees from the private or manufacturing or services field, we certainly have to change essential things. One of the ways of doing this is to use some advanced management methods in the Romanian organizations. [2]

The evaluations of the local Business Intelligence (BI) market, carried out by means of the main local implementers indicate an unusual tendency, namely, in the context of a severe reduction of the volume of global sales regarding company informational applications, worldwide business intelligence (BI) software revenue will reach $13.8 billion in 2013, a 7 percent increase from 2012, according to Gartner, Inc. The market is forecast to reach $17.1 billion by 2016. [13]

The paper is organized in four chapters and deals with the essential characteristics of the BI concept regarding two aspects, namely the spatial aspect by analyzing the worldwide and national BI market and the temporary aspect regarding design, implementation and maintenance, corroborated with the continuous update to the customer's needs by the analyzed feedback of the support decision systems.

The first chapter highlights the status of the main competitors on the BI market as it is presented in the studies of the prestigious companies IDC and Gartner regarding the years 2011 and 2012. There are presented both the software tools used and the fields of implementation of the informational systems in the BI category. The second chapter expounds the degree of involvement of the four main worldwide players on the BI market, (respectively SAP, Oracle, IBM and Microsoft), on the Romanian market these great actors owning 60% of the entire BI market worldwide. Moreover there is punctuated a problem specific to Romanian economy, valid for all the small and incompletely globalized markets, namely the importance of Small & Medium Enterprises (SMEs). Taking into account the fact that the large Romanian companies represent companies with public capital and subsidiaries of large multinational companies, and the large actors have already managed to implement the assisted managerial decision software within these companies, a market that is very attractive for the companies that create BI and DSS software in the management field is still that of SMEs. This market is the target of both large actors mentioned before as well as the local developers such as SIVECO by SIVECO Business Analyzer, S&T by Micro Strategy, Spectrum Solutions by Oracle BI and Wizrom by Panorama, out of which SIVECO represents by far an autochthonous leader of the Enterprise Resource Planning (ERP) and BI market.[14] The third chapter is dedicated to the BI application of SIVECO, application fully designed in Romania, and describes the main modules and types of analysis proposed by the SBA component such as what if and top bottom analysis and various types of simulations, as well as some design characteristics of the application. The last chapter is a research based on a questionnaire, and represents the essence of the effects induced by the implementation of the BI applications in the Romanian organizations. When quantifying these effects we relied on a questionnaire directed to a population made up of representative companies in the public and private field where SIVECO carried out implementations of company applications. The study is limited to the SIVECO companies portfolio and deals with the dependence between existing computers, software programs and the level of business software implementation of 11 Romanian organizations. Even data were collected only from 11 organizations, these are representative for the 2010 Romanian economy, as we have demonstrate it in paragraph 5.1 (respondents).

2. The performance of leading worldwide BI tools vendors in 2012

The main objective of any company is to earn profit. In order to achieve it, the company has to be efficient, to bring financial benefit. To that effect, there are necessary models of decision making in order to measure, manage and optimize the performance of different activities in the company. These decisions are made because of information, and BI is exactly the concept that, if implemented correctly, leads to capitalization of the information necessary to measure, manage, make decisions and optimize performances. “Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making. It allows business users to make informed business decisions with real-time data that can put a company ahead of its competitors." [5]

We can say that the Business Intelligence (BI) has an important role for the performance and competitiveness of the organization rather than marketing or the quality of its products. The actions of the BI are not unknown any more for the current, modern, competitive and efficient organization. Such an organization is considered to be ”intelligent” not because it has an intelligent management, where the economic intelligence activities are correctly and realistically exploited in setting and putting the decisions into practice. Business Intelligence is a concept which refers to the way in which decision can be made faster and easier. In the current society the companies collect huge quantities of data daily: information about orders, inventories, and transactions from work sites and of course information about customers. Companies also collect data as demographic data and email lists from external sources. Unfortunately more than 93% of the data are not used in the decision making process. The better consolidation and organization of the date to take a better decision can lead to advantages in what the competence is concerned and these aspects refer to Business Intelligence. The quantity of information and data is growing exponentially, and it is actually doubling each two, three years. More information means more competition. In the informational era due to the explosion of information, the executors, managers, professionals and workers, they all need to be able to make fast decision because now time means money. The BI applications transform information into knowledge. The concept of BI allows us to put the right information in the hands of the right user, at the right moment to be able to take a fair decision.

The essence of BI is to gain information and knowledge from data in order to enable people to make better decisions. Especially in times of economic turmoil, the quality and efficiency of this process can be a crucial advantage for any company. BI is a highly dynamical field and builds an interface of growing importance between IT and management.

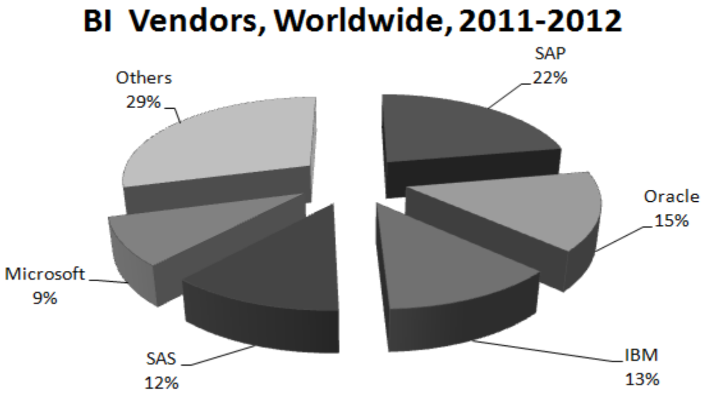


Fig. 1. Top 5 BI, CPM and Analytic Applications/Performance Management Vendors, Worldwide, 2011-2012 (Source: Gartner, June 2013)

In first place, SAP once again had significantly higher revenue than any other vendor at $2.9 billion with 22.1 percent of the market, although this was up by just 0.6 percent from 2011. Second-place Oracle's revenue grew by 2.0 percent from 2011 to reach $1.9 billion. Fifth-place Microsoft enjoyed the highest growth of the top five vendors in 2012, with revenue rising by 12.2 percent compared with 2011, to reach $1.2 billion. [14]

3. Romania Enterprise Application Software Market 2013–2017 Forecast and 2012 Vendor Shares

We tried to identify the main obstacles hindering the penetration of BI solutions on the Romanian market. Therefore, we reached the conclusion that poor technical abilities, poor education in the IT&C field regarding the top management in Romanian companies and the limited budgetary resources are impediments that we are frequently confronted with, when we are talking to a potential customer. This is not, however, a particular situation – it does not happen only in Romania, it is a situation that all the other markets are confronted with. We consider these impediments as opportunities for the solutions we promote. Nevertheless, this does not mean that there are no real obstacles, and the most frequent problem consists of the fact that the direction towards a BI solution is not seen as a business decision, but as one prevailingly technical. Or a successful deployment of a BI solution is always realized when it benefits from the support of those decision persons that are the main beneficiaries and direct users of these types of solutions, that understand the business vision, business strategy and motivation of that precise company.

There are two distinct situations on the Romanian market: the case of local companies that enter the category “large enterprise” that are aware of the necessity for BI tools, and the sector represented by Small & Medium Enterprise, a case in which is difficult to assert that a Business Intelligence solution becomes a “must have”. As for the segment of large and very large companies and multinational corporations, we can say that they have reached an awareness threshold. There are some market verticals that reached maturation very rapidly and that progressed up to the international standards such as banking, mobile telephony, big retail chains, oil industry, services etc. These verticals (as well as others), where competition is very strong, generate especially the demand for BI. [5]

Admittedly, BI materialization, under the pressure of the economic crisis, is no longer only a simple "sales argument". The concept of operational BI outran the theoretic level, gaining materiality under the pressure of a well defined objective – making the business efficient, starting with the operational level, but without neglecting the tactic level of approaching business policy. Still, remains to be seen how much interested in the new offer is the local market. Traditionally, Business Intelligence was viewed as a complex technology, in most cases destined for large companies, whose main purpose is to provide to the top management the analysis tools that granted strategic or tactic decision making. In other words, it was an abstract concept, having an action area for an unlimited time span, placed in an area hardly accessible for "commoners", that is medium enterprises that do not possess a solid investment budget and/or an extended IT&C department. Lately this view has changed alertly, the worldwide economic crisis constraining vendors to turn their attention towards medium enterprises that need analysis tools in real time in order to cope with the challenges of the market. The worldwide economic context brought about a visible turn of the Business Intelligence solution editors who, during the last two years, forcefully attacked the sector represented by Small & Medium Enterprises (SME or SMB), trying to destroy the myth according to which BI applications are accessible only for the large enterprises that propagate huge volumes of data and own extended IT&C departments. In order to achieve this, the supply addressed to SMBs was considerably diversified, trying to cater for the main "reproaches" that the potential customers in the category of medium enterprises made.

In order to broad the spectra of potential customers, there was a reorientation of the concept of BI, from the traditional "data-centric" to the more pragmatic "process-centric", meant to enable a quicker reaction to the challenges on the market that appeared as a consequence of the economic crisis. Thus, BI applications are not only for business analysts in top management any more, but they are also accessible for executive directors, managers and final users that have the decision making power who can rapidly interpret and analyze relevant information, using it as a basis to be able to make viable decisions. All the great actors of the world on the BI market – Oracle, by Hyperion, IBM by Cognos, SAP by Business Objects, Microsoft etc. – joined this trend, trying to bring supplies as attractive as possible for the SMB sector.

This IDC study provides a detailed overview of the Romanian market for integrated enterprise application software (EAS) suites. EAS vendors tracked in this study include SAP, Siveco, TotalSoft, Oracle, Microsoft, Wizrom, Epicor, Infor, QAD, Transart, EBS, Endava, and SeniorSoftware, among others. "As most clients have limited IT budgets, they prefer to spend them on IT solutions that have a direct impact on business growth, such as cost efficiency, optimization, and consolidation. As a result, vendors should focus on targeting decision makers with solutions that offer fast ROI and provide an immediate competitive advantage." – Research Analyst Dana Samson, IDC CEMA

The double perspective would enable speculating the assumption that the Romanian BI market extends due to the large enterprises, or through what is presented as a BI solution in a more elevated reference solution. In the first case we are confronted with a limited manipulation platform, which will not be able to maintain the trend up for a long time. In the second case, there is the hope that the excessive reported data, at some point, will call for a complex analysis tool. Even if "our IMMs are smaller than their SMBs", the recipe applies for Romania as well. [5]

4. SIVECO Business Analyzer a high performance Information Management System

BI Solutions of SIVECO Romania monitors and correlates all the levels of company`s activities, positively influencing its performances. SIVECO Business Analyzer (SBA) is a high performance Information Management System, capable to be adapted and customized according to the business particularities of any company. SBA supports the decision making process during the activities of planning monitoring, controlling and providing information support for the adoption of strategies for cost control and the identification of sources to increase profit. It also provides synthesis, coherent, consistent and real-time information, and represents the requested analyses under the form of graphs and tables in an appealing manner, easy to manage and customize.

The second application in the field of BI is SIVECO Balanced Score Card (SBSC) is a software solution for strategic management and a latest generation information product launched on the Romanian market, able to monitor, analyze and compare the organization's performance in order to improve it. Placing the strategy at the centre, the proposed solution provides the beneficiary with relevant information regarding the manner in which the organization is heading towards reaching its strategic objectives.

We exemplified four simulation analysis and models of SBA and SBSC that are useful in the management activity and in the research, and which are highly used by the managers of Romanian enterprises that purchased this type of product. The following Fig. 2 illustrates the analysis of the evolution of an indicator, and allows the representation of the values in the database under the form of diagrams or charts.

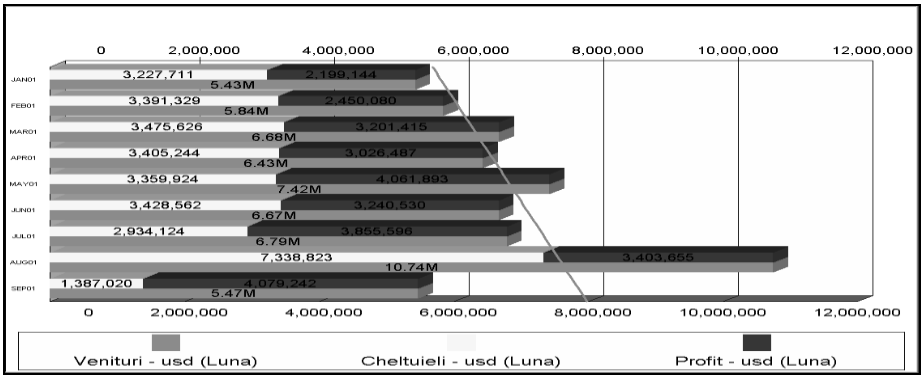


Fig. 2. The status of revenues, costs and profit regarding the entire enterprise, in which the user may change the selection of the displayed values while the program is running (Source: SIVECO Business Analyzer 3.2.90)

"What if?" Analysis consists of the possibility to modify the entrance parameters of an indicator and to see how this indicator varies after this modification. One can easily imagine potential situations when this type of analysis would be useful. For example, what happens with costs and profit if there is a certain percentage increase of the salary or power costs? What happens if the number of employees decreases or increases? On the other hand, if the value of certain taxes is modified (VAT, income taxes).

Top/Bottom Analysis. This analysis enables the selection of one of the most important types of incomes or the most important customers. The existence of this possibility is necessary especially in case the dimensions have many values (Customers in the following Fig. 3) that cannot be monitored each at large, but you want to monitor the extreme values: the highest ones and the lowest ones.

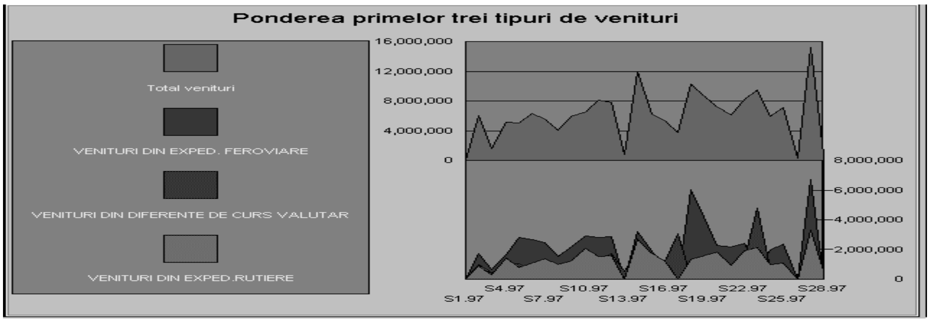


Fig. 3. Exemplification of Top/Bottom Analysis by the first three types of incomes in comparison to total incomes (Source: SIVECO Business Analyzer 3.2.90)

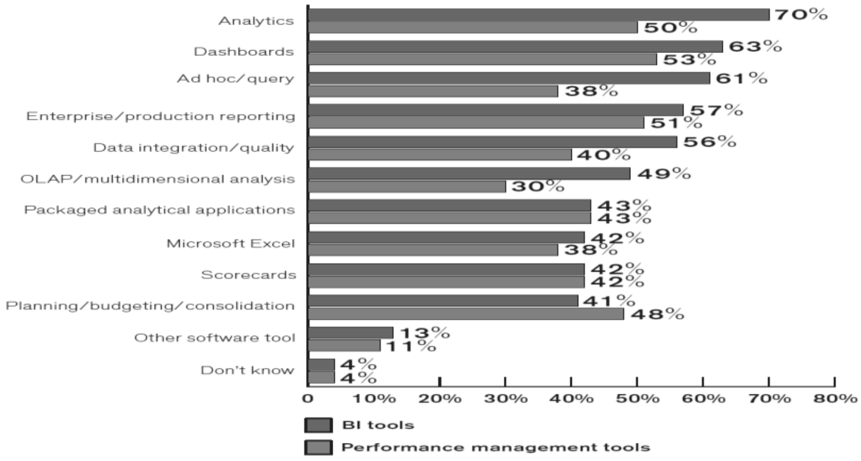
Forecasting is a simulation process in which the forecast values are based on the existing data history and the richer this history is, higher will be the accuracy of the forecast.

5. Effects induced by implementing BI applications in Romanian organizations

To start with, we tried to identify the applicability field of different applications and BI tools, the way they are perceived in the view of managers as beneficiaries of BI solutions. For this first questionnaire data were collected from 156 IT respondents. [12]

We asked the respondent companies a series of questions regarding the main obstacles in the penetration of BI solutions on the local market. Taking into account the “scores” they provided, the results are the following (ranking according to the degree of importance):

- Defective knowledge of the “concept” and of the advantages offered by the BI solutions;
- Romanian companies’ prevailingly focusing on the operational aspect (to the detriment of the strategic one);
- Low technical “appetite” of decision factors;
- Low financial power of potential customers;
- Low technical abilities of potential customers;
- Inadequate IT&C infrastructure. [12]



Base: BI tools – 156 IT respondents; performance management tools – 142 IT respondents

Fig. 4. Software tools considered part of BI and performance management

When asked which tools they consider to be part of their BI solution (see fig. 4), IT respondents most frequently mentioned analytics, dashboards, ad hoc query, enterprise production reporting and data integration quality. Dashboards, enterprise production reporting, analytics and planning budgeting consolidation are the software tools that IT respondents most frequently reported to be part of performance management.

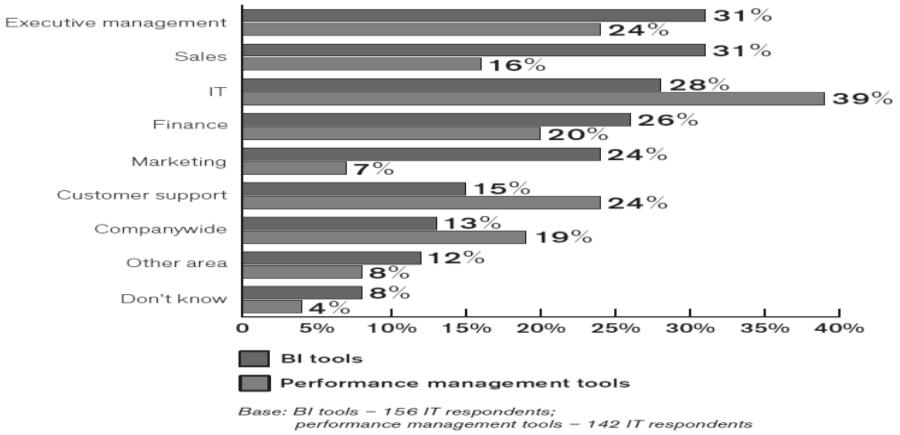


Fig. 5. Departmental implementation of BI and performance management (areas in which a company first implement a BI/performance management tool/solution)

Sales and executive management were reported as the first departments to implement BI, at 31% each, probably in response to pressure from those areas for improved data availability. The IT&C department, a group that can test the tools before the rest of the organization sees them was cited most for initial implementation of performance management tools (39%), followed by executive management and customer support, at 24% each. (see fig. 5)

The third field of investigation was the key benefit that is derived or expected to be derived from BI and performance management tools is related to the improvement of decision-making process, such as the quality and relevance of decisions made. Seventy percent of IT&C respondents indicated this as the top BI benefit, and 55% indicated this as the top performance management benefit. Producing a single, unified view of enterprise wide information (57%), better aligning resources with strategies (56%), speeding up the decision-making process (53%), and responding to user needs for availability of data on a timely basis (52%) are the other top BI benefits reported by IT respondents.

As a conclusion for BI implementation we can say that the nature of the top benefits and challenges cited by IT&C respondents make it clear that today's technology purchasers demand comprehensive and integrated BI and performance management solutions that can overcome challenges related to integrating data from multiple sources and data quality. What are companies hoping to achieve with BI and performance management solutions? Seven out of 10 BI respondents and more than half of performance management respondents cited the desire to improve the decision-making process, including the quality and relevancy of decisions. Vendors that have the technical expertise to deliver these benefits can expect to appear on the short lists of BI and performance management technology purchasers. [12]

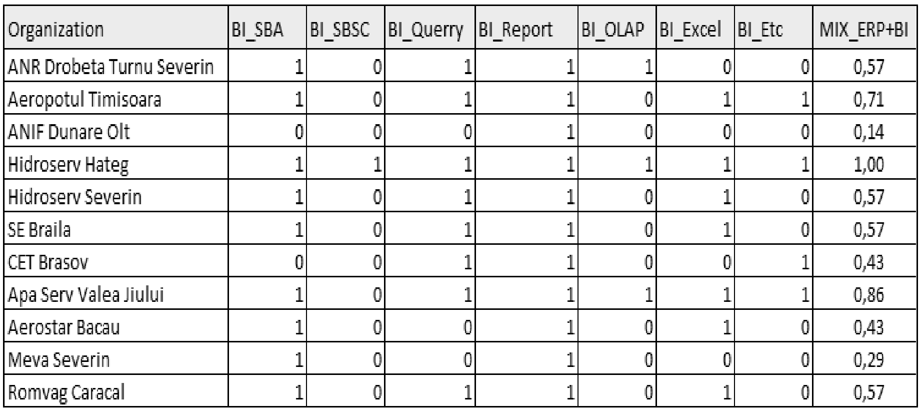
5.1. Respondents

Still in the virtue of the questionnaire, we achieved the results and we were able to formulate and validate research assumptions. Even data were collected only from 11 organizations, these are representative for the 2010 Romanian economy, because in this economical moment Romania has only 5,000 companies that need an ERP and a BI software instrument as a advanced management method. So we have only 2,000 big companies having more than 250 employees which can afford to implement a SAP, Oracle or SIVECO ERP software. But these 2,000 companies generate incomes two times higher than the other 10,000 SMB, and equal those of the 500,000 small Romanian companies, that have under 50 employees. From these 2,000 big organizations most of them, about 1000 are branches from transnational companies, and have mostly implemented ERP existing in their main organization, usually SAP or Oracle. So, are likely to be investigated public organizations and private Romanian capital organizations. These two categories have a hundred percent Romanian management, and had to optimize it.

The data were collected during January and June 2010, (see Table 1 and 2) with the help of Sorin Dimofte Implementing and Consultancy Manager of SIVECO Romania.[3]

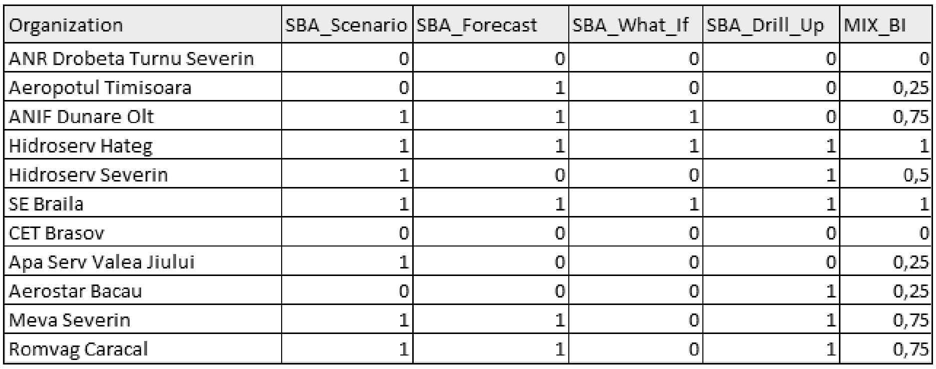
The regression analysis using the method of least squares requires a dataset of n pairs (xi,yi), where yi are the dependent variables and xi are the independent ones. There were analyzed over 500 Romanian enterprises, that implement a SIVECO ERP and were identified 11 distinct groups without any significant deviations among the enterprises forming each group. In order to apply this method there were chosen 11 companies that were included in the present study. Each of these 11 companies is representative for the enterprises group, having similar trend. This simplified version of the method is suitable for our problem because this way there were obtained very clear results.

Table 1. ERP & BI instruments implementation



Organization	BI_SBA	BI_SBSC	BI_Querry	BI_Report	BI_OLAP	BI_Excel	BI_Etc	MIX_ERP+BI
ANR Drobeta Turnu Severin	1	0	1	1	1	0	0	0,57
Aeropotul Timisoara	1	0	1	1	0	1	1	0,71
ANIF Dunare Olt	0	0	0	1	0	0	0	0,14
Hidroserv Hateg	1	1	1	1	1	1	1	1,00
Hidroserv Severin	1	0	1	1	0	1	0	0,57
SE Braila	1	0	1	1	0	1	0	0,57
CET Brasov	0	0	1	1	0	0	1	0,43
Apa Serv Valea Jiului	1	0	1	1	1	1	1	0,86
Aerostar Bacau	1	0	0	1	0	1	0	0,43
Meva Severin	1	0	0	1	0	0	0	0,29
Romvag Caracal	1	0	1	1	0	1	0	0,57

Table 2. BI instruments implementation



Organization	SBA_Scenario	SBA_Forecast	SBA_What_If	SBA_Drill_Up	MIX_BI
ANR Drobeta Turnu Severin	0	0	0	0	0
Aeropotul Timisoara	0	1	0	0	0,25
ANIF Dunare Olt	1	1	1	0	0,75
Hidroserv Hateg	1	1	1	1	1
Hidroserv Severin	1	0	0	1	0,5
SE Braila	1	1	1	1	1
CET Brasov	0	0	0	0	0
Apa Serv Valea Jiului	1	0	0	0	0,25
Aerostar Bacau	0	0	0	1	0,25
Meva Severin	1	1	0	1	0,75
Romvag Caracal	1	1	0	1	0,75

5.2. Results

5.2.1. Research Hypothesis

Although in elaborating the research hypotheses we have had in view a series of works belonging to American literature[10],[7],[1],[6], we have made these hypotheses particular for Romanian organizations according to our own experience of almost 20 years in the field of planning, implementing, and of training in the field of informational systems.

H01 The number of employees in an organization influences the role of the BI applications within the respective organizations. The organization dimension is directly connected with the role of the BI applications within the respective organization.[10]

H02 In the private domain there is a more efficient usage of money than in the public domain.[7]

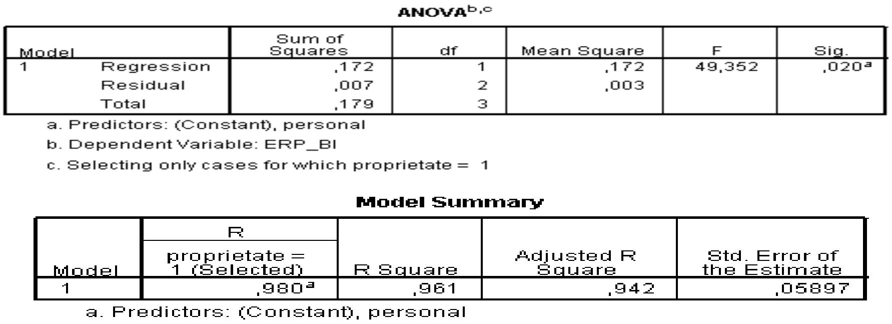
5.2.2. Testing the Hypothesis

H01 The number of employees in an organization influences the role of the BI applications within the respective organizations. The organization dimension is directly connected with the role of the BI applications within the respective organization.

We used regression analysis, as a statistical method to evaluate the relation between one independent variable (personal - size of organization) and another continuous dependent variable (ERP_BI given to the ERP and BI level of implementation). With this analysis tool we have performed a linear regression analysis using the method of the least square in order to plot a line by a set of observations. Thus we have perform the analysis of the dependence and we have appreciated the extent to which the independent variable influence the dependent. With linear regression we output the regression coefficients necessary to predict one variable ERP_BI from the other personal. The model has been confirmed to be valid because the F test value were 49,35, with significant Sig. <0,05 (0,02). The regression coefficient R=0,980 shows a very strong link between the variable ERP_BI given to the ERP and BI level of implementation and the independent variable personal showing the size of the organization, for the private sector. The model explains 96,1% from the total variation of the variable personal (R^2= 0,961). The rest of 3,9% is influenced by other residual factors not included in the model. (Table 3)

In conclusion hypothesis H01 has been confirmed.

Table 3. Linear regression analysis between an independent variable called personal and a dependent variable called ERP_BI for private cases (proprietate=1)



ANOVA[b,c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,172	1	,172	49,352	,020[a]
	Residual	,007	2	,003		
	Total	,179	3			

a. Predictors: (Constant), personal

b. Dependent Variable: ERP_BI

c. Selecting only cases for which proprietate = 1

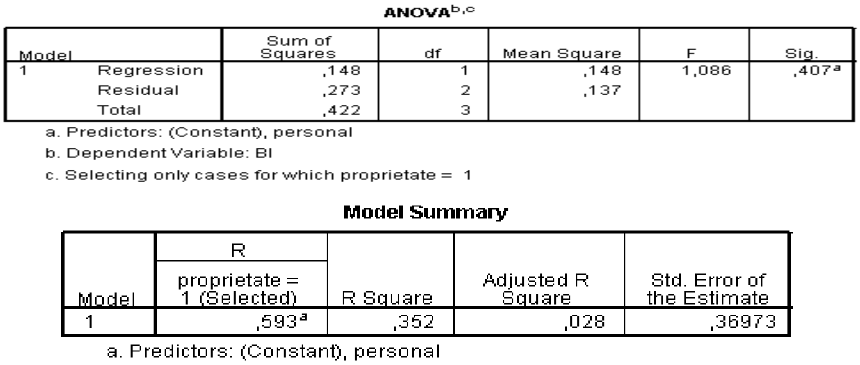
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	proprietate = 1 (Selected)			
1	,980[a]	,961	,942	,05897

a. Predictors: (Constant), personal

But in BI methods we found a weak link (R=0,167) and also for the private sector we found R=0,593 < 0,63. This regression coefficient R=0,593 shows an intermediate link in these case. (Table 4)

Table 4. Linear regression analysis between an independent variable called personal and a dependent variable called BI for private cases (proprietate=1)



ANOVA[b,c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,148	1	,148	1,086	,407[a]
	Residual	,273	2	,137		
	Total	,422	3			

a. Predictors: (Constant), personal

b. Dependent Variable: BI

c. Selecting only cases for which proprietate = 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	proprietate = 1 (Selected)			
1	,593[a]	,352	,028	,36973

a. Predictors: (Constant), personal

H02 In the private domain there is a more efficient usage of money than in the public domain.

This hypothesis would be explained by the following specifications:

In private companies hardware facility is correlated with the number of employees

In private companies hardware facility is correlated with software facility

We used statistical techniques in order to define the differences between the groups, using T test. For this purpose, we tested the average equality of two paired samples

(Paired Samples T Test) for six variables grouped in four pairs each. Out of the six variables three pertain to the companies with public capital (pers_public, calc_public and BI_public), respectively three to companies with private capital (pers_privat, calc_privat and BI_privat).

