

Darina Dicheva  
Zdravko Markov  
Eliza Stefanova (Eds.)

**Third International Conference  
on Software, Services and  
Semantic Technologies  
S3T 2011**

 Springer



# Advances in Intelligent and Soft Computing

## Editor-in-Chief

Prof. Janusz Kacprzyk  
Systems Research Institute  
Polish Academy of Sciences  
ul. Newelska 6  
01-447 Warsaw  
Poland  
E-mail: [kacprzyk@ibspan.waw.pl](mailto:kacprzyk@ibspan.waw.pl)

---

Further volumes of this series can be found on our homepage: [springer.com](http://springer.com)

Vol. 89. J.B. Pérez, J.M. Corchado, M.N. Moreno, V. Julián, P. Mathieu, J. Canada-Bago, A. Ortega, and A.F. Caballero (Eds.)  
*Highlights in Practical Applications of Agents and Multiagent Systems, 2011*  
ISBN 978-3-642-19916-5

Vol. 90. J.M. Corchado, J.B. Pérez, K. Hallenborg, P. Golinska, and R. Corchuelo (Eds.)  
*Trends in Practical Applications of Agents and Multiagent Systems, 2011*  
ISBN 978-3-642-19930-1

Vol. 91. A. Abraham, J.M. Corchado, S.R. González, J.F. de Paz Santana (Eds.)  
*International Symposium on Distributed Computing and Artificial Intelligence, 2011*  
ISBN 978-3-642-19933-2

Vol. 92. P. Novais, D. Preuveneers, and J.M. Corchado (Eds.)  
*Ambient Intelligence - Software and Applications, 2011*  
ISBN 978-3-642-19936-3

Vol. 93. M.P. Rocha, J.M. Corchado, F. Fernández-Riverola, and A. Valencia (Eds.)  
*5th International Conference on Practical Applications of Computational Biology & Bioinformatics 6-8th, 2011*  
ISBN 978-3-642-19913-4

Vol. 94. J.M. Molina, J.R. Casar Corredera, M.F. Cátedra Pérez, J. Ortega-García, and A.M. Bernardos Barbolla (Eds.)  
*User-Centric Technologies and Applications, 2011*  
ISBN 978-3-642-19907-3

Vol. 95. Robert Burduk, Marek Kurzyński, Michał Woźniak, and Andrzej Żołnierek (Eds.)  
*Computer Recognition Systems 4, 2011*  
ISBN 978-3-642-20319-0

Vol. 96. A. Gaspar-Cunha, R. Takahashi, G. Schaefer, and L. Costa (Eds.)  
*Soft Computing in Industrial Applications, 2011*  
ISBN 978-3-642-20504-0

Vol. 97. W. Zamojski, J. Kacprzyk, J. Mazurkiewicz, J. Sugier, and T. Walkowiak (Eds.)  
*Dependable Computer Systems, 2011*  
ISBN 978-3-642-21392-2

Vol. 98. Z.S. Hippe, J.L. Kulikowski, and T. Mroczek (Eds.)  
*Human – Computer Systems Interaction: Backgrounds and Applications 2, 2011*  
ISBN 978-3-642-23186-5

Vol. 99. Z.S. Hippe, J.L. Kulikowski, and T. Mroczek (Eds.)  
*Human – Computer Systems Interaction: Backgrounds and Applications 2, 2011*  
ISBN 978-3-642-23171-1

Vol. 100. Shoumei Li, Xia Wang, Yoshiaki Okazaki, Jun Kawabe, Toshiaki Murofushi, and Li Guan (Eds.)  
*Nonlinear Mathematics for Uncertainty and its Applications, 2011*  
ISBN 978-3-642-22832-2

Vol. 101. Darina Dicheva, Zdravko Markov, and Eliza Stefanova (Eds.)  
*Third International Conference on Software, Services and Semantic Technologies S3T 2011, 2011*  
ISBN 978-3-642-23162-9

Darina Dicheva, Zdravko Markov,  
and Eliza Stefanova (Eds.)

---

Third International Conference  
on Software, Services and  
Semantic Technologies  
S3T 2011

## Editors

Dr. Darina Dicheva  
Winston-Salem State University  
Department of Computer Science  
3206 E.J. Jones Computer Science Bldg  
601 S. Martin Luther King Jr. Drive  
Winston Salem, NC 27110  
USA  
<http://myweb.wssu.edu/dichevad/>  
E-mail: [dichevad@wssu.edu](mailto:dichevad@wssu.edu)

Eliza Stefanova  
Sofia University St. Kl. Ohridski  
Faculty of Mathematics and Informatics  
5, James Bourchier Blvd.  
1164 Sofia  
Bulgaria  
E-mail: [eliza@fmi.uni-sofia.bg](mailto:eliza@fmi.uni-sofia.bg)

Dr. Zdravko Markov  
Central Connecticut State University  
1615 Stanley Street  
New Britain, CT 06050  
USA  
<http://www.cs.ccsu.edu/~markov/>  
E-mail: [MarkovZ@mail.ccsu.edu](mailto:MarkovZ@mail.ccsu.edu)

ISBN 978-3-642-23162-9

e-ISBN 978-3-642-23163-6

DOI 10.1007/978-3-642-23163-6

Advances in Intelligent and Soft Computing

ISSN 1867-5662

Library of Congress Control Number: 2011934145

© 2011 Springer-Verlag Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

*Typeset & Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India

Printed on acid-free paper

5 4 3 2 1 0

[springer.com](http://springer.com)

# Preface

This volume contains the Proceedings of The Third International Conference on Software, Services & Semantic Technologies (S3T) held in Bourgas, Bulgaria on September 1–3, 2011. It is the third S3T conference in a series of annually organized events supported by the F7 EU SISTER Project and hosted by Sofia University.

The conference is aimed at providing a forum for researchers and practitioners to discuss the latest developments in the area of Software, Services and Intelligent Content and Semantics. The special focus of this forum is on Intelligent Content and Semantics, and Technology Enhanced Learning. Particular emphasis is placed on applying intelligent semantic technologies in educational and professional environments. In order to emphasize the multidisciplinary nature of S3T and still keep it focused, the conference topics have been organized in four tracks. The conference sessions and the contents of this volume are also structured according to the track themes:

- Intelligent Content and Semantics
- Knowledge Management, Business Intelligence and Innovation
- Software and Services
- Technology Enhanced Learning

The S3T 2011 conference attracted a large number of submissions from many different countries. The papers, selected after a rigorous blind reviewing process, were organized in three categories: full papers, short papers and poster presentations. The papers published in this volume cover a wide range of topics related to the track themes and address a broad spectrum of issues within the announced conference topics and related areas of application.

The conference program is complemented by the presentations of four distinguished invited speakers: “What Computer Can Do When It Knows Learning/Instructional Theories” by Riichiro Mizoguchi from University of Osaka (Japan), “The Web and its Users: Engineering the Personal and Social Web” by Geert-Jan Houben from Delft University of Technology (The Netherlands), “Networked Learning in Learning Networks” by Peter B. Sloep from Open University (The Netherlands), and “Unite to Triumph and

Divide to Conquer: Intuitive, Iterative, and Modular Ontology Authoring” by Vania Dimitrova from University of Leeds (UK).

We would like to thank the many people who have helped to make this conference possible. First and foremost, we would like to express our sincere appreciation to the S3T Track Chairs - Ivan Koychev (Intelligent Content and Semantics), Krassen Stefanov (Technology Enhanced Learning), Sylvia Ilieva (Software and Services), and Elisaveta Gurova (Knowledge Management, Business Intelligence and Innovation) for their extraordinary efforts and invaluable help in the conference organization, and to all members of the Program Committee and other reviewers for their dedication in the review process.

Many people helped us make the conference as convenient as possible for all participants. The local organizing committee recruited from the Faculty of Mathematics and Informatics at Sofia University has done an excellent job. We wish to mention Eugenia Kovatcheva, Victoria Damyanova, Marin Barzakov, Stanimira Yordanova, and Atanas Georgiev. They deserve full credit for their hard work.

Last, but not least, we would like to offer our special thanks to all the authors who contributed to this event.

We hope that you all enjoy the S3T conference and find it illuminating and stimulating.

September, 2011

Darina Dicheva  
Zdravko Markov  
Eliza Stefanova

# Organization

## Program Chairs

Darina Dicheva  
Zdravko Markov

Winston-Salem State University, USA  
Central Connecticut State University, USA

## Local Organization Chair

Eliza Stefanova

Sofia University, Bulgaria

## Track Chairs

Elisaveta Gurova (Knowledge Management, Business Intelligence and Innovation)

Sylvia Ilieva (Software and Services)

Ivan Koychev (Intelligent Content and Semantics)

Krassen Stefanov (Technology Enhanced Learning)

## Program Committee

Galia Angelova

Bulgarian Academy of Science, Bulgaria

Soren Auer

University of Leipzig, Germany

Nick Bassiliades

Aristotle University, Thessaloniki, Greece

Maria Bielikova

Slovak University of Technology, Slovakia

Ig Ibert Bittencourt

Federal University of Alagoas, Brazil

Boyan Bontchev

Sofia University, Bulgaria

Pavel Boytchev

Sofia University, Bulgaria

Svetla Boytcheva

State University of Library Studies and

Info Technologies, Bulgaria

Federica Cena

University of Turin, Italy



Antonina Dattolo	University of Udine, Italy
Mariana Damova	Ontotext, Bulgaria
Ralph Deters	University of Saskatchewan, Canada
Christo Dichev	Winston-Salem State University, USA
Vania Dimitrova	University of Leeds, UK
Milena Dobreva	University of Strathclyde, UK
Radu Dobrin	Malardalen University, Sweden
Danail Dochev	Bulgarian Academy of Science, Bulgaria
Gabriella Dodero	Free University of Bozen-Bolzano, Italy
Avram Eskenazi	Bulgarian Academy of Sciences, Bulgaria
Albert Esterline	NC A&T State University, USA
Ricardo Jardim-Goncalves	UNINOVA Institute, Portugal
Panos Hahamis	University of Westminster, UK
Geert-Jan Houben	Delft University of Technology, The Netherlands
Elva Jones	Winston-Salem State University, USA
Jelena Jovanovic	University of Belgrade, Serbia
Franz Lehner	Universitat Passau, Germany
Lutz Maicher	University of Leipzig, Germany
Riichiro Mizoguchi	University of Osaka, Japan
Preslav Nakov	National University of Singapore
Roumen Nikolov	State University of Library Studies and Info Technologies, Bulgaria
Maria Nisheva	Sofia University, Bulgaria
Alexandre Passant	DERI, National University of Ireland, Ireland
Atanas Radenski	Chapman University, USA
Allan Ramsay	University of Manchester, UK
Demetrios Sampson	Center for Research and Technology, Greece
Olga C. Santos	National University for Distance Education, Spain
Giovanni Semeraro	Universita' di Bari, Italy
Kiril Simov	Bulgarian Academy of Sciences, Bulgaria
Sergey Sosnovsky	DFKI, Germany
Peter Stanchev	Kettering University, USA
Slavi Stoyanov	Open University, The Netherlands
Kamelia Stefanova	University of National and World Economy, Bulgaria
Bedir Tekinerdogan	Bilkent University, Turkey
Stefan Trausan-Matu	"Politehnica" University of Bucharest, Romania
Tania Tudorache	Stanford University, USA
Julita Vassileva	University of Saskatchewan, Canada
Denny Vrandecic	AIFB, Germany
Vladimir Zanev	Columbus State University, USA