Table 5. The result of the T test for six variables grouped in four pairs each

	N	Correlation	Sig.
Pair 1 pers_public & calc_public	4	0.997	0.003
Pair 2 pers_privat & calc_privat	4	0.632	0.368
Pair 3 calc_public & BI_public	4	0.613	0.387
Pair 4 calc_privat & BI_privat	4	0.979	0.021

We noticed (see table 5) that if in the public domain exists a better correlation between the number of employees and the number of computer-related equipment (0.997 in comparison with 0.632), the situation is different regarding the correlation hardware – software which we defined as using computers for BI applications. (In this case, the figures in the private domain are 0.979 in comparison with 0.613 in the public domain). We reach the conclusion that providing with excessive hardware is not necessarily a useful thing to do, if that hardware is not then correlated with the software devices. There was also validated the hypothesis referring to the better efficiency of investing money in the private domain, and also a better correlation between the efficiency of the hardware and employee resources and the implementation of a BI in the private domain in comparison to the public one.

6. Conclusion

The conclusion of the authors is that in the 2010-2012 years, Romania, a country that has lately managed to make itself known worldwide, most often by means of local achievements in the field of software industry, and less in other fields of industry, should take advantage of this appetite of the Romanians for the computer-science field and try to improve the management of the organizations. This could be made by prevailingly using the management methods from the category of DSS and BI that are based on informational tools; and if possible the product has to originate in the Romanian market and at the same time, we should always monitor especially the efficiency of the implementation and the efficiency of the investment in informational software.

Thus, modeling and simulation prove to be useful, reaching a maximum level of efficiency, consistency and importance by means of the BI tools. The applications in the category of informational systems for the support of managerial decision that the team of authors studied, promoted and implemented are especially BI solutions made in Romania. In this way, more and more Romanian organizations benefit from the advantages of special Romanian decision-making oriented software, and here we refer to the applications of SIVECO software house. The implementation of the BI applications is studied worldwide as well as at a national level, and the efficiency of local BI software usage is assessed and quantified especially by means of managerial research carried out on the SIVECO customers.

Our scientific and teaching activities over the past few years have attempted to point out certain essential elements of integrated information systems, used as decision and management instruments available for managers. Thus for more than seven years we have tried to induce the managers of the organizations, with which we have collaborated, the awareness that the ERP systems are most mere accounting applications. Actually accounting is not exactly the most remarkable advantage of the implementation of an integrated resource planning system, as it is mostly perceived. On the contrary an ERP is a corporate instrument which connects users with distinct responsibilities on the same platform, and involves an information exchange between department and the individual in charge, the decision maker.

Another direction of our activities was the awareness of both managers and end-users that information technology is an instrument that supposed to be use with maximum benefits, not a dangerous element for the employees and for the organization. Future research will also include Small and Medium Business (SMB's) organizations which in the present economic environment probably represent one of the element to re launch the Romanian economy. This sector will be analyzed by the implementations that Senior Software company, through the Senior ERP application has carried out on the Romanian SMB's market.

References

1. Dai, Z., Duserick, F.: Toward Process-oriented ERP Implementation, Competition Forum Journal, 1, January, 148. (2006)
2. Edelhauser, E., Ionică,A.: Enterprise Resource Planning and Business Intelligence, Advanced Management Methods for Romanian Companies, in Proc.1st Int.Conf. Management: Twenty Years After, How Management Theory Works, Todesco Publishing House, Romania, Cluj Napoca, 63-72. (2010)
3. Edelhauser, E.: IT&C Impact on the Romanian Business and Organizations. The Enterprise Resource Planning and Business Intelligence Methods Influence on Manager's Decision: A Case Study, Informatica Economica Journal, 2 (15), 16-28. (2011)
4. Evelson, B.: Topic Overview: Business Intelligence, November 21, (2008) Available: http://www.forrester.com/rb/Research/topic_overview_business_intelligence/q/id/39218/t/2 . (current September 2013)
5. Hurbean, L., Danaiata, D., and Hurbean,C.: Enterprise Applications Implementation in Romanian SMEs, Working Paper Series at Social Science Research Network, (2008) Available: http://ssrn.com/abstract=1360722
6. Jones, M., Young, R.: ERP Usage in Practice: An Empirical Investigation, Information Resources Management Journal, 19 (1) January-March, 23-42. (2006)
7. King, R.W.: Ensuring ERP Implementation Success, Information Systems Management Journal, 3 (22), 83-84. (2005)
8. Leba, M., Pop, E.: Adaptive Software Oriented Automation for Industrial Elevator, in Proc. 11th Int. Conf. on Automatic Control, Modeling & Simulation, Turkey, Istanbul, 128-133. (2009)

9. Pop, E., Leba,M.: Software Engineering Approach on Administration Management, in Proc. 8th Int. Int. Conf. on Circuits, Systems, Electronics, Control & Signal Processing, Spain, Puerto de la Cruz, Tenerife, 186-191. (2009)
10. Princely, I.: Interactions Between Organizational Size, Culture, and Structure and Some IT Factors in the Context of ERP Success Assessment: An Exploratory Investigation, The Journal of Computer Information Systems, 4 (47), 28-44. (2007)
11. Velicanu, M, Lungu, I.: Business Intelligence Technology, ASE Publishing House, Bucharest. (2009)
12. ***,: Implementation of Business Intelligence and Performance Management Tools and Solutions, Research conducted Computerworld, White Paper, (2007) Available: http://www.pmi.it/file/whitepaper/000184.pdf (current September 2013)
13. ***,: Gartner Says Worldwide Business Intelligence Software Revenue to Grow 7 Percent in 2013, Stamford, Conn., February 19, 2013, Available: http://www.gartner.com/newsroom/id/2340216 (current September 2013)
14. ***,: Gartner Says Worldwide Business Intelligence, CPM and Analytic Applications/Performance Management Software Market Grew Seven Percent in 2012, Available: http://www.gartner.com/newsroom/id/2507915\ (current September 2013)

Eduard Edelhauser is Associate Professor at the University of Petrosani, has a BSc in engineering (1991), a MSc in computer science (2002) and a Bachelor's Degree in economics (2009). He has also a PhD in Industrial Engineering (2004) and a PhD in Management (2011). His areas of interest are information systems and management. He teaches courses of Management Information Systems, Computer Aided Design and Management and is the author of 10 books and over 100 scientific papers.

Andreea Ionica is Associate Professor at the University of Petrosani, has Graduated the University of Petrosani as engineer (1992), as economist (2002) and PhD in Industrial Engineering (2004). Her research interests include: Quality Management Systems (QMS) and TQM implementation. She participated as coordinator or member in about 10 national and international research projects, two of them having eLearning related theme, and grants and published about 100 papers.

Received: December 7, 2012; Accepted: January 16, 2014

Computer Science and Information Systems 11(2):825–838 DOI: 10.2298/CSIS121225045A

Fuzzy Claim Reserving In Non-Life Insurance

Jorge de Andrés-Sánchez

Social and Business Research Laboratory
Rovira i Virgili University
Avinguda de la Universitat 1, 43204, REUS, SPAIN
jorge.deandres@urv.cat

Abstract. This paper develops several expressions to quantify claim provisions to account in financial statements of a non-life insurance company under the hypothesis of a fuzzy environment. Concretely, by applying the expected value of a fuzzy number and the more general concept of value of a fuzzy number to the ANOVA claim predicting model [2] we estimate claim reserves to account in insurer's balance sheet and income account.

Key-words: Fuzzy logic; Fuzzy numbers; Value of a Fuzzy Number, Expected Value of a Fuzzy Number; Insurance; Claims provisions

1. Introduction

As Dubois and Prade [15] point out Fuzzy Sets Theory (FST) and its extensions are applied, basically, in the following three circumstances: gradualness, epistemic uncertainty and bipolarity. So, although actuarial quantitative analysis is essentially based on statistical methods, academics and practitioners now tend to believe that FST is a useful complement to statistics in the cases that require a great deal of actuarial subjective judgement and problems for which the information available is scarce or vague. An extended survey can be found in [29]. So, the Encyclopaedia of Actuarial Science in [13] dedicated a chapter to FST. Any case, as it is pointed out in [27], fuzziness does not figure as central to any science and, of course, Actuarial Mathematics is not an exception.

One of the most interesting areas of FST for actuaries is Fuzzy Data Analysis (FDA). As Statistics, FST provides several techniques for searching and ordering the information contained in empirical data (e.g. for grouping elements, to find relations between variables, etc.). Within an actuarial context, Fuzzy Regression (FR) has been used intensively in several areas. In life insurance field, [1] and [22] use two different Fuzzy Regression (FR) methods to fit the temporal structure of interest rates whereas [21] develops a FR methodology to forecast mortality with a Lee-Carter model. In non-life insurance Berry-Stölze et al. in [4] use FR to evaluate solvency requirements for property-liability insurers.

In non-life claim reserving, the use of FR is motivated, basically, because of it is not advisable to use a wide data-base since data too far from the present can lead to unrealistic estimates. In this context, Andrés-Sánchez in [2] proposes a claim reserving method that mixes FR with the ANOVA reserving method [23], which has been used intensively in actuarial literature (see e.g. [8, 28]). Andrés-Sánchez's method estimates future liabilities not only as a point values but also their variability with the use of Fuzzy Numbers (FNs) instead of random variables.

The fuzzy estimation of claiming costs finally needs to be transformed into a crisp equivalent in order, for example, to compute them in financial statements. So, [2] takes into account only non-discounted reserves (i.e. does not considers the profit of the assets that support liabilities) and uses the concept of the expected value [7]. This paper extends those results in two ways. Firstly, we also introduce in the analysis the expected profit of assets and so we obtain discounted reserves. We consider that the interest rate for finding the present value of future claims will be estimated subjectively by the actuary and so, it is a very natural to quantify that magnitude with a FN. Financial pricing with fuzzy parameters has been widely developed [5, 16, 20, 25]. Subsequently, fuzzy financial mathematics has been extended to life insurance pricing [3, 26] and non-life insurance pricing [1, 9, 12, 23]. Secondly, to transform fuzzy liabilities into crisp estimates we will use the concept of value of FNs [11] which generalises the expected value of a FN [7]. It will allow us weighting all the possible values of fuzzy estimates of liabilities taking into account their actual reliability.

The structure of the paper is as follows. In the next section we shall describe the aspects of fuzzy arithmetic and deffuzyfication that are used in this paper. Section 3 exposes claim reserving prediction exposed in [2] and develops several expressions for the value of reserves to be accounted. Those expressions will depend on the weighting function used for defuzzifying and if we discount future liabilities with an interest rate. Subsequently we develop a numerical application. Finally, we state the most important conclusions of the paper.

2. Fuzzy Numbers and related concepts

A Fuzzy Number (FN) is a fuzzy subset $\tilde{a}$ defined over real numbers which is normal (i.e. $\sup_{\forall x \in X} \mu_{\tilde{a}}(x) = 1$) and convex (i.e. its α-cuts must be convex sets). For practical purposes, Triangular Fuzzy Numbers (TFNs) are widely used FNs since they are easy to handle arithmetically and they can be interpreted intuitively. We shall symbolise a TFN $\tilde{a}$ as $\tilde{a}=(a, l_a, r_a)$ where a is the centre and l_a and r_a are the left and right spreads, respectively. Analytically, a TFN is characterised by its membership function $\mu_{\tilde{a}}(x)$ and its α-cuts, a_α:

$$\mu_{\tilde{a}}(x) = \begin{cases} \dfrac{x-a+l_a}{l_a} & a-l_a < x \leq a \\ \dfrac{a+r_a-x}{r_a} & a < x \leq a+r_a \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(1)}$$

$$a_\alpha = \left[\underline{a}(\alpha), \overline{a}(\alpha)\right] = \left[a - l_a(1-\alpha), a + r_a(1-\alpha)\right] \quad \textbf{(2)}$$

In many actuarial analyses, it is often necessary to evaluate functions (e.g. the net present value of an annuity), which we shall name $y=f(x_1, x_2, \dots,x_n)$. Then, if $x_1,x_2,\dots,x_n$ are not crisp numbers but the FNs $\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_n$, $f(\cdot)$ induces the FN $\tilde{b} = f(\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_n)$. Buckley and Qu in [6] demonstrate that if the function $f(\cdot)$ that induces $\tilde{b}$ is increasing with respect to the first m variables, where $m \leq n$, and decreasing with respect to the last n-m variables, then b_α is:

$$b_\alpha = \left[\underline{b}(\alpha), \bar{b}(\alpha)\right] = \left\lfloor f\left(\underline{a_1}(\alpha), \dots, \underline{a_m}(\alpha), \overline{a_{m+1}}(\alpha), \dots, \overline{a_n}(\alpha)\right), f\left(\overline{a_1}(\alpha), \dots, \overline{a_m}(\alpha), \underline{a_{m+1}}(\alpha), \dots, \underline{a_n}(\right. \right. \quad \textbf{(3)}$$

It is very usual in real insurance situations to estimate magnitudes as approximate quantities, for example, by means of a sentence like "the claim provisions must be around 2000 monetary units". Clearly, FNs can be used to represent these magnitudes. However, these magnitudes also often need to be quantified with crisp values. For example, in our context, this will occur when the definitive amount of claim provisions needs to be specified in financial statements. This paper proposes using the concept of value of a FN by Delgado et al. in [11], but adapted in order to introduce decision maker's risk aversion in the sense of Hurwitz. For a FN $\tilde{a}$, it will be we symbolised as $V[\tilde{a}, \beta]$. If we use a decision-maker risk aversion β, where $0 \leq \beta \leq 1$ we obtain:

$$V[\tilde{a}, \beta] = (1-\beta)\int_0^1 w(\alpha)\underline{a}(\alpha)\mathrm{d}\alpha + \beta\int_0^1 w(\alpha)\bar{a}(\alpha)\mathrm{d}\alpha \quad \textbf{(4a)}$$

where $w(\alpha)$ is the weighting function. In this paper, following [11], we will consider $w(\alpha)=2\alpha$. Likewise, as it is done in [2], we will also consider $w(\alpha)=1$ (there is no weighting). Notice that under this hypothesis, (3) is now the expected value of a FN (Campos and González, 1989). In this case, the value will be symbolised as $EV[\tilde{a}, \beta]$.

Notice that the value of a FN is an additive measure, and so:

$$V\left[\sum_{i=1}^{n} \tilde{a}_i, \beta\right] = \sum_{i=1}^{n} V[\tilde{a}_i, \beta] \quad \textbf{(4b)}$$

3. Fuzzy claim reserving

3.1. Fitting future claiming with fuzzy ANOVA

Let us symbolise as $s_{i,j}$ the claim cost of the insurance contracts originated in the ith period (i=0,1,...,n) within the jth claiming period (j=0,1,...,n). Given that the past claiming costs of the ith occurrence period are $s_{i,j}$, j=0,1,...,n-i, calculating claim reserves implies forecasting and adding the claim costs of future development periods: $s_{i,j}$, i=1,2,...,n; j=n-i+1, n-i+2,..., n.

Andrés-Sánchez in [2] extends ANOVA claim reserving method [23], to the use of Fuzzy Regression. ANOVA chain ladder, as classical chain ladder, supposes that $s_{i,j}$ can be represented by the product $C_i{\cdot}p_j$ where C_i is the total claiming cost in the ith origin period, whereas p_j is the proportion of this cost paid in the jth development period. Therefore, the parameters C_i, i=0,1,...,n and p_j, j=0,1,...,n can be obtained by using linear regression since ln $s_{i,j}$ = ln C_i+ln p_j. To introduce uncertainty[23] supposes that:

$$s_{i,j} = C_i p_j \varepsilon_{i,j} \qquad \textbf{(5)}$$

where $\varepsilon_{i,j}$ is a random variable whose mean is 1. The theoretical model (5) can be transformed into the following linear regression equation:

$$\ln s_{i,j} = a+b_i+c_j+\varepsilon_{i,j}\ ,\ i=1,2,...,n;\ j=1,2,...,n \qquad \textbf{(6)}$$

where here, $\varepsilon_{i,j}$, i=0,1,...,n; j=0,1,...,n-i are identical and uncorrelated distributed normal random variables with mean 0 and variance σ^2. Notice that in this model the incremental cost of claims $s_{i,j}$ is a log-normal random variable, because (6) lets us deduce:

$$s_{i,j} = e^{a+b_i+c_j+\varepsilon_{i,j}} \qquad \textbf{(7)}$$

Here, e^a can be interpreted as the incremental claim cost in the initial origin and development periods (i=0, j=0). Thus, b_i can be understood as the logarithmical growth rate of total claiming during the origin period i, (i=1,2,...,n) with respect to the initial origin period. Analogously, for a given origin period, c_j is the logarithmical growth rate of the incremental cost of claims during the development period j (j=1,2,...,n) with respect to the cost of claims declared during the development period j=0.

On the other hand [2] considers that the uncertainty about incremental claims is due to fuzziness and it is not exogenous to total claiming cost, C_i, and proportions p_j, j=1,2,…,n. So, we will obtain an estimate of the incremental claim cost $s_{i,j}$ by means of the fuzzy number $\tilde{s}_{i,j}$:

$$\tilde{s}_{i,j} = \tilde{C}_i \tilde{p}_j \qquad \textbf{(8a)}$$

then, the linear regression model to fit is analogous to (6):

$$\ln \tilde{s}_{i,j} = \tilde{a} + \tilde{b}_i + \tilde{c}_j \qquad \textbf{(8b)}$$

Andrés-Sánchez in [2] supposes that $\tilde{a}, \tilde{b}_i$ and $\tilde{c}_j$ are TFNs and therefore $\ln \tilde{s}_{i,j}$ is also a TFN. If we symbolise $\tilde{a} = (a, l_a, r_a)$, $\tilde{b}_i = \left(b_i, l_{b_i}, r_{b_i}\right)$, i=1,2,...,n and $\tilde{c}_j = \left(c_j, l_{c_j}, r_{c_j}\right)$, j=1,2,...,n, we obtain from (6) and (8b):

$$\begin{aligned}\ln \tilde{s}_{i,j} &= \left(\ln s_{i,j}, l_{\ln s_{i,j}} r_{\ln s_{i,j}}\right) = (a, l_a, r_a) + \left(b_i, l_{b_i}, r_{b_i}\right) + \left(c_j, l_{c_j}, r_{c_j}\right) \\ &= \left(a + b_i + c_j, l_a + l_{b_i} + l_{c_j}, r_a + r_{b_i} + r_{c_j}\right)\end{aligned} \qquad \textbf{(8c)}$$

Notice that in the fuzzy regression model (8b) $\tilde{a}$ is the independent term whereas $\tilde{b}_i, \tilde{c}_j$ are the coefficients of dichotomous explanatory variables, in such a way that their observations are equal to one when the period in which they are located is equal to the parameter period, and zero in the other case. Therefore, although in the FR model that we use, the spreads of the estimate response are increasing with respect to the magnitude of the explanatory variables, this is not a problem since we are handling dummy independent variables.

After fitting the model (8a)-(8c) that is done with the FR model described in [19], we can now predict the claiming cost of all origin periods in the development periods in which they are unknown. Given that the logarithm of incremental claim costs $\ln \tilde{s}_{i,j}$, i=1,2,...,n; j≥n-i+1 are the TFN in (8b) and that the incremental claim cost $\tilde{s}_{i,j}$ is:

$$\tilde{s}_{i,j} = e^{\ln \tilde{s}_{i,j}} \tag{9a}$$

Then, the α-cuts of $\tilde{s}_{i,j}$, s_{i,j_α} are obtained from (2) and (3):

$$s_{i,j_\alpha} = \left[\underline{s_{i,j}}(\alpha), \overline{s_{i,j}}(\alpha)\right] = \left[e^{a+b_i+c_j-\left(l_a+l_{b_i}+l_{c_j}\right)(1-\alpha)}, e^{a+b_i+c_j+\left(r_{a'}+r_{b_i}+r_{c_j}\right)(1-\alpha)}\right] \tag{9b}$$

3.2. Fitting accounting value for claim reserves

Firstly, we fit a crisp quantification for non-discounted claim reserves, i.e., without taking into account the return of the assets that support these reserves. In this case, we must obtain the value of $\tilde{s}_{i,j}$, $V\left[\tilde{s}_{i,j};\beta\right]$ by applying (3) to (9b):

$$V\left[\tilde{s}_{i,j};\beta\right]=(1-\beta)\int_0^1 w(\alpha)e^{a+b_i+c_j-\left(l_a+l_{b_i}+l_{c_j}\right)(1-\alpha)}\mathrm{d}\alpha+\beta\int_0^1 w(\alpha)e^{a+b_i+c_j+\left(r_a+r_{b_i}+r_{c_j}\right)(1-\alpha)}\mathrm{d}\alpha \tag{10a}$$

In the case where $w(\alpha)$=2α, (10a) must be integrated by parts and so:

$$V\left[\tilde{s}_{i,j};\beta\right]=2\left[(1-\beta)e^{A}\frac{(B-1)e^{B}+1}{B^2}+\beta e^{D}\frac{1-(C+1)e^{-C}}{C^2}\right] \tag{10b}$$

where:

$A=a+b_i+c_j-\left(l_a+l_{b_i}+l_{c_j}\right); B=l_a+l_{b_i}+l_{c_j}; C=r_a+r_{b_i}+r_{c_j}; D=a+b_i+c_j+\left(r_a+r_{b_i}+r_{c_j}\right)$

Likewise if w(α)=1 we obtain $EV\left[\tilde{s}_{i,j};\beta\right]$ in such a way that:

$$EV\left[\tilde{s}_{i,j};\beta\right]=(1-\beta)\frac{e^{A}}{B}\left[e^{B}-1\right]+\beta\frac{e^{D}}{C}\left[1-e^{-C}\right] \tag{11}$$

where the parameters A, B, C and D are equal to the case (10b). Of course, in both cases, (10b) and (10c), if $l_a = l_{b_i} = l_{c_j} = 0$, the first summand is simply $(1-\beta)e^{a+b_i+c_j}$ and when $r_a = r_{b_i} = r_{c_j} = 0$, the second summand must be $\beta e^{a+b_i+c_j}$.

The non-discounted provision corresponding to the ith origin period is obtained by doing:

$$\widetilde{PRO}_i = \sum_{j=n-i+1}^{n} \tilde{s}_{i,j} \tag{12}$$

therefore, the provision for all the occurrence periods is:

$$\widetilde{PRO} = \sum_{i=1}^{n} \widetilde{PRO}_i = \sum_{i=1}^{n} \sum_{j=n-i+1}^{n} \tilde{s}_{i,j} \tag{13}$$

From the α-cuts of $\tilde{s}_{i,j}$ in (9b) we can derive the exact value for the α-cuts of $\widetilde{PRO}_i$ and $\widetilde{PRO}$ by applying (3):

$$PRO_{i\alpha} = \left[\underline{PRO_i}(\alpha), \overline{PRO_i}(\alpha)\right] = \left[\sum_{j=n-i+1}^{n} e^{a+b_i+c_j-\left(l_a+l_{b_i}+l_{c_j}\right)(1-\alpha)}, \sum_{j=n-i+1}^{n} e^{a+b_i+c_j+\left(r_a+r_{b_i}+r_{c_j}\right)(1-\alpha} \right. \tag{14}$$

$$PRO_{\alpha} = \left[\underline{PRO}(\alpha), \overline{PRO}(\alpha)\right] = \left[\sum_{i=1}^{n} \sum_{j=n-i+1}^{n} e^{a+b_i+c_j-\left(l_a+l_{b_i}+l_{c_j}\right)(1-\alpha)}, \sum_{i=1}^{n} \sum_{j=n-i+1}^{n} e^{a+b_i+c_j+\left(r_a+r_{b_i}+r_{c_j}\right)(1-\alpha} \right. \tag{15}$$

To account for the claim provision in the financial statements, we must transform $\widetilde{PRO}_i$ and $\widetilde{PRO}$ into the crisp numbers PRO_i^* and PRO^* respectively. To do so, we will use the concept of value of a FN which we have described in (3) and (4). In our problem, to fix β, we must bear in mind that actuarial decisions must be prudent, that is, β>0.5. Given that the expected value of FN is an additive measure, we can obtain PRO_i^* and PRO^* by aggregating $V\left[\tilde{s}_{i,j};\beta\right]$, i=1,2,...,n; j=n-i+1,...,n, that we have obtained in (10a)-(10c). Thus:

$$PRO_i^* = V\left[\widetilde{PRO}_i;\beta\right] = \sum_{j=n-i+1}^{n} V\left[\tilde{s}_{i,j};\beta\right] \tag{16}$$

Therefore, the crisp provision for all the occurrence years is:

$$PRO^* = V\left[\widetilde{PRO};\beta\right] = \sum_{i=1}^{n} \sum_{j=n-i+1}^{n} V\left[\tilde{s}_{i,j};\beta\right] \tag{17}$$

However, these developments do not take into account that the assets that support future liabilities produce financial returns and so, non-discounted provisions overrate the real value of reserves. To avoid this problem, the regulation of many countries allows discounting future liabilities to quantify claim reserves. We consider that the interest rate for finding the present value of future claims will be estimated subjectively by the actuary and so, it is a very natural to quantify that magnitude with a FN. This is a relatively common hypothesis in non-life insurance, [1, 9, 12, 23], that extend financial pricing with fuzzy parameters [5, 16, 20, 25] to non-life insurance pricing.

This paper supposes a constant force of interest throughout the temporal horizon estimated with the TFN $\tilde{\rho} = \left(\rho, l_\rho, r_\rho\right)$, that the periodicity of claiming is annual and that its distribution is uniform within the year. As is pointed out in [18], this last hypothesis is applied in practice by supposing that the cost of claims is paid in the middle of the year. Therefore, the discounted value of the incremental claiming of the *i*th origin year during the *j*th development period will be symbolised as $d\tilde{s}_{i,j}$, *i*=1,2,...,*n*; *j*≥*n*-*i*+1 and it is obtained as:

$$d\tilde{s}_{i,j} = \tilde{s}_{i,j}e^{\tilde{\rho}[n+1/2-(i+j)]} = e^{\tilde{a}+\tilde{b}_i+\tilde{c}_j+\tilde{\rho}[n+1/2-(i+j)]} \qquad \textbf{(14a)}$$

Then, the α-cuts of $d\tilde{s}_{i,j}$ are obtained by using (2) and (3):

$$ds_{i,j_\alpha} = \left[\underline{ds_{i,j}}(\alpha), \overline{ds_{i,j}}(\alpha)\right] = \left[e^{a+b_i+c_j+\rho[n+1/2-(i+j)]-\{l_a+l_{b_i}+l_{c_j}-r_\rho[n+1/2-(i+j)]\}(1-\alpha)},\right.$$
$$\left. e^{a+b_i+c_j+\rho[n+1/2-(i+j)]+\{r_a+r_{b_i}+r_{c_j}-l_\rho[n+1/2-(i+j)]\}(1-\alpha)}\right] \qquad \textbf{(18)}$$

So, the value of the discounted incremental cost of claims $V\left[d\tilde{s}_{i,j};\beta\right]$ is obtained by applying (3) to (14b):

$$V\left[d\tilde{s}_{i,j};\beta\right] = (1-\beta)\int_0^1 w(\alpha)e^{a+b_i+c_j+\rho[n+1/2-(i+j)]-\{l_a+l_{b_i}+l_{c_j}-r_\rho[n+1/2-(i+j)]\}(1-\alpha)}\mathrm{d}\alpha$$
$$+\beta\int_0^1 w(\alpha)e^{a+b_i+c_j+\rho[n+1/2-(i+j)]+\{r_a+r_{b_i}+r_{c_j}-l_\rho[n+1/2-(i+j)]\}(1-\alpha)}\mathrm{d} \qquad \textbf{(19)}$$

In the case where $w(\alpha)$=2α, (15a) is also (10b) but in this case the parameters *A*, *B*, *C* and *D* are:

$$A = a+b_i+c_j+\rho[n+1/2-(i+j)]-\left(l_a+l_{b_i}+l_{c_j}-r_\rho[n+1/2-(i+j)]\right),$$
$$B = l_a+l_{b_i}+l_{c_j}-r_\rho[n+1/2-(i+j)]$$
$$C = r_a+r_{b_i}+r_{c_j}-l_\rho[n+1/2-(i+j)]$$
$$A = a+b_i+c_j+\rho[n+1/2-(i+j)]-\left(r_a+r_{b_i}+r_{c_j}-l_\rho[n+1/2-(i+j)]\right) \qquad \textbf{(15b)}$$

Likewise if $w(\alpha)$=1 we obtain $EV\left[\tilde{s}_{i,j};\beta\right]$ and also the result is (10c) but the values of *A*, *B*, *C* and *D* are in (15b).

The discounted reserve to account for the *i*th origin period is then:

$$dP\tilde{R}O_i = \sum_{j=n-i+1}^{n} d\tilde{s}_{i,j} \qquad \textbf{(20)}$$

and therefore, the provision for all the occurrence years is:

$$d\widetilde{PRO} = \sum_{i=1}^{n} \widetilde{PRO}_i = \sum_{i=1}^{n} \sum_{j=n-i+1}^{n} d\widetilde{s}_{i,j} \tag{21}$$

From the α-cuts of $d\widetilde{s}_{i,j}$ in (16b) we derive the value for the α-cuts of $dP\widetilde{R}O_i$ and $dP\widetilde{R}O$ with (3):

$$dPRO_{i\alpha} = \left[\underline{dPRO_i}(\alpha), \overline{dPRO_i}(\alpha)\right] = \left[\sum_{j=n-i+1}^{n} e^{a+b_i+c_j+\rho[n+1/2-(i+j)]-\left\{l_a+l_{b_i}+l_{c_j}-r_\rho[n+1/2-(i+j)]\right\}(1-\alpha)}, \right.$$
$$\left. \sum_{j=n-i+1}^{n} e^{a+b_i+c_j+\rho[n+1/2-(i+j)]+\left\{r_a+r_{b_i}+r_{c_j}-l_\rho[n+1/2-(i+j)]\right\}(1-\alpha)} \right] \tag{22}$$

$$dPRO_\alpha = \left[\underline{dPRO}(\alpha), \overline{dPRO}(\alpha)\right] = \left[\sum_{i=1}^{n} \sum_{j=n-i+1}^{n} e^{a+b_i+c_j+\rho[n+1/2-(i+j)]-\left\{la+lb_i+lc_j-r\rho[n+1/2-(i+j)]\right\}(1-\alpha)} \right.$$
$$\left. \sum_{i=1}^{n} \sum_{j=n-i+1}^{n} e^{a+b_i+c_j+\rho[n+1/2-(i+j)]+\left\{ra+rb_i+rc_j-l\rho[n+1/2-(i+j)]\right\}(1-\alpha)} \right] \tag{23}$$

Finally, to fit the claim provision to account in financial statements we must transform the fuzzy value of the discounted reserves into the crisp values. Analogously to non-discounted reserves, we obtain:

$$dPRO_i^{*} = V\left[dP\widetilde{R}O_i ; \beta\right] = \sum_{j=n-i+1}^{n} EV\left[d\widetilde{s}_{i,j} ; \beta\right] \tag{24}$$

Therefore, the provision for all the occurrence years is:

$$dPRO^{*} = V\left[dP\widetilde{R}O; \beta\right] = \sum_{i=1}^{n} \sum_{j=n-i+1}^{n} EV\left[d\widetilde{s}_{i,j} ; \beta\right] \tag{25}$$

4. Numerical application

Our numerical example is developed over the run-off triangle in Table 2 [8, p. D5.4] that was also used in [2]. From Table 1, we can immediately obtain the log-incremental payments in Table 2.

Table 1. Run off triangle. The amounts are the cost of the claims from the *i*th origin year paid at the *j*th development year ($s_{i,j}$, *i*=0,1,2,3, *j*=0,1,...,3-*i*)

		Development year			
	i\j	0	1	2	3
Origin year	0	11073	6427	1839	766
	1	14799	9357	2344	
	2	15636	10523		
	3	16913			

Table 2. Run off triangle in which the quantities are ln$s_{i,j}$, *i*=0,1,2,3, *j*=0,1,...,3-*i*

		Development year			
	i\j	0	1	2	3
Origin year	0	9.312	8.768	7.517	6.641
	1	9.602	9.144	7.760	
	2	9.657	9.261		
	3	9.736			

Now, we develop our numerical with certain detail.

a) First we need to fit the parameters $\tilde{a}=(a,l_a,r_a)$, $\tilde{b}_i=(b_i,l_{b_i},r_{b_i})$, *i*=1,2,3 and $\tilde{c}_j=(c_j,l_{c_j},r_{c_j})$, *j*=1,2,3. Thus, following [2] we implement two steps.

a.1) We obtain the estimates of the centres of $\tilde{a}$, $\tilde{b}_i$, *i*=1,2,3 and $\tilde{c}_j$, *j*=1,2,3, with the algorithm used in [23]. See the results in Table 3.

Table 3. Least Squares estimates for the centres of $\tilde{a}$, $\tilde{b}_i$, *i*=1,2,3 and $\tilde{c}_j$, *j*=1,2,3

Coefficient	*a*'	b_1'	b_2'	b_3'	c_1'	c_2'	c_3'
Value	9.288	0.303	0.04	0.447	-0.466	-1.801	-2.647
Standard deviation	0.040	0.043	0.050	0.066	0.043	0.050	0.066

a.2) We fit the spreads of $\tilde{a}$, $\tilde{b}_i$, *i*=1,2,3 and $\tilde{c}_j$, *j*=1,2,3, being the result:

$\tilde{a}$ =(9.288, 0.024, 0.000)

$\tilde{b}_1$=(0.303, 0.000, 0.000)

$\tilde{b}_2$ =(0.404, 0.011, 0.016)

$\tilde{b}_3$ =(0.447, 0.000, 0.000)

$\tilde{c}_1$ =(-0.466, 0.030, 0.019)

$\tilde{c}_2$ =(-1.801, 0.006, 0.030)

$\tilde{c}_3$=(-2.647, 0.000, 0.000)

b) Now, we must evaluate (9a) to calculate the future cost of claims $\tilde{s}_{i,j}$, i=1,2,3; j=3-i+1,...,3. Table 4 indicates the values of α-cuts s_{i,j_α} for α=0, 0.5, 1. The value s_{i,j_1} gives us a prediction of the most feasible point estimate of the future incremental claim whereas the 0-cut quantifies its estimated range. Subsequently, by applying (12a) and (12b) to the results of Table 4 we calculate the α-cuts of provisions (see Table 5).

Table 4. Values of s_{i,j_α} i=1,2,3,j=3-i+1,...,3 for the α-levels α=0, 0.5, 1.

	$s_{1,3_\alpha}$	$s_{2,2_\alpha}$	$s_{2,3_\alpha}$
α=1	[1036.86, 1036.86]	[2672.95, 2672.95]	[1147.35, 1147.35]
α=0.5	[1012.38, 1036.86]	[2564.96, 2799.47]	[1107.81, 1166.06]
α=0	[988.48, 1036.86]	[2461.33, 2931.96]	[1069.64, 1185.08]
	$s_{3,1_\alpha}$	$s_{3,2_\alpha}$	$s_{3,3_\alpha}$
α=1	[10611.44, 10611.44]	[2791.62, 2791.62]	[1198.29, 1198.29]
α=0.5	[10054.03, 10813.80]	[2708.94, 2876.83]	[1170.00, 1198.29]
α=0	[9525.91, 11020.01]	[2628.71, 2964.63]	[1142.37, 1198.29]

Table 5. Values of the reserves for the α-levels α=0, 0.5, 1.

	$PROV_{1\alpha}$	$PROV_{2\alpha}$	$PROV_{3\alpha}$	$PROV_\alpha$
α=1	[1036.86, 1036.86]	[3820.30, 3820.30]	[14601.35, 14601.35]	[19458.51, 19458.51]
α=0.5	[1012.38, 1036.86]	[3672.77, 3965.53]	[13932.97, 14888.91]	[18618.12, 19891.30]
α=0	[988.48, 1036.86]	[3530.97, 4117.05]	[13296.99, 15182.94]	[17816.43, 20336.84]

To account non-discounted provisions in financial statements it is necessary to defuzzify the FNs that estimate future claiming. By using (10c), (13a) and (13b) and taking a risk aversion coefficient β=1, we obtain:

$$V[\tilde{s}_{1,3};1]=1036.86;\ V[\tilde{s}_{2,2};1]=2759.29;\ V[\tilde{s}_{2,3};1]=1159.93;$$

$$V[\tilde{s}_{3,1};1]=10747.63;\ V[\tilde{s}_{3,2};1]=2849.29;\ V[\tilde{s}_{3,3};1]=1198.29$$

$$PRO_1^{*}=V[\widetilde{PRO}_1;1]=1036.86$$

$$PRO_2^{*}=V[\widetilde{PRO}_2;1]=3919.21$$

$$PRO_3^{*}=V[\widetilde{PRO}_3;1]=14795.21$$

$$PRO^{*}=V[\widetilde{PRO}_i;1]=19751.28$$

On the other hand, by using a risk aversion coefficient β=1 but the expected value of claims (10c), we find:

$$EV\left[\tilde{s}_{1,3};1\right]=1036.86;\ EV\left[\tilde{s}_{2,2};1\right]=\ 2800.46;\ EV\left[\tilde{s}_{2,3};1\right]=\ 1166.11;$$

$$EV\left[\tilde{s}_{3,1};1\right]=\ 10814.44;\ EV\left[\tilde{s}_{3,2};1\right]=\ 2877.26;\ EV\left[\tilde{s}_{3,3};1\right]=1198.29$$

$$PRO_1^{\ *}=EV\left[P\tilde{R}O_1;1\right]=1036.86$$

$$PRO_2^{\ *}=EV\left[P\tilde{R}O_2;1\right]=3966.58$$

$$PRO_3^{\ *}=EV\left[P\tilde{R}O_3;1\right]=14889.99$$

$$PRO^{*}=EV\left[P\tilde{R}O_i;1\right]=19893.42$$

c) To obtain discounted reserves, we will suppose that the actuary predicts for the coming years that the returns of the insurer's investments will be approximately 3%, and that deviations over that value may be about 0.5%. That rate can be quantified by the TFN $\tilde{\rho}=(0.03,0.005,0.005)$. Subsequently, by applying (14b) we calculate the α-cuts of the discounted value of incremental claims (see Table 6). Likewise, the discounted value of reserves in Table 7 is calculated with (17a) and (17b).

Table 6. Values of $ds_{i,j_{\alpha}}$ i=1,2,3,j=2,3 for the α-levels α=0, 0.5, 1.

	$ds_{1,3_{\alpha}}$	$ds_{2,2_{\alpha}}$	$ds_{2,3_{\alpha}}$
α=1	[1021.42, 1021.42]	[2633.16, 2633.16]	[1096.86, 1096.86]
α=0.5	[996.06, 1022.70]	[2523.62, 2761.24]	[1055.10, 1118.94]
α=0	[971.33, 1023.98]	[2418.63, 2895.54]	[1014.93, 1141.46]
	$ds_{3,1_{\alpha}}$	$ds_{3,2_{\alpha}}$	$ds_{3,3_{\alpha}}$
α=1	[10453.46, 10453.46]	[2668.78, 2668.78]	[1111.70, 1111.70]
α=0.5	[9891.98, 10666.13]	[2580.05, 2760.57]	[1078.69, 1118.67]
α=0	[9360.65, 10883.12]	[2494.26, 2855.52]	[1046.66, 1125.69]

Table 7. Estimated values of the discounted reserves for the α-levels α=0, 0.5, 1.

	$dPROV_{1\alpha}$	$dPROV_{2\alpha}$	$dPROV_{3\alpha}$	$dPROV_{\alpha}$
α=1	[1021.42, 1021.42]	[3730.02, 3730.02]	[14233.94, 14233.94]	[18985.39, 18985.39]
α=0.5	[996.06, 1022.70]	[3578.72, 3880.18]	[13550.72, 14545.37]	[18125.49, 19448.25]
α=0	[971.33, 1023.98]	[3433.56, 4037.01]	[12901.58, 14864.33]	[17306.47, 19925.31]

To obtain a crisp value for the discounted value of incremental claims and discounted provisions, we calculate $d\tilde{s}_{i,j}$, i=1,2,3; j≥3-i+1 and apply (10b) and (15b). Subsequently we apply (18a) and (18b), considering again that β= 1:

$$V\left[d\tilde{s}_{1,3};1\right]=1022.27;\ V\left[d\tilde{s}_{2,2};1\right]=2720.62;\ V\left[d\tilde{s}_{2,3};1\right]=1111.73;$$

$$V\left[d\tilde{s}_{3,1};1\right]=10596.68;\ V\left[d\tilde{s}_{3,2};1\right]=2731.03;\ V\left[d\tilde{s}_{3,3};1\right]=1116.36$$

$$PRO_1{}^* = V\left[d\tilde{PRO}_1;1\right] = 1022.27;$$

$$PRO_2{}^* = V\left[d\tilde{PRO}_2;1\right] = 3832.35$$

$$PRO_3{}^* = V\left[d\tilde{PRO}_3;1\right] = 14444.07$$

$$PRO^* = V\left[d\tilde{PRO}_i;1\right] = 19298.69$$

From (10c) and (15b) we obtain $EV\left[d\tilde{s}_{i,j};\beta\right]$, i=1,2,3; j≥3-i+1. If we establish β= 1, those values are:

$$EV\left[d\tilde{s}_{1,3};1\right] = 1022.70;\ EV\left[d\tilde{s}_{2,2};1\right] = 2762.27;\ EV\left[d\tilde{s}_{2,3};1\right] = 1119.02;$$

$$EV\left[d\tilde{s}_{3,1};1\right] = 10666.85;\ EV\left[d\tilde{s}_{3,2};1\right] = 2761.10;\ EV\left[d\tilde{s}_{3,3};1\right] = 1118.68$$

And so, we obtain the following crisp values for provisions:

$$PRO_1{}^* = EV\left[d\tilde{PRO}_1;1\right] = 1022.70$$

$$PRO_2{}^* = EV\left[d\tilde{PRO}_2;1\right] = 3881.29$$

$$PRO_3{}^* = EV\left[d\tilde{PRO}_3;1\right] = 14546.63$$

$$PRO^* = EV\left[d\tilde{PRO}_i;1\right] = 19450.62$$

5. Conclusions

Few data must be used to adjust claim reserves. In this context we think that Fuzzy Set Theory is a suitable alternative to the usual statistical methods. Specifically, to quantify claim provisions we have used the method developed by Andrés-Sánchez in [12].

This paper deals with the fact that fuzzy estimation of claim reserves needs to be transformed into a crisp equivalent in order, for example, to compute them in balance sheet and income account. In this paper, we propose using the concept of value of a FN [11] because this defuzzification method makes it possible to introduce the actuarial risk aversion easily and intuitively and, likewise, to graduate the weights of the values embedded in fuzzy quantification. In a first approach we have fitted non-discounted reserves, whose value overrates real present value of liabilities. So, subsequently we have introduced in the analysis the return of the assets that cover future claims. By using financial mathematics with fuzzy parameters we calculate fuzzy discounted reserves and propose several expressions that allow their crisp quantification.

Any case, we think that other alternative FDA instruments like FR method [32] or fuzzy transforms [14] may give a viable quantitative basis to calculate claim reserves within the framework of Fuzzy Sets Theory. Likewise, in our opinion non-classical statistical methods like grey models [24] or gray correlation methods may also offer interesting solutions approaches to quantify claim reserves.

Acknowledgements. We thank the helpful comments of anonymous reviewers on previous versions of this paper.

References

1. Andrés, J. de, Terceño, A.: Applications of Fuzzy Regression in Actuarial Analysis. Journal of Risk and Insurance, Vol. 70, No. 4, 665-699. (2003)
2. Andrés-Sánchez, J. de: Claim reserving with fuzzy regression and the two ways of ANOVA. Applied Soft Computing, Vol. 12, No. 8, 2435-2441. (2012)
3. Andrés-Sánchez, J. de, González-Vila, L. : Using fuzzy random variables in life annuity pricing. Fuzzy Sets and Systems, Vol. 188, No. 1, 27-44. (2012)
4. Berry-Stölzle, T. R., Koissi, M.-C., Shapiro, A. F.: Detecting fuzzy relationships in regression models: The case of insurer solvency surveillance in Germany. Insurance: Mathematics and Economics, Vol. 46, No. 3, 554-567. (2009)
5. Buckley, J.J.**:** The fuzzy mathematics of finance. Fuzzy Sets and Systems , Vol. 21, No. 1, 57-73. (1987)
6. Buckley, J.J., Qu, Y. **:** On using α-cuts to evaluate fuzzy equations. Fuzzy Sets and Systems, Vol. 38, No. 3, 309-312. (1990)
7. Campos, L.M., González, A.**:** A subjective approach for ranking fuzzy numbers. Fuzzy Sets and Systems, Vol. 29, No. 2, 145-153. (1989)
8. Christofides, S.: Regression models based on log-incremental payments. Claims reserving manual, Volume 2, Institute of Actuaries, London. (1989)
9. Cummins, J.D., Derrig, R.A.: Fuzzy financial pricing of property-liability insurance. North American Actuarial Journal, Vol. 1, No.4, 21-44. (1997)
10. Da, L., Dongxiao, N., Yuanyuan, L., Guanjuan, C.: Combined models for day-ahead electricity price forecasting based on improved gray correlation methodology. Kybernetes, Vol. 38, No. 3/4, 354 – 361. (2009)
11. Delgado, M.; Vila, M.A.; Woxman, W.: On a canonical representation of a fuzzy number. Fuzzy Sets and Systems, Vol. 93, No. 1, 125-135. (1998)
12. Derrig, R.A., Ostaszewski, K.: Managing the tax liability of a property liability insurance company. Journal of Risk and Insurance, Vol. 64, No. 4, 695-711. (1997)
13. Derrig, R.A., Ostaszewski, K.: Fuzzy Sets. Encyclopedia of Actuarial Science, Volume 2, John Wiley & Sons, Chichester, 745-750. (2004)
14. Di Martino, F., Loia, V., Sessa S:. Fuzzy transforms method in prediction data analysis. Fuzzy Sets and Systems, Vol. 180, No. 1, 146-156. (2011)
15. Dubois, D., Prade, H.: Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets. Fuzzy Sets and Systems, Vol. 192, No. 1, 3-24. (2012)
16. Gil-Aluja, J.: Investment on uncertainty, Kluwer Academic Publishers, Dordretch. (1998)
17. Gil-Lafuente, A.M.: Fuzzy logic in financial analysis, Springer-Verlag, Berlin. (2005)
18. Institute Of Actuaries: Claims reserving manual, Volume 1, Institute of Actuaries, London. (1989)
19. Ishibuchi, H., Nii, M.: Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks. Fuzzy Sets and Systems, Vol. 119, No. 2, 273-290. (2001)
20. Kaufmann, A.; Gupta, M.M.: Fuzzy mathematical models in engineering and management science, North-Holland, Amsterdam (Netherlands) (1988)
21. Koissi, M.C. Shapiro, A.F.: Fuzzy formulation of the Lee–Carter model for mortality forecasting. Insurance: Mathematics and Economics, Vol. 39, No. 3, 289-307. (2006)
22. Koissi, M.C. Shapiro, A.F.: Fuzzy regression and the term structure of interest rates- A least squares approach. Actuarial Research Clearing House Newsletters, Vol. 2009, No. 1. (2009)
23. Kremer, E.: IBNR claims and the two way model of ANOVA. Scandinavian Actuarial Journal, Vol. 1, No. 1, 47-55. (1982)

23. Lai, L.-H.: Underwriting Underwriting Systematic Risk and Profit Margin in Fuzzy CAPM and ICAPM Models: The Case of Aviation Coverage. Contemporary Management Research, Vol. 5, No. 4, 387-396. (2009),
24. Li, Q-X., Liu, S.-F.: The grey input-occupancy-output analysis. Kybernetes, Vol. 38, No. 3/4, 306 – 313. (2009)
25. Li Calzi, M.: Towards a general setting for the fuzzy mathematics of finance. Fuzzy Sets and Systems, Vol. 35, No. 3, 265-280. (1990)
26. Ostaszewski, K.: An investigation into possible applications of fuzzy sets methods in actuarial science. Society of Actuaries, Schaumburg (USA). (1993)
27. Negoita, C.V.: Fuzzy systems and cybernetics. Kybernetes, Vol. 38, No. 1/2, 58-60. (2009)
28. Renshaw, A.E.: Chain-ladder and interactive modelling (claims reserving and GLIM). Journal of the Institute of Actuaries, Vol. 116, 559-587. (1989)
29. Shapiro, A.F.: Fuzzy logic in insurance. Insurance: Mathematics and Economics, Vol. 35, No. 2, 399-424. (2004)
30. Tanaka, H.: Fuzzy data analysis by possibilistic linear models. Fuzzy Sets and Systems, Vol. 24, No. 3, 363-375. (1987)
31. Tanaka, H., Ishibuchi H.: A possibilistic regression analysis based on linear programming, In: Kacprzyk, J.; Fedrizzi, M. (eds.): Fuzzy Regression Analysis, Physica-Verlag, Heildelberg, 47-60. (1992)
32. Xu, S.-Q., Luo, Q.-Y. Xu, G.-H. Lei, Z.: Asymmetrical interval regression using extended ε-SVM with robust algorithm. Fuzzy Sets and Systems, Vol. 160, No. 7, 988-1002. (2009)

Jorge de Andrés-Sánchez is an Associate Professor and a member of the Social and Business Research Laboratory of the Rovira i Virgili University, Spain. He is MSc degree in Actuarial Sciences from University of Barcelona and PhD in Business Administration from Rovira i Virgili University. Hehas published more than 100 papers in journals, books and conference proceedings including journals such as /Fuzzy Sets and Systems/, /International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems/, /Applied Soft Computing, Insurance: Mathematics and Economics/. He is currently interested in Fuzzy Mathematics applications in Finance and Insurance.

Received: December 25, 2012; Accepted: February 22, 2014

Computer Science and Information Systems 11(2):839–857 DOI: 10.2298/CSIS130109046S

Induced intuitionistic fuzzy ordered weighted averaging - weighted average operator and its application to business decision-making

Zeng Shouzhen[1], Wang Qifeng[2*], José M. Merigó[3], and Pan Tiejun[1]

[1]College of Computer and Information, Zhejiang Wanli University,
Ningbo, 315100, China
{zszzxl, Pantiejun}@163.com
[2]College of Modern Logistics, Zhejiang Wanli Univeristy, Ningbo, 315100, China
Lhywqf@163.com
[3]Manchester Business School, University of Manchester, Booth Street West, M15 6PB
Manchester, UK
jose.merigolindahl@mbs.ac.uk

Abstract. We present the induced intuitionistic fuzzy ordered weighted averaging-weighted average (I-IFOWAWA) operator. It is a new aggregation operator that uses the intuitionistic fuzzy weighted average (IFWA) and the induced intuitionistic fuzzy ordered weighted averaging (I-IFOWA) operator in the same formulation. We study some of its main properties and we have seen that it has a lot of particular cases such as the IFWA and the intuitionistic fuzzy ordered weighted averaging (IFOWA) operator. We also study its applicability in a decision-making problem concerning strategic selection of investments. We see that depending on the particular type of I-IFOWAWA operator used, the results may lead to different decisions.

Keywords: Intuitionistic fuzzy sets, weighted average, IOWA operator, decision-making.

1. Introduction

~~Different types of~~ aggregation operators are found in the literature for aggregating numerical data information [1-4].The weighted average (WA) is one of the most common aggregation operators found in the literature. It can be used in a wide range of different problems including statistics, economics and engineering. Another interesting aggregation operator is the ordered weighted averaging (OWA) operator introduced by Yager [5], whose prominent characteristic is the reordering step. The OWA operator provides a parameterized family of aggregation operators that includes as special cases the maximum, the

[*] **Corresponding author:** Wang Qifeng. E-mail: Lhywqf@163.com, zszzxl@163.com

minimum and the average criteria. Since its appearance, the OWA operator has been used in a wide range of applications [6-16].

A very practical extension of the OWA operator is the induced OWA (IOWA) operator [17]. The IOWA operator differs in that the reordering step is not developed with the values of the arguments but can be induced by another mechanism such that the ordered position of the arguments depends upon the values of their associated order-inducing variables. The IOWA operator has been studied by different authors in recent years [18-27].

Recently, some authors have tried to unify the OWA operator with the WA in the same formulation. It is worth mentioning the work developed by Torra [9] with the introduction of the weighted OWA (WOWA) operator and the work of Xu and Da [4] about the hybrid averaging (HA) operator. Both models arrived to an unification between the OWA and the WA because both concepts were included in the formulation as particular cases. However, as it has been studied by Merigó [20], these models seem to be a partial unification but not a real one because they can unify them but they cannot consider how relevant these concepts are in the specific problem considered. For example, in some problems we may prefer to give more importance to the OWA operator because we believe that it is more relevant and vice versa. To overcome this issue, Merigó [20] has developed the ordered weighted averaging weighted averaging (OWAWA) operator that unifies the OWA and the WA in the same formulation considering the degree of importance that each concept may have in the problem. Merigó [20] also presented a new operator that unifies the OWA operator with the WA when we assess the information with induced aggregation operators called the induced ordered weighted averaging–weighted average (IOWAWA) operator. The main advantage of this approach is that it unifies the IOWA and the WA taking into account the degree of importance that each concept has in the formulation. Thus, we are able to consider situations where we give more or less importance to the IOWA and the WA depending on our interests on the problem analyzed. Note that by using the IOWA we are considering complex reordering processes affected by different factors such as the degree of optimism, time pressure and psychological aspects. By using the WA we are considering the subjective probability of the decision-maker against the possibility that each state of nature will occur.

Usually, when using the IOWAWA and the above aggregation operators, it is assumed that the available information is clearly known and can be assessed with exact numbers. However, in the real-life world, due to the increasing complexity of the socio-economic environment and the lack of knowledge or data about the problem domain, exact numerical is sometimes unavailable. Thus, the input arguments may be vague or fuzzy in nature. Atanassov [28] defined the notion of an intuitionistic fuzzy set (IFS), whose basic elements are intuitionistic fuzzy numbers (IFNs) [10, 24, 29], each of which is composed of a membership degree and a non-membership degree. In many practical situations, particularly in the process of group decision making under uncertainty, the experts may come from different research areas and thus have different backgrounds and levels of knowledge, skills, experience, and personality. The experts may not have enough expertise or possess a sufficient level of knowledge to precisely express their preferences over the objects, and then, they usually have some uncertainty in providing their preferences, which makes the results of cognitive performance exhibit the characteristics of affirmation, negation, and hesitation. In such cases, the data or preferences given by the experts may be appropriately expressed in IFNs. For example, in multi-criteria decision-making problems, such as personnel evaluations, medical diagnosis, project investment analysis, etc., each IFN provided by the expert can be used to express both the degree that an alternative should satisfy a criterion and the degree that the alternative should not satisfy the criterion. The IFN is highly useful in depicting uncertainty and vagueness of an object, and thus can be used as a powerful tool to express data information under various different fuzzy environments which has attracted great attentions [29-39].

The aim of this paper is to extend the IOWAWA operator to accommodate the intuitionistic fuzzy situations. For doing so, we present a new intuitionistic fuzzy aggregation operator called the induced intuitionistic fuzzy ordered weighted averaging weighted average (I-IFOWAWA) operator, which unifies the IOWA operator with the WA when the available information is uncertain and can be assessed with IFNs. The main advantage of this approach is that it unifies the OWA and the WA taking into account the degree of importance of each case in the formulation and considering that the information is given with IFN. Thus, we are able to consider situations where we give more or less importance to the IFOWA and the IFWA depending on our interests and the problem analyzed. Furthermore, by using the I-IFOWAWA, we are able to use a complex reordering process in the OWA operator in order to represent complex attitudinal characters. We also study different properties of the I-IFOWAWA operator and different particular cases.

We also analyze the applicability of the new approach and we see that it is possible to develop an astonishingly wide range of applications. For example, we can apply it in a lot of problems about statistics, economics and decision theory. In this paper, we focus on a decision-making problem concerning strategic selection of investments. The main advantage of the I-IFOWAWA in these problems is that it is possible to consider the subjective probability (or the degree of importance) and the attitudinal character of the decision maker at the same time.

This paper is organized as follows. In Section 2 we briefly review some basic concepts. Section 3 presents the I-IFOWAWA operator and analyzes different families of I-IFOWAWA operators in Section 4. In Section 5 we develop an application of the new approach. Section 6 summarizes the main conclusions of the paper.

2. Preliminaries

In this Section, we briefly review some basic concepts about the intuitionistic fuzzy sets, the IOWA and the IOWAWA operator.

2.1. Intuitionistic fuzzy sets

Intuitionistic fuzzy set (IFS) introduced by Atanassov [28] is an extension of the classical fuzzy set, which is a suitable way to deal with vagueness. It can be defined as follows.

Definition 1. Let a set $X=\{x_1,x_2,...,x_n\}$ be fixed, an IFS A in X is given as following:

$$A=\{\langle x,\mu_A(x),v_A(x)\rangle | x\in X\}. \quad \textbf{(1)}$$

The numbers $\mu_A(x)$ and $v_A(x)$ represent, respectively, the membership degree and non-membership degree of the element x to the set A, $0\le \mu_A(x)+v_A(x)\le 1$, for all $x\in X$. The pair $(\mu_A(x),v_A(x))$ is called an intuitionistic fuzzy number (IFN) and each IFN can be simply denoted as $\alpha=(\mu_\alpha,v_\alpha)$, where $\mu_\alpha\in[0,1]$, $v_\alpha\in[0,1]$, $\mu_\alpha+v_\alpha\le 1$.

Additionally $S(\alpha)=\mu_\alpha-v_\alpha$ and $H(\alpha)=\mu_\alpha+v_\alpha$ are called the score and accuracy degree of α respectively.

For any three IFNs $\alpha=(\mu_\alpha,v_\alpha),\alpha_1=(\mu_{\alpha_1},v_{\alpha_1})$ and $\alpha_2=(\mu_{\alpha_2},v_{\alpha_2})$, the following operational laws are valid [10, 29].

(1) $\alpha_1+\alpha_2=(\mu_{\alpha_1}+\mu_{\alpha_2}-\mu_{\alpha_1}\cdot\mu_{\alpha_2},v_{\alpha_1}\cdot v_{\alpha_2})$

(2) $\lambda\alpha=(1-(1-\mu_\alpha)^\lambda,{v_\alpha}^\lambda)$

To compare two IFNs α_1 and α_2, Xu and Yager [29] introduced a simple method as below:

- If $S(\alpha_1)<S(\alpha_2)$, then $\alpha_1<\alpha_2$;
- If $S(\alpha_1)=S(\alpha_2)$, then

(1) If $H(\alpha_1)<H(\alpha_2)$, then $\alpha_1<\alpha_2$;

(2) If $H(\alpha_1)=H(\alpha_2)$, then $\alpha_1=\alpha_2$.

2.2. The Induced OWA Operator

The IOWA operator is an extension of the OWA operator. The main difference is that the reordering step is not carried out with the values of the argument a_i. In this case, the reordering step is developed with order-inducing variables that reflect a more complex reordering process. The IOWA operator also includes as particular cases maximum, minimum and average criteria. The IOWA operator can be defined as follows:

Definition 2. An IOWA operator of dimension n is a mapping *IOWA:* $R^n \times R^n \rightarrow R$ that has an associated weighting W with $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$ such that:

$$IOWA\left(\langle u_1, a_1\rangle, \ldots, \langle u_n, a_n\rangle\right) = \sum_{j=1}^{n} w_j b_j . \qquad (2)$$

where b_j is a_i value of the IOWA pair $\langle u_i, a_i\rangle$ having the j th largest u_i, u_i is the order inducing variable and a_i is the argument variable.

2.3. The intuitionistic fuzzy OWA operator

The intuitionistic fuzzy OWA (IFOWA) operator was introduced by Xu.[26] It is an extension of the OWA operator for uncertain situations where the available information can be assessed with IFNs. Let Ω be the set of all IFNs, it can be defined as follows:

Definition 3. Let $\alpha_i = \left(\mu_i, v_i\right)(i = 1, 2, \ldots, n)$ be a collection of IFNs, an IFOWA operator of dimension n is a mapping *IFOWA*: $\Omega^n \rightarrow \Omega$ that has an associated weighting W with $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$ such that:

$$IFOWA\left(\alpha_1, \alpha_2, \ldots, \alpha_n\right) = \sum_{j=1}^{n} w_j \beta_j = \left(1 - \prod_{j=1}^{n}\left(1 - \mu_j\right)^{w_j}, \prod_{j=1}^{n} v_j^{w_j}\right) . \qquad (3)$$

where $\beta_j = \left(\mu_j, v_j\right)$ is the jth largest of the α_i and α_i is the argument variable represented in the form of IFN.