# Contents

## Intelligent Content and Semantics

<b>News Article Classification Based on a Vector Representation Including Words' Collocations</b> . . . . .	1
<i>Michal Kompan, Mária Bielíková</i>	
<b>Classification of Online Reviews by Computational Semantic Lexicons</b> . . . . .	9
<i>Boris Kraychev, Ivan Koychev</i>	
<b>Reason-Able View of Linked Data for Cultural Heritage</b> . . . . .	17
<i>Mariana Damova, Dana Dannells</i>	
<b>Semantic Retrieval Approach to Factoid Question Answering for Bulgarian</b> . . . . .	25
<i>Stanislav Peshterliev, Ivan Koychev</i>	
<b>Using Semantic Relations for Representing Long-Term User Interests</b> . . . . .	33
<i>Muhammad Ali Zeb, Maria Fasli</i>	
<b>Integrating Internet Protocol Television and Social Networking</b> . . . . .	41
<i>Afshana S. Hussain, Julita Vassileva</i>	
<b>Clustering and Visualization of ECG Signals</b> . . . . .	47
<i>Anikó Vágner, László Farkas, István Juhász</i>	
<b>Visualizing and Presenting Story Maps in TM4Book</b> . . . . .	53
<i>Lars Johnsen, Darina Dicheva, Christo Dichev</i>	
<b>A Fuzzy Bat Clustering Method for Ergonomic Screening of Office Workplaces</b> . . . . .	59
<i>Koffka Khan, Alexander Nikov, Ashok Sahai</i>	

**The Influence of Global Constraints on DTW and LCS Similarity Measures for Time-Series Databases** ..... 67  
*Vladimir Kurbalija, Miloš Radovanović, Zoltan Geler, Mirjana Ivanović*

**Knowledge Management, Business Intelligence and Innovation**

**Social Network Analysis in Professional e-Recruitment** ..... 75  
*Kostadinka Toteva, Elissaveta Gourova*

**A System Integrating Agent Services with JMS for Shared Battlefield Situation Awareness** ..... 81  
*Albert Esterline, William Wright, Srinivas Banda*

**A Model for Fuzzy Logic Assessment of Real Estate Investment Risks** ..... 89  
*Plamena Zlateva, Dimiter Veleu, George Zabunov*

**Towards the Foundation for Read-Write Governance of Civilizations** ..... 95  
*Alois Paulin*

**Software and Services**

**An Analysis and Forecast of Software and Services Research in Bulgaria – Comparing the Situation Two Years Later** ..... 103  
*Irena Pavlova, Desislava Petrova-Antonova, Avram Eskenazi, Sylvia Ilieva, Aleksandar Dimov*

**A Fuzzy System Model for Task Implementation in Extreme Programming Process** ..... 111  
*Ensar Gul*

**Integration of Natural Language Processing Chains in Content Management Systems** ..... 119  
*Diman Karagiozov*

**Supporting Interactive IPTV Apps with an Enterprise Resource Bus** ..... 127  
*Sunny Sharma, Ralph Deters*

**Service Oriented Architecture Applied to Provide Inter-operability between Macedonian MoD and NATO** ..... 133  
*Zoran Gacovski, Goko Vancov*

<b>Empirical Analysis of Software Reliability Models Predictive Ability</b> .....	139
<i>Aleksandar Dimov</i>	

## Technology Enhanced Learning

<b>Living Labs in E-Learning, E-Learning in Living Labs and Living Labs for E-Learning</b> .....	147
<i>Albena Antonova, Korneliya Todorova</i>	

<b>An Ontology-Based Model for Driving the Building of Software Product Lines in an ITS Context</b> .....	155
<i>Alan Pedro da Silva, Ig Ibert Bittencourt, Walker Ataide, Olavo Holanda, Evandro Costa, Thyago Tenório, Patrick H.S. Brito</i>	

<b>A Proposal for Storysculpturing</b> .....	161
<i>Pavel Boytchev</i>	

<b>Ad-Hoc Radial Visualization of SCORM Objects with Extended Search and Keywords Handling</b> .....	167
<i>Ivan Angelov, Ivan Ganchev</i>	

<b>Applied Cognitive Task Analysis in the Context of Serious Games Development</b> .....	175
<i>Albena Antonova, Krassen Stefanov</i>	

<b>Using Robotics to Achieve Meaningfully Engaged Learning</b> .....	183
<i>Rebecca Caldwell, Elva Jones</i>	

<b>Using a Fuzzy Emotion Model in Computer Assisted Speech Therapy</b> .....	189
<i>Ovidiu A. Schipor, Stefan G. Pentiu, Maria D. Schipor</i>	

<b>Towards Better Discoverability and Use of Open Content</b> . . . .	195
<i>Christo Dichev, Bandana Bhattacharai, Curan Clonch, Darina Dicheva</i>	

<b>Teachers' Training Design Model and Decision Support System</b> .....	205
<i>Eliza Stefanova, Svetla Boytcheva</i>	

## Posters

<b>A Framework for Scaling Up Distributed Minds</b> .....	211
<i>Oisín Mac Fhearaí, Mark Humphrys, Ray Walshe</i>	

<b>Ontology-Based Electronic Test Results Evaluation</b> . . . . .	213
<i>Veselina Jecheva, Daniela Orozova</i>	
<b>Self-Adaptive Software Systems through Exploratory Changes</b> . . . . .	215
<i>Stavros Stavru, Sylvia Ilieva</i>	
<b>Semantic Annotation in SINUS Project</b> . . . . .	217
<i>Kamenka Staykova, Gennady Agre, Kiril Simov, Petya Osenova</i>	
<b>A Multi-agent Approach to Development of E-Learning Facilities</b> . . . . .	219
<i>Valentina Terzieva, Petia Kademova-Katzarova, Rumen Andreev</i>	
<b>Author Index</b> . . . . .	221

# News Article Classification Based on a Vector Representation Including Words' Collocations

Michal Kompan and Mária Bieliková

Slovak University of Technology, Faculty of Informatics and Information Technologies  
Ilkovičova 3, 842 16 Bratislava, Slovakia  
e-mail: {kompan,bielik}@fiit.stuba.sk

**Abstract.** In this paper we present a proposal including collocations into the pre-processing of the text mining, which we use for the fast news article recommendation and experiments based on real data from the biggest Slovak newspaper. The news article section can be predicted based on several article's characteristics as article name, content, keywords etc. We provided experiments aimed at comparison of several approaches and algorithms including expressive vector representation, with considering most popular words collocations obtained from Slovak National Corpus.

**Keywords:** text pre-processing, news recommendation, news classification, vector representation.

## 1 Introduction and Related Work

Nowadays no one is discussing the need for the web personalization. One of the ways in which personalization is performed represents recommendations. Recommendation task can be defined as follows:

$$\forall c \in C, s'_c = \operatorname{argmax}_{s \in S} u(c, s)$$

where  $C$  represent users,  $S$  represents objects of recommendation and  $u$  is the usefulness function (usefulness of an object for specific user).

Recommender systems had become important part of well-known web portals in several domains as online shops, libraries or news portals for years. News portals are characteristic with thousands of daily added articles with high information decrease degree, so the one from most relevant recommender systems' attributes are reaction time of processing and the start of recommending new articles.

There are two widely used approaches to the recommendation: collaborative filtering based on an assumption that interest is shared between similar users and the content based recommendation where computed "associations" between entities (generally similarity relation) based on extracted useful information on the entity content are used. We focus on the content based recommendation in news domain.

Several recommender systems in the news domain (OTS, NewsMe, Pure, Google News, NewsBrief [1]) have been proposed in the last decade. The collaborative recommender on SME.SK [10] and the content based recommenders TRecom [12] and Titrer [7] have been proposed for the Slovak language.

## 1.1 Classification Task

Content based recommendation obviously includes recommended entities classification aimed at finding relations between entities. Considering recommending news articles, we are often limited to a specific source of articles – news portal. In such a portal often every article has its own category (mostly assigned by a human), which is rich and reliable source of important data (in the context of recommendation and similarity search). Nowadays researchers focus on aggregating recommender systems like NewsBrief or Google News, where several news portals over the world are monitored and used for generating recommendations. One of the possible solutions for the aggregation news from several portals is the classification based on articles categories from various portals. Several classifiers for news articles have been proposed respectively [8].

Main goal of the document classification is to assign one or more (probability) categories of the classified document. In the literature the classification task is often divided into supervised and unsupervised methods [4], where an unsupervised method refers to the document clustering. Unsupervised methods are based on an assumption that documents having similar content should be grouped into the one cluster. Hierarchical clustering algorithms have been intensively studied as well as neural network models. Several methods are used as Support vector machines, Naïve Bayes or Neural networks [11], where Naïve Bayes outperforms the others [2].

## 1.2 Text Representation for Classification

Because of the high information value decrease and the high dynamics in the news domain, there is need to process objects of the recommendation (articles) in a fast manner. For this purpose several text representations have been proposed [5].

The simplest method for representation of an article is the Bag of Words (BoW). As the best unit for the text representation is a term [5], BoW consists of term from the text. Other often used method of text representation is Vector Space Model, which adds weights (term frequency) to terms from BoW. In this way we obtain vectors representing text documents. It is clear that these representations have a huge problem with dimensionality and thus with the performance of any information retrieval method applied on. Various enhancements have been proposed as binary representation, ontology models or N-Grams [5].

Some methods do not consider all terms extracted from texts, but only the relevant. For the keywords or relevant term extraction Latent Semantic Indexing is often used. When extracting terms from semi-structured documents (such as HTML) additional information is used as HTML tags [5] for the relevant terms recognition.

## 2 News Article Representation Proposal

In the domain of news recommendation, the time complexity of the classification process is critical. To reduce the space of words and to extract relevant information from articles often a vector representation of text is used. This brings

usual the words' space reduction and accuracy improvements (if information extraction is included). We have proposed a vector representation (Table 1) based on the extraction of important (distinctive) terms from the article, in order to reduce the dimension of the space of words.