2.4. The induced ordered weighted averaging-weighted average (IOWAWA) operator

The induced ordered weighted averaging–weighted average (IOWAWA) operator is a new model that unifies the IOWA operator and the WAs in the same formulation and considering a complex reordering process based on order-inducing variables. Therefore, both concepts can be seen as a particular case of a more general one. This approach seems to be complete, at least as an initial real unification between IOWA operators and WA. It can also be seen as a unification between decision-making problems under uncertainty (with IOWA operators) and under risk (with probabilities). It can be defined as follows.

Definition 4. An IOWAWA operator of dimension n is a mapping *IOWAWA*: $R^n \times R^n \to R$ that has an associated weighting W with $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$ such that:

$$IOWAWA\left(\langle u_1, a_1 \rangle, \ldots, \langle u_n, a_n \rangle\right) = \sum_{j=1}^{n} \hat{\upsilon}_j b_j \,. \qquad (4)$$

where b_j is a_i value of the IOWAWA pair $\langle u_i, a_i \rangle$ having the j th largest u_i, u_i is the order inducing variable and a_i is the argument variable, each argument a_i has an associated weight (WA) υ_j with $\sum_{j=1}^{n} \upsilon_j = 1$ and $\upsilon_j \in [0,1]$, $\hat{\upsilon}_j = \lambda w_j + (1-\lambda)\upsilon_j$ with $\lambda \in [0,1]$ and υ_j is the weight (WA) υ_i ordered according to b_j, that is, according to the j th largest u_i.

Note that it is also possible to formulate the IOWAWA operator separating the part that strictly affects the IOWA operator and the part that affects the WAs. This representation is useful to see both models in the same formulation but it does not seem to be as a unique equation that unifies both models.

Definition 5. An IOWAWA operator of dimension n is a mapping *IOWAWA*: $R^n \times R^n \to R$ that has an associated weighting W with $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$, and a weighting vector V that affects the WA, with $\sum_{i=1}^{n} \upsilon_i = 1$ and $\upsilon_i \in [0,1]$, such that:

$$IOWAWA\left(\langle u_1, a_1\rangle, \ldots, \langle u_n, a_n\rangle\right) = \lambda \sum_{j=1}^{n} w_j b_j + (1-\lambda)\sum_{i=1}^{n} \upsilon_i a_i \,. \qquad \textbf{(5)}$$

where b_j is a_i value of the IOWAWA pair $\langle u_i, a_i\rangle$ having the j th largest u_i, u_i is the order inducing variable and $\lambda \in [0,1]$.

3. Induced intuitionistic fuzzy ordered weighted averaging-weighted average (I-IFOWAWA) operator

The induced intuitionistic fuzzy ordered weighted averaging-weighted average (I-IFOWAWA) operator is an extension of the IOWAWA operator that uses uncertain information in the aggregation represented in the form of IFNs. Note that the I-IFOWAWA can also be seen as an aggregation operator that uses the IOWA operator, the WA and IFNs in the same formulation. The reason for using this operator is that sometimes, the uncertain factors that affect our decisions are not clearly known and in order to assess the problem we need to use IFNs. This operator can be defined as follows.

Definition 6. An I-IFOWAWA operator of dimension n is a mapping *I-IFOWAWA*: $R^n \times \Omega^n \rightarrow \Omega$ that has an associated weighting W with $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$ such that:

$$I\text{-}IFOWAWA\left(\langle u_1, \alpha_1\rangle, \ldots, \langle u_n, \alpha_n\rangle\right) = \sum_{j=1}^{n} \hat{\upsilon}_j \beta_j \,. \qquad \textbf{(6)}$$

where β_j is α_i value of the I-IFOWAWA pair $\langle u_i, \alpha_i\rangle$ having the j th largest u_i, u_i is the order inducing variable and α_i is the argument variable, each argument α_i has an associated weight (WA) υ_j with $\sum_{j=1}^{n} \upsilon_j = 1$ and $\upsilon_j \in [0,1]$, $\hat{\upsilon}_j = \lambda w_j + (1-\lambda)\upsilon_j$ with $\lambda \in [0,1]$ and υ_j is the weight (WA) υ_i ordered according to β_j, that is, according to the j th largest u_i.

Note that it is also possible to formulate the I-IFOWAWA operator separating the part that strictly affects the IOWA operator and the WA.

Definition 7. An I-IFOWAWA operator of dimension n is a mapping *I-IFOWAWA:* $R^n \times \Omega^n \to \Omega$ that has an associated weighting W with $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$, and a weighting vector V that affects the WA, with $\sum_{i=1}^{n} \upsilon_i = 1$ and $\upsilon_i \in [0,1]$, such that:

$$I-IFOWAWA(\langle u_1, \alpha_1 \rangle, \ldots, \langle u_n, \alpha_n \rangle) = \lambda \sum_{j=1}^{n} w_j \beta_j + (1-\lambda) \sum_{i=1}^{n} \upsilon_i \alpha_i . \tag{7}$$

where β_j is α_i value of the I-IFOWAWA pair $\langle u_i, \alpha_i \rangle$ having the j th largest u_i, u_i is the order inducing variable and $\lambda \in [0,1]$.

Note that if $\gamma = 1$, we get the I-IFOWA operator and if $\gamma = 0$, the IFWA operator.

In the following, we are going to give a simple example on how to aggregate with the I-IFOWAWA operator. We consider the aggregation with both definitions.

Example 1. Assume the following arguments in an aggregation process: $((0.5, 0.3), (0.4, 0.5), (0.8, 0.1), (0.6, 0.3))$ with the following order-inducing variables $U = (4, 7, 1, 9)$. Assume the following weighting vector *W*= (0.2, 0.2, 0.3, 0.3) and the following probabilistic weighting vector *V* = (0.3, 0.2, 0.4, 0.1). Note that the WA has a degree of importance of 70% while the weighting vector *W* of the IOWA a degree of 30%. If we want to aggregate this information by using the I-IFOWAWA operator, we will get the following result. The aggregation can be solved either with Eq. (6) or Eq. (7). With Eq. (6) we calculate the new weighting vector as:

$$\hat{\upsilon}_1 = 0.3 \times 0.2 + 0.7 \times 0.1 = 0.13$$
$$\hat{\upsilon}_2 = 0.3 \times 0.2 + 0.7 \times 0.2 = 0.2$$
$$\hat{\upsilon}_3 = 0.3 \times 0.3 + 0.7 \times 0.3 = 0.3$$
$$\hat{\upsilon}_4 = 0.3 \times 0.3 + 0.7 \times 0.4 = 0.37$$

And then, we calculate the aggregation process as follows:

$$I-IFOWAWA = 0.13 \times (0.6, 0.3) + 0.2 \times (0.4, 0.5) +$$
$$+0.3 \times (0.5, 0.3) + 0.37 \times (0.8, 0.1) = (0.64, 0.22)$$

With Eq. (7), we aggregate as follows:

$$I-IFOWAWA = 0.3\times(0.2\times(0.6,0.3)+0.2\times(0.4,0.5)+ \\ 0.3\times(0.5,0.3)+0.3\times(0.8,0.1))+0.7\times(0.3\times(0.5,0.3)+ \\ 0.2\times(0.4,0.5)+0.4\times(0.8,0.1)+0.1\times(0.6,0.3)) = (0.64,0.22)$$

Obviously, we get the same results with both methods.

From a generalized perspective of the reordering step, we can distinguish between the descending I-IFOWA (DI-IFOWA) operator and the ascending I-IFOWA (AI-IFOWA) operator by using $w_j = w^*_{n-j+1}$, where w_j is the jth weight of the DI-IFOWA and w^*_{n-j+1} the jth weight of the AI-IFOWA operator.

Note that if the weighting vector is not normalized, i.e., $\hat{V} = \sum_{j=1}^{n} \hat{\upsilon}_j \neq 1$, then, the I-IFOWAWA operator can be expressed as:

$$I-IFOWAWA(\langle u_1,\alpha_1\rangle,...,\langle u_n,\alpha_n\rangle) = \frac{1}{\hat{V}}\sum_{j=1}^{n}\hat{\upsilon}_j\beta_j . \quad \textbf{(8)}$$

Similar to the IOWAWA operator, the I-IFOWAWA operator is monotonic, bounded and idempotent. These properties can be proved with the following theorems.

Theorem 1 (Monotonic). Assume f is the I-IFOWAWA operator, if $\alpha_i \geq \alpha_i'$ for all i, then:

$$f(\langle u_1,\alpha_1\rangle,...,\langle u_n,\alpha_n\rangle) \geq f(\langle u_1,\alpha_1'\rangle,...,\langle u_n,\alpha_n'\rangle) . \quad \textbf{(9)}$$

Theorem 2 (Idempotency). Assume f is the I-IFOWA operator, if $\alpha_i = \alpha$ for all i, then

$$f(\langle u_1,\alpha_1\rangle,...,\langle u_n,\alpha_n\rangle) = \alpha . \quad \textbf{(10)}$$

Theorem 3 (Bounded). Assume f is the I-IFOWA operator, then

$$\min\{\alpha_i\} \leq f(\langle u_1,\alpha_1\rangle,...,\langle u_n,\alpha_n\rangle) \leq \max\{\alpha_i\} . \quad \textbf{(11)}$$

Note that the proofs of these theorems are straightforward and thus omitted.

A further interesting issue is the problem of ties in the reordering process of the order inducing variables. In order to solve this problem, we recommend following the policy explained by Yager and Filev [17] about replacing the tied arguments by their average. Note that in this case, it would mean that we are replacing the tied arguments by their intuitionistic fuzzy average.

Another interesting issue to consider is the measures for characterizing the weighting vector $\hat{V}$ (V and W) of the I-IFOWA operator such as the attitudinal character, the entropy of dispersion, the divergence of W and the balance operator. As these features do not depend upon the intuitionistic fuzzy numbers arguments, the formulation is the same as the IOWAWA operator. The entropy of dispersion is defined as follows:

$$H\left(\hat{V}\right)=-\left(\gamma\sum_{j=1}^{n}w_j\ln(w_j)+(1-\gamma)\sum_{i=1}^{n}\upsilon_i\ln(\upsilon_i)\right). \tag{12}$$

Note that υ_i is the ith weight of the WA aggregation. As we can see, if $\gamma=1$, we obtain the Yager entropy of dispersion and if $\gamma=0$, we get the classical Shannon entropy. Note that strictly speaking, the Shannon entropy is extended by using the I-IFOWAWA operator as follows:

$$H\left(\hat{V}\right)=-\left(\gamma\sum_{j=1}^{n}w_j\log_2(w_j)+(1-\gamma)\sum_{i=1}^{n}\upsilon_i\log_2(\upsilon_i)\right). \tag{13}$$

Thus, it is easy to see that we can extend all the analysis developed by Shannon and others in information theory by using this new approach. Note also that it is possible to consider other entropy measures in the I-IFOWAWA operator.

If we extend the analysis of the orness–andness measure to the I-IFOWAWA operator, we get the following expressions. For the degree of orness:

$$\alpha\left(\hat{V}\right)=\gamma\sum_{j=1}^{n}w_j^*\left(\frac{n-j}{n-1}\right)+(1-\gamma)\sum_{j=1}^{n}\upsilon_j^*\left(\frac{n-j}{n-1}\right). \tag{14}$$

Note that w_j^* and υ_j^* are the w_j and υ_j weights of the I-IFOWAWA aggregation ordered according to the values of the arguments α_i. As we can see, if $\gamma=1$, we get the usual orness measure of Yager[7] and if $\gamma=0$, we obtain the orness measure of the weighted average. It is straightforward to calculate the andness measure by using the dual. That is, $Andnesse\left(\hat{V}\right)=1-\alpha\left(\hat{V}\right)$.

If we extend the divergence of *W* to the I-IFOWAWA operator, we get the following divergence of $\hat{V}$:

$$Div\left(\hat{V}\right)=\sum_{j=1}^{n}\hat{\upsilon}_j\left(\frac{n-j}{n-1}-\alpha\left(\hat{V}\right)\right)^2. \tag{15}$$

or also as:

$$Div\left(\hat{V}\right)=\gamma\left(\sum_{j=1}^{n}w_j\left(\frac{n-j}{n-1}-\alpha(W)\right)^2\right)+ \tag{16}$$

$$(1-\gamma)\left(\sum_{j=1}^{n}\upsilon_j\left(\frac{n-j}{n-1}-\alpha(W)\right)^2\right).$$

Note that if $\gamma=1$, we get the I-IFOWA divergence and if $\gamma=0$, the IFWA divergence.

The balance operator can be defined as:

$$BAL(W)=\sum_{j=1}^{n}w_j\left(\frac{n+1-2j}{n-1}\right). \tag{17}$$

And the divergence of $\hat{V}$:

$$\begin{aligned}Bal\left(\hat{V}\right)&=\gamma g^{-1}\left(\sum_{j=1}^{n}g\left(\frac{n+1-2j}{n-1}\right)w_j\right)\\&+(1-\gamma)g^{-1}\left(\sum_{j=1}^{n}g\left(\frac{n+1-2j}{n-1}\right)\upsilon_j\right)\ .\end{aligned} \tag{18}$$

Note that $\hat{\upsilon}_j=\lambda w_j+(1-\lambda)\upsilon_j$ is the *j*th weight of the I-IFOWAWA aggregation. And if *g(b)* = *b*, then, we get the usual balance operator applied to the I-IFOWAWA operator as follows:

$$Bal\left(\hat{V}\right)=\sum_{j=1}^{n}\hat{\upsilon}_j\left(\frac{n+1-2j}{n-1}\right). \tag{19}$$

Note that if $\gamma=1$, we get the classic balance operator and if $\gamma=0$, the IFWA divergence.

Analyzing the applicability of the I-IFOWAWA operator, we can see that it is applicable to similar situations already discussed in other types of induced aggregation operators where it is possible to use intuitionistic fuzzy numbers information. For example, we could use it in different decision-making problems, statistics, fuzzy set theory, soft computing, operational research, business administration, economics etc.

4. Families of I-IFOWAWA Operators

In this Section we analyze different families of I-IFOWAWA operators. The main advantage is that we can consider a wide range of particular cases that can be used in the I-IFOWAWA operator leading to different results. Thus, we are able to provide a more complete representation of the aggregation process. Basically, we distinguish between the families found in the weighting vector $\hat{V}$ (V and W) and those found in the parameter γ.

Remark 1. If we analyze the parameter γ, we are going to consider the two main cases of the I-IFOWAWA operator. Basically:

- If $\gamma = 0$, we get the IFWA [10].
- If $\gamma = 1$, we get the I-IFOWA [33].
- If $\gamma = 1$ and the ordered position of u_i is the same as the ordered position of β_j such that β_j is the jth largest α_i, the IFOWA [10].

Remark 2. Another group of interesting families are the maximum-IFWA, the minimum-IFWA, the step-I-IFOWAWA operator and the usual average.

- The maximum-IFWA is found when $w_p = 1$, $w_j = 0$, for all $j \neq p$, and $u_p = \text{Max}\{\alpha_i\}$.
- The minimum-IFWA is found when $w_p = 1$, $w_j = 0$, for all $j \neq p$, and $u_p = \text{Min}\{\alpha_i\}$.
- More generally, the step- I-IFOWAWA is formed when $w_k = 1$ and $w_j = 0$ for all $j \neq k$.
- The intuitionistic fuzzy average is obtained when $w_j = 1/n$, and $\upsilon_j = 1/n$, for all α_i.

Remark 3. The arithmetic-IFWA is obtained when $w_j = 1/n$ for all *j*, and the weighted average is equal to the IFOWA when the ordered position of *i* is the same as the ordered position of *j*. The arithmetic-IFWA (A-IFWA) can be formulated as follows:

$$A - IFWA(\langle u_1, \alpha_1 \rangle, \ldots, \langle u_n, \alpha_n \rangle) = \frac{1}{n} \lambda \sum_{i=1}^{n} \alpha_i + (1-\lambda) \sum_{i=1}^{n} \upsilon_i \alpha_i . \quad \textbf{(20)}$$

Note that if $\upsilon_i = 1/n$, for all *i*, then, we get the unification between the intuitionistic fuzzy arithmetic mean (or simple intuitionistic fuzzy average) and the I-IFOWA operator, that is, the arithmetic-I-IFOWA (A-I-IFOWA). The A-I-IFOWA operator can be formulated as follows:

$$A - I - IFOWA(\langle u_1, \alpha_1 \rangle, \ldots, \langle u_n, \alpha_n \rangle) = \lambda \sum_{j=1}^{n} w_j \beta_j + (1-\lambda) \frac{1}{n} \sum_{i=1}^{n} \alpha_i . \quad \textbf{(21)}$$

Remark 4. Following the OWA literature [4, 16, 20], we can develop many other families of I-IFOWAWA operators such as:

- The olympic-I-IFOWAWA operator ($w_1 = w_n = 0$ and for all others $w_j = 1/(n-2)$).
- The general olympic-I-IFOWAWA operator ($w_j = 0$ for $j = 1, 2, ..., k, n,$ $n-1,$ ..., $n-k+1$; and for other $w_{j*} = 1/(n-2k)$ where $k < n/2$).
- The centered- I-IFOWAWA (if it is symmetric, strongly decaying from the center to the maximum and the minimum, and inclusive).

*Remark 5***.** Using a similar methodology, we could develop numerous other families of I-IFOWAWA operators. For more information, refer to Ref. 16, 20-23, 27.

5. Application in business decision-making

The I-IFOWAWA operator can be applied in a wide range of problems including statistics, economics and engineering.

In the following, we are going to develop a brief example where we will see the applicability of the new approach. We will focus on a decision-making problem about selection of strategies. Note that other business decision making applications could be developed such as financial decision making and human resource selection.

Assume a company that operates in North America and Europe is analyzing the general policy for the next year and they consider five possible strategies to follow.

- A_1 = Expand to the Asian market.
- A_2 = Expand to the South American market.
- A_3 = Expand to the African market.
- A_4 = Expand to the three continents.
- A_5 = Do not develop any expansion.

In order to evaluate these strategies, the experts consider that the key factor is the economic situation of the next year. Thus, depending on the situation, the expected benefits will be different. The experts have considered five possible situations for the next year:

- S_1 = Very bad.
- S_2 = Bad.
- S_3 = Regular.
- S_4 = Good.
- S_5 = Very good.

The group of experts of the company gives its opinion about the uncertain expected results that may occur in the future. The expected results depending on the situation S_i and the alternative A_i are shown in Tables 1. Note that the results are IFNs.

In this problem, the decision maker assumes the following degrees of importance (IFWA) of the characteristics: V = (0.1, 0.2, 0.2, 0.2, 0.3). He assumes that the IFOWA weight is: W = (0.1, 0.1, 0.2, 0.2, 0.4); with the following order-inducing variables: U = (8, 4, 9, 3, 2). Note that IFWA has an importance of 70% and the IFOWA an importance of 30% (γ = 0.3) because he believes that the IFWA is more relevant in the problem.

Table 1. Available strategies

	S_1	S_2	S_3	S_4	S_5
A_1	(0.5,0.4)	(0.6,0.2)	(0.4,0.4)	(0.7,0.1)	(0.5,0.5)
A_2	(0.6,0.2)	(0.9,0.1)	[(0.7,0.2)	(0.3,0.5)	(0.4,0.5)
A_3	(0.4,0.4)	(0.6,0.3)	(0.8,0.1)	(0.5,0.4)	(0.6,0.3)
A_4	(0.7,0.2)	(0.4,0.6)	(0.5,0.3)	(0.3,0.6)	(0.7,0.1)
A_5	(0.8,0.2)	(0.5,0.3)	(0.6,0.3)	(0.6,0.4)	(0.4,0.5)

Due to the fact that the attitudinal character of the entrepreneurs is very complex because it involves a lot of complexities, the experts use order inducing variables to express it. Table 2 shows the results.

Table 2. Order inducing variables

	S_1	S_2	S_3	S_4	S_5
U	12	13	16	15	19

With this information, we can make an aggregation to make a decision. In this example, we will consider the Max-IFWA, the Min- IFWA, the IFWA, the IFOWA, the I-IFOWA, the A-IFWA, the A-I-IFOWA, the I-IFOWAWA. The results are shown in Table 3.

As we can see, depending on the aggregation operator used, the ordering of the strategies may be different. Therefore, the decision about which strategy select may be also different.

If we establish an ordering of the investments, a typical situation if we want to consider more than one alternative, we will get the following orders shown in Table 4. Note that the first alternative in each ordering is the optimal choice.

Table 3. Aggregated results

	Max-IFWA	Min-IFWA	IFWA	IFOWA	I-IFOWA	I-IFOWAWA
A_1	(0.7,0.1)	(0.4,0.4)	(0.55,0.28)	(0.50,0.34)	(0.56,0.29)	(0.55,0.28)
A_2	(0.9,0.1)	(0.3,0.5)	(0.64,0.28)	(0.54,0.32)	(0.61,0.30)	(0.63,0.28)
A_3	(0.8,0.1)	(0.4,0.4)	(0.62,0.26)	(0.54,0.32)	(0.59,0.29)	(0.61,0.27)
A_4	(0.7,0.1)	(0.3,0.6)	(0.55,0.27)	(0.46,0.39)	(0.57,0.24)	(0.55,0.26)
A_5	(0.8,0.2)	(0.4,0.5)	(0.56,0.36)	(0.54,0.37)	(0.54,0.37)	(0.55,0.36)

Table 4. Ordering of the Strategies

	Ordering
Max-IFWA	$A_2 \succ A_3 \succ A_5 \succ A_1 = A_4$
Min-IFWA	$A_1 = A_3 \succ A_5 \succ A_2 \succ A_4$
IFWA	$A_2 \succ A_3 \succ A_4 \succ A_1 \succ A_5$
IFOWA	$A_2 = A_3 \succ A_5 \succ A_1 \succ A_4$
I-IFOWA	$A_4 \succ A_2 \succ A_3 \succ A_1 \succ A_5$
I-IFOWAWA	$A_2 \succ A_3 \succ A_1 \succ A_4 \succ A_5$

As we can see, depending on the particular type of I-IFOWAWA operator used, the results may lead to different decisions. Note that the main advantage of using the I-IFOWAWA operator is that we can consider a wide range of scenarios in order to form a general picture. As each case may give different results, the decision maker will select the particular case that it is closest to his interests.

6. Conclusions

We have introduced the I-IFOWAWA operator. It is an aggregation operator that uses the IOWA operator, the WA and uncertain information represented in the form of IFNs. It is very useful for uncertain situations where the decision maker can not assess the information with exact numbers or singletons but it is possible to assess it with

IFNs. By using the I-IFOWAWA, we are able to deal with complex situations that require a complex reordering process of the arguments before the aggregation step. Moreover, by using the I-IFOWAWA, we get a generalization that includes a wide range of intuitionistic fuzzy operators such as the IFWA, the IFOWA, the I-IFOWA, the A-IFWA and the A-I-IFOWA.

We have also developed an application of the new approach in a strategic decision-making problem. We have seen that the I-IFOWAWA is very useful because it represents very well the uncertain information by using IFNs. We have also seen that depending on the particular case of the I-IFOWAWA operator used the results may lead to different decisions.

In future research we expect to develop further extensions by adding new characteristics in the problem such as the use of probabilistic aggregations. We will also consider other decision making applications such as human resource management, investment selection and product management.

Acknowledgements. Acknowledgements. This paper is supported by the MOE Project of Humanities and Social Sciences No.14YJC910006), Zhejiang Province Natural Science Foundation (No. LQ14G010002), Statistical Scientific Key Research Project of China (No.2013LZ48), Key Research Center of Philosophy and Social Science of Zhejiang Province CModern Port Service Industry and Creative Culture Research Center, Zhijiang Young Talents Project of Social Science of Zhejiang (G210), Zhejiang Provincial Key Research Base for Humanities and Social Science Research (Statistics), Projects in Science and Technique of Ningbo Municipal (No. 2012B82003) and Ningbo Natural Science Foundation (No. 2013A610286).

References

1. Beliakov, G., Pradera, A., Calvo, T.: Aggregation Functions: A Guide for Practitioners. Springer-Verlag, Berlin. (2007)
2. Calvo, T., Mayor, G., Mesiar, R.: Aggregation operators: New trends and applications. Physica-Verlag, New York. (2002)
3. Torra, V., Narukawa, Y.: Modelling decisions: Information fusion and aggregation operators. Springer-Verlag, Berlin. (2007)
4. Xu, Z. S., Da, Q. L.: An overview of operators for aggregating information. International Journal of Intelligent Systems, Vol. 18, No. 9, 953–968. (2003)
5. Yager, R. R.: On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Transactions on Systems, Man and Cybernetics B, Vol. 18, No. 1, 183–190. (1988)
6. Liu, P. D., Zhang X., Jin, F.: A multi-attribute group decision-making method based on interval-valued trapezoidal fuzzy numbers hybrid harmonic averaging operators. Journal of Intelligent and Fuzzy Systems, Vol. 23, No. 5, 159–168. (2012)
7. Chen, H. Y., Zhou, L. G.: An approach to group decision making with interval fuzzy preference relations based on induced generalized continuous ordered weighted averaging operator. Expert Systems with Applications, Vol. 38, No. 10, 13432–13440. (2011)
8. Su, W. H., Yang, Y., Zhang, C. H., Zeng, S. Z.: Intuitionistic fuzzy decision-making with similarity measures and OWA Operator. International Journal of Uncertainty, Fuzziness and Knowledge-Based System, Vol. 21, No. 2, 245–262. (2013)

9. Torra, V.: The weighted OWA operator. Journal of Intelligent Systems, Vol. 12, No. 2, 153–166. (1997)
10. Xu, Z. S.: Intuitionistic fuzzy aggregation operators. IEEE Transactions on Fuzzy Systems, Vol. 15, No. 6, 1179-1187. (2007)
11. Yager, R. R.: Centered OWA operators. Soft Computing, Vol. 11, No. 7, 631-639. (2007)
12. Yager, R. R., Kacprzyk, J., Beliakov, G.: Recent developments on the ordered weighted averaging operators: Theory and practice. Springer-Verlag, Berlin. (2011)
13. Zhang, Z., Xu, Z. S.: On continuity of ordered aggregation operators. International Journal of Intelligent Systems, Vol. 28, No. 4, 307-318. (2013)
14. Zeng, S. Z., Merigó, J. M., Su, W. H.: The uncertain probabilistic OWA distance operator and its application in group decision making. Applied Mathematical Modelling, Vol. 37, No. 9, 6266-6275. (2013)
15. Zhou, L. G., Chen, H. Y., Liu, J. B.: Generalized power aggregation operators and their applications in group decision making. Computers & Industrial Engineering, Vol. 62, No. 4, 989-999. (2012)
16. Merigó, J. M., Gil-Lafuente, A. M.: New decision-making techniques and their application in the selection of financial products. Information Sciences, Vol. 180, No. 11, 2085-2094. (2010)
17. Yager, R. R., Filev, D. P.: Induced ordered weighted averaging operators. IEEE Transactions on Systems, Man and Cybernetics B, Vol. 29, No. 1, 141–150. (1999)
18. Chiclana, F., Herrera-Viedma, E., Herrera, F., Alonso, S.: Some induced ordered weighted averaging operators and their use for solving group decision-making problems based on fuzzy preference relations. European Journal of Operational Research, Vol. 182, No. 1, 383-399. (2007)
19. Liu, H. C., Mao, L. X, Zhang, Z. Y., Li, P.: Induced aggregation operators in the VIKOR method and its application in material selection. Applied Mathematical Modelling, Vol. 37, No. 9, 6325–6338. (2013)
20. Merigó, J. M.: A unified model between the weighted average and the induced OWA operator. Expert Systems with Applications, Vol. 38, No. 9, 11560-11572. (2011)
21. Merigó, J. M., Casanovas, M.: Decision making with distance measure and induced aggregation operators. Computers & Industrial Engineering, Vol. 60, No. 11, 66-76. (2011)
22. Merigó, J. M., Casanovas, M.: Induced aggregation operators in the Euclidean distance and its application in financial decision-making. Expert Systems with Applications, Vol. 38, No. 6, 7603-7608. (2011)
23. Su, W. H., Peng, W. Z., Zeng, S. Z., Peng, B., Pan, T. J.: A method for fuzzy group decision making based on induced aggregation operators and Euclidean distance. International Transactions in Operational Research, Vol. 20, No. 4, 579-594. (2013)
24. Wei, G. W., Zhao, X. F.: Some induced correlated aggregating operators with intuitionistic fuzzy information and their application to multiple attribute group decision making. Expert Systems with Applications, Vol. 39, No. 2, 2026-2034. (2012)
25. Wei, G. W., Zhao, X. F.: Induced hesitant interval-valued fuzzy Einstein aggregation operators and their application to multiple attribute decision making. Journal of Intelligent and Fuzzy Systems, Vol. 24, No. 4, 789-803. (2013).
26. Xu, Z. S.: Induced uncertain linguistic OWA operators applied to group decision making. Information Fusion, Vol. 7, No. 2, 231–238. (2006)
27. Zeng, S. Z., Su, W. H.: Linguistic induced generalized aggregation distance operators and their application to decision making. Economic Computer and Economic Cybernetics Studies and Research, Vol. 46, No. 2, 155-172. (2012)
28. Atanassov, K.: Intuitionistic fuzzy sets. Fuzzy Sets and Systems, Vol. 20, No. 1, 87-96. (1986)

29. Xu, Z. S., Yager, R. R.: Some geometric aggregation operators based on intuitionistic fuzzy sets. International Journal of General Systems, Vol.35, No.4, 417-433. (2006)
30. Liu, P. D., Jin, F.: Methods for aggregating intuitionistic uncertain linguistic variables and their application to group decision making. Information Sciences, Vol. 205, No. 1, 58–71. (2012)
31. Wu, J. Z., Zhang, Q.: Multicriteria decision making method based on intuitionistic fuzzy weighted entropy. Expert Systems with Applications, Vol. 38, No. 1, 916-922. (2011)
32. Xu, Y. J., Wang, H. M.: The induced generalized aggregation operators for intuitionistic fuzzy sets and their application in group decision making. Applied Soft Computing, Vol. 12, No. 3, 1168–1179. (2012)
33. Xu, Z. S., Xia, M. M.: Induced generalized intuitionistic fuzzy operators. Knowledge-Based Systems, Vol. 24, No. 2, 197-209. (2011)
34. Yu, D. J.: Intuitionistic fuzzy geometric Heronian mean aggregation operators. Applied Soft Computing, Vol. 13, No. 2, 131235-1246. (2013)
35. Yu, X. H., Xu, Z. S.: Prioritized intuitionistic fuzzy aggregation operators. Information Fusion, Vol. 14, No. 1, 108-116. (2013)
36. Zeng, S. Z.: Some intuitionistic fuzzy weighted distance measures and their application to group decision making. Group decision and Negotiation, Vol. 22, No. 2, 281-298. (2013)
37. Zeng, S. Z, Su, W. H.: Intuitionistic fuzzy ordered weighted distance operator. Knowledge-Based Systems, Vol. 24, No. 8, 1224–1232. (2011)
38. Zeng, S. Z., Su, W. H., Sun, L. R.: A method based on similarity measures for interactive group decision-making with intuitionistic fuzzy preference relations. Applied Mathematical Modelling, Vol. 37, No. 10, 6909–6917. (2013)
39. Zeng, S. Z., BALEŽENTIS, T., Su, W. H.: The multi-criteria hesitant fuzzy group decision making with multimoora method. Economic Computer and Economic Cybernetics Studies and Research, Vol. 47, No.3, 171-184. (2013)

Zeng Shouzhen was born in 1981. He has a MS degree in applied mathematics obtained in 2007 from Tianjin University, China. He received a PhD degree in statistics from Zhejiang Gongshang University. At present, he is an instructor in the College of Computer and Information in Zhejiang Wanli University. He has published more than 40 papers in journals, books and conference proceedings including journals such as *Statistics Research, Knowledge-Based Systems* and *Group Decision and Negotiation*. His main research fields are aggregation operators, decision making, comprehensive evaluation and uncertainty.