**Table 1** Vector representation for an article.

Vector part	Weights
Title	<i>transplantation_0.5</i> <i>face_0.5</i>
TF of title words in the content	<i>transplantation_0.0178571428571429</i> <i>face_0.0714285714285714</i>
Category	<i>Sme.sk_0.5</i> <i>PRESS_FOTO_1.0</i>
Keywords	<i>clinc_0.0357142857142857</i> <i>surface_0.0178571428571429</i> <i>nose_0.0178571428571429</i> <i>tooth_0.0178571428571429</i> <i>nerve_0.0178571428571429</i> <i>masculature_0.0178571428571429</i> <i>patient_0.0178571428571429</i> <i>scale_0.0178571428571429</i>
Names/Places	<i>Cleveland_1</i>
CLI	<i>0.2543</i>

The article vector consists of six parts:

- *Title* - Article vector comprises lemmatized words from article title. It consists of approximately 5 words (150 000 Slovak article dataset). We suggest that article title should be in most occurrences good describing attribute.
- *Term Frequency of title words in the content* - We used TF to estimate the article name confidence. If the article name is abstract and do not correspond to article content, we can easily discover this situation.
- *Keywords* – We store 10 most relevant keywords. News portals have a list of keywords for every article usually. These are unfortunately at different abstraction level over various portals thus we have our own keywords list, which is based on TF-IDF list calculated over the dataset (100 000 Slovak news articles SME.SK).
- *Category* - This category part is constructed based on the portal specific category hierarchy (optional), while the hierarchy is represented by a tree structure.
- *Names/Places* - In this step we extract list of names and places obtained from the article content - as words starting with upper letter and with no full-stop before (precision = 0.934, recall = 0.863).
- *CLI* - Coleman-Liau readability index provides information about the level of the understandability of the text.



Most of the vector parts do not depend on a particular news portal, and can be easily extracted from a standard article. The only one dependent part in our representation is the Category. If we want to abstract of this, it is necessary to find similar articles over various portals and then respectively create corresponding virtual categories (the text classification task).

## 2.1 Text Pre-processing and Collocations

The text pre-processing plays critical role in the text classification process. It can significantly reduce space of words, but on the other hand, it can easily decrease the information value. In our experiments we work with texts in Slovak language.

The article pre-processing can be divided into several steps [9]:

- *Tokenization* – A simple strategy is to just split the text on all non-alphanumeric characters. As far as this step is language depending, some information (special addresses, names etc.) can be lost. Thus, advanced techniques are need for text tokenization, considering local habits.
- *Dropping common terms: stop-words* – For every language we can easily identify most common words without any or only with small information value (and, is, be, in etc.) By removing these words we are able to significantly reduce the words' space while in the most cases the information value of processed texts remains.
- *Normalization* – In other words the process of creating equivalence classes of terms. The goal is to map words with the same sense to the one class (e.g. "USA" and "U.S.A.").
- *Stemming and lemmatization* – Documents contain different words' forms and there are families of related words with similar meanings (car, cars, car's, cars' - car). For the English language the most common stemmer is Porter Stemmer, for flexive languages such as Slovak language it is more complicated.

As a result of the pre-processing step (in connection to the needs of our vector representation) we obtain:

- lemmatized article title (without stop words and punctuation),
- 10 most relevant keywords, the list of Names and Places.

The most frequent words occurred in specific language together are considered as collocations (bigrams). In order to improve the pre-processing step and to increase the information gain, we introduce words collocations into the text pre-processing step. We expect that enhancing the pre-processing step with words collocations will lead to the article similarity or classification tasks improvement.

We extracted word collocations from the Slovak national corpus (Ľ. Štúr Institute of Linguistic, Slovak Academy of Sciences). The example of collocations for word "conference" in Slovak are "central", "OSN", "focused". The most frequent collocations in general are stop words or punctuations. We do not consider such words ("in the", "does not" etc.).

In other words we enhanced pre-processing step while not only stop words, but collocations are removed. This leads to word space reduction. Our hypothesis is that

after removing the collocations the information value of the pre-processed text remains the same – the information gain (words with distinctive characteristics) will remain and classification task accuracy will not decrease.

### 3 Hypothesis and Design of Experiments

Our hypothesis is that introducing words' collocations (removing collocations in the pre-processing step) can improve classification task results over the dataset. Thus we suggest several experiments, with various initial settings and classification algorithms.

For the classification task experiments we use SME.SK dataset from the project SMEFIIT [3]. We have total of 1 387 articles from 20 categories (extracted directly from news portal) in our dataset. Each article consists of the title, the article content and the real section in which was assigned by the article author. For each article we constructed representative article vector as described in Section 2. For the implementation and experiments we used RapidMiner [6] as one of the well-known and widely used information discovery environment.

First, we investigated which one from weighting techniques performs best. For this purpose we modelled a standard classification task with Naïve Bayes, K-NN and Decision trees as a classificatory. Weights for words were step by step calculated as TF-IDF, Term frequencies, Term occurrences and Binary term occurrences. We also used a pruning method where all words with the weight below 3.0 or above 30.0 percent were pruned. The pruning has negative impact on the whole process computation complexity. However, this can be compensated by the proposed vector representation.

In the second experiment we investigated which of the classifiers perform best for the classification task. We considered K-nearest neighbour, Naïve Bayes and Decision trees and their implementations in RapidMiner. For these methods we also evaluated the best weight function and these best-performers were used in next experiments.

Our aim was to evaluate properties of proposed article representation for classification task. So we performed all the experiments for both classical (TF-IDF, Term Frequency, Term Occurrences, Binary Term Occurrences) and our proposed representations. We do not use whole vector representation - the Category part was excluded (it is used as a learner for supervised learning).

Similarly, all experiments were performed for standard pre-processing as we described in the section 2 (without collocation remove/add). As the next step we added words' collocations to both representations and we studied performance changes. In the next experiment collocation were removed instead of added. Because most frequent collocations are stop words or words with a small information value, we decided to pre-process these words' collocations too and not to include most frequent collocations.

### 4 Results of Experimental Evaluation and Discussion

For each experiment we measured the classification accuracy (ration between correctly and incorrectly classified articles) for pruned and not pruned data

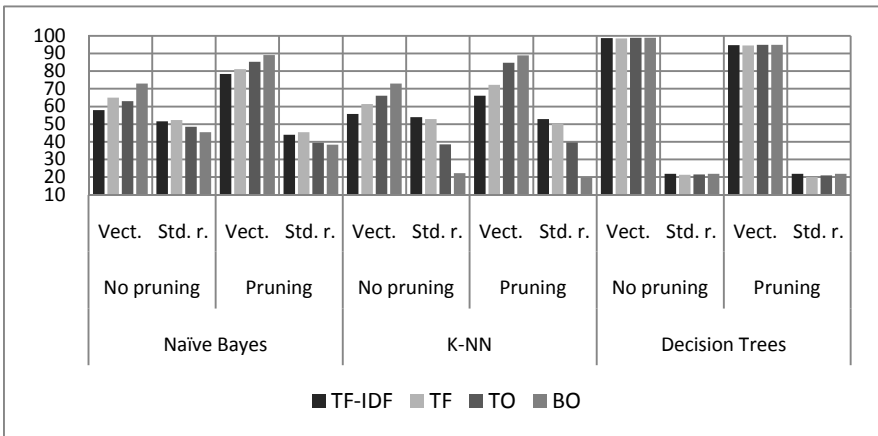
respectively. The evaluation was performed as “X-validation” ( $x=10$ ) with stratified sampling. In Table 2 we present the weight functions’ comparison for three classifiers.

As we can see in almost all cases classification performs the best when pruning method was active. Only in the classification with Naïve Bayes and with the standard representation no pruning outperformed the pruned data. The difference between pruned and not pruned data is significant. It is important to note, that while not pruning do not brings better results as pruning in general it also takes almost 10x longer as a classification with pruning.

**Table 2** Naïve Bayes, K-NN and Decision Trees classifiers comparison for various weight functions (TF-IDF, Term Frequency, Term Occurrences, Binary Term Occurrences) considering vector (Vec.) and standard (Std.) representations (Classification accuracy).

	Naïve Bayes				K-NN				Decision Trees			
	No pruning		Pruning		No pruning		Pruning		No pruning		Pruning	
	Vec.	Std.	Vec.	Std.	Vec.	Std.	Vec.	Std.	Vec.	Std.	Vec.	Std.
TF-IDF	57,9	51,6	78,3	44,0	55,6	54,0	66,0	52,7	98,7	21,9	94,6	21,9
TF	64,9	52,2	81,1	45,5	61,2	52,8	72,1	50,2	98,4	21,4	94,5	20,1
TO	63,0	48,6	85,2	39,5	66,1	38,5	84,6	39,6	98,8	21,5	94,8	21,0
BO	72,8	45,3	89,0	38,3	72,9	22,2	88,9	19,8	98,7	21,9	94,8	21,9

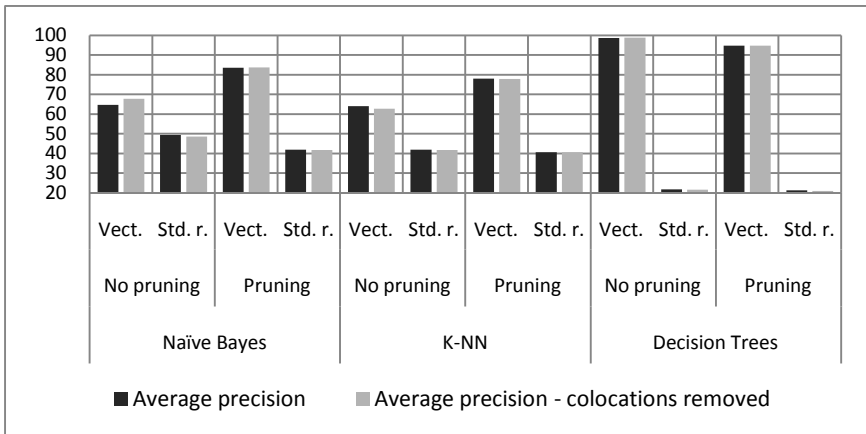
Our proposed vector representation significantly outperforms standard article representation in the classification accuracy (Fig. 1). We can say that our proposed representation extracts relevant information (words with distinctive characteristics) and can be used for various information retrieval tasks not only for the similarity computation.



**Fig. 1** Classifiers comparison (Classification accuracy) with various pre-processing methods (TF-IDF, Term Frequency, Term Occurrences, Binary Term Occurrences).

The highest accuracy increase, that we can observe, is using Decision Trees, which seems to be the best classifier for our task (Average improvement 75,22%). On the other hand classification with Decision trees takes the longest time even in case of vector representation and pruning included. Results when using Decision tree are “flat” in comparison to other approaches. This can be explained by the used approach, when the computed weights were not used.

We provided experiments for every possible combination of a weight function, the classifier, the representation and also considering words’ collocations. Because of similar patterns we do not provide complete results for the collocations excluded. Aggregated results (mean of 4 weight functions) for the classification with collocations consideration can be seen in Fig. 2.



**Fig. 2** Classifiers comparison with collocation consideration (average classification accuracy for 4 weight functions).

Our hypothesis appears to be wrong, i.e. excluding collocations in the pre-processing step does not significantly improve classification task. However, we can see, that removing collocations did not degrade the classification accuracy while it reduces words’ space, in other words, correlated words were removed. For the similarity computation task it can be interesting to experiment with word collocations adding. In this case the sub-group without stop words and low information gain words should be carefully selected.

## 5 Conclusion

In this work we compared several classification methods applied to news articles considering proposed vector representation. The article classification allows us to abstract from concrete news portal and to start recommending and aggregating articles from various portals.

We enhanced pre-processing process by introducing words’ collocations excluding. The proposed vector representation outperforms standard representation not only in the way of the classification accuracy (the best improvement 77,27%) but

it reduces the computation complexity of the classification process which is strictly connected to the computation time. Such a representation and category classification also are language independent. When we will replace collocations statistics and stop words list, we are able to use our proposed method for other languages.

Introducing collocations to the process of pre-processing does not bring improvement of pre-processing. On the other hand, collocations reduce the word space while do not decrease the information value at the same time.

Proposed vector representation can be used for content-based news recommendation and also to aggregate news articles from various news portals (using category classification) in a fast and effective way. As there are only few language depending steps during the pre-processing process, various languages can be included respectively.

**Acknowledgements.** This work was supported by the grants VEGA1/0508/09, VEGA1/0675/11, APVV-0208-10 and it is a partial result of the Research & Development Operational Program for the project Support of Center of Excellence for Smart Technologies, Systems and Services II, ITMS 25240120029, co-funded by ERDF.

## References

1. Billsus, D., Pazzani, M.J.: Adaptive news access. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) *Adaptive Web 2007*. LNCS, vol. 4321, pp. 550–570. Springer, Heidelberg (2007)
2. Burgin, R.: The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *Journal of the American Soc. for Inf. Science* 46(8), 562–572 (1995)
3. Barla, M., Kompan, M., Suchal, J., Vojtek, P., Zeleník, D., Bieliková, M.: News recommendation. In: *Proc. of the 9th Znalosti, Jindrichuv Hradec*, pp. 171–174 (2010)
4. Göker, A., Davies, J.: *Inform. Ret.: searching in the 21st cent.* John Wiley & Sons, Chichester (2009)
5. Harish, B.S., Guru, D.S., Manjunath, S.: Representation and Classification of Text Documents: A Brief Review. *IJCA, Special Issue on RTIPPR* (2), pp. 110–119 (2010)
6. Jungermann, F.: Information Extraction with RapidMiner. In: *Proc. of the GSCL Symposium Sprachtechnologie und eHumanities 2009, Duisburg* (2009)
7. Kompan, M., Bieliková, M.: Content-Based News Recommendation. In: *Proc. of the 11th Conf. EC-WEB*, pp. 61–72. Springer, Berlin (2010)
8. Kroha, P., Baeza-Yates, R.A.: A Case Study: News Classification Based on Term Frequency. In: *Proceedings of DEXA Workshops*, pp. 428–432 (2005)
9. Manning, D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)
10. Suchal, J., Návrat, P.: Full text search engine as scalable k-nearest neighbor recommendation system. In: *Bramer, M. (ed.) IFIP AI 2010. IFIP Advances in Information and Communication Technology*, vol. 331, pp. 165–173. Springer, Heidelberg (2010)
11. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann Ser. in Data Manag. Sys. Morgan Kaufmann, San Francisco (2005)
12. Zeleník, D., Bieliková, M.: News Recommending Based on Text Similarity and User Behaviour. In: *WEBIST 2011: Proc. of the 7th Int. Conf. On Web Informations Systems and Technologies, Noordwijkerhout, Holland*, pp. 302–307 (2011)

# Classification of Online Reviews by Computational Semantic Lexicons

Boris Kraychev<sup>1</sup> and Ivan Koychev<sup>1,2</sup>

<sup>1</sup>Faculty of Mathematics and Informatics, University of Sofia, Sofia, Bulgaria

<sup>2</sup>Institute of Mathematics and Informatics at Bulgarian Academy of Sciences

email: {kraychev, koychev}@fmi.uni-sofia.bg

**Abstract.** The paper presents a method for opinion polarity classification of online reviews, which includes a web crawler, part of speech tagger, constructor of lexicons of sentiment aware words, sentiment scoring algorithm and training of opinion classifier. The method is tested on 500 000 online reviews of restaurants and hotels, which are relatively simple and short texts that are also tagged by their authors and limited to a set of topics like service, food quality, ambience, etc. The results from conducted experiment shows that the presented method achieves an accuracy of up to 88%, which is comparable with the best results reported from similar approaches.

**Keywords:** Opinion Polarity Classification, Sentiment Analysis, Natural Language Processing.

## 1 Introduction

The task of large scale sentiment analysis draws increasing research interest in recent years. With the rise of the social networks and different types of web media like forums, blogs, video sharing, it became very important to develop methods and tools that are able to process the information flow and automatically analyse opinions and sentiment from online texts and reviews. Such analysis has various applications in the business and government intelligence and the online public relationships.

The paper presents a method that builds semantic lexicons for online review polarity classification. It includes building a sentiment aware dictionary, morphological approaches for feature extraction, label sequential rules, opinion orientation identification by scoring and linear regression algorithms. The method was implemented and tested with 500 000 recent online user reviews about restaurants and hotels.

The domain of restaurant and hotel reviews suggests the usage of feature oriented analysis because customers are discussing few aspects like food, service, location, price, and general ambience. Our goal is to estimate the sentiment polarity using multiple approaches. We built two independent lexicons: the first consisting of sentiment aware parts of speech and the second one representing evaluation pairs of adjectives and nouns extracted from the reviews. The second lexicon actually represents a set of extracted features from the online reviews.

We implemented the above method and conducted experiments with 500 000 recent online user reviews about restaurants and hotels. As a result our sentiment classifier achieves an 88% of accuracy, which can be considered as very good result, given that the raw online data contains spam reviews and human errors in the self assessment (the number of ‘stars’ assigned by the author to the review).

## 2 Related Work

Recently, the area of automated sentiment analysis has been very actively studied. Two major streams of research can be distinguished: The first relates to the building of sentiment aware lexicons and the second group consists of the work on complete sentiment analysis systems for documents and texts.

The early works in this field has been initiated by psychological researches in the second half of twentieth century (Deese, 1964; Berlin and Kay, 1969; Levinson, 1983) which postulated that words can be classified along semantic axes like “big-small”, “hot-cold”, “nice-unpleasant”, etc. This enabled the building of sentiment aware lexicons with explicitly labelled affect values.

The recent work on this subject involves the usage of statistical corpus analysis (Hatzivassiloglou and McKeown, 1997) which expands manually built lexicons by determining the sentiment orientation of adjectives by analyzing their appearance in combination with adjectives from the existing lexicon. Usually adjectives related with “and” like the clause “*The place is awesome and clean*” suppose that both adjectives have the same orientation, while the conjunction with “but” supposes that the adjectives have opposite orientation.

Other recent research is made by Grefenstette, Shanahan, Evans and Qu [4] [7] with exploration of the number of findings by search engines where an adjective, supposed to enter the lexicon is being examined towards a set of other well determined adjectives over several semantic axes. The authors consider that adjectives would appear more frequently closer to their synonyms and their sentiment orientation can be determined statistically by the number of search engine hits where the examined word appears close to any of the seed words.

The movie reviews have been a subject of research for Pang, Lee and Vathyanathan [8] and Yang Liu [2]. The first system achieves an accuracy of roughly 83% and shows that machine learning techniques perform better than simple counting techniques. The second system implements linear regression approaches, (an interesting introduction in that area is presented by C. Bishop[1]) and combines the box office revenues from previous days, together with the people’s sentiments about the movie to predict the sales performance of the current day. The best results of the algorithm achieve an accuracy of 88%.

Some of the authors as Pang [8] try to separate the text on factual and opinion propositions, while other as Godbole [6] considers that both mentioned facts and opinions contribute to the sentiment polarity of a text.

Other approach for product reviews is the feature-based sentiment analysis explored by B. Liu, Hu and Cheng [9] which extracts sentiment on different features of the subject. The techniques used are Label Sequential Rules (LSR) and

Pointwise Mutual Information (PMI) score, introduced by Tourney [10]. General review of the sentiment analysis methods is made by Pang and Lee [3] in 2008.

A recent approach is proposed by Hassan and Radev [13] in 2010 which determines the sentiment polarity of words by applying Markov random walk model to a large word relatedness graph where some of the words are used as seeds and labelled with their sentiment polarity. To determine the polarity of a word the authors generate Markov random chains, supposing that walks started from negative words would hit first a word labelled as negative. The algorithm has excellent performance and does not require large corpus.

Our approach for the current experiment is to use scoring algorithms, enhanced by sequential rules in order to improve the sentiment extraction for the different estimation axes for restaurants and perform the polarity classification by standard machine learning algorithms, based on numerical attributes, issued from the scoring process.

### 3 Sentiment Lexicon Generation and Sentiment Analysis

We apply two algorithms which, to our knowledge, have not been explored until now. The first one is the expansion of the dictionary through WordNet by keeping the sentiment awareness and positivity value by applying a histogram filter from the learning set of text. The second is the discovery of propositional patterns, determined as label sequential rules using relatively large test set of online reviews (250 000).

The major processing steps of our sentiment analysis system are:

1. *Construction of lexicons of sentiment aware words.* Actually all major sentiment analysis systems rely on a list of sentiment aware words to build initial sentiment interpretation data. We developed the following dictionaries of sentiment aware words and pairs of words.
  - (a) *Lexicon of sentiment aware adjectives and verbs* - a manually built list of seed words, expanded with databases of synonyms and antonyms to a final list of sentiment aware words.
  - (b) *Lexicon of sentiment aware adjective-noun pairs.* It is obtained with feature extraction techniques using propositional models and Label Sequential Rules (LSR) introduced by [9]. LSR discover sequential patterns of parts of speech. They are very effective extracting the sentiment for specific features, mentioned in the review.
2. *Sentiment scoring algorithms.* We are using scoring techniques to calculate a list of attributes per review. The aim is to build numerical depiction of the sentiment attributes of the text, taking care of negation, conditionality and basic pronoun resolution. The reviews represented in this attribute space are passed to the machine learning module.
3. *Opinion polarity classification.* We trained Machine learning algorithms based on attributes provided by the scoring algorithm then we evaluated the performance of the learned classifiers on new reviews.