Wang Qifeng graduated from Chongqing University of the People's Republic of China and obtained the PHD degree in Management Science and Engineering in 2009. He has published more than 50 papers in journals, books and conference proceedings. His main research fields are in management science and information systems, etc. At present, he is a fulltime associate professor at Zhejiang Wanli University.

José M. Merigó is a Senior Research Fellow at the Manchester Business School, University of Manchester, UK. Previously, he was an Assistant Professor in the Department of Business Administration at the University of Barcelona, Spain. He received a MSc and a PhD degree (with Extraordinary Award) in Business Administration from the University of Barcelona. He also holds a Bachelor Degree in Economics from Lund University, Sweden. He has published more than 250 papers in journals, books and conference proceedings including journals such as *Information Sciences*, *International Journal of Intelligent Systems*, and *Computers & Industrial Engineering*. He is on the editorial board of several journals and scientific committees and serves as a reviewer in a wide range of journals. He is interested in Aggregation Operators, Decision Making and Uncertainty.

Pan Tiejun received the MS and PhD degrees in modern mechanical engineering from Zhejiang University, Hangzhou, China, in 1997 and 2001, respectively. He is currently a Master Tutor professor in the Department of Computer Science and Information Engineering at Zhejiang Wanli University. His research interests include software engineering, embedded system, theory and application of networked control system.

Received: January 09, 2013; Accepted: November 25, 2014

Computer Science and Information Systems 11(2):859–880 DOI: 10.2298/CSIS121222047T

Stability of Beta Coefficients of Sector and Subsector Portfolios in an Uncertain Environment

Antonio Terceño[1], M. Glòria Barberà-Mariné[1], Hernán Vigier[2], and Yanina Laumann[3]

[1] Department of Business, Universitat Rovira i Virgili
Av. de la Universidad 1, Reus, 43204, Spain
{antonio.terceno, gloria.barbera}@urv.cat
[2] Universidad Provincial del Sudoeste /
Department of Economics, Universidad Nacional del Sur / CIC - Bs As,
Alvarado 332, Bahía Blanca, 8000, Argentina
hvigier@uns.edu.ar
[3] Universidad Abierta Interamericana /
Department of Economics, Universidad Nacional del Sur,
Av. Pellegrini 1957, Rosario, 2000, Argentina
ylaumann@criba.edu.ar

Abstract. This paper is a first approach to the study of beta coefficients using fuzzy regression. We intend to improve the calculation of the sector and subsector betas of the Spanish Stock Market using fuzzy regression in an attempt to incorporate all future inaccuracies and the subjectivity associated with decision making. Our objective is to use all the information provided by the market to determine the systematic risk.

Keywords: risk, beta coefficient, fuzzy regression, CAPM.

1. Introduction

As is well known, the Capital Asset Pricing Model (CAPM) is a model for establishing optimal portfolios. It sets the expected return on any asset as a positive linear function of its systematic risk measured by means of the beta coefficient (β). This concept emphasizes the importance of systematic risk as a measure of non-diversifiable risk, the only risk that is remunerated in financial markets.

Since betas are non-observable, approximations that are typically based on historical data must be used. The basic underlying notion of this model is that every asset is affected by the market's general movements, assuming that the market factor is a systematic force. Other effects are assumed to be specific or unique to an individual asset and they diversify in a portfolio. One measure of the response of assets to changes in the markets could be obtained by relating the asset performance, R_j, to the performance of the market index, R_M, according to the following expression:

$$Rjt=aj+\beta j\ RMt+\varepsilon jt \quad j=1, 2, \ldots, N;t=1, \ldots, T$$

In practice, beta is the ordinary least squares estimator (OLS) of the return on asset j on the portfolio return over a period of time. This estimation, besides using historical returns, requires other practical assumptions. Each assumption can significantly affect results.

The aim of this study is to determine what impact the size of the portfolio and periodicity of the data have on the stability of betas. To do so we shall use fuzzy regression, taking into account all the stock quotes information during the trading days. Results are consistent with the literature but, nevertheless, there is an excess of information. Data was taken from Madrid´s stock exchange in the period 2005–2009, sector and subsector portfolios were studied and Tanaka & Ishibuchi's model was used (see [37]).

This paper begins by discussing related work. In Section 3 we present a brief description of the Tanaka e Ishibuchi fuzzy regression model that we will use later. In Section 4, we will use fuzzy regression to estimate the sector and subsector beta from the Spanish Stock Market in Madrid and analyze obtained data. Finally, we will present the findings.

2. Related Work

An essential requirement for using beta to obtain the future risk of a financial asset is that it has predictive power. Since future values are calculated from past data, they must be stable over time so that the estimation is correct and precise. Therefore, the more stable a value is over time, the more useful it will be. Although beta is an indicator of risk, its value is not unique and its result will depend on the hypothesis and data that are used. Many authors have studied beta´s historical evolution, and analyzed its capacity to make predictions from empirical and theoretical points of view.

The first decision that must be considered when calculating betas is the length of the sample period. A longer period provides more data, but the company itself could have changed its risk characteristics.

A conceptual problem arises when we try to determine the return on an asset. Financial theory does not specify if returns should be considered on a daily, weekly or even monthly basis. Several studies have shown that beta coefficients can vary substantially depending on the possession period by which their performances have been determined. The magnitude of such changes provides a measure of beta stability. Moreover, the calculation of betas will depend on which price is considered: closing price, average daily price, etc.

Various studies [1], [2], [3], [4], [5], [6] analyze the relation between the length of the estimation period and beta stationarity. They find that the prediction ability of betas (and consequently their stationarity) increases with the length of the period. However, this increase decreases in more diversified portfolios.

Beta assets vary from one period to another because, in the first place, the risk measured by the beta coefficient of a value can vary over time. In the second place, each period's beta is calculated with a random error which increases as the coefficient goodness and the prediction power decrease. If we consider a portfolio, random errors committed in the calculation of individual betas will tend to cancel each other out, so a portfolio beta is more stable than a single beta value.

Two studies, [1] and [2], analyze the seasonality of betas of individual securities and portfolios. They observe that, whereas betas of portfolios with a high number of securities provide a considerable amount of information about future betas, the betas of individual securities provide much less. This result suggests that a portfolio's beta is more stable than a single security's beta. The same direct relationship between the portfolio size and the beta stationarity has been observed in various studies [3], [4], [7], [8]. Another study [9] states that sector betas vary very little and, therefore, recommends using the calculated beta of one sector.

Traditional studies on the stability of portfolio betas differ from each other mainly in two areas: the portfolio construction method and the stability test.

Some of the traditional studies mentioned above have used the same portfolio construction method. This involves classifying portfolio securities according to their historical beta. In this way, they produce portfolios of N assets each. The assets with the highest beta are assigned to the first, the assets with the next highest beta to the second and so on until the smallest N is included in the last beta. This procedure is questioned by [10], which attributes the results of [1] and [2] to the portfolio selection method. If it had been random, there would have been no significant increase in the stationarity of the portfolio beta, even when it increased in size. They conclude, in short, that the results of traditional tests are a direct result of the portfolio construction method, and not of the increase in securities. However, [11] shows that both methods are valid and, therefore, that they lead to consistent results.

The most plausible explanation for the results obtained by [10] is the combination in this study of the random method of portfolio construction and the particular stability test used. This test was also used in pioneering stability work and involves calculating the portfolio betas for every two consecutive assessment periods and obtaining the correlation coefficients between them. If these coefficients in the different periods have high values (close to 1), the betas would be significantly stable. Otherwise, they would not be.

As well as indicating to what extent beta values change over time, this procedure makes it possible to detect the extent to which betas remain in the same group in successive time periods (see [4]). Like portfolios constructed on the basis of securities ranked by their beta, it will be difficult to produce changes in the beta value that are big enough to make them change their risk class. In fact, many studies have shown that high or low portfolio betas are more stable than intermediate portfolio betas. Instead, with randomly grouped securities, portfolios change their risk class more often.

In short, it is reasonable to expect correlation coefficients to be higher with prior holdings of securities ranked by their beta than with randomly selected securities portfolios. This explains the results obtained by [10] and the observation made by [11] that beta stabilities improved, regardless of the portfolio construction method employed, when the stability test used was the calculation of absolute deviations in betas rather than correlation coefficients. Using the mean absolute deviation as a measure of beta stationarity, these studies observed that it decreased as the number of securities in the portfolio increased.

In the light of the above, and in order to measure the instability of betas not the risk classes, [7] considers that it is much more appropriate to construct portfolios with randomly selected titles and some measure of deviation or change in those values over time, rather than correlation coefficients.

[7] (p. 46) stresses that “in the real world, investors are more worried that their portfolios do not change their risk class than they are of the changes in the order of their

portfolios in relation to other portfolios. Seasonality, in this way, should be an absolute measure and not a relative one". For this reason, the author proposes the simultaneous satisfaction of two conditions so that it can be said that beta is stationary. First, historical or ex-post betas should be an adequate approximation of future or ex-ante betas. This condition must be fulfilled if betas are to be used for predictive purposes. Second, the value of the future beta must not exceed certain limits that are acceptable to investors, so that the portfolio can remain within the same risk class on the considered horizon. This condition will be satisfied if the standard deviation (or variance) of the ex-ante beta is small, as this will mean that expected beta values have low dispersion around an expected mean value.

In short, stationarity improves when the number of portfolio securities increases if the average ex-post beta provides a better approximation of the average ex-ante beta as the size of the portfolio increases and if the standard deviation of the ex-ante beta decreases when the number of securities in the portfolio increases.

Another line of analysis is the hypothesis that betas vary over time. In [12] a conditional CAPM is specified, on the basis that the beta and expected returns vary over time. The results are better than those of the static model. Similarly, [13] uses 6 different techniques to make a study of 18 sectors in Europe, and shows that variable betas estimate the profitability of the sector, explained in terms of market movements, more efficiently than OLS. Similar results can be found in [14] and [15].

The previous empirical and theoretical literature on factors that can influence beta stability usually focused on a risk environment. This perspective highlights the instability of betas. In an attempt to incorporate all the underlying future uncertainty and the subjectivity related to the decision making process, we propose a further step that uses elements of the Theory of Fuzzy Sets. In particular, we propose to estimate the market model using fuzzy regression methods.

The objective of fuzzy regression is to determine a functional relationship between a dependent variable and a set of independent variables. As we will show, fuzzy regression is in many ways more versatile than conventional linear regression because functional relationships can be obtained when the independent variables, dependent variables, or both, are not crisp values but intervals or fuzzy numbers.

In contrast to ordinary regression, which is based on probability theory, fuzzy regression can be based on possibility theory and fuzzy set theory. In ordinary regression analysis, the unfitted errors between a regression model and observed data are generally assumed to be observation error, which is a random variable with a normal distribution, constant variance, and a zero mean. In fuzzy regression analysis, the same unfitted errors are viewed as the fuzziness of the model structure, as was initially developed in [16]. Subsequently, [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] and [27] made other contributions by applying different optimization criteria for a linear or curved adjustment. The literature on fuzzy regression applied to finance is growing. Some of the most recent contributions are [28], [29], [30], [31], [32], [33] and [34].

This modelling technique has some advantages over the traditional regression technique. It enables all the available information on prices to be incorporated. It is not limited to a single or an average price. In financial markets the same asset is traded at different prices during market hours. When econometric techniques are used, a single number must quantify observations (closing prices, average prices, etc.). In this process, a great deal of information is lost. The selection of one value or another is arbitrary. Fuzzy regression methods, on the other hand, make it possible to adjust the functional relation using all the information available about the observed values. In addition, the results of estimations are fuzzy numbers, not random variables, so they are simpler to treat and less demanding in terms of assumptions. For a more rigorous analysis of this issue see [35] and [36].

3. Fuzzy Regression using the Tanaka and Ishibuchi Model

The goal of fuzzy regression is to determine a functional relation between a dependent variable and several explanatory variables, where the estimated parameters are confidence intervals (CI). For a more rigorous analysis of this issue, see [35], [36], [38] and [39].

A CI A is represented by its upper and lower bounds as A=[a_1, a_2]; or by its centre and its radius A=<a_C, a_R> where:

$$a_C = (a_1 + a_2)/2 \qquad a_R = (a_1 - a_2)/2$$

If we have a sample {(Y_1, X_1), (Y_2, X_2),… , (Y_j, X_j), …, (Y_n, X_n)} where:

- Y_j is the j-th observation of the dependent variable, j=1,2,…,n, expressed by a confidence interval $Y_j = [Y_j^1, Y_j^2] = \langle Y_{jC}, Y_{jR} \rangle$
- X_j is the vector of the j-th observation of the independent variables, with j=1,2,…,n. Then, X_j is an m-dimensional variable $X_j = (X_{0j}, X_{1j}, X_{2j}, \ldots, X_{ij}, \ldots, X_{mj})$ where $X_{0j}=1$ ∀j, and X_{ij} is the value of the j-th observation for the i-th variable. We assume that observations are crisp.

The relation between the dependent and independent variables is linear:

$$Y = A_0 + A_1 X_1 + A_2 X_2 + \cdots + A_m X_m$$

where A_i, i =0,1,…,m are CI:

$$A_i = \langle a_{iC}, a_{iR} \rangle \quad i = 0, 1, \ldots, m$$

The final goal is to determine centres and radiuses of A_i that are compatible with the available observations.

In order to estimate the value of the j-th independent variable, $\hat{Y}_i = \langle \hat{Y}_{iC}, \hat{Y}_{iR} \rangle$, we do the sum:

$$\langle \hat{Y}_{iC}, \hat{Y}_{iR} \rangle = \sum \langle a_{iC}, a_{iR} \rangle X_{ii} = \left(\sum a_{iC} X_{ii}, \sum a_{iR}, |X_{ii}| \right), \quad j = 1, 2, \ldots, n$$

The goodness of fit is inversely related to the uncertainty (width) of the estimations of the observations $Y_j, \hat{Y}_i$. The width of $\hat{Y}_i$ is the radius of the confidence interval $\hat{Y}_{iR}$, which is obtained in the following way:

$$\hat{Y}_{iR} = \sum a_{iR} |X_{ii}| = a_{0R} + a_{1R} |X_{1i}| + \cdots + a_{mR} |X_{mi}|$$

Then, the total uncertainty of all the sample estimations, z, is the sum of the radiuses of the estimations:

$$z = \sum \hat{Y}_{iR} = \sum \sum a_{iR} |X_{ii}|$$

The parameters A_i must achieve not only the least possible uncertainty of $\hat{Y}_j$, but also that $\hat{Y}_j$, be as close as possible to the observation of the explained variable Y_j. In this context, we define two approximations of $\hat{Y}_i$ congruent with Y_j. [37] postulates that the observation must be included within its estimation: $Y_i \subseteq \hat{Y}_i \ \forall j$. In other words:

$$Y_{iC} - Y_{iR} \geq \hat{Y}_{iC} - \hat{Y}_{iR}$$

and

$$Y_{iC} + Y_{iR} \leq \hat{Y}_{iC} + \hat{Y}_{iR}$$

In order to determine the parameters, A_i, the following linear program must be solved:

$$\text{Min}\, z = \sum \hat{Y}_{iR} = \sum \sum a_{iR} |X_{ii}|$$

s. t:

$$\hat{Y}_{iC} - \hat{Y}_{iR} = \sum a_{iC} X_{ii} - \sum a_{iR} |X_{ii}| \leq Y_{iC} - Y_{iR} \quad j = 1, \ldots, n$$

$$\hat{Y}_{iC} + \hat{Y}_{iR} = \sum a_{iC} X_{ii} + \sum a_{iR} |X_{ii}| \geq Y_{iC} + Y_{iR} \quad j = 1, \ldots, n$$

$$a_{iR} \geq 0 \quad i = 0, 1, \ldots, m$$

The first restriction ensures that the lower bounds of the estimations are lower than the lower bounds of the observations. The second restriction guarantees that the upper bounds of the estimations are higher than the upper bounds of the observations. The third restriction imposes a positive radius for the CI.

4. Estimation of Sector and Subsector Betas

In order to perform our study, we took daily values from the General Index of the Madrid Stock Exchange (IGBM) and from the sector and subsector indices.

The specification and rationale for the temporal dimension of the rate of return is sometimes forgotten. Some studies use daily performance rates, while others use weekly, monthly or annual rates. There are, however, theoretical grounds for believing that the historical rate must be calculated from weekly quotes. First, [40] concludes that the beta coefficient estimated from weekly data is a response to the concept of instantaneous systematic risk of the market model.

Secondly, the use of daily quotes presents the problem of asynchronous or infrequent trading ([41], [42], among others), whereby the stock dynamics is not identical for all securities. This lack of synchronization in stock movements makes it advisable to space out the frequency of observations.

Thirdly, [43] argues that the week can be considered as the possession horizon for the investor for reasons of imperfection in information processing. The detection of the "weekend effect" or the "day-of-the–week effect" has led to proposals that investors should have a weekly possession horizon and objective considerations whose frequency is weekly (ordinary press releases, financial newsletters and securities, etc.).

Market return is proxied by the closing prices of the General Index of the Madrid Stock Exchange (IGBM). According to [31], in order to calculate betas using fuzzy regression models, $\left[\beta_1 , \beta_2\right]$, we express weekly returns by means of a confidence interval, $[R_1, R_2]$, whose bounds are given by:

- the lowest return, R_1, that the investor can achieve. This happens if the investor buys the asset at its highest price on day t-1 ($P_{\max,\ t-1}$) and sells the asset on day t at the lowest price ($P_{\min,\ t}$):

$$R_1 = (P_{\min,\ t} - P_{\max,\ t-1}) / P_{\max,\ t-1}$$

- the highest return, R_2. The investor obtains this return if he buys the asset at the lowest price and sells it at the highest price:

$$R_2 = (P_{\max,\ t} - P_{\min,\ t-1}) / P_{\min,\ t-1}$$

This way of calculating returns enables all the information contained in the different prices of each trading day to be included. With these data, we proceed with the estimation of the market model by using the Tanaka and Ishibuchi fuzzy regression model for the period between 01-01-2005 and 06-31-2009.

In order to verify if the number of securities and the length of the holding period influence the stability of the beta coefficient, as reported in studies with traditional techniques, we make estimations using fuzzy regression. Results are shown in Appendix A and B. Each table shows the result for a sector covered by IGBM and includes the result of the subsectors into which they are divided. So for each sector and subsector the table gives the minimum and maximum value (β_1, β_2) of the resulting interval. To analyze the importance of the length of time, the results are computed using quarterly (Table 3 to 7) and biannual data (Table 8 to 14).

According to [31] in order to verify if the number of securities and the length of the holding period influence the stability of the beta coefficient as reported in studies with traditional techniques, we use the standard deviation of the estimation of calculated betas as a comparison measure. We calculate the deviation of the lower betas, σ_{β_1}, upper betas, σ_{β_2}, the sum of both $\sigma_{\beta_1} + \sigma_{\beta_2}$, and the joint deviation of the β_1 and β_2, σ_{β_1,β_2}. The results for quarterly (biannual) betas are presented in Table 1 (Table 2).

Table 1. Quarterly beta coefficients of sector and subsector portfolios

Sector	Subsector	β average	σ_{B_1}	σ_{B_2}	$\sigma_{B_1}+\sigma_{B_2}$	σ_{B_1,B_2}
Oil And Energy		[0.82, 0.93]	0.51	0.47	0.98	0.49
	Oil	[0.58, 0.76]	1.16	0.95	2.12	1.05
	Electricity and Gas	[0.61, 0.70]	0.61	0.72	1.33	0.66
	Water and Others	[0.67, 0.76]	0.95	0.93	1.88	0.93
Basic Mat., Industry and Construction		[0.99, 1.00]	0.39	0.39	0.78	0.38
	Mineral/Metals/Transf	[0.66, 1.00]	0.49	1.51	2.00	1.12
	Capital Goods	[0.76, 0.80]	0.57	0.61	1.18	0.58
	Construction	[0.97, 1.00]	0.72	0.71	1.43	0.71
	Construct. Materials	[0.93, 0.97]	0.95	0.98	1.93	0.95
	Chemical	[0.78, 1.28]	1.46	1.52	2.98	1.49
	Engineering and Others	[0.97, 1.01]	0.73	0.84	1.57	0.78
	Aerospace	[0.58, 0.70]	1.22	1.02	2.24	1.11
Consumer Goods		[0.54, 0.63]	0.27	0.29	0.56	0.28
	Food and Beverages	[0.57, 0.59]	0.42	0.47	0.89	0.44
	Textiles/Clothing/Shoes	[0.63, 0.83]	0.55	0.66	1.21	0.61
	Paper and Graphic Arts	[0.59, 0.68]	0.63	0.47	1.10	0.55
	Other Consumer Goods	[0.03, 0.45]	0.66	0.93	1.59	0.82
	Pharmacy Products	[0.68, 0.79]	0.88	0.78	1.67	0.83
Consumer Services		[0.79, 0.82]	0.29	0.32	0.61	0.30
	Leisure/Tourism/Hotel	[0.85, 0.99]	0.62	0.84	1.46	0.73
	Retailing	[1.07, 1.19]	1.11	1.07	2.18	1.08
	Communicaction and Publicity	[0.79, 0.83]	0.62	0.68	1.30	0.64
	Car Parks/Motorways	[0.76, 0.84]	0.50	0.46	0.96	0.48
	Transport/Distribution	[0.88, 1.07]	0.81	0.63	1.44	0.72
	Other Services	[0.55, 0.62]	0.91	0.82	1.73	0.85
Financial Serv. & Real Estate		[0.93, 1.11]	0.33	0.41	0.74	0.38
	Banks	[0.92, 1.13]	0.34	0.45	0.79	0.41
	Insurance	[0.82, 0.96]	0.37	0.52	0.89	0.45
	Portfolio and Holding	[0.88, 0.98]	0.51	0.56	1.07	0.53
	Real Estate and Others	[0.47, 0.47]	0.78	0.78	1.56	0.77
Technology & Telecommunications		[0.65, 0.87]	0.50	0.85	1.35	0.70
	Electronics/Software	[0.85, 0.92]	0.51	0.54	1.05	0.52
	Telecommunications & Others	[0.64, 0.86]	0.52	0.87	1.39	0.72

Table 2. Biannual beta coefficients of sector and subsector portfolios

Sector	Subsector	β average	σ_{B_1}	σ_{B_2}	$\sigma_{B_1}+\sigma_{B_2}$	σ_{B_1,B_2}
Petrol And Power		[1,16, 1,16]	0.31	0.31	0.62	0.30
	Oil	[0.48, 0.48]	0.76	0.76	1.53	0.74
	Electricity and Gas	[0.93, 0.93]	0.74	0.74	1.48	0.72
	Water and Others	[0.83, 0.83]	0.83	0.83	1.67	0.81
Basic Mat., Industry And Construction		[1,01, 1,01]	0.38	0.38	0.76	0.37
	Mineral/Metals/Transf	[0.91, 0.98]	0.68	0.77	1.45	0.71
	Capital Goods	[0.57, 0.57]	0.52	0.52	1.04	0.51
	Construction	[0.67, 0.67]	0.90	0,90	1,81	0,88
	Construct. Materials	[0,67, 0,67]	0,90	0,90	1,81	0,88
	Chemical	[1,17, 1,17]	1,58	1,58	3,16	1,53
	Engineering and Others	[0.95, 0.95]	0.82	0.82	1.65	0.80
	Aerospace	[0.51, 0.51]	1.11	1.11	2.22	1.08
Consumer Goods		[0.61, 0.61]	0.22	0.22	0.44	0.21
	Food and Beverages	[0.52, 0.52]	0.30	0.30	0.61	0.29
	Textiles/Clothing/Shoes	[0.70, 0.70]	0.51	0.51	1.03	0.50
	Paper and Graphic Arts	[0.77, 0.77]	0.46	0.46	0.93	0.45
	Other Consumer Goods	[0.13, 0.13]	0.51	0.51	1.01	0.49
	Pharmacy Products	[0.73, 0.73]	0.67	0.67	1.34	0.65
Consumer Services		[0.92, 0.92]	0.24	0.24	0.48	0.23
	Leisure/Tourism/Hotel	[0.69, 0.69]	0.34	0.34	0.69	0.33
	Retailing	[1.41, 1.41]	1.04	1.04	2.07	1.00
	Communicaction and Publicity	[0.66, 0.66]	0.23	0.23	0.46	0.22
	Car Parks/Motorways	[1.09, 1.09]	0.46	0.46	0.92	0.45
	Transport/Distribution	[0.87, 0.87]	0.55	0.55	1.10	0.53
	Other Services	[0.74, 0.76]	0.88	0.87	1.75	0.85
Financial Serv. & Real Estate		[1.01, 1.01]	0.29	0.29	0.58	0.28
	Banks	[1.04, 1.04]	0.31	0.31	0.62	0.30
	Insurance	[1.21, 1.21]	0.49	0.49	0.98	0.48
	Portfolio and Holding	[0.85, 0.85]	0.51	0.51	1.03	0.50
	Real Estate and Others	[0.84, 0.84]	0.48	0.48	0.95	0.46
Technology & Telecommunications		[0.89, 0.89]	0.60	0.60	1.19	0.58
	Electronics/Software	[0.74, 0.74]	0.36	0.36	0.72	0.35
	Telecommunications & Others	[0.90, 0.90]	0.63	0.63	1.27	0.61

In this way, we verify that all quarterly sector betas are much more stable than those from their corresponding subsector. Similar conclusions are drawn from biannual betas.

To study the extent to which the length of the estimation period affects beta stability, portfolio betas are estimated for different intervals but the same holding period (weekly). An analysis of the beta's standard deviations for different estimation periods shows that the longer the period is the greater the stability. We verify that, in the Spanish market, the biannual beta is more stable than the quarterly beta using both ways of measuring deviation.

Some sectors are more stable than others, independently of the estimation period. Consumer Goods and Consumer Services are the most stable sectors, whereas Technology and Communications is a highly unstable sector.

Our results, obtained using a fuzzy methodology, are consistent with results of studies based on traditional econometric techniques.

5. Conclusions

In recent decades many academic studies have questioned beta stability. While earlier studies were based on simple methodologies, the development of models, algorithms and computational systems has led to more sophisticated testing techniques. Nevertheless, all these contributions take place within a risk environment. We consider that decision making processes, especially those using betas as a risk measure, take place in an uncertain environment. Therefore, in this paper we made a preliminary approach to studying this coefficient stability by using fuzzy regression models.

This modelling approach has some advantages over the traditional regression technique. In the first place, the estimations that we obtain after adjusting the fuzzy coefficients are not random variables, which are often difficult to treat numerically, but fuzzy numbers, which are easier to treat.

The fuzzy regression method presents different results from ordinary regression. The differences between fuzzy regression and ordinary regression are due to the different meanings of the deviations between the observed values and estimated values. In ordinary regression, deviations are viewed as random errors due to observation inconsistency. In fuzzy regression, the deviations are viewed as fuzzy errors due to system fuzziness. In ordinary regression analysis, probability theory is used to model random errors, and the result is presented as an ordinary regression equation. On the other hand, fuzzy set theory can be used to model fuzzy errors, and the result can be presented using a fuzzy regression equation.

If the phenomenon under study is economic or social, observations are a consequence of the interaction between the beliefs and expectations of the agents who take part in the phenomenon. We have already stated that, in our opinion, such a phenomenon should not be modelled using probability theory. For example, the security prices that are negotiated in financial markets are the consequence of the participants' expectations about the economic future, the trust that the security issuers generate in operators, etc.

In this case the linearity between the explained variable and the explanatory variables, which is assumed using both conventional and fuzzy regression, is oversimplified. However, we believe it is more realistic to model the bias that can arise between the realizations of the dependent variable and their theoretical values on the assumption that the relationship between the dependent variable and the explanatory variables is fuzzy, and not on the assumption that this bias is of a random nature. With respect to the prices of financial assets, we will be assuming, at least, that there is a strong subjective component in their determination.

Moreover, in many circumstances the observations of the dependent variable, the independent variable or both do not come from a particular number, but from a confidence interval. For example, the price that is negotiated in financial markets during a trading day for a particular security is hardly unique, but it is usually negotiated within a range limited by a maximum price and a minimum price. When the minimum square techniques—or the most sophisticated likelihood—are used, the observations of the explained (and explanatory) variable must be quantified using a unique number (for example, the average price negotiated or the last price in the model which will be implemented). This procedure clearly involves considerable information loss. When fuzzy regression models are implemented, the value of the observed variables does not need to be reduced to and represented by a single real number so we can work with all the information available.