### 3.1 Determining Lexicon Seeds and Lexicon Expansion through WordNet

We sorted the parts of speech from the training set to find out the most frequently used ones. Then we manually classified adjectives and verbs as seeds for future classification expansion. This forms our seeds for future lexicon development.

We used WordNet to expand the dictionary with synonyms and antonyms. It is well known that WordNet offers a very large set of synonyms and there are paths that connect even good and bad as synonyms, so we limited the expansion to two levels and applied a percentage to decrease the confidence weight of words found by that method.

Significance weight for lexicon expansion through WordNet is calculated with a method proposed by Godbole [6]. The significance weight of a word is equal to  $w = 1/c^d$ , where  $c$  is a constant  $> 1$  and  $d$  is the distance from the considered to the original word. The expansion is planned in two stages – the first stage is to simply enlarge the dictionary by the 1<sup>st</sup> and 2<sup>nd</sup> level synonyms of words, then as a second stage – apply a filter on the resulting words to eliminate words ending in contradictory positivity assessment. This can happen by building a histogram for each word over the sentiment tagged reviews from the learning set. We exclude the words having different histogram than their corresponding seeds. The final polarity weight is calculated as follows: for a given term we can mark with  $p$  the appearances in positive texts, with  $n$  the appearances in negative texts and with  $P$ ,  $N$  and  $U$  the total number of positive, negative and neutral texts, respectively. The polarity weight is then calculated by the equation

$$polarity\_weight = \frac{p - n}{P + N + U} w.$$

Unknown words which are not mentioned in the learning set are kept with the weight of their first ancestor with calculated weight, multiplied by a coefficient between 0 and 1 following the formula above. In our case the value chosen was 0.8 e.g.  $c = 1.25$  and words without clear evidence in the learning set were kept with decreased weight by 20%.

### 3.2 Lexicon Generation with Label Sequential Rules

The label sequential rules [9] provide a method for feature extraction and discovery of common expression patterns. Our targeted area of short online reviews suggests that people would follow similar expression models. The label sequential rules are mapping sequences of parts of speech and are generated in the following form:

```
{ $feature, noun } { (be), verb } { $quality, adjective }
[ { and, conjunction } { $quality, adjective } ] => 90%
{ $actor, pronoun } { *, verb }
[ { *, determiner } ] { $feature, noun } => 90%
```

where the square brackets indicate that the part is non mandatory and each rule has a confidence weight to be considered further. The conjunctions ‘and’ and ‘but’ in the phrases were used to enlarge the lexicon with adjectives having similar or

opposite sentiment orientation. It is important to note that the LSR method allows splitting the analysis to features and further summarize and group the reviews by features.

The construction of LSR patterns is important part of the learning algorithm. By sorting all N-term part-of-speech sequences, the ones which frequency is over a pre-defined threshold are kept and added to LSR knowledge base, declaring the nouns as features and the adjectives and verbs as sentiment positivity evaluators.

### 3.3 Methods for Sentiment Analysis

Our sentiment analysis algorithm is based on sentiment aware term scoring which is then evaluated by machine learning algorithms.

The scoring algorithm determines sentiment aware terms in text and assigns their sentiment weight in the dictionary of sentiment aware words. The weight values are real numbers, positive or negative according to the determined sentiment orientation. The algorithm takes into account negation like “not, don’t, can’t” and inverses the relative weight value. It also takes care of simple conditional propositions like *‘if the staff was polite, I would...’* and applies a simple technique for pronoun resolution. For our results we rely on the fact that short online reviews are kept simple and the lack of profound conditionality and pronoun resolution analysis would not impact our final results. We have to admit that these modules could be improved further.

The final result of the scoring algorithm is a set of weight sums, counts and expression of previously estimated values that would facilitate further machine learning classification.

With this set of attributes, we obtained a regular problem for machine learning which we explored in our experiments.

## 4 The Sentiment Analysis Experiment

### 4.1 Design

Our experiment involves the following steps:

1. Web crawling to collect online reviews and their self assessment by their authors.
2. Part of speech analysis to all acquired texts using MorphAdorner [11].
3. Sorting the data from the test set to determine the seed words and LSR patterns for the generation of the lexicons.
4. Generation of the lexicons by expansion through WordNet [5] and LSR extraction [12].
5. Numerical representation of the texts by scoring sentiment aware words.
6. Experiments with machine learning algorithms over the attributes’ space.

The goal of the experiment is first to extract live data from the web, then analyze the contents and extract seed words and patterns for lexicon generation.

The final sentiment analysis consists of calculating numerical attributes like sum of weighted positive/negative items, count of contradiction related words and mathematical expressions using previously calculated parameters. The expressions are actually forming the scores that can be assessed. The sentiment polarity classification is then performed in the environment for machine learning benchmarking WEKA.

## 4.2 Determining the Positive and Negative Weights of the Text

The sum of the weights of positive and negative items in the text forms the first two classification attributes:  $PosW$  and  $NegW$  respectively. We obtain these sums by the scoring algorithm which identifies the sentiment aware words and phrases from both lexicons. It also counts the negations, conditionality and pronoun resolution, and procedure the  $Contr$  attribute. For example if the word is preceded by negation like ‘not’, ‘don’t’, ‘can’t’ the polarity of the item is exchanged. For example ‘not good’ goes to the sum of negative words instead of the one for positive, with its default weight. The Table 1 describes the final list of attributes.

**Table 1** The list of attributes passed to the machine learning algorithm.

Attribute	Description	Implementation
$PosW$	$\Sigma$ of the weights of positive items	Scoring algorithm
$NegW$	$\Sigma$ of the weights of negative items	Scoring algorithm
$Contr$	Count of contradiction elements	Scoring algorithm
Score1	$f(posw, negw)$	{posw} + {negw}
score2	$f(posw, negw)$	{posw} + 2* {negw}
score3	$f(posw, negw)$	2* {posw} + {negw}
score4	$f(posw, negw, contr)$	{posw} + {negw} - {contr}

## 4.3 Results of Sentiment Polarity Classification with WEKA

In order to be able to experiment with more machine learning algorithms we added supplementary attributes, formed by the original three ones. The most evident one is a simple addition of the positive weight and the negative weight (they have indeed opposite signs) which forms a simple score of positive minus negative items in the text. We also experimented with doubling the value of negative or positive items to handle the fact that reviewers might tend to give more strength on one of these groups.

The classification through three machine learning algorithms gives the results shown in. The accuracy of 87-88% is satisfying our expectation because our raw review data contains classification errors. The estimation of the classification errors should be explored further and requires voluminous manual data revision.

**Table 2** Results by different machine learning algorithms

<b>Algorithm</b>	<b>Accuracy</b>	<b>Precision</b>
NaiveBayes	87%	87%
VotedPerceptron	83%	69%
ADTree	88%	87%

## 5 Discussion: Thumbs Up or Thumbs Down for Restaurants

The sentiment classification tasks vary for different domains. In the current experiment we showed that sentiment analysis algorithms can perform better when it is restricted to particular domain, where it is easier to perform feature extraction algorithm. Interesting results can be obtained by examining the expressed sentiment over all scanned reviews of UK restaurants by features as food, staff, ambiance, etc.

We should note that restaurants are a very competitive domain and reviewers are attentive to all details. The feature that annoys most of the clients is the non-politeness of the staff. Next to it stands the quality of the food and the price comes as the third most bothering feature.

If we count the general customer sentiment about all evaluated restaurants we should conclude ‘Thumbs up’ because the bigger part of expressed reviews and features are positive.

## 6 Conclusion

In the present work we built method for online review classification, which was tested on a large data set of UK restaurant reviews. The approach constructs a lexicon of sentiment aware words and phrases over the application domain. Then it estimates the sentiment polarity by applying scoring techniques over the reviews and providing the results to machine learning algorithms. The final classification is made using machine learning algorithms from the WEKA environment.

The results are showing a clear path to follow – topic related sentiment analysis is a prominent area where automatic sentiment classification can be considered as effective and robust monitoring tool. Future researches could include demographic and geographic data to show peoples’ preferences and provide deeper analysis.

Future work might include improvement of the scoring algorithm – better pronoun resolution, improvement in the detection of conditional propositions. The generation of the lexicon of sentiment aware words could be improved in the area of feature extraction by implementing more sequential rules and detecting more part-of-speech patterns. Last but not least the lexicon building algorithm could be applied on different topic areas like sentiment analysis of reviews of movies, books, news stories, and certainty identification in text.

**Acknowledgements.** This research is supported by the SmartBook project, subsidized by the Bulgarian National Science Fund, under Grant D002-111/15.12.2008.

## References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
2. Liu, Y.: Review Mining from Online Media: Opinion Analysis and Review Helpfulness Prediction for Business Intelligence. VDM Verlag (2010)
3. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now Publishers Inc. (2008)
4. Shanahan, J.G., Qu, Y., Wiebe, J. (eds.): Computing Attitude and Affect in Text: Theory and Applications. Springer, Heidelberg (2006)
5. Witten, I.H., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam (2005)
6. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale Sentiment Analysis for News and Blogs. In: Int. Conf. on Weblogs and Social Media ICWSM (2007)
7. Grefenslette, G., Qu, Y., Evans, D.A., Shanahan, J.G.: Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. Springer, Heidelberg (2006)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods of Natural Language Processing, EMNLP (2002)
9. Liu, B., Hu, M., Cheng, J.: Opinion Observer: Analyzing and comparing opinions on the web. In: Proceedings of WWW (2005)
10. Tourney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Association for Computational Linguistics, ACL (2002)
11. MorphAdorner Part of Speech Tagger,  
<http://morphadorner.northwestern.edu/morphadorner/postagger/>
12. Miller, G.A.: WordNet: A lexical database. Communications of the ACM 38(11) (1995)
13. Hassan, A., Radev, D.: Identifying Text Polarity Using Random Walks. In: Proceedings of the Association for Computational Linguistics

# Reason-Able View of Linked Data for Cultural Heritage

Mariana Damova<sup>1</sup> and Dana Dannells<sup>2</sup>

<sup>1</sup> Ontotext, Tsarigradsko Chausse 135, Sofia 1784, Bulgaria

<sup>2</sup> University of Gothenburg, Lennart Torstenssonsgatan 8, 405 30 Gothenburg, Sweden  
email: mariana.damova@ontotext.com,  
dana.dannells@svenska.gu.se

**Abstract.** This paper presents a novel approach that relies on the innovative idea of Reason-able View of the Web of linked data applied to the domain of cultural heritage. We describe an application of data integration based on Semantic Web technologies and the methods necessary to create an integrated semantic knowledge base composed of real museum data that are interlinked with data from the Linked Open Data (LOD) cloud. Thus, creating an infrastructure to allow for easy extension of the domain specific data, and convenient querying of multiple datasets. Our approach is based on a model of schema level and an instance level alignment. The models use several ontologies, e.g. PROTON and CIDOC-CRM, showing their integration by using real data from the Gothenburg City Museum.