The fuzzy regression method uses linear programming to estimate the fuzzy coefficients in the resulting models. As pointed out in [44], as the number of data sets increases, so it may be more difficult to use linear programming to estimate fuzzy beta. Each data set results in two constraints on the fuzzy regression formulation. As the number of data sets increases, the number of constraints increases proportionally. This increase might result in computational difficulties when using linear programme software or computers.

Taking into account that econometric fuzzy models mean that all the stock quotes information can be incorporated, and that there is no need to make assumptions on the basis of the random term which is difficult to apply, this method makes it possible to improve the prediction of future stock quotes.

The empirical evidence obtained from fuzzy regressions is consistent with that reported in traditional econometric studies on beta stability. The relevance of this verification is that the more stable β is, the more confident the predictions are. We observe that sector betas are more stable than subsector betas. Additionally, betas are more stable if the estimation periods are longer. Moreover, some sectors are more stable than others, independently of the estimation period. Consumer goods and Consumer Services are the most stable sectors, whereas Technology and Communications is highly unstable.

References

1. Levy, R. A.: On the Short-Term Stationarity of Beta Coefficients. Financial Analysts Journal, Vol. 27, No. 5, 55-62. (1971)
2. Blume, M. E.: On the Assessment of Risk. The Journal of Finance, Vol. 26, No.1, 1-10. (1971)
3. Altman, E. I., Jacquillat, B., Levasseur M.: Comparative Analysis of Risk Measures: France and the United States. The Journal of Finance, Vol. 29, No. 5, 1495-1511. (1974)
4. Eubank, A. A., Zumwalt, J. K.: Impact of Alternative Lengh Estimation and Prediction Periods on the Stability of Security and Portfolio Betas. Journal of Business Research, Vol. 9, No. 3, 321-325. (1981)
5. Armitage, S., Brzeszczynski, J.: Heteroscedasticity and Interval Effects in Estimating Beta: UK Evidence. 22nd Australasian Finance and Banking Conference 2009, Available at SSRN: http://ssrn.com/abstract=1100573. (2011)
6. KiHoon H. J., Satchell, S.: The sensitivity of beta to the time horizon when log prices follow an Ornstein–Uhlenbeck process. The European Journal of Finance ahead-of-print, 1-27. (2012)
7. Tole, T. M.: How to Maximize Stationarity of Beta. Journal of Portfolio Management, Vol. 7, No. 2, 45-49. (1981)
8. Iglesias Antelo, S.: Nuevas Evidencias acerca de la Estabilidad de las Betas de las Carteras.

In Proccedings of the XIII Congreso Nacional, IX Congreso Hispano-Francés de Asociación Europea de Dirección y Economía de la Empresa (Logroño, Spain), 59-62. (1999)

9. Damodaran, A.: The dark side of valuation. Financial Times-Prentice Hall, Nueva Jersey. (2001)
10. Porter, R. B., Ezzell, J. R.: A Note on the Predictive Ability of Beta Coefficients. Journal of Business Research, Vol. 3, No. 4, 365-372. (1975)
11. Alexander, G. J., Chervany, N. L.: On the Estimation and Stability of Beta. Journal of Financial and Quantitative Analysis, Vol. 15, No. 1, 123-137. (1980)
12. Jagannathan, R., Wang, Z.: The conditional CAPM and the cross-section of expected returns. The Journal of Finance, Vol. 51, No. 1, 3-53. (1996)
13. Mergner, S., Bullab, J.: Time-varying beta risk of Pan-European industry portfolios: A comparison of alternative modeling techniques. The European Journal of Finance, Vol. 14, No. 8, 771-802. (2008)
14. Lettau, M., Ludvigson, S.: Resurrecting the (C) CAPM: A crosssectional test when risk premia are time-varying. Journal of Political Economy, Vol. 109, No. 6, 1238-1287. (2001)
15. Nieto, B.: Evaluating multi-beta pricing models: An empirical analysis with spanish data. Revista de Economía Financiera, Vol. 2, 80-108. (2004)
16. Tanaka, H., Uejima, S., Asai, K.: Linear regression analysis with fuzzy model. IEEE Transactions on Systems, Man and Cybernetics, Vol.12, No. 6, 903-907. (1982)
17. Celminš, A.: Least squares model fitting to fuzzy vector data. Fuzzy Sets and Systems, Vol. 22, No. 3, 245-269. (1987)
18. Celminš, A.: Multidimensional least-squares model fitting of fuzzy models. Mathematical Modelling, Vol. 9, No. 9, 669-690. (1987)
19. Diamond, P.: Fuzzy least squares. Information Sciences, Vol. 46, No. 3, 141-157. (1988)
20. Tanaka, H.: Fuzzy data analysis by possibilistic linear models. Fuzzy Sets and Systems, Vol. 24, No. 3, 363-375. (1987)
21. Tanaka, H., Ishibuchi, H.: Identification of possibilistic linear systems by quadratic membership functions of fuzzy parameters. Fuzzy Sets and Systems, Vol. 41, No. 2, 145-160. (1991)
22. Savic, D.A., Pedrycz, W.: Evaluation of fuzzy regression models. Fuzzy Sets and Systems, Vol. 39, No. 1, 51-63. (1991)
23. González-Rodríguez, G., Blanco, A., Colubi, A., Lubiano, M. A.: Estimation of a simple linear regression model for fuzzy random variables. Fuzzy Sets and Systems, Vol. 160, No. 3, 357-370. (2009)
24. Tsaur, R. C., Wang, H. F.: Necessity analysis of fuzzy regression equations using fuzzy goal programming model. International Journal Fuzzy Systems, Vol. 11, No. 2, 107-115. (2009)
25. Li, X., Er, M. J., Lim, B. S., Zhou, J. H., Gan, O. P., Rutkowski, L.: Fuzzy regression modeling for tool performance prediction and degradation detection. International Journal of Neural Systems, Vol. 20, No. 05, 405-419. (2010)
26. Bisserier, A., Boukezzoula, R., Galichet, S.: Linear fuzzy regression using trapezoidal fuzzy intervals. In Foundations of Reasoning under Uncertainty. Springer Berlin Heidelberg, 1-22. (2010)
27. D'Urso, P., Massari, R., Santoro, A.: Robust fuzzy regression analysis. Information Sciences, Vol.181, No. 19, 4154-4174. (2011)
28. Barberà, M.G., Garbajosa, M.J., Guercio, M.B.: Term Structure of Interest Rate Analysis in the Spanish Market. Fuzzy Economic Review, Vol. 13, No. 2, 53-62. (2008)
29. Khemchandani, R., Jayadeva, Chandra, S.: Regularized least squares fuzzy support vector regression for financial time series forecasting. Expert Systems Applications, Vol. 36, No. 1, 132-138. (2009)
30. Chen, C. W., Wang, M. H. L., Lin, J. W.: Managing target the cash balance in construction firms using a fuzzy regression approach. International Journal of Uncertainty, Fuzziness and

Knowledge-Based Systems, Vol. 17, No. 5, 667-684. (2009)

31. Terceño, A., Barberà, M. G., Vigier, H., Laumann, Y.: Beta coefficient in the Spanish market sectors. Fuzzy regression versus crisp regression. Cuadernos de CIMBAGE, Vol. 13, 79-105. (2011)
32. Kocadagli, O.: A novel nonlinear programming approach for estimating CAPM beta of an asset using fuzzy regression. Expert Systems with Applications, Vol. 40, No. 3, 858-865. (2013)
33. Smimou, K.: On the significance testing of fuzzy regression applied to the CAPM: Canadian commodity futures evidence. International Journal of Applied Management Science, Vol. 5, No. 2, 144-171. (2013)
34. de Andres Sanchez, J., Terceño Gomez, A.: Applications of fuzzy regression in actuarial analysis. Journal of Risk and Insurance, Vol. 70, No.4, 665-699. (2003)
35. de Andrés, J., Terceño, A.: Estimating a Term Structure of Interest Rates for Fuzzy Financial Princing by Using Fuzzy Regression Methods. Fuzzy Sets and Systems, Vol 139, No. 2, 313-331. (2003)
36. de Andrés, J., Terceño, A.: Estimating a Fuzzy Term Structure of Interest Rates Using Fuzzy Regression Techniques. European Journal of Operational Research, Vol. 154, 804-818. (2004)
37. Tanaka, H., Ishibuchi, H.: A possibilistic regression analysis based on linear programming. In Fuzzy regression analysis, eds. J. Kacprzyk and M. Fedrizzi (Physica-Verlag, Heildelberg), 47-60. (1992)
38. Wang, H-F., Tsaur, R-C.: Insight of a fuzzy regression model. Fuzzy Sets and Systems, Vol. 112, No. 3, 355-369. (2000)
39. Yen, K. K, Ghoshray, S., Roig, G.: A linear regression model using triangular fuzzy number coefficients. Fuzzy Sets and Systems, Vol. 106, No. 2, 167-177. (1999)
40. Chen, P. L., Deets, M. K.: Systematic Risk and the Horizon Problem. Journal of Financial and Quantitative Analysis, Vol. 8, No. 2, 299-316. (1973)
41. Scholes, M., Williams, J.: Estimating Betas from Nonsynchronous Data. Journal of Financial Economics, Vol. 5, No. 3, 309-327. (1977)
42. Dimson, E.: Risk Measurement when Shares are Subject to Infrequent Trading. Journal of Financial Economics, Vol. 7, No. 2, 197-226. (1979)
43. Grande, I.: Efecto de las Ampliaciones de Capital y Pagos de Dividendos sobre la Evaluación Histórica del Coeficiente Beta. Revista Española de Financiación y Contabilidad, Vol. 13, No. 44, 273-295. (1984)
44. Chang, Y., Ayyub, B.: Fuzzy regression methods – a comparative assessment. Fuzzy Sets and Systems, Vol. 119, No. 2, 187-203. (2001)

Appendix A. Quarterly beta

Table 3. Oil and Energy

Quarter	Oil		Electricity and Gas		Water and Others		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	0.69	0.69	1.53	1.53	-0.15	0.87	1.23	1.23
2 2005	1.48	1.48	0.95	0.95	1.17	1.17	1.27	1.27
3 2005	-0.62	-0.62	1.20	1.29	1.90	1.90	0.65	0.65
4 2005	2.30	2.30	1.32	1.32	0.87	0.87	1.46	1.46
1 2006	-2.32	0.02	0.74	2.20	3.68	3.68	-0.28	1.62
2 2006	1.78	1.78	0.80	0.80	0.35	0.35	1.02	1.02
3 2006	0.51	1.12	1.74	1.74	0.75	0.75	1.55	1.55
4 2006	1.84	2.12	0.46	0.46	-0.22	-0.22	0.62	0.62
1 2007	1.13	1.13	0.32	0.32	1.21	1.21	0.25	0.25
2 2007	1.13	1.13	1.04	1.04	1.04	1.29	1.22	1.22
3 2007	2.02	2.02	0.81	0.81	0.44	0.44	1.39	1.39
4 2007	1.00	1.00	0.24	0.24	0.33	0.80	0.31	0.31
1 2008	-0.52	-0.52	0.41	0.41	-0.01	-0.01	1.34	1.34
2 2008	0.29	0.29	0.03	0.03	-0.10	-0.10	0.57	0.57
3 2008	0.36	0.36	-0.35	-0.35	0.03	0.03	0.22	0.22
4 2008	-0.41	-0.41	-0.15	-0.15	0.16	0.16	0.80	0.80
1 2009	0.14	0.14	0.02	0.02	0.69	0.69	0.58	0.58
2 2009	-0.41	-0.41	-0.09	-0.09	-0.12	-0.12	0.62	0.62

Table 4. Basic Material, Industry and Construction

Quarter	Mineral/ Metals/ Transf		Capital Goods		Construction		Construct. Materials		Chemical		Engineering and Others		Aerospace		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	1.17	2.09	0.49	0.49	0.93	0.93	0.92	0.92	1.59	1.59	0.66	0.66	0.95	0.95	0.70	0.81
2 2005	1.41	1.41	0.02	0.02	1.39	1.39	-0.37	-0.37	4.06	4.06	0.59	0.59	1.16	1.16	1.13	1.13
3 2005	0.79	0.79	0.86	1.36	0.49	0.93	-0.07	0.33	-2.41	4.65	1.45	1.45	0.27	0.27	0.79	0.84
4 2005	0.66	0.66	1.57	1.57	1.88	1.88	1.06	1.06	1.64	1.64	0.33	0.33	2.08	2.08	1.76	1.76
1 2006	1.49	6.65	1.61	1.88	1.19	1.19	3.54	3.83	0.39	0.39	2.29	3.12	-0.79	-0.79	1.69	1.69
2 2006	0.84	0.84	1.13	1.13	1.88	1.88	1.40	1.40	0.25	0.25	2.95	2.95	1.57	1.57	1.24	1.24
3 2006	0.26	0.26	0.66	0.66	1.03	1.27	0.05	0.05	-0.17	-0.17	1.54	1.54	-0.06	-0.06	1.01	1.12
4 2006	0.64	0.64	0.47	0.47	1.23	1.23	0.65	0.65	0.50	2.39	0.95	0.95	-2.31	-0.25	0.88	0.88
1 2007	0.05	0.05	1.02	1.02	0.91	0.91	1.34	1.34	1.00	1.00	0.44	0.44	0.04	0.04	0.87	0.87
2 2007	1.00	1.00	0.55	0.55	1.73	1.73	1.08	1.08	0.60	0.60	1.30	1.30	1.19	1.19	1.12	1.12
3 2007	0.84	0.84	1.28	1.28	2.04	2.04	1.49	1.49	1.45	1.45	1.16	1.16	1.41	1.41	1.45	1.45
4 2007	0.11	0.11	1.78	1.78	0.71	0.71	2.36	2.36	3.52	3.52	0.58	0.58	-0.30	-0.30	0.87	0.87
1 2008	0.74	0.74	0.18	0.18	1.39	1.39	1.38	1.38	0.37	0.37	0.51	0.51	2.13	2.13	1.19	1.19
2 2008	0.10	0.10	0.25	0.25	0.16	0.16	0.42	0.42	0.62	0.62	0.36	0.36	1.72	1.72	0.43	0.43
3 2008	0.71	0.71	0.28	0.28	0.92	0.92	0.64	0.64	0.61	0.61	0.40	0.40	1.78	1.86	0.62	0.62
4 2008	0.07	0.07	0.14	0.14	-0.04	-0.04	-0.19	-0.19	0.09	0.09	0.17	0.17	-0.20	-0.20	0.94	0.94
1 2009	1.00	1.00	1.00	1.00	-0.47	-0.47	0.40	0.40	-1.13	-1.13	0.61	0.61	-1.11	-1.11	0.27	0.27
2 2009	0.59	0.59	0.78	0.78	0.03	0.03	0.59	0.59	1.11	1.11	1.11	1.11	0.93	0.93	0.85	0.85

Table 5. Consumer goods

Quarter	Food and Beverages		Textiles/Clothing/ Shoes		Paper and Graphic Arts		Other Consumer Goods		Pharmacy Products		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	0.63	0.63	1.37	1.37	0.49	0.49	0.67	0.67	1.36	1.36	1.01	1.01
2 2005	0.93	0.93	-0.09	-0.09	0.46	0.46	0.70	0.70	1.74	1.74	0.37	0.37
3 2005	-0.32	-0.32	0.23	0.61	-0.18	-0.18	-0.79	-0.79	-1.26	0.34	0.16	0.78
4 2005	1.15	1.15	0.05	0.05	0.29	0.29	1.00	1.09	-0.36	-0.36	0.76	0.76
1 2006	1.23	1.69	1.23	1.23	-1.07	0.56	-2.09	3.52	0.94	0.94	0.42	0.42
2 2006	0.76	0.76	0.32	0.32	0.79	0.79	0.06	0.81	0.50	0.50	0.55	0.55
3 2006	0.41	0.41	0.68	0.68	1.01	1.01	0.28	0.28	0.82	0.82	0.60	0.60
4 2006	0.59	0.59	1.23	2.39	0.17	0.17	0.31	1.49	-0.46	-0.46	1.12	1.40
1 2007	0.69	0.69	0.62	1.33	2.07	2.07	0.07	0.07	1.48	1.96	0.51	0.51
2 2007	0.82	0.82	1.24	1.24	0.64	0.64	-0.06	-0.06	0.75	0.75	0.59	0.59
3 2007	0.49	0.49	1.11	1.11	1.15	1.15	-0.18	-0.18	0.95	0.95	0.55	0.55
4 2007	0.85	0.85	0.97	0.97	1.10	1.10	-0.04	-0.04	2.63	2.63	0.50	0.50
1 2008	0.91	0.91	0.04	0.04	0.68	0.68	0.00	0.00	0.58	0.58	0.31	0.31
2 2008	0.16	0.16	0.27	0.27	0.48	0.48	0.04	0.04	-0.15	-0.15	0.11	0.11
3 2008	0.53	0.53	-0.13	-0.13	0.67	0.67	-0.22	-0.22	0.34	0.34	0.46	0.46
4 2008	0.33	0.33	0.95	0.95	0.56	0.56	0.32	0.32	0.55	0.55	0.60	0.60
1 2009	0.20	0.20	1.25	1.25	0.53	0.53	0.16	0.16	0.78	0.78	0.91	0.91
2 2009	-0.18	-0.18	-0.04	1.31	0.71	0.71	0.30	0.30	0.98	0.98	0.31	0.86

Table 6. Consumer services

Quarter	Leisure/Tour . Hotel		Retailing		Communication and Publicity		Car Parks/ Motorways		Transport/ Distribution		Other Services		SECTOR	
	β1	β2	β1	β2	β1	β2	β1	β2	β1	β2	β1	β2	β1	β2
1 2005	0.18	0.48	3.54	3.54	-0.42	-0.42	0.11	0.69	-1.37	1.16	-1.22	-0.53	0.4	0.45
2 2005	0.29	0.29	0.28	0.28	1.83	1.83	1.43	1.43	1.09	1.09	0.99	0.99	0.9	0.91
3 2005	0.13	0.13	-0.31	0.77	0.58	0.58	-0.25	-0.25	1.49	1.49	0.82	0.91	0.3	0.32
4 2005	0.68	0.68	0.23	0.23	2.31	2.31	1.70	1.70	0.52	0.52	2.25	2.25	1.5	1.54
1 2006	1.33	2.48	2.13	2.13	0.92	0.92	0.50	0.50	0.92	0.92	-1.14	-0.85	0.6	0.61
2 2006	0.43	0.43	1.78	1.78	0.78	0.78	0.68	0.68	0.47	0.47	0.46	0.46	0.8	0.81
3 2006	0.56	0.56	1.98	1.98	0.65	0.92	0.36	0.36	0.64	0.64	1.64	1.64	0.6	0.81
4 2006	0.48	0.48	-0.31	-0.31	0.47	0.47	0.47	1.03	0.77	0.77	-0.23	-0.23	0.5	0.57
1 2007	1.21	1.21	0.99	0.99	0.31	0.31	0.60	0.60	1.41	2.17	0.95	0.95	0.8	0.81
2 2007	0.54	0.54	2.28	2.28	0.99	0.99	0.90	0.90	0.98	0.98	0.87	0.87	0.7	0.73
3 2007	1.40	1.40	2.23	2.23	0.60	0.60	1.73	1.73	0.46	0.46	1.05	1.05	1.2	1.26
4 2007	1.97	3.06	1.45	1.45	0.58	0.58	1.03	1.03	1.85	1.85	1.15	1.15	0.9	0.94
1 2008	0.51	0.51	-0.54	-0.54	0.45	0.45	0.86	0.86	0.76	0.76	1.33	1.33	1.0	1.01
2 2008	0.35	0.35	-0.14	-0.14	0.48	0.48	0.79	0.79	0.63	0.63	0.03	0.03	0.6	0.65
3 2008	1.11	1.11	0.65	0.65	0.50	0.50	0.84	0.84	2.81	2.81	-0.06	-0.06	0.8	0.85
4 2008	0.84	0.84	0.91	0.91	0.49	0.49	0.70	0.70	0.83	0.83	0.48	0.75	0.6	0.65
1 2009	0.98	0.98	1.25	1.25	1.02	1.02	0.78	0.78	0.75	0.75	-0.43	-0.43	0.5	0.56
2 2009	2.39	2.39	0.88	1.93	1.59	2.21	0.48	0.74	0.88	0.88	0.93	0.93	0.9	1.37

Table 7. Financial service & real estate

Quarter	Banks		Insurance		Portfolio And Holding		Real Estate And Others		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	0.96	0.96	0.98	1.48	-0.11	-0.11	0.89	0.89	0.97	0.97
2 2005	1.17	1.17	1.37	1.37	1.07	1.07	0.43	0.43	1.23	1.23
3 2005	0.75	1.50	1.08	1.08	0.51	1.30	-0.53	-0.53	0.79	1.42
4 2005	0.81	0.81	0.78	0.78	1.79	1.79	0.92	0.92	0.88	0.88
1 2006	1.18	1.53	0.67	1.07	1.01	1.93	-1.77	-1.77	1.38	1.38
2 2006	0.78	0.78	0.78	0.78	1.16	1.26	0.95	0.95	0.90	0.90
3 2006	0.79	0.79	0.62	0.62	1.20	1.20	0.24	0.24	0.81	0.81
4 2006	0.96	1.66	1.05	1.05	0.57	0.57	1.45	1.45	0.82	1.48
1 2007	1.05	1.05	0.63	0.63	1.22	1.22	1.46	1.46	1.01	1.01
2 2007	0.88	0.88	0.66	0.66	0.45	0.45	1.39	1.39	0.89	0.89
3 2007	0.87	0.87	0.75	0.75	1.46	1.46	0.18	0.18	0.82	0.82
4 2007	0.29	1.10	1.52	2.06	1.09	1.09	0.73	0.73	0.33	1.12
1 2008	0.73	0.88	0.54	0.54	1.59	1.59	0.59	0.59	0.83	0.91
2 2008	1.10	1.10	-0.11	-0.11	0.79	0.79	0.10	0.10	1.03	1.03
3 2008	0.31	0.31	0.50	0.50	0.05	0.05	-0.22	-0.22	0.25	0.25
4 2008	0.98	0.98	0.91	0.91	0.59	0.59	0.56	0.56	0.96	0.96
1 2009	1.84	1.84	1.22	1.22	0.91	0.91	0.38	0.38	1.75	1.75
2 2009	1.04	2.18	0.89	1.93	0.43	0.52	0.76	0.76	1.00	2.09

Table 8. Technology & Telecommunications

Quarter	Electronics/Software		Telecommunications & Others		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	0.30	0.30	0.89	0.89	0.90	0.90
2 2005	1.15	1.15	1.34	1.34	1.36	1.36
3 2005	1.59	1.59	0.55	0.55	0.60	0.60
4 2005	1.36	1.36	-0.89	-0.89	-0.76	-0.76
1 2006	0.75	1.49	0.97	1.68	1.01	1.74
2 2006	0.91	0.91	0.84	0.84	0.82	0.82
3 2006	0.86	1.23	0.93	0.93	0.95	0.95
4 2006	0.60	0.60	0.96	0.96	0.98	0.98
1 2007	1.08	1.09	0.95	0.95	0.94	0.94
2 2007	0.87	0.87	0.67	0.74	0.68	0.75
3 2007	1.17	1.17	0.95	0.95	0.93	0.93
4 2007	2.20	2.20	0.31	3.54	0.35	3.46
1 2008	0.49	0.49	1.23	1.23	1.22	1.22
2 2008	0.55	0.55	0.33	0.33	0.33	0.33
3 2008	0.37	0.37	0.13	0.13	0.12	0.12
4 2008	0.54	0.54	0.81	0.81	0.82	0.82
1 2009	0.18	0.18	0.15	0.15	0.15	0.15
2 2009	0.41	0.41	0.34	0.40	0.33	0.40

Appendix B. Biannual beta

Table 9. Oil and energy

Semester	Oil		Electricity And Gas		Water And Others		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	1.30	1.30	1.53	1.53	0.24	0.24	1.27	1.27
2 2005	1.12	1.12	1.29	1.29	1.90	1.90	1.27	1.27
1 2006	-0.64	-0.64	2.03	2.03	1.83	1.83	1.60	1.60
2 2006	0.79	0.79	1.56	1.56	1.70	1.70	1.43	1.43
1 2007	1.13	1.13	0.88	0.88	1.29	1.29	1.04	1.04
2 2007	1.08	1.08	0.81	0.81	0.44	0.44	1.17	1.17
1 2008	-0.57	-0.57	0.41	0.41	-0.10	-0.10	1.27	1.27
2 2008	0.00	0.00	-0.15	-0.15	0.16	0.16	0.80	0.80
1 2009	0.11	0.11	0.02	0.02	0.02	0.02	0.58	0.58

Table 10. Basic material, industry and construction

Semester	Mineral/ Metals/Transf		Capital Goods		Construction		Construct. Materials		Chemical		Engineering and Others		Aerospace		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	1.41	2.09	0.15	0.15	-0.37	-0.37	-0.37	-0.37	1.59	1.59	0.59	0.59	0.95	0.95	0.70	0.70
2 2005	0.66	0.66	1.33	1.33	0.69	0.69	0.69	0.69	4.65	4.65	0.42	0.42	0.70	0.70	1.47	1.47
1 2006	2.40	2.40	1.13	1.13	2.33	2.33	2.33	2.33	-0.13	-0.13	2.95	2.95	1.57	1.57	1.24	1.24
2 2006	0.21	0.21	0.15	0.15	0.05	0.05	0.05	0.05	0.13	0.13	1.13	1.13	-0.67	-0.67	0.88	0.88
1 2007	0.91	0.91	0.70	0.70	-0.30	-0.30	-0.30	-0.30	1.41	1.41	0.44	0.44	1.19	1.19	1.12	1.12
2 2007	0.84	0.84	0.70	0.70	1.49	1.49	1.49	1.49	2.40	2.40	1.16	1.16	-0.30	-0.30	1.45	1.45
1 2008	0.47	0.47	-0.18	-0.18	1.38	1.38	1.38	1.38	0.37	0.37	0.51	0.51	2.26	2.26	1.06	1.06
2 2008	0.25	0.25	0.19	0.19	0.31	0.31	0.31	0.31	0.40	0.40	0.29	0.29	-0.03	-0.03	0.94	0.94
1 2009	1.00	1.00	1.00	1.00	0.40	0.40	0.40	0.40	-0.28	-0.28	1.10	1.10	-1.11	-1.11	0.27	0.27

Table 11. Consumer goods

Semester	Food and Beverage		Textiles/ Clothing/Shoes		Paper and Graphic Arts		Other Consumer Goods		Pharmacy Products		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	0.80	0.80	0.09	0.09	0.69	0.69	0.70	0.70	1.36	1.36	0.66	0.66
2 2005	-0.09	-0.09	0.23	0.23	0.01	0.01	0.41	0.41	0.34	0.34	0.40	0.40
1 2006	0.58	0.58	0.32	0.32	0.74	0.74	-1.00	-1.00	0.50	0.50	0.55	0.55
2 2006	0.59	0.59	1.23	1.23	0.90	0.90	0.61	0.61	-0.44	-0.44	0.99	0.99
1 2007	0.74	0.74	1.18	1.18	1.68	1.68	0.07	0.07	1.96	1.96	0.51	0.51
2 2007	0.49	0.49	1.11	1.11	1.15	1.15	-0.18	-0.18	0.95	0.95	0.55	0.55
1 2008	0.87	0.87	0.04	0.04	0.48	0.48	0.00	0.00	0.58	0.58	0.31	0.31
2 2008	0.52	0.52	0.95	0.95	0.56	0.56	0.32	0.32	0.55	0.55	0.60	0.60
1 2009	0.20	0.20	1.13	1.13	0.71	0.71	0.22	0.22	0.78	0.78	0.91	0.91

Table 12. Consumer services

Semester	Leisure/ Tour. Hotel		Retail		Communic. Publicity		Car Parks /Motorways		Transport/ Distribution		Other Services		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	0.32	0.32	2.76	2.76	0.72	0.72	1.43	1.43	1.16	1.16	-0.53	-0.53	1.20	1.20
2 2005	0.43	0.43	0.66	0.66	0.88	0.88	1.90	1.90	0.42	0.42	2.25	2.25	1.07	1.07
1 2006	0.43	0.43	2.13	2.13	0.55	0.55	0.72	0.72	0.47	0.47	0.08	0.08	0.78	0.78
2 2006	0.79	0.79	0.98	0.98	0.65	0.65	0.47	0.47	0.64	0.64	1.49	1.49	0.60	0.60
1 2007	0.54	0.54	2.28	2.28	0.48	0.48	0.90	0.90	2.17	2.17	1.00	1.00	0.81	0.81
2 2007	1.40	1.40	2.23	2.23	0.60	0.60	1.55	1.55	1.03	1.03	1.15	1.15	1.26	1.26
1 2008	0.51	0.51	-0.54	-0.54	0.45	0.45	1.20	1.20	0.43	0.43	0.90	0.90	1.01	1.01
2 2008	0.84	0.84	0.91	0.91	0.44	0.44	0.84	0.84	0.75	0.75	0.49	0.75	0.61	0.61
1 2009	0.98	0.98	1.25	1.25	1.14	1.14	0.78	0.78	0.75	0.75	-0.22	-0.22	1.00	1.00

Table 13. Financial service & real estate

Semester	Banks		Insurance		Portfolio And Holding		Real Estate And Others		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	0.96	0.96	1.10	1.10	-0.11	-0.11	0.89	0.89	0.97	0.97
2 2005	0.85	0.85	1.61	1.61	0.37	0.37	1.45	1.45	0.82	0.82
1 2006	0.87	0.87	1.07	1.07	1.16	1.16	0.95	0.95	0.90	0.90
2 2006	0.96	0.96	0.92	0.92	1.11	1.11	1.31	1.31	0.82	0.82
1 2007	1.05	1.05	2.03	2.03	0.92	0.92	1.39	1.39	1.04	1.04
2 2007	0.94	0.94	1.09	1.09	1.46	1.46	0.18	0.18	0.91	0.91
1 2008	0.88	0.88	0.42	0.42	1.43	1.43	0.41	0.41	0.91	0.91
2 2008	0.98	0.98	0.91	0.91	0.59	0.59	0.56	0.56	0.95	0.95
1 2009	1.84	1.84	1.70	1.70	0.72	0.72	0.38	0.38	1.75	1.75

Table 14. Technology & telecommunications

Semester	Electronics/Software		Telecommunications & Others		SECTOR	
	β_1	β_2	β_1	β_2	β_1	β_2
1 2005	0.34	0.34	1.19	1.19	1.18	1.18
2 2005	1.36	1.36	-0.10	-0.10	-0.02	-0.02
1 2006	0.75	0.75	0.97	0.97	1.01	1.01
2 2006	0.60	0.60	0.93	0.93	0.95	0.95
1 2007	1.02	1.02	0.83	0.83	0.81	0.81
2 2007	1.17	1.17	2.12	2.12	2.04	2.04
1 2008	0.49	0.49	1.12	1.12	1.09	1.09
2 2008	0.54	0.54	0.84	0.84	0.80	0.80
1 2009	0.41	0.41	0.15	0.15	0.15	0.15

Antonio Terceño is Professor in the Department of Business Management at the University Rovira I Virgili, Spain. He received a MSc and a PhD degree in Economics and Business from the University of Barcelona. He has published more than 140 papers in journals, books and conference proceedings including journals such as *Fuzzy Sets and System, European Journal of Operational Research, Journal of Risk and Insurance, Emerging Market Review, The International Journal of Uncertainty, Fuzziness and Knowledge-based Systems. Economic Computation and Economic Cybernetics Studies and Research. He is* Coordinator Doctoral with Mention to the Excellence in Economics and Business. It has been: Vice rector, Dean of the Faculty of Business and Economics and Director of Department. He is interested in Finance, Decision Making and Uncertainty.