**Keywords:** linked open data, reason-able view, cultural heritage, museum, ontology, data integration, Semantic Web.

## 1 Introduction

Being able to obtain useful information from Linked Open Data (LOD) [12], i.e. combining knowledge and facts from different datasets is the ultimate goal of the Semantic Web. Although clear, the vision of LOD and the Semantic Web is still looking for convincing real life use cases demonstrating the benefits of these technologies. MacManus in [13] defines one exemplar test for the Semantic Web. He formulates a conceptual query about cities around the world which have “Modigliani artwork”, and states that the vision of the Semantic Web will be realized when an engine will return an answer to it. Actually, the answer to this question can be found in the LOD; where different facts about the artist, his artwork and the museums or galleries that host them are to be found in different datasets. To our knowledge FactForge [4], a public service provided by Ontotext, is the only engine capable of passing this test (cf. Fig. 1).<sup>1</sup> FactForge is based on the method of *Reason-able Views* of the web of data [9], [10].

---

<sup>1</sup> The SPARQL query to obtain this information can be run at <http://factforge.net/sparql>.

SPARQL Query

Results for PREFIX: <http://rdf.berkeley.edu/> (172)

View as Table | Download in CSV | SPARQL Results in XML | SPARQL Results in JSON

painting_f	owner_f	city_f_con	city_4b_loc	city_4b_cit
Head@gen	Museum of Modern Art	Manhattan		
Head@gen	Museum of Modern Art	New York City		
Anna Zorovska@gen	Museum of Modern Art	Manhattan		
Anna Zorovska@gen	Museum of Modern Art	New York City		
Portrait of Diego Rivera@gen	The São Paulo Museum of Art@gen		São Paulo@gen	
Portrait of Diego Rivera@gen	The São Paulo Museum of Art@gen		São Paulo	
Woman with a Necklace@gen	School of the Art Institute of Chicago@gen			Chicago
Portrait of a Woman@gen	School of the Art Institute of Chicago@gen			Chicago
Reclining Nude@gen	Museum of Modern Art	Manhattan		
Reclining Nude@gen	Museum of Modern Art	New York City		
Madam Pompadour@gen	School of the Art Institute of Chicago@gen			Chicago
Jeanne Hébert@gen	Barnes Foundation@gen	Philadelphia		

Fig. 1 Results of the Modigliani test

The “Modigliani artwork” example gives evidence for the potential of the cultural heritage domain to become a useful use case for the application of the semantic technologies. Our work is a step in this direction showing a Reason-able View of the web of data integrating museum and LOD cloud data. In this paper we present a Reason-able View of the web of data, using real museum data that are integrated with data from the LOD cloud.

## 2 Linked Open Data - The Vision

The notion of “linked data” is defined by Tim Berners-Lee, [1] as RDF [14] graphs, published on the WWW and explorable across servers in a manner similar to the way the HTML web is navigated. Linked Open Data (LOD) is a W3C SWEO community project aiming to extend the Web by publishing open datasets as RDF and by creating RDF links between data items from different data sources, cf. Fig. 2.

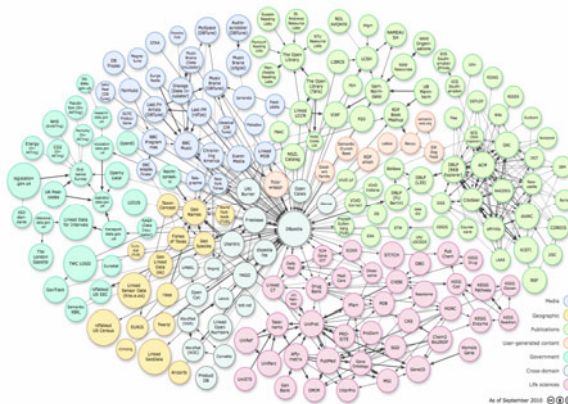


Fig. 2 The LOD Cloud

### 3 Linked Open Data Management

Using linked data for data management is considered to have great potential in view of the transformation of the web of data into a giant global graph [3]. Kiryakov et al., [10] present a *Reason-able View (RAV)* approach for reasoning with and managing linked data. RAV is an assembly of independent datasets, which can be used as a single body of knowledge with respect to reasoning and query evaluation. It aims at lowering the cost and the risks of using specific linked datasets for specific purposes. The linkage between the data is made at the schema-level by mapping ontologies [3], [8], and at the instance level with the predicate `owl:sameAs`, i.e. the common method of connecting data in the LOD cloud. *Reason-able Views* are accessible via a SPARQL [15] end-point and keywords. Because each Reason-able View is a compound dataset, i.e. it consists of several datasets, and one can formulate queries, combining predicates from different datasets and ontologies in a single SPARQL query. The results from such queries return instances which also come from different datasets in the Reason-able View.

### 4 Museum Reason-Able View

The datasets in each Reason-able View depend on the underlying purpose of use of the compound dataset. In our case, the Museum Reason-able View has to be constructed in a way to provide adequate content for the two following requirements:

- the ability to handle generic knowledge, such as people, institutions, and locations
- the ability to handle specific subject domains, such as the cultural heritage and museums

The Museum Reason-able View, presented in this paper, comprises a heterogeneous dataset reflecting a combination of generic knowledge, and domain specific knowledge. It includes the following datasets from the LOD cloud:

- **DBpedia**<sup>2</sup> - the RDF-ized version of Wikipedia, describing more than 3.5 million things and covers 97 languages.
- **Geonames**<sup>3</sup> - a geographic database that covers 6 million of the most significant geographical features on Earth.
- **PROTON**<sup>4</sup> - an upper-level ontology, 542 entity classes and 183 properties.

These datasets cover the generic knowledge of the Museum Reason-able View.

The next sections introduce the Museum specific knowledge integrated into the Museum Reason-able View.

---

<sup>2</sup> DBpedia, structured information from Wikipedia: <http://dbpedia.org>.

<sup>3</sup> Geonames, a geographical database: <http://www.geonames.org>.

<sup>4</sup> PROTON, a lightweight upper-level ontology: <http://proton.semanticweb.org/>.



## 5 Museum Data Models

The CIDOC-CRM is an object oriented ontology developed by the International Council of Museum's Committee for Documentation (ICOM-CIDOC)<sup>5</sup>, with overall scope of curated knowledge of museums. The model provides the level of details and precision necessary for museum professionals to perform their work well. The CIDOC-CRM ontology consists of about 90 classes and 148 properties. It represents an upper-level ontology view for cultural and natural history. Its higher level concepts are general concepts, e.g. Entity, Temporal Entity, Time Span, Place, Dimension, and Persistent Item. Physical items and non-material products produced by humans are described as Man-made-thing, and Conceptual Object. The concept Event of CIDOC-CRM covers through its sub concepts the entire lifecycle of an artifact, e.g. Production, Creation, Dissolution, Acquisition, Curation, etc. Some of these concepts have more than one immediate superclass.

The integration of CIDOC-CRM into the Museum Reason-able View takes place at the schema level by providing mappings between the CIDOC-CRM concepts and PROTON concepts, cf. Fig. 4. The CIDOC-CRM concepts are linked to PROTON concepts with the built-in property `owl:equivalentClass`. Six classes from CIDOC-CRM and PROTON are being interlinked in this way.

K-samsök [11], the Swedish Open Cultural Heritage (SOCH), is a Web service for applications to retrieve data from cultural heritage institutions or associations with cultural heritage information. The idea behind K-samsök is to harvest any data format and structure that is used in the museum sector in Sweden and map it into K-samsök's categorization structure available in an RDF compatible form. It includes features which are divided in the following categories:

- (a) Identification of the item in the collection
- (b) Internet address, and thumbnail address
- (c) Description of the item
- (d) Description of the presentation of the item, including a thumbnail
- (e) Geographic location coordinates
- (f) Museum information about the item
- (g) Context, when was it created, to which style it belongs, etc.
- (h) Item specification, e.g. size, and type of the item – painting, sculpture and the like.

Fig. 3 presents a painting item from The History Museum in Sweden described according to this categories available at the following URL:

<http://mis.historiska.se/mis/sok/fid.asp?fid=96596&g=1>

---

<sup>5</sup> CIDOC CRM webpage: <http://www.cidoc-crm.org/>.

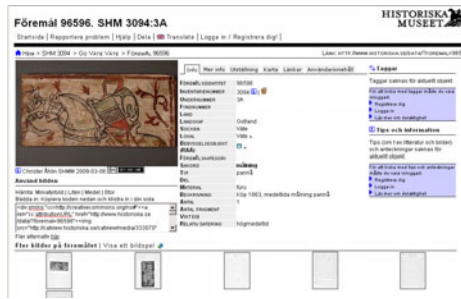


Fig. 3 Website of a painting item from the History Museum in Sweden

The CIDOC-CRM schema is not enough to cover all the information that K-samsök tends to capture. In order to provide the necessary infrastructure to load the complete information about a museum item, it is required to integrate the schema of K-samsök into the Museum Reason-able View. This is possible by defining a new intermediary layer described in a specific ontology, which we will call the Museum Artifacts Ontology (MAO).<sup>6</sup> The MAO ontology was developed for mapping between museum data and the K-samsök schema. The ontology includes concepts reflecting the K-samsök schema to allow integrating the data from the Swedish museums. It has about 10 concepts and about 20 new properties.

It is important to note, that this Museum Artifacts Ontology can be further specified with descriptions of additional concepts covering a specific type of museum artifacts, like for example paintings.

## 6 The Gothenburg Museum Data

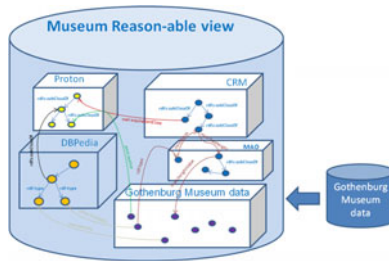
The Gothenburg museum [5] preserves 8900 museum objects described in its database. These objects correspond to two museum collections (GSM and GIM) and are placed in two tables of the museum database. 39 concept fields display each museum object, including its identification, its type - a painting, a sculpture, etc.-, its material, its measurements, its location, etc. All concept fields are described in Swedish.

The Gothenburg City Museum database structure follows the structure of the CIDOC-CRM, and the part of its data described above is used as experimental data for the Museum Reason-able View. The data is mapped to concepts from PROTON and MAO in the cases when the concepts available in the data are not available in CIDOC-CRM.

Fig. 4 shows the architecture of the integration of the Gothenburg City Museum data into the Museum Reason-able View by representing and linking them with elements from different schemata, e.g. PROTON, CIDOC-CRM and MAO. Additionally, the linkage with external to the Gothenburg City Museum data,

<sup>6</sup> It is just a coincidence that this ontology has the same as the Finish MAO [6], which also describes museum artifacts for the Finish museums.

e.g. DBpedia, is provided by connecting the MAO concepts to DBpedia instances, or by connecting the Gothenburg museum data with the corresponding DBpedia instances using the predicate owl:sameAs.



**Fig. 4** Dataset interconnectedness in the Museum Reason-able View

The process of the Gothenburg City Museum data integration into the Museum Reason-able View consists in transforming the information from the museum database into RDF triples based on the described ontologies. Each museum item is given an unique URI, and the concept fields from the database are interpreted as describing concepts or properties from one of the three ontologies, e.g. PROTON, CIDOC-CRM or MAO. The objects of the triples are derived from the columns of the database.

The triple generation goes through a process of localization, e.g. using English words for the naming of the properties and URIs in the Museum Reason-able View.

Loading the Gothenburg City Museum data into the Museum Reason-able View enables queries of the following nature:

- Museum artefacts preserved in the museum since 2005
- Paintings from the GSM collection
- Inventory numbers of the paintings from the GSM collection
- Location of the objects created by Anders Hafren
- Paintings with length less than 1 meter
- etc.

## 7 Museum Reason-Able View Environment

The Museum Reason-able View environment is built as an instance of BigOWLIM triple store. It provides the knowledge to query Gothenburg City Museum data in a structured way. It contains: DBpedia 3.6, Geonames, PROTON, CIDOC-CRM and MAO ontologies, and their mappings, and the triplified Gothenburg City Museum data. BigOWLIM performs full materialization during loading. It was expected that the available retrievable statements after loading will exceed the loaded explicit statements by about 20%. The loading statistics confirmed this expectation, e.g. the number of the loaded explicit statements was 257,774,678 triples, whereas the overall number of triples available for querying was 16% more, e.g. 305,313,536.

## 8 Related Work

Museum Data Integration with semantic technologies as proposed in this paper is intended to enable efficient sharing of museum and cultural heritage information. Initiatives about developing such sharing museum data infrastructures have increased in the recent years. Only few of them rely on semantic technologies. Similar project has been carried out for the Amsterdam Museum, developed by VUA.<sup>7</sup> This project aims at producing Linked Data within the Europeana<sup>8</sup> data model. To our knowledge ours is the first attempt of using CIDOC-CRM to produce museum linked data with connections to external sources from the LOD cloud like DBpedia and Geonames. Schema-level alignment is a new method of achieving interoperability in LOD [3], [8]. This method has not been applied on data in the cultural heritage domain, which we propose in this paper.

## 9 Conclusion

We presented the methods of using a knowledge representation infrastructure to build a knowledge base in the cultural heritage domain according to the described above innovative methods and models. The Museum Reason-able View provides an easy path to extension of the knowledge base with data from other Swedish museums or generally museum data, and allows to query and obtain results not only about artifacts belonging to different museum collections but also general knowledge about them from DBpedia and Geonames.

Our future work includes detailed experiments with the Museum Reason-able View regarding querying and navigation, extensions of the data models to cover detailed museum artifacts descriptions, like paintings, and using the interlinked ontologies as an interface for access to and presentation of the structured museum data in natural language.

**Acknowledgments.** This work is supported by MOLTO European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement FP7-ICT-247914.

## References

1. Berners-Lee, T.: Design Issues: Linked Data,  
<http://www.w3.org/DesignIssues/LinkedData.html>
2. Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., Velkov, R.: OWLIM: A family of scalable semantic repositories. Semantic Web Journal, Special Issue: Real-World Applications of OWL,  
<http://www.semantic-web-journal.net/content/new-submission-owlim-family-scalable-semantic-repositories>

---

<sup>7</sup> VUA: <http://www.vu.nl/en/>.

<sup>8</sup> Europeana portal: <http://www.europeana.eu/portal/>.

3. Damova, M., Kiryakov, A., Grinberg, M., Bergman, M., Giasson, F., Simov, K.: Creation and Integration of Reference Ontologies for Efficient LOD Management. In: Semi-Automatic Ontology Development: Processes and Resources. IGI Global, Hershey (to appear, 2011)
4. FactForge, a Reason-able View to the Web of Data, <http://factforge.net>, <http://www.ontotext.com/factforge>
5. Gothenburg City Museum, <http://www.stadsmuseum.goteborg.se/prod/kultur/stadsmuseum/dalis2.nsf/0/3955ae522e06a42bc12573950046ec10?OpenDocument>
6. Hyvonen, E., Maekelae, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: Museum Finland. Finnihs Museum on the Semantic Web (2008), <http://www.websemanticsjournal.org/index.php/ps/article/download/74/72>
7. International Council of Museums (ICOM), <http://icom.museum>
8. Jain P, Yeh P Z, Verma K, Vasquez R G, Damova M, Hitzler P, Sheth A P (2011). Contextual Ontology Alignment of LOD with an Upper Ontology: A Case Study with Proton. In G. Antoniou (Ed.), Proceedings of 8th Extended Semantic Web Conference. Heraklion, Crete.
9. Kiryakov, A., Momtchev, V.: Two Reason-able Views to the Web of Linked Data. Presentation at the Semantic Technology Conference 2009, San Jose (2009), <http://www.slideshare.net/ontotext/two-reasonable-views-to-the-web-of-linked-data>
10. Kiryakov, A., Ognyanoff, D., Velkov, R., Tashev, Z., Peikov, I.: LDSR: Materialized Reason-able View to the Web of Linked Data. In: Hoekstra, R., Patel-Schneider, P.F. (eds.) Proceedings of OWLED 2009: Chantilly, USA (2009)
11. K-samsök, <http://www.ksamsok.se/om-k-samsok/anslutna-institutioner/>
12. Linking Open Data, W3C Semantic Web Education and outreach community project, <http://linkeddata.org/>
13. MacManus, R.: The Modigliani Test: The Semantic Web's Tipping Point, [http://www.readwriteweb.com/archives/the\\_modigliani\\_test\\_semantic\\_web\\_tipping\\_point.php](http://www.readwriteweb.com/archives/the_modigliani_test_semantic_web_tipping_point.php) (retrieved April 2010)
14. RDF Primer, <http://www.w3.org/TR/rdf-syntax/>
15. SPARQL. The Query Language for RDF, <http://www.w3.org/TR/rdf-sparql-query/>

# Semantic Retrieval Approach to Factoid Question Answering for Bulgarian

Stanislav Peshterliev<sup>1</sup> and Ivan Koychev<sup>1,2</sup>

<sup>1</sup> Faculty of Mathematics and Informatics,  
University of Sofia "St. Kliment Ohridski", Sofia, Bulgaria

<sup>2</sup> Institute of Mathematics and Informatics at Bulgarian Academy of Sciences  
email: {speshterliev, koychev}@fmi.uni-sofia.bg

**Abstract.** With the abundance of information available today, we need efficient tools to explore it. Search engines attempt to retrieve the most relevant documents for a given query, but still require users to look for the exact answer. Question Answering (Q&A) systems go one step further by trying to answer users' questions posed in natural language. In this paper we describe a semantic approach to Q&A retrieval for Bulgarian language. We investigate how the usage of named entity recognition, question answer type detection and dependency parsing can improve the retrieval of answer-bearing structures compared to the bag-of-words model. Moreover, we evaluate nine different dependency parsing algorithms for Bulgarian, and a named entity recognizer trained with data automatically extracted from Wikipedia.

**Keywords:** question answering, information retrieval, semantic annotation.

## 1 Introduction

Most Q&A systems use a search engine to find relevant information for a given question and then process the results for answer extraction. We support the opinion that the improvement of the retrieval of answer-bearing structures is critical for the overall performance of Q&A systems. Consider the question *Who wrote Macbeth?*, a simple bag-of-words query will look like *wrote AND macbeth*, and it will match both sentences *Shakespeare wrote Macbeth* and *Macbeth wrote four poems*, as possible answer-bearing structures. In order to improve the query precision we can pre-process the sentences for named entity recognition and the question for expected answer type. Then we can use this information to formulate the query (*wrote AND macbeth*) *NEAR author* that will match sentences with the question's keywords close to its expected answer type *author*. Additionally, we can improve the retrieval precision by specifying that *author* is the subject and *Macbeth* is the object of the verb *wrote*.

What is the improvement to the retrieval precision that the described semantic approach brings compared to bag-of-words model? In this paper we evaluate how question answer type detection, named entity recognition and dependency parsing affects the retrieval of answer-bearing structures for Bulgarian language. For the

evaluation, we developed a test set of questions for Bulgarian that is based on the test set available for TREC 2007 Q&A Track. We also investigate the application of the available Bulgarian linguistic resources for the training of Malt dependency parser. Additionally, we present the performance of a named entity recognizer built with data extracted from Wikipedia (<http://wikipedia.org>) and DBpedia (<http://dbpedia.org>).

## 2 Related Work

This paper is largely influenced by the work of Mihalcea and Moldovan [3], and Bilotti et al. [1], both evaluated for English with data from TREC Q&A track (<http://trec.nist.gov>). Mihalcea and Moldovan have shown how question answer type detection and named entity recognition reduce two times the number of candidates for answer extraction. Bilotti et al. have researched how semantic role labeling can contribute to even more precise query formulation, and measured the impact of the annotation quality on Q&A retrieval. Other important Bilotti et al. contribution is that they demonstrated that scoring based on term dependency is better than proximity keyword occurrences used by Mihalcea and Moldovan. Tiedemann [11] used dependency parsing in combination with genetic based algorithm for query features selection for Dutch Q&A retrieval and answer extraction, achieving considerable difference compared to bag-of-words model.