M^a Glòria Barberà-Mariné is an Associate Professor in the Department of Business Management and dean of de Faculty of Business and Economics at the University Rovira I Virgili, Spain. She received a MSc and a PhD degree in Business Administration from the University of Rovira I Virgili. She has published more than 90 papers in journals, books and conference proceedings including journals and editorials such as *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Fuzzy Economic Review, World Scientific, Kluwer*. She is currently interested in Decision Making and Uncertainty, Finance and Taxation.

Yanina Laumann is an Associate Professor in the Faculty of Business Administration at the Universidad Abierta Interamericana, Argentina. She is completing her PhD degree in Business Administration from the University of Rovira I Virgili, Spain. She has participated in several research projects at the University Rovira I Virgili, and also at the University Nacional del Sur, including publications in national and international congresses and specialized journals. She is currently interested in Uncertainty and Finance.

Hernan Vigier. Is a Full Professor in Department of Economics at National University of South, Bahia Blanca, Argentina. Also is Full Professor of faculty of Entrepreneurship and Local Development of Provincial South Western University, Pigue, Argentina. He received a PhD degree in Business Administration from the University of Rovira I Virgili, Reus, Spain. He is an adjunct researcher of the Committee of Scientific Research in the Province of Buenos Aires. He has published more than 120 papers in journals, books and conference proceedings including journals and editorials such as *Fuzzy Sets and System, Fuzzy Economic Review, Journal of Economic Computation and Economic Cybernetics Studies and Research, World Scientific,* Journal of Business and Entrepreneurship. He is currently interested in Decision Making and Uncertainty, Finance and Entrepreneurship.

Received: December 22, 2012; Accepted: December 03, 2014

Computer Science and Information Systems 11(2):881–904 DOI: 10.2298/CSIS121215048G

Joint Propagation of Ontological and Epistemic Uncertainty across Risk Assessment and Fuzzy Time Series Models

Vasile Georgescu [1]

[1] Department of Statistics and Informatics, University of Craiova, 13 A.I.Cuza,
Craiova, 200375, Romania
v_geo@yahoo.com

Abstract. This paper discusses hybrid probabilistic and fuzzy set approaches to propagating randomness and imprecision in risk assessment and fuzzy time series models. Stochastic and Computational Intelligence methods, such as Probability bounds analysis, Fuzzy α-levels analysis, Fuzzy random vectors, Wavelets decomposition and Wavelets Networks are combined to capture different kinds of uncertainty. Their most appropriate applications are probabilistic risk assessments carried out in terms of probability distributions with imprecise parameters and stochastic processes modeled in terms of fuzzy time series.

Keywords: risk assessment, fuzzy time series, probability bounds analysis, fuzzy random vectors, wavelets, Hukuhara difference.

1. Introduction: Challenges, Criticism and Rationales for a Novel Approach

Two kinds of uncertainty are contrasted in this paper (ontological, vs. epistemic uncertainty) and several techniques are addressed in order to capture and propagate both of them jointly across a specific model. The most relevant areas of applications carried out with such hybrid approaches are presented in what follows.

The first application area is addressed when attempting to extend the classical *probabilistic risk assessment* (PRA) modeling framework in such a way that allows probability distribution parameters to be imprecisely defined. Basically, PRA uses probability models to represent the likelihood of different risk levels in a population (i.e., randomness). In the standard probabilistic approach, inputs to the risk equation are described as *random variables* that can be defined mathematically by a probability distribution. The CDF for risk can be especially informative for illustrating the percentile corresponding to a particular risk level of concern. However, the presence of imprecision in a risk model adds another dimension to that of randomness and requires more comprehensive approaches to capture and represent the range within which the risk distribution might vary. Analytic methods for propagating the uncertainty (such as Probability bounds analysis, Fuzzy α-levels analysis) as well as stochastic simulation techniques, can be combined or integrated to reach synergy. This may lead to hybrid simulation frameworks, such as Fuzzy Monte Carlo, in an attempt to find the output of a model that has both random variables (given by probability distributions) and fuzzy variables for the inputs.

The second application area is addressed when attempting to capture the inherent fuzzy and random nature of some stochastic processes, expressed in terms of fuzzy time series. Unfortunately, modeling, estimating and forecasting fuzzy time series faces the problem of non-invertibility of the standard Minkowski addition and multiplication by scalars in a fuzzy framework.

In contrast with the case of real numbers, for some set-defined quantities, such as intervals and fuzzy sets, the opposite of A is not the inverse of A in Minkovsky addition (unless $A = \{a\}$ is a singleton). This implies that, in general, additive simplification is not valid, i.e., $(A + C = B + C) \not\Rightarrow A = B$, or $(A + B) - B \neq A$.

To partially overcome this situation, the Hukuhara difference has been proposed instead of fuzzy subtraction, assuming that there exists a set C for which $C = A \sigma_{\mathrm{H}} B \Leftrightarrow A = B + C$.

Let $\mathcal{K}_C(\Re^p)$ be the class of the non-empty compact convex subsets of $\Re^p$. An important property of "σ_{H}" is that $A \sigma_{\mathrm{H}} A = \{0\} \quad \forall A \in \mathcal{K}_C(\Re^p)$ and $(A + B) \sigma_{\mathrm{H}} B = A \quad \forall A, B \in \mathcal{K}_C(\Re^p)$. The H-difference is unique, but it does not always exist. A necessary condition for $A \sigma_{\mathrm{H}} B$ to exist is that A contains a translate $\{c\} + B$ of B.

Unfortunately, even if the Hukuhara difference exists, some distortions may still appear when applying least squares estimation.

Several generalizations of Hukuhara difference have been proposed in an attempt to obtain a more tractable way to deal with fuzzy regression analysis, assuming that fuzzy estimates can be still obtained from the condition of minimizing the sum of square residuals, expressed as a difference between two fuzzy quantities: the response of a system and its model based estimation.

The standard assumption is to consider square-integrable random variables defined on a Hilbert space equipped with a suitable L_2-metric that allows the projection theorem to be still valid. However, the projection cannot be properly applied as usually onto a subspace, but rather onto cones (i.e., subject to some constraints), due to the lack of a general additive inverse in the space of fuzzy variables, which is only a semi-linear space. This may lead to distorted results such as obtaining fuzzy least squares estimates with negative spreads.

The first and most notable generalizations of Hukuhara difference proposed in the context of fuzzy linear regression was that of Diamond [2], or Diamond and Körner [4], as the least squares solution of equation $A + X = B$, i.e., $C = B \sigma_{\mathrm{H}} A$, if and only if $d(A+C, B) = \inf_{X \in \mathcal{Y}} d(A+X, B)$, in some L_2-type metric space $(\mathcal{Y}, d)$.

When the usual Hukuhara difference $B \sigma_{\mathrm{H}} A$ exists, it coincides with the least squares solution defined above.

However, the extension of Hukuhara difference to an L_2-approximant, as proposed by Diamond, has some important limitations. The main criticism I can address to this approach is that it attempts to project the fuzzy data onto a projection cone as a whole, thus conserving a rigid structure and imposing constraints that are still too strong. Instead, I propose a suitable new method, based upon a partial decoupling principle. This exploits and extends an idea I introduced in 1998 (see [10]). It allows decomposing the monolithic fuzzy model into several more tractable crisp estimation sub-problems, starting from that one corresponding to modal values ($\alpha = 1$) in fuzzy data, and then proceeding in a decremental way for left and right α-level bounds, with α progressively decreasing towards 0. The estimates of modal values are not subject to any constraints, thus being obtained by applying the Hilbert space projection theorem directly onto the corresponding subspace. However, the estimates for the left and right α-level bounds can only be obtained by applying the projection theorem onto cones, in such a way to obtain least squares estimates without negative spreads. This leads to constrained quadratic programs, conveniently defined.

When applying to fuzzy time series estimation, the proposed approach is not only able to help decomposing, but also to make the process invertible, by recomposing a non-stationary fuzzy time series from its components, such as trend, cycle, seasonality and the simulated residuals, all of them properly defined as LR-fuzzy sets.

As an alternative to fuzzy estimation methods, computational intelligence techniques, based on wavelet decomposition and wavelet networks for nonlinear model fitting have been proposed to address fuzzy time series estimation and prediction.

2. Ontological versus Epistemic Uncertainty

Although the terms *imprecision* and *uncertainty* are often contrasted by fuzzy sets theorists, the latter is commonly used in a broad range of contexts with confusing connotations, varying from very specific to rather generic ones. According to EPA's guidance [6], an essential distinction is to be made between ontological and epistemic uncertainty.

Ontological uncertainty (also called randomness, variability, or aleatory / objective / irreducible uncertainty) arises from natural stochasticity, environmental variation across space or through time, genetic heterogeneity among individuals, and other sources of randomness.

Although randomness can often be better characterized by further specific study, it is not generally reducible by empirical effort. Randomness can be translated into risk (i.e., probability of some adverse consequence) by the application of an appropriate probabilistic model.

Epistemic uncertainty (also called subjective / reducible uncertainty) arises from incomplete knowledge about the world. Sources of epistemic uncertainty include measurement uncertainty, small sample sizes, detection limits and data censoring, ignorance about the details of the mechanisms and processes involved and other imperfections in scientific understanding. Epistemic uncertainty can in principle be reduced by focused empirical effort. It cannot be translated into probability, but it can be used in hybrid approaches to generate bounds on probability distributions. Such bounds may be either crisp (e.g., probability bounds, where the interval arithmetic is used to propagate the imprecision regarding distribution parameters) or fuzzy (e.g., fuzzy randomness, where both the machineries of probability theory and fuzzy sets theory are combined).

3. Hybrid Representations of Ontological and Epistemic Uncertainty

3.1. Computing with Probability Bounds

Probability bounds analysis combines probability theory and interval arithmetic to produce probability boxes (p-boxes), structures that allow the comprehensive propagation of both randomness and epistemic uncertainty through calculations in a rigorous way.

If we have only partial information about the probability distribution, then we cannot compute the exact values *F*(*x*) of the CDF. Instead, we can circumscribe *F*(*x*) by a pair of functions $\underline{F}(x)$ and $\overline{F}(x)$, each one representing a CDF, which bounds the (unknown) actual CDF. Such a pair of CDFs is called a *probability bound*, or a *p-bound*, for short. For every *x*, the possible values of the probability *F*(*x*) belongs to the interval [$\underline{F}(x)$, $\overline{F}(x)$].

In computations, it is often convenient to express a *p*-box in terms of its inverse functions ℓ and u defined on the interval of probability levels [0,1]. The function u is the inverse function of the upper bound on the distribution function and ℓ is the inverse function of the lower bound. These monotonic functions are bounds on the inverse of the unknown distribution function F

$$\ell(p) \geq F^{-1}(p) \geq u(p) \quad \textbf{(1)}$$

where p is probability level. Note that ℓ corresponds to $\overline{F}$ and u to $\underline{F}$.

It is simple to compute probability bounds for many cases in which the distribution family is specified, but only interval estimates can be given for the parameters. For instance, suppose that, from previous knowledge, it is assumed that a distribution is normal, but the precise values of the parameters that would define this distribution are uncertain. If there exist bounds on μ and σ (mean and standard deviation), bounds on the distribution can be obtained by computing the envelope of all normal distributions that have parameters within the specified intervals. These bounds are

$$\ell(p) = \max_{\theta} F_{\theta}^{-1}(p); \quad u(p) = \min_{\theta} F_{\theta}^{-1}(p) \quad \textbf{(2)}$$

where

$$\theta \in \{(\mu, \sigma) \mid \mu \in [\mu_\ell, \mu_u], \sigma \in [\sigma_\ell, \sigma_u]\} \quad \textbf{(3)}$$

and F is the CDF of a normal distribution with such parameters. In principle, making these calculations might be a difficult task since θ indexes an infinite set of distributions. However, in practice, finding the bounds requires computing the envelope over only four distributions: those corresponding to the parameter sets (μ_ℓ, σ_ℓ), (μ_ℓ, σ_u), (μ_u, σ_ℓ), and (μ_u, σ_u), as is shown in Figure 1. This simplicity is the result of how the family of distributions happens to be parameterized by μ and σ.

Nevertheless, it is just as easy to find probability bounds for cases with other commonly used distribution families such as lognormal, uniform, exponential, Cauchy, and many others.

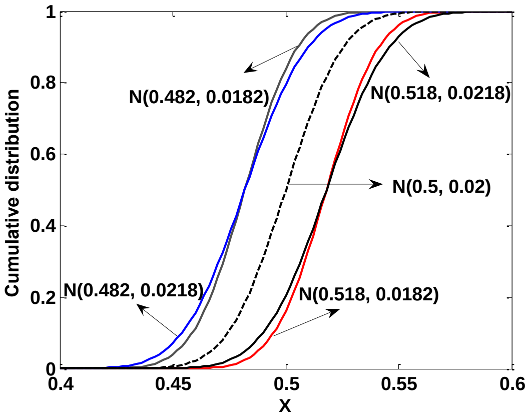


Fig. 1. Bounds on the CDF of a normal distribution with μ = [0.482, 0.518] and σ = [0.0182, 0.0218]. The dotted line is the CDF for the normal distribution with μ=0.5 and σ=0.02.

3.2. Fuzzy Random Variables and Processes

I consider an extension of the probability space $[\Omega; \mathcal{A}; P]$ by the dimension of fuzziness, i.e., by introducing a membership scale. This enables the consideration of imprecise observations as fuzzy realizations $\tilde{x}(\omega) = (\tilde{x}_1, \ldots, \tilde{x}_n) \subseteq X$ of each elementary event $\omega \in \Omega$. The attention will be restricted to the class $\mathcal{F}_C(\Re)$ of normal convex fuzzy sets on $\Re$, whose α-level sets are in the class $\mathcal{K}_C(\Re)$ of nonempty compact real intervals.

A *fuzzy* random variable $\tilde{X}$ is the fuzzy result of the uncertain mapping $\tilde{X} : \Omega \to \mathcal{F}_C(\Re)$, such that for each $\alpha \in [0, 1]$ and $\omega \in \Omega$, the α-level intervals $X_\alpha(\omega) = \left[\inf(X(\omega))_\alpha, \sup(X(\omega))_\alpha\right]$, generated by the mapping $X_\alpha : \Omega \to \mathcal{K}_C(\Re)$, are (compact convex) random sets. In other words, $X_\alpha(\omega)$ are Borel-measurable w.r.t. the Borel σ-field generated by the topology associated with a suitable metric on $\mathcal{K}_C(\Re)$, usually the Hausdorff metric d_H:

$$d_H(K,K') = \max\left\{ \sup_{k\in K} \inf_{k'\in K'} |k-k'|, \sup_{k'\in K'} \inf_{k\in K} |k-k'| \right\} \\ = \max\left\{ \left|\inf K - \inf K'\right|, \left|\sup K - \sup K'\right| \right\} \tag{4}$$

The fuzzy probability distribution function $\widetilde{F}(x)$ of $\widetilde{X}$ is the set of probability distribution functions of all originals X_j of $\widetilde{X}$ with the membership values $\mu(F(x))$. The quantification of fuzziness by fuzzy parameters leads to the description of the fuzzy probability distribution function $\widetilde{F}(x)$ of $\widetilde{X}$ as a function of the fuzzy bunch parameter $\widetilde{s}$.

$$\widetilde{F}(x) = F(\widetilde{s}, x)\,. \tag{5}$$

For the purposes of numerical evaluation, α-discretization is advantageously applied.

$$F(\widetilde{s},x) = \{F_\alpha(x);\, \mu(F_\alpha(x)) \mid F_\alpha(x) = [\underline{F}_\alpha(x), \overline{F}_\alpha(x)], \\ \mu(F_\alpha(x)) = \alpha, \forall \alpha \in [0,1]\}\,. \tag{6}$$

with $\underline{F}_\alpha(x) = \inf\{F(s,x) \mid s \in s_\alpha\}$ and $\overline{F}_\alpha(x) = \sup\{F(s,x) \mid s \in s_\alpha\}$.

With the aid of α-discretization a fuzzy random function may be formulated as a set of α-level sets of ordinary random functions

$$\widetilde{X}(t) = \left\{X_\alpha(t);\, \mu\left(X_\alpha(t)\right) \mid X_\alpha(t) = \left[\underline{X}_\alpha(t), \overline{X}_\alpha(t)\right], \\ \mu\left(X_\alpha(t)\right) = \alpha, \forall \alpha \in [0,1] \right\}. \tag{7}$$

A *fuzzy random process* $(\widetilde{X}_t)_{t\in T}$ is defined as a family of fuzzy random variables $\widetilde{X}_t$ over the space T of the time coordinate t.

A *fuzzy time series* $(\widetilde{x}_t)_{t\in 1,2,\ldots,N}$ is a realization of a fuzzy random process $(\widetilde{X}_t)_{t\in T}$ and consists of a temporally ordered sequence of fuzzy variables $\widetilde{x}_t$, each one assigned to each discrete observation time.

4. Propagating Uncertainty in Risk Assessment Models

4.1. Propagating Randomness in Risk Models with no Parameter Uncertainty: One-dimensional Monte Carlo Analysis (1D MCA)

A Monte Carlo analysis that characterizes either uncertainty or variability in each input variable can be described as a one-dimensional Monte Carlo analysis (1D MCA). In its general form, the risk equation can be expressed as a function of multiple risk exposure variables (V_i) : $\text{Risk} = f(V_1, \ldots, V_n)$.

Often the input distributions are assumed to be independent. The value of one variable has no relationship to the value of any other variable. In this case, a value for each variable (V_i) is selected at random from a specified PDF and the corresponding risk is calculated. This process is repeated many times (e.g., 10,000). Each iteration of a Monte Carlo simulation should represent a plausible combination of input values. A unique risk estimate is calculated for each set of random values. Repeatedly sampling (V_i) results in a frequency distribution of risk, which can be described by a PDF or a CDF. A sufficient number of iterations should be run to obtain numerical stability in percentiles of the output (e.g., risk distribution). The risk distributions derived from a PRA allow for inferences to be made about the likelihood or probability of risks occurring within a specified range of the input variables.

More complex Monte Carlo simulations can be developed that quantify some dependence between one or more input distributions by using conditional distributions or correlation coefficients.

4.2. Propagating Randomness and Epistemic Uncertainty Simultaneously: Two-dimensional Monte Carlo Analysis

A two-dimensional Monte Carlo Analysis (2D MCA) is a term used to describe a model that simulates both uncertainty and randomness in one or more input variables. Uncertainty in the parameter estimates can be represented in a PRA model as follows. Consider a random input variable whose parameter estimates are affected by uncertainty. Assume normal PDFs can be specified for both uncertain parameters: the mean and the standard deviation. Uncertainty in the mean is described by the normal PDF with parameters (μ_{mean}=5, σ_{mean}=0.5); similarly, uncertainty in the standard deviation is described by the normal PDF with parameters (μ_{SD} =1, σ_{SD} =0.5). A variable described in this way is called a second order random variable.

A two-dimensional Monte Carlo simulation is a nesting of two ordinary Monte Carlo simulations. Typically, the inner simulation represents natural variability of the underlying processes, while the outer simulation represents the analyst's uncertainty about the particular parameters that should be used to specify inputs to the inner simulation. This structure means that each iterate in the outer simulation entails an entire Monte Carlo simulation, which can lead to a very large computational burden.

4.3. Probability Bounds Analysis Compared to Monte Carlo Simulation

When one or more point estimates defined in the risk model are uncertainty, a Monte Carlo analyst might employ a two-dimensional Monte Carlo simulation that includes an uncertainty (inner) loop for each uncertain point estimate.

In a probability bounds analysis, the same interval of possible values used in the Monte Carlo analyst's uncertainty loop replaces the point estimate; however the semi-analytic nature of the probability bounds analysis results in an exact representation of the stated uncertainty. As the number of times the inner loop is called in the Monte Carlo simulation approaches infinity, the result of the Monte Carlo analysis converges on the probability bounds result.

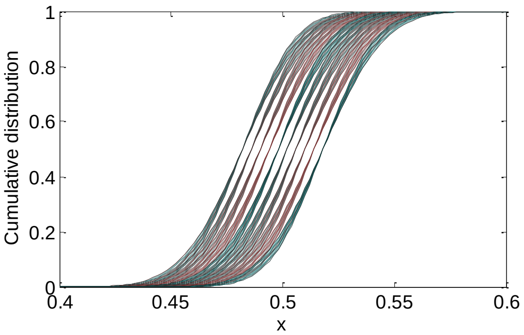


Fig. 2. Monte Carlo vs. Probability Bounds (see also Fig.1): uncertainty regarding the exact value of the parameters of a probability distribution, where $\mu = [0.482, 0.518]$ and $\sigma = [0.0182, 0.0218]$.

4.4. Hybrid Approaches to Propagating Randomness and Fuzziness in Risk Assessment

The idea is to find the output of a model $g(X_1, \ldots, X_n, \widetilde{X}_1, \ldots, \widetilde{X}_m)$ that has both random variables $X_1, \ldots, X_n$, given by probabilistic distributions, and fuzzy variables $\widetilde{X}_1, \ldots, \widetilde{X}_m$, for the inputs. To estimate the output of this generalized model, most researchers attempt to eliminate or transform one type of uncertainty to another before performing a simulation (e.g. possibility to probability transformation). Guyonnet et al. (2003) first proposed a "hybrid approach" with both fuzzy and random types of uncertainty without transforming one type to another. They calculated the Inf and Sup values of the model g considering all the values that are located within the $\alpha -$cuts of the input fuzzy sets and suggested that minimization and maximization algorithm can be used for finding Inf and Sup values of a general model. However, in their application, the model was a simple monotonic function, and the Inf and Sup values were identified directly without using minimization or maximization algorithms.

A more tractable way to propagating both randomness and fuzziness is based on a fuzzy generalization of the Monte Carlo (FMC) simulation framework, which integrates fuzzy arithmetic method with Monte Carlo simulation to find the output of a model with both fuzzy and probabilistic inputs.

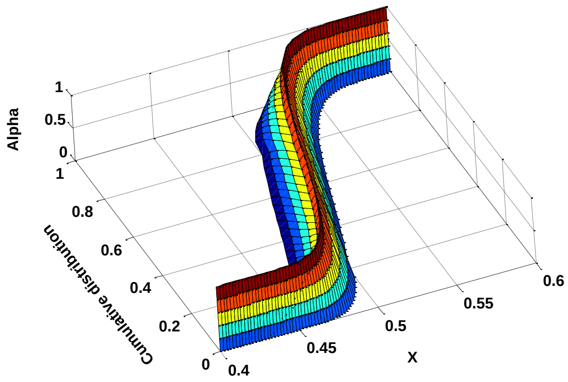


Fig. 3. 3D view of fuzzy CDF resulting from the output of FMC by aggregating α-CDF bounds.

Since in FMC, fuzzy arithmetic (in α-cut form) is performing for each sample set, the output of FMC is represented as a number of fuzzy sets with random variation. This randomness results from random sampling of random input parameters. The fuzzy CDF is used for finding the *fuzzy probability* of not exceeding a given threshold and a *fuzzy quantile* corresponding to a given probability.

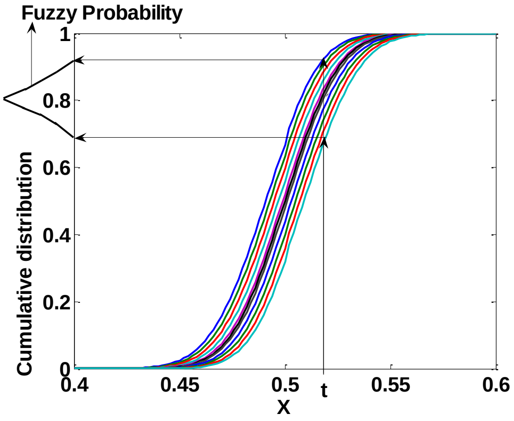


Fig. 4. The fuzzy probability of not exceeding a specific threshold $t \in X$.

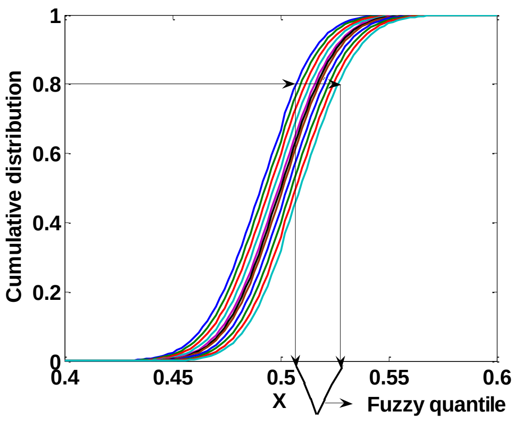


Fig. 5. A fuzzy quantile corresponding to a given probability.

5. Propagating Randomness and Fuzziness Jointly Across Fuzzy Time Series Models

The estimation technique proposed in this paper is based upon a partial decoupling principle, an early form of which I proposed in [10]. It allows decomposing the monolithic fuzzy model into several crisp models, starting from that one corresponding to modal values ($\alpha = 1$) in fuzzy data, and then proceeding in a decremental way for left and right α-level bounds, with α progressively decreasing towards 0. The estimates of modal values are not subject to any constraints, thus being obtained by projecting the 1-level data directly onto the corresponding subspace. However, the estimates for the left and right α-level bounds can only be obtained by projecting the α-level data onto cones, in such a way to obtain least squares estimates without negative spreads. This leads to constrained quadratic programs, conveniently defined.

5.1. Fuzzy Data: Minimum, Average and Maximum Daily Temperatures Registered at a Local Weather Station

The observed sequence consists of the minimum, average and maximum daily temperatures registered at a local weather station. The fuzzy time series is represented in figure 6 and the corresponding empirical fuzzy cumulative distribution function in figure 7.

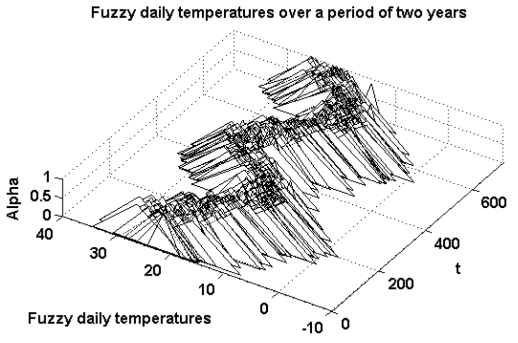


Fig. 6. Fuzzy daily temperatures over a period of two years (730 days).

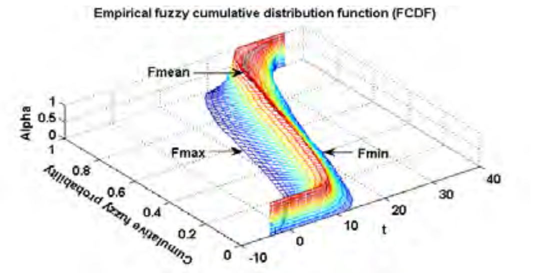


Fig. 7. Empirical Fuzzy Cumulative Distribution Function (FCDF), showing the fuzzy probability of not exceeding a given temperature, or reversely, the fuzzy quantile corresponding to a given probability.

5.2. The Fuzzy Time Series Model for Daily Temperatures, with Fuzzy Trend and Fuzzy Cyclical Component

In what follows, we will exemplify some suitable methods for modelling and forecasting non-stationary fuzzy time series, based on the fuzzy component model, which decomposes the fuzzy time series into a trend component, a cyclical component (or, alternatively, a seasonal component) and a fuzzy residual component:

$$\widetilde{Y}(t) = \widetilde{T}(t) + \widetilde{C}(t) + \widetilde{u}(t) \quad \textbf{(8)}$$

If we restrict to the class of triangular fuzzy numbers, which are a special case of LR fuzzy numbers, we can decompose the fuzzy model into 3 crisp models: one model for the modal (average) values $Y^C(t)$ and two models for the minimum values $Y^L(t)$ and maximum values $Y^R(t)$, respectively.

$$\begin{pmatrix} Y^L(t) \\ Y^C(t) \\ Y^R(t) \end{pmatrix} = \begin{pmatrix} T^L(t) \\ T^C(t) \\ T^R(t) \end{pmatrix} + \begin{pmatrix} C^L(t) \\ C^C(t) \\ C^R(t) \end{pmatrix} + \begin{pmatrix} u^L(t) \\ u^C(t) \\ u^R(t) \end{pmatrix}. \qquad \textbf{(9)}$$

Due to the properties of minimum, mean and maximum, the following order relations hold (implying positive spreads for $Y(t)$, $T(t)$, $C(t)$ and $u(t)$):

$$Y^L(t) \le Y^C(t) \le Y^R(t); \qquad T^L(t) \le T^C(t) \le T^R(t); \qquad \textbf{(10)}$$

$$C^L(t) \le C^C(t) \le C^R(t); \qquad u^L(t) \le u^C(t) \le u^R(t).$$

The model for the modal values is estimated without any constrains, i.e., by orthogonal projection of the observed values onto the appropriate subspace. The other two models for the minimum and maximum values are estimated subject to some non-negativity restrictions on spreads, corresponding to the time series components: trend ($\hat{T}^C - \hat{T}^L \ge 0$, $\hat{T}^R - \hat{T}^C \ge 0$), cyclical component ($\hat{C}^C - \hat{C}^L \ge 0$, $\hat{C}^R - \hat{C}^C \ge 0$) and residuals ($\hat{u}_1^C - \hat{u}_1^L \ge \xi > 0$, $\hat{u}_1^R - \hat{u}_1^C \ge \xi > 0$). This leads to constrained quadratic programs, i.e., to the projection of the observed values onto some cones.

The regressors for the linear fuzzy trend are defined by the matrix:

$$T = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \cdots & \cdots \\ 1 & N \end{pmatrix}; \quad N = 730. \qquad \textbf{(11)}$$

The simplest way for representing $C(t)$ as a periodic function, with $C(t) = C(t-p)$, is to assume harmonic functions, such as the sine or cosine: $\sin(2\pi t/p)$ and $\cos(2\pi t/p)$, where p (=365 in our case) is called period, its inverse $f = 1/p$ is called frequency, and $\omega = 2\pi f = 2\pi/p$ is the angular frequency. Thus, the regressors for $C(t)$ are combined in the matrix:

$$C = \begin{pmatrix} \cos(1 \cdot 2\pi/365) & \sin(1 \cdot 2\pi/365) \\ \cos(2 \cdot 2\pi/365) & \sin(2 \cdot 2\pi/365) \\ \cdots & \cdots \\ \cos(\mathrm{N} \cdot 2\pi/365) & \sin(\mathrm{N} \cdot 2\pi/365) \end{pmatrix}. \qquad \textbf{(12)}$$

5.3. Estimating the Fuzzy Trend (De-trending)

Step1: First, we estimate the trend corresponding to the average daily temperature, without any constraints, i.e., as a projection of Y^C onto the subspace Im(T) generated by the columns of matrix T :

$$\hat{T}^C = P \cdot Y^C, \quad \text{where } P = T \cdot (T' \cdot T)^{-1} \cdot T'. \tag{13}$$

Step 2: Second, we estimate the trend corresponding to the minimum daily temperature, as a solution of a constrained quadratic program:

$$\min_{b^L_{trend}} \left\{ \left(b^L_{trend}\right)' T' T \, b^L_{trend} - 2\left(Y^L\right)' T \, b^L_{trend} \right\}. \tag{14}$$

subject to:

$$\begin{cases} \hat{T}^L \le \hat{T}^C \\ \hat{u}^L_1 + \xi \le \hat{u}^C_1 \end{cases} \Leftrightarrow \begin{cases} T \cdot b^L_{trend} \le \hat{T}^C \\ -T \cdot b^L_{trend} \le Y^C - \hat{T}^C - Y^L - \xi \end{cases}. \tag{15}$$

$\hat{T}^L \le \hat{T}^C$ means that the left spread of the fuzzy trend must be non-negative, i.e., $\hat{T}^C - \hat{T}^L \ge 0$, with $\hat{T}^C$ already estimated at step 1.

$\hat{u}^L_1 + \xi \le \hat{u}^C_1$ means that the left spread of the intermediary fuzzy residuals (after de-trending) must be strictly positive, i.e., $\hat{u}^C_1 - \hat{u}^L_1 \ge \xi > 0$. The reason for this is that the intermediary fuzzy residuals will be further decomposed into a cyclical component and final residuals (those obtained after removing both trend and cyclical component). At this step, we recommend for ξ a value between 0.1 and 0.5.

Step 3: Third, we estimate the trend corresponding to the maximum daily temperature, as a solution of a constrained quadratic program:

$$\begin{cases} \hat{T}^L \le \hat{T}^C \\ \hat{u}^L_1 + \xi \le \hat{u}^C_1 \end{cases} \Leftrightarrow \min_{b^R_{trend}} \left\{ \left(b^R_{trend}\right)' T' T \, b^R_{trend} - 2\left(Y^R\right)' T \, b^R_{trend} \right\}. \tag{16}$$

subject to:

$$\begin{cases} -T \cdot b^R_{trend} \le -\hat{T}^C & \Leftrightarrow \quad \hat{T}^R \ge \hat{T}^C \\ T \cdot b^R_{trend} \le -Y^C + \hat{T}^C + Y^R - \xi & \Leftrightarrow \quad \hat{u}^R_1 - \hat{u}^C_1 \ge \xi > 0 \end{cases}. \tag{17}$$

$\hat{T}^R \ge \hat{T}^C$ means that the right spread of the fuzzy trend must be non-negative, i.e., $\hat{T}^R - \hat{T}^C \ge 0$

$\hat{u}^R_1 - \hat{u}^C_1 \ge \xi > 0$ means that the right spread of the intermediary fuzzy residuals (after de-trending) must be strictly positive.

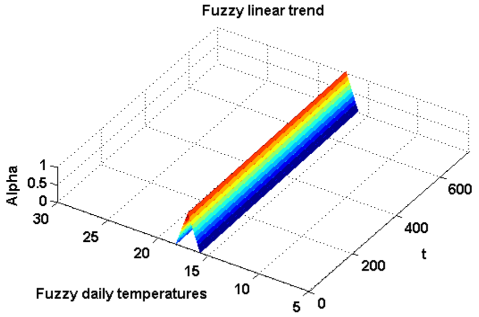


Fig. 8. Fuzzy linear trend.

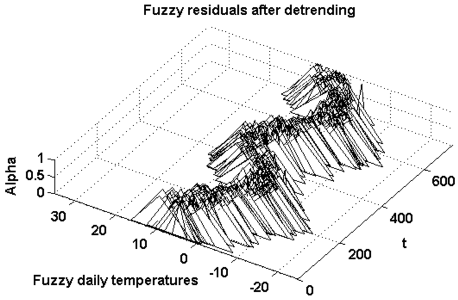


Fig. 9. Intermediary fuzzy residuals after detrending (the residuals are fuzzy sets in a proper sense – with non-negative spreads).

5.4. Estimating the Fuzzy Cyclical Component

The periodogram or sample spectrum shows the variation of the peak points of empirical daily temperature data. The maximum value of the periodogram is about 365 days.

Step1: First, we estimate the cyclical component corresponding to the average daily temperature, without any constraints, i.e., as a projection of $\hat{u}_1^C = Y^C - \hat{T}^C$ (the intermediary residuals after de-trending) on the subspace $\mathrm{Im}(C)$ generated by the columns of matrix C :

$$\hat{C}^C = P \cdot \hat{u}_1^C \text{, where } P = C \cdot (C' \cdot C)^{-1} \cdot C' . \tag{18}$$

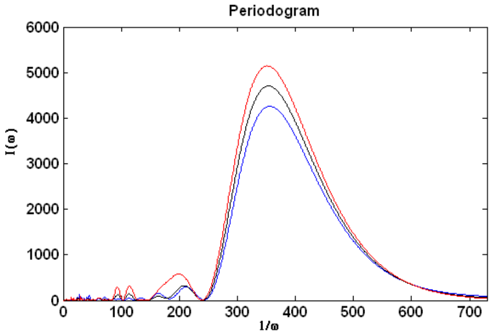


Fig. 10. Periodogram with a maximum corresponding to about 365 days (one year).

Step 2: Second, we estimate the cyclical component corresponding to the minimum daily temperature, as a solution of a constrained quadratic program:

$$\min_{b^L_{cycle}} \left\{ \left(b^L_{cycle}\right)' C' C\, b^L_{cycle} - 2\left(\hat{u}^L_1\right)' C\, b^L_{cycle} \right\} . \tag{19}$$

subject to:

$$\begin{cases} \hat{C}^L \le \hat{C}^C \\ \hat{u}^L_2 + \xi \le \hat{u}^C_2 \end{cases} \Leftrightarrow \begin{cases} C \cdot b^L_{cycle} \le \hat{C}^C \\ -C \cdot b^L_{cycle} \le \hat{u}^C_1 - \hat{C}^C - \hat{u}^L_1 - \xi \end{cases} . \tag{20}$$

$\hat{C}^L \le \hat{C}^C$ means that the left spread of the fuzzy cyclical component must be non-negative, i.e., $\hat{C}^C - \hat{C}^L \ge 0$.

$\hat{u}^C_2 - \hat{u}^L_2 \ge \xi > 0$ means that the left spread of the final fuzzy residuals (after removing both trend and cyclical component) must be strictly positive. The reason for this is that the final fuzzy residuals will be further decomposed as a multivariate auto-regressive process (VAR). At this step, we recommend for ξ a value of about 0.1.

Step 3: Third, we estimate the cyclical component corresponding to the maximum daily temperature, as a solution of a constrained quadratic program:

$$\min_{b^R_{cycle}} \left\{ \left(b^R_{cycle}\right)' C' C\, b^R_{cycle} - 2\left(\hat{u}^R_1\right)' C\, b^R_{cycle} \right\} . \tag{21}$$

subject to:

$$\begin{cases} \hat{C}^R \ge \hat{C}^C \\ \hat{u}^R_2 - \hat{u}^C_2 \ge \xi > 0 \end{cases} \Leftrightarrow \begin{cases} -C \cdot b^R_{cycle} \le -\hat{C}^C \\ C \cdot b^R_{cycle} \le -\hat{u}^C_1 + \hat{C}^C + \hat{u}^R_1 - \xi \end{cases} . \tag{22}$$

$\hat{C}^R \geq \hat{C}^C$ means that the right spread of the fuzzy cyclical component must be non-negative, i.e., $\hat{C}^R - \hat{C}^C \geq 0$

$\hat{u}_2^R - \hat{u}_2^C \geq \xi > 0$ means that the right spread of the final fuzzy residuals (after removing both trend and cyclical component) must be strictly positive.

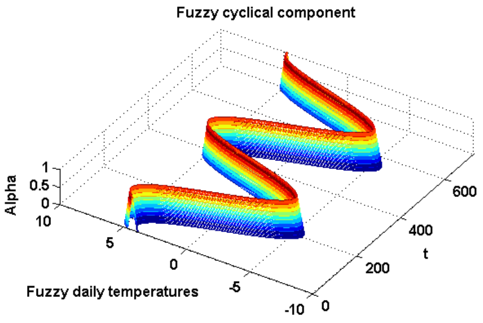


Fig. 11. Fuzzy cyclical component.

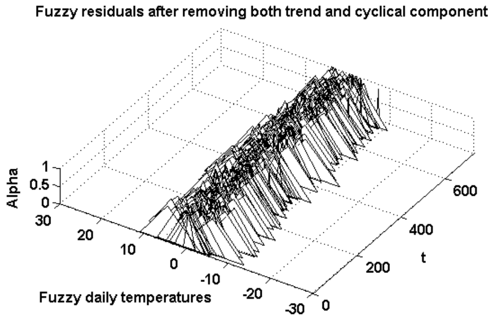


Fig. 12. Fuzzy residuals after removing both trend and cyclical component (the residuals are fuzzy sets in a proper sense – with non-negative spreads).

5.5. Modeling and Forecasting the Fuzzy Residuals as a VAR(4) Process, after Removing both Trend and Cyclical Component

The fuzzy residuals obtained after removing both trend and cyclical component can now be modeled as a multivariate auto-regressive process. A VAR(4) model has been chosen (among some other candidate models) based upon likelihood ratio tests and Akaike Information Criterion. This allows forecasting or simulating the residuals, starting from a sequence of the latest 10% observed historical temperatures.

Afterwards, based on the inversion property of the generalized Hukuhara difference, we can forecast the series of fuzzy daily temperatures by recomposing them from its components: trend component + cyclical component + the simulated residuals.

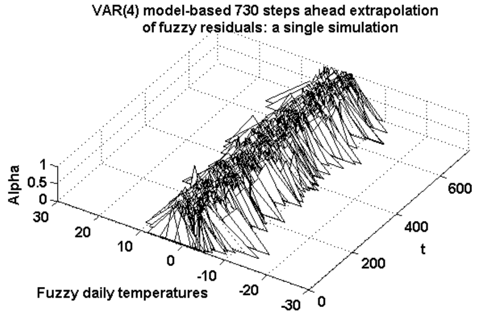


Fig. 13. VAR(4) model-based 730 steps ahead extrapolation of fuzzy residuals: a single simulation (the residuals are fuzzy sets in a proper sense – with non-negative spreads).

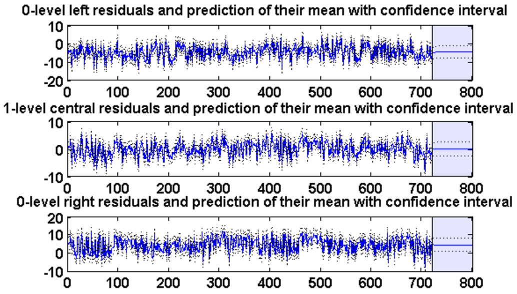


Fig. 14. Fuzzy residuals (left, central and right) and prediction of their mean with confidence interval ($\pm\sigma$).

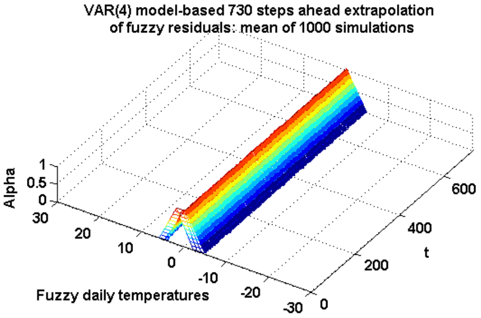


Fig. 15. VAR(4) model-based 730 steps ahead extrapolation of fuzzy residuals: mean of 1000 simulations.

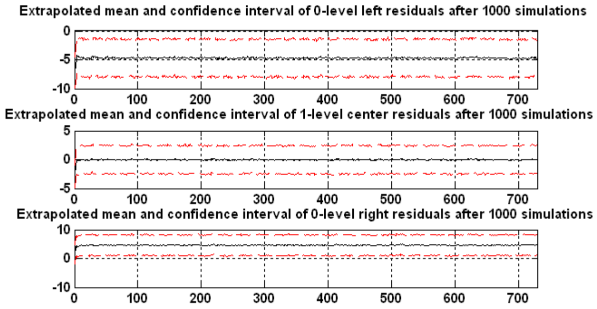


Fig. 16. Extrapolated mean and confidence interval of 0-level left / 1-level center / 0-level right fuzzy residuals, after 1000 simulations.

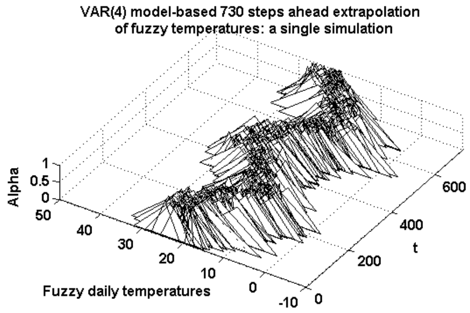


Fig. 17. VAR(4) model-based 730 steps ahead extrapolation of fuzzy temperatures, additively recomposed from trend, cyclical component and simulated fuzzy residuals.

5.6. Fuzzy Time Series Wavelet Decomposition and Nonlinear Model Fitting with Wavelet Networks

Another alternative to removing disturbances from a time series is de-noising data by wavelet decomposition.

The Discrete Wavelet Transform (DWT) [17] uses scaled and shifted versions of a mother wavelet function, usually with compact support, to form either an orthonormal basis (Haar wavelet, Daubechies) or a bi-orthonormal basis (Symlets, Coiflets). Wavelets allow cutting up data into different frequency components (called approximations and details), and then studying each component with a resolution matched to its scale. They can help de-noise inherently noisy data through wavelet shrinkage and thresholding methods, developed by David Donoho [5]. The idea is to set to zero all wavelet coefficients corresponding to details in the data set that are less than a particular threshold. These coefficients are used in an inverse wavelet transformation to reconstruct the data set. An important advantage is that the de-noising is carried out without smoothing out the sharp structures and thus can help to increase the predictive performance.

We start with the de-trended fuzzy time series shown in figure 13. A level 5 decomposition with Sym8 wavelets and a fixed form soft thresholding is first performed (see figures 18 and 19).

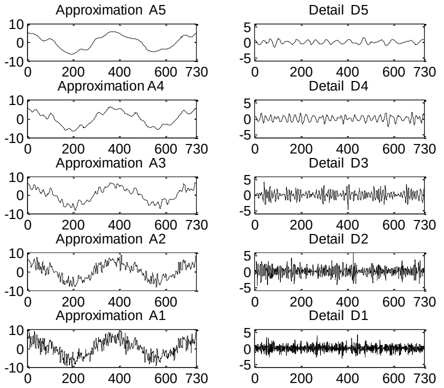


Fig. 18. A level 5 decomposition of the average temperature time series using Sym8 wavelets: approximations and details. The successive approximations appear less and less noisy; however, they also lose progressively more high-frequency information.

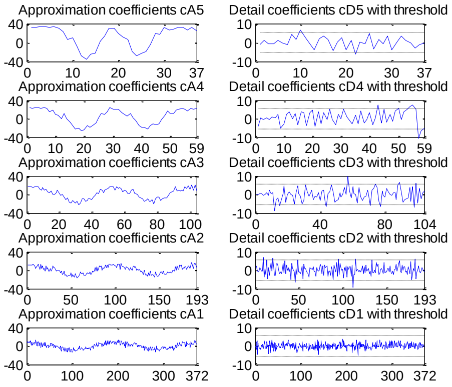


Fig. 19. Approximation coefficients and detail coefficients with a global threshold.

The initially de-trended average, minimum and maximum temperature time series are now de-noised in turn (see figures 20, 21 and 22).

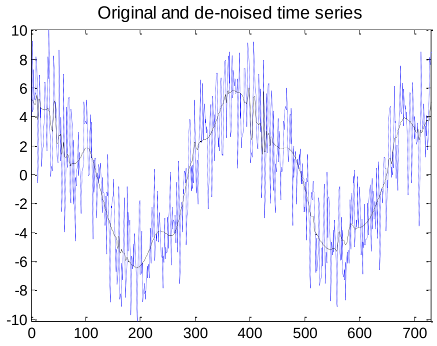


Fig. 20. De-trended vs. de-trended & de-noised average temperature time series.

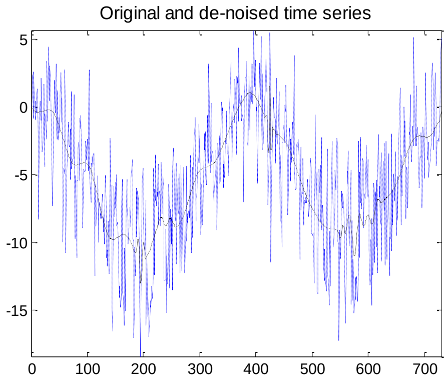


Fig. 21. De-trended vs. de-trended & de-noised minimum temperature time series.

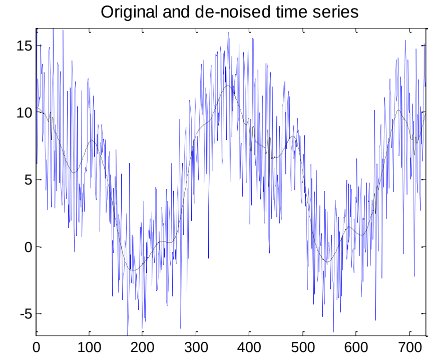


Fig. 22. De-trended vs. de-trended & de-noised maximum temperature time series.

However, the representations are not smooth enough, because of some weather turbulences that occur in certain time intervals. In order to produce smoothed representations, the time series obtained after de-noising can be further fitted to some nonlinear approximation functions by using wavelet networks to learn them.

Wavelet networks attempt to combine the properties of the Wavelet decomposition previously described, along with the learning capabilities of feedforward neural networks. They employ wavelets instead of sigmoidal activations functions, are trained with a backpropagation-like algorithm and behave as universal approximators, being capable of estimating almost any computable function on a compact set arbitrarily closely. Their rigorous mathematical foundations and better localization and approximation properties allow hierarchical and multi-resolution learning as well as transparent design of the network. Wavelet networks can be easily generalized to the case of multidimensional nonlinear function approximation in order to approximate functions in $L_2(\Re^n)$ and their representation can be extended with radial wavelets that are better suited for approximation problems of large dimensions. This results in the following network structure:

$$g(x)=\sum_{i=1}^{N} w_i\,\psi(diag(d_i)(x-t_i))+c'x+b \quad \textbf{(23)}$$

where Ψ is a radial wavelet function, $d_i \in \Re^n$ are dilatation parameters, $t_i \in \Re^n$ are translation parameters, $w_i \in \Re$ are linear weights, N is the number of wavelets, $c \in \Re^n$ is the additional direct linear combination parameters (direct connection parameters), and $b \in \Re$ is the bias parameter.

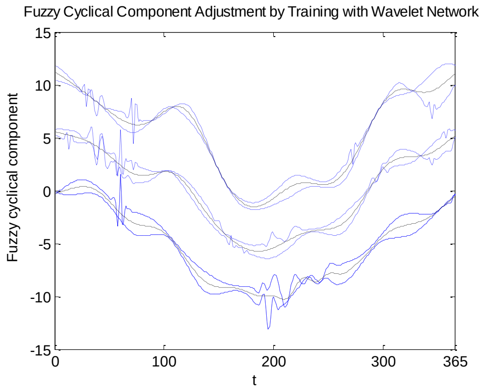


Fig. 23. 2D representation of the adjusted fuzzy cyclical component after training with wavelet network.

The input space is the set $T=\{1,2,\ldots,365\}$ of discrete time values. We train the wavelet network three times, for each set of minimum, average and maximum daily temperatures in turn. Each time, the output space is the set of de-trended and de-noised minimum, average and maximum daily temperatures, i.e., $S_{\min}$, S_{avg} and $S_{\max}$, respectively. 10 wavelets in the hidden layer and 10 iterations (epochs) are used for training. Finally, the adjusted fuzzy cyclical component is obtained by mapping T

onto $S_{\min}$, S_{avg} and $S_{\max}$, i.e. $\hat{S}_{\min}(t):T \to S_{\min}$, $\hat{S}_{avg}(t):T \to S_{avg}$ and $\hat{S}_{\max}(t):T \to S_{\max}$, respectively. 2D and 3D representations of the adjusted fuzzy cyclical component after training with wavelet network, starting from the de-trended and de-noised data, are shown in figures 23 and 24.

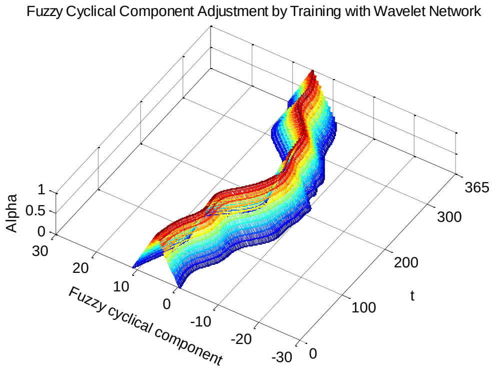


Fig. 24. 3D representation of the adjusted fuzzy cyclical component after training with wavelet network.

Finally, the fuzzy trend + fuzzy cyclical component can be re-compounded (figure 25).

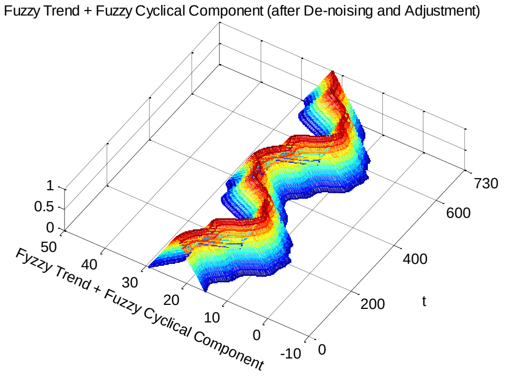


Fig. 25. Fuzzy trend + fuzzy cyclical component (after de-noising and adjustment).

6. Conclusion

Two kinds of uncertainty were contrasted in this paper (ontological, vs. epistemic uncertainty) and several techniques were addressed in order to capture and propagate both of them jointly across a specific model. The first application area is concerned with extending the classical *probabilistic risk assessment* modeling framework in such a way that allows probability distribution parameters to be imprecisely defined. The presence of imprecision in a risk model adds another dimension to that of randomness and requires more comprehensive approaches to capture and represent the range within which the risk distribution might vary. Analytic methods for propagating the uncertainty (such as Probability bounds analysis, Fuzzy α-levels analysis) as well as stochastic simulation techniques (such as Monte Carlo Analysis), can be combined or integrated, in an attempt to find the output of a model that has both random and fuzzy variables for the inputs.

The second application area is addressed when attempting to capture the inherent fuzzy and random nature of some stochastic processes, expressed in terms of fuzzy time series. Suitable new methods for fuzzy time series estimation and prediction, using both the estimation theory and Computational Intelligence techniques were proposed.

We combined a generalized Hukuhara difference, which allows the fuzzy estimation problem to be handled in some L_2-type metric space, with a partial decoupling principle, which allows the monolithic fuzzy model to be broken in several more tractable crisp estimation sub-problems. This approach was proved to provide an efficient solution to the problem of non-invertibility of the standard Minkovsky addition and multiplication in a fuzzy feature space, while enabling to obtain fuzzy estimations in a proper sense (i.e., with non-negative spreads).

Alternatively, wavelet decomposition, a Computational Intelligence based technique, has been also used to de-noising fuzzy time series. Finally, starting from the de-trended and de-noised time series, wavelet networks have been employed as universal approximators to adjust the fuzzy cyclical component and thus to produce smoothed representations of the fuzzy time series components

References

1. Baudrit C., Guyonnet D., Dubois D.: Postprocessing the Hybrid Method for Addressing Uncertainty in Risk Assessments. Journal of Environmental Engineering 131(12), 1750-1754. (2005)
2. Diamond P.: Fuzzy least squares. Inform. Sci. 46, 141-157. (1988)
3. Diamond P., Kloeden P.: Metric Spaces of Fuzzy Sets: Theory and Applications. World Scientific, Singapore, (1994)
4. Diamond P., Körner R.: Extended fuzzy linear models and least squares estimates. Comput. Math. Appl. 33, 15-32. (1997)
5. Donoho D.: Nonlinear Wavelet Methods for Recovery of Signals, Densities, and Spectra from Indirect and Noisy Data. In Daubechies, I. (eds). Different Perspectives on Wavelets Amer. Math. Soc., Providence, R.I., 173-205. (1993)
6. EPA's guidance: Risk Assessment Guidance for Superfund (RAGS). Volume III - Part A: Process for Conducting Probabilistic Risk Assessment, (2001). [Online]. Available: www.epa.gov/oswer/riskassessment/rags3adt/index.htm
7. Ferson S., Tucker W.T.: Probability Bounds Analysis in Environmental Risk Assessments (Technical report). Applied Biomathematics, Setauket, New York (2003). [Online]. Available: www.ramas.com/pbawhite.pdf.

8. Ferson S., Ginzburg L., Kreinovich V., Nguyen H.T., Starks S.A.: Uncertainty in risk analysis: towards a general second-order approach combining interval, probabilistic, and fuzzy techniques. In Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, 1342-1347. (2002)
9. Georgescu V.: Estimation of fuzzy regression models using quadratic programming. Economic Computation and Economic Cybernetics Studies and Research, 4, 105-124. (1997)
10. Georgescu V.: New estimation methods in fuzzy regression analysis, based on projection theorem and decoupling principle. Fuzzy Economic Review, III/1, 21-38. (1998)
11. Georgescu V.: On the Foundations of Granular Computing Paradigm, Fuzzy Economic Review, VIII/2, 73-105. (2003)
12. Georgescu V.: A Generalization of Symbolic Data Analysis Allowing the Processing of Fuzzy Granules, Lecture Notes in Artificial Intelligence, 3131, Springer, 215-226. (2004)
13. Georgescu V.: Fuzzy Time Series Estimation and Prediction: Criticism, Suitable New Methods and Experimental Evidence. Studies in Informatics and Control, Volume 19, Issue 3. (2010)
14. Guyonnet D., Bourgigne B., Dubois D., Fargier H., Côme B., Chilès J.: Hybrid approach for addressing uncertainty in risk assessments. Journal of environmental engineering 129(1), 68-78. (2003)
15. Hukuhara M.: Integration des applications measurables dont la valeur est un compact convexe, Funkcialaj Ekvacioj 10, 205–223. (1967)
16. Körner R.., Näther W.: Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates, Inform. Sci. 109, 95–118. (1998)
17. Mallat S. G., Peyré G.: A Wavelet Tour of Signal Processing: The Sparse Way, Academic Press, 3rd Edition (2009)
18. Möller B., Beer M.: Fuzzy Randomness – Uncertainty Computational Mechanics, Springer, Berlin and New York (2004)
19. Möller B., Reuter U.: Uncertainty Forecasting in Engineering, Springer-Verlag, Berlin Heidelberg. (2007)
20. Puri M.L., Ralescu D.A., Fuzzy random variables, J. Math. Anal. Appl. 114, 409–422. (1986)
21. Stefanini L.: A generalization of Hukuhara difference and division for interval and fuzzy arithmetic, Fuzzy Sets and Systems, Vol. 161, Issue 11, 1564-1584. (2010)
22. Terán P.: Probabilistic foundations for measurement modeling with fuzzy random variables, Fuzzy Sets and Systems 158(9), 973-986. (2007)

Vasile Georgescu is currently a professor of econometrics at University of Craiova, Romania and director of the Research Center for Computational Intelligence in Business and Economics. He has published more than 150 papers in fields such as computational intelligence, computational statistics, econometrics, cybernetics, pattern recognition and more than 20 books.

Received: December 15, 2012; Accepted: December 09, 2013

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

004

COMPUTER Science and Information Systems : the International journal / Editor-in-Chief Mirjana Ivanović. – Vol. 11, No 2 (2014) - . – Novi Sad (Trg D. Obradovića 3): ComSIS Consortium, 2014 - (Belgrade : Sigra star). –30 cm

Polugodišnje. – Tekst na engleskom jeziku

ISSN 1820-0214 = Computer Science and Information Systems
COBISS.SR-ID 112261644

Cover design: V. Štavljanin
Printed by: Sigra star, Belgrade

ComSIS Computer Science and Information Systems

Volume 11, Number 2, June 2014

Contents

Regular papers

Special Section: Contemporary ICT Research Trends Emerging from South-East Europe

Special Section: Computational Intelligence in Business Administration

ISSN: 1820-0214

Computer Science and Information Systems

Vol 11, No 2, June 2014