Simov and Osenova described the architecture of BulQA [9], a Q&A system for Bulgarian language with a question analysis module, an interface module and an answer extraction module, which was evaluated at CLEF 2005. The main difference between BulQA and our work is that we focus on the retrieval for Q&A, whereas Simov and Osenova focus on the answer extraction and other problems that require sophisticated natural language processing. As a part of the work on BulQA the team also investigated the adoption of available Bulgarian language resources [6] to Q&A. Moreover, Georgiev, Nakov, Osenova, and Simov [2] evaluated the application of Maximum Entropy models for sentence splitting, tokenizing, part-of-speech tagging, chunking and synthetic parsing.

## 3 Retrieval for Question Answering

In this section we investigate the problems of the retrieval for Q&A for Bulgarian language. To give context for further discussions let's consider the following question and its typical bag-of-words query:

*Question 1: Koj e napisal Makbet? (in English: Who wrote Macbeth?)*

```
Query 1: #combine[sentence](napisa makbet)
```

Note that we are using Indri query language [10] for Query 1, where the `#combine [sentence]` clause specifies that the keywords should be in one sentence. Also, pay attention that the keywords are stemmed *napisal* → *napisa*, and question words are not included because they are part of the stop words list.

### 3.1 Query Formulation

How can we improve Query 1? The simplest way is to elaborate it by adding extra information. This is often done by users when they get poor results and try to add more keywords to the query to make it more precise. For Question 1 we can include *avtor* (*author*), which is the question's expected answer.

Query 2: #combine[sentence](napisa makbet avtor)

However, not all questions have one word or phrase answers as in the example. Questions are categorized into two groups by their expected answer type: factoid questions that have short specific answers and procedural questions, with answers several sentences or paragraphs. For example, in English the questions *When*, *Who* and *Where*, expect answers like time, person and location, whereas *Why* and *How* expect long answers such as tutorial or manual. The same examples are valid for Bulgarian questions: *Koga* (*When*), *Koj*, *Koja*, *Koe* (*Who*), *Kyde* (*Where*), *Zashto* (*Why*) and *Kak* (*How to*). In this paper we focus on the problems of factoid question answering for Bulgarian.

### 3.2 Retrieval

Let's consider Query 2 that contains the expected answer. There are cases when directly injected expected answer can result in lower recall. For instance, from the following two sentences, Query 2 will match only Sentence 2 as a relevant sentence because Sentence 1 does not contain the keyword *avtor* (*author*).

*Sentence 1: Shekspir e napisal Makbet.*

(in English: Shakespeare wrote Macbeth.)

*Sentence 2: Izvesten avtor e napisal piesa za Makbet.*

(in English: Famous author wrote a play for Macbeth.)

On other hand, Query 2 can match irrelevant sentences due to the lack of knowledge about predicate-arguments structures. To illustrate the problem we will use the following sentences:

*Sentence 3: Izvestnijat avtor Shekspir e napisal Makbet.*

(in English: The famous author Shakespeare wrote Macbeth.)

*Sentence 4: Izvestnijat avtor Makbet e napisala kniga za John.*

(in English: The famous author Macbeth wrote a book about John.)

Here both sentences contain the keywords from Query 2, although only Sentence 3 is relevant, because, in Sentence 4, *Makbet* (*Macbeth*) is the subject, but we are searching for sentences where *Makbet* (*Macbeth*) is the object of the verb *napisal* (*wrote*).

## 4 Employing the Semantic Information

Here we describe how question answer type detection, named entity recognition and dependency parsing can solve the introduced difficulties in retrieving relevant sentences for answer extraction.



## 4.1 Named Entities and Answer Types

Named entities are words such as the names of persons, organizations, locations, expressions of times and quantities. Finding named entities in unstructured text is one of the main information extraction tasks. It is common to use ontology classes for named entity classes because you can use the ontology data to train recognition tools and take advantage of well defined ontology hierarchies as *Person*  $\rightarrow$  *Writer*. For example if we can recognize *Writer* named entity, Sentence 1 can be annotated as follows:

*Sentence 5: <Writer>Shekspir</Writer> e napisal Makbet.*

Although in some cases using only named entity annotations can improve the precision of the retrieval for Q&A, they alone are not enough. We can automatically detect factoid questions expected answer type, and combine it with the named entities information to formulate more precise queries. To achieve this expected answer type classes and named entity classes should be mapped to the same ontology. We can classify question's expected answer type by its question word, for example:

Koj, Koja, Koe (Who)  $\rightarrow$  Person, Kyde (Where)  $\rightarrow$  Location, Koga (When)  $\rightarrow$  Date.

By leveraging named entity recognition and answer type detection we can change Query 2 to:

Query 3: #combine[sentence](napisa makbet #any:writer)

To see the difference between Query 2 and Query 3, let's consider the plain Sentence 2 and the annotated Sentence 5 as possible retrieval candidates. Sentence 5 is more relevant answer to Question 1 than Sentence 2 because it contains expected type of answer-bearing named entity for the given factoid question. In this case structured Query 3 will retrieve the more relevant Sentence 5 because the clause #any:writer will match the annotation <Writer>...<Writer>, whereas the bag-of-words Query 2 will retrieve Sentence 2. This example illustrates that bag-of-words queries lack the necessary constraints required for Q&A.

## 4.2 Dependency Parsing

We have seen that Query 3 is better than Query 2, but it is still not good enough to retrieve the relevant sentence between Sentence 3 and Sentence 4. The problem is that Query 3 does not specify that the writer is the subject and *Makbet* (*Macbeth*) is the object of the sentence. The problem can be solved by annotating the sentences with predicate-arguments information, and then use it in the query. To do this, we use dependency parsing - an approach to automatic syntactic analysis inspired by theoretical linguistics. For instance, the result from the dependency parsing for Sentence 3 will be: mod(Izvestnijat, avtor) subj(avtor, e) mod(Shekspir, avtor) ROOT(e) comp(napisal, e) obj(Makbet, e), which is a tree structure with the head verb as a root, and the

other sentence structures (i.e. subject, object and complements) as sub-trees. However, to make this information useful for our problem we join the dependencies into larger groups, which are then added as annotations. The predicate-argument annotations for Sentence 3 and Sentence 4 will be:

*Sentence 6:* <Subj>Izvestnijat avtor <Writer>Shekspir</Writer> </Subj> <Root>e napisal</Root> <Obj>Makbet</Obj>.

*Sentence 7:* <Subj>Izvestnijat avtor <Writer>Makbet</Writer> </Subj> <Root>e napisala</Root> <Obj>kniga</Obj> <Prepcomp>za Dzhon</Prepcomp>.

To formulate the query we are processing the question with the dependency parser, and then we are constructing the query using predicate-argument structure annotations. For example Question 1 will look like:

```
Query 4: #combine[sentence](
  #combine[./subj] (#any:writer)
  #combine[./root] (napisa) #combine[./obj] (makbet)
)
```

Query 4 will match only the relevant Sentence 7 that contains the answer as a subject. Using this approach we can retrieve sentences for questions like *Koga Shekspir e napisal Makbet?* (in English: *When did Shakespeare wrote Macbeth?*). Here the expected answer type is *Date*, which will be a prepositional complement of the possible answer-bearing sentence. So the query will be:

```
Query 5: #combine[sentence](
  #combine[./subj] (shekspir) #combine[./root] (napisa)
  #combine[./obj] (maklbet) #combine[./prepcomp] (#any:date)
)
```

## 5 Experiments

A similar approach to Q&A retrieval was evaluated for English by Bilotti et al. We believe that with the current state of tools and linguistic resources available for Bulgarian language, we can achieve similar results.

### 5.1 Information Retrieval System

For the experiments we used Galago toolkit (<http://galagosearch.org>), one of the components of Lemur project. It includes the distributed computation framework TupleFlow that manages the difficult parts of the text processing. The retrieval system supports variant of the Indri query language that provides the necessary constrain checking for the task.

### 5.2 Linguistic Processing

We use OpenNLP (<http://opennlp.sourceforge.net>) tools for tokenization, sentence detection, part of speech tagging. Georgiev, Nakov, Osenova and

Simov [2] performed the evaluation of the tools for Bulgarian language. For dependency parsing we use MaltParser (<http://maltparser.org>) with LIBLINEAR machine learning package, trained for Bulgarian with the data from Bultreebank project [7]. We have implemented a stemmer based on BulStem [4] Bulgarian language stemmer developed by Preslav Nakov. Stop word list is also available from Bultreebank project.

For named entity recognition, we have built a training corpus based on the data from Wikipedia and DBpedia formatted for training OpenNLP named finder component. Named entity classes are mapped to the DBpedia Ontology, which allow us to add richer annotation.

Question answer types detection is implemented with a set of hand-coded rules for Bulgarian, which rule on the question word and of other keywords.

### 5.3 Testing Corpus

The testing corpus is based on the Bulgarian version of Wikipedia dump from March 2011, which contains 173459 articles. The data from the dump is extracted in separated XML les which are stripped from the wiki mark-up, and enriched with annotations for paragraphs, sentence boundaries, named entities and predicate-argument structures.

We have developed a test set of 100 factoid questions, based on the test set available for TREC 2007 Q&A Track. For each question we have made manual relevance judgments in the testing corpus. The questions cover diverse topics and are divided in two types single and multiple answers.

### 5.4 Results

We used Galago compared evaluation tool to measure the difference between the described semantic approach and bag-of-word model. We achieved an average precision improvement of 9.3%, which is a less than that that reported for English [1] and Dutch [11]. The difference is not considerable because the testing corpus is not big and the content consists only of encyclopedic articles.

Table 1 contains the results from the evaluation of MaltParser with different dependency parsing algorithms on the data from Bultreebank project. From the available machine learning packages, we have used only LIBLINEAR because both training and parsing with LIBSVN were too slow in our experiments. For all other parameters, we kept the default values. Our results are in the range between 80%-90% reported by Nivre et al. [5] for various other languages, when no language specific optimizations are applied. Stack lazy performed best with accuracy of 89.8551%, but the difference compared to the other parsing algorithms is not significant, and it may change depending on the data.

**Table 1** MaltParser evaluation result

<b>Parsing Algorithm</b>	<b>Accuracy</b>
Nivre arc-eager	89.1810
Nivre arc-standard	89.2990
Covington non-projective	88.4395
Covington projective	87.8834
Stack projective	89.4169
Stack eager	89.7371
Stack lazy	<b>89.8551</b>
Planar eager	88.5912
2-Planar eager	89.4506

In Table 2 we provide evaluation for the top seven largest classes from the first level of DBpedia ontology. The low recall for persons, places and species shows that classes with entities from many different languages are hard to generalize. Another observation is that the larger one class is the harder is to detect its entities. In practice we use a combination of these models, dictionaries and regular expressions to perform named entity recognition.-

**Table 2** OpenNLP Name Finder evaluation results

	<b>Sentences</b>	<b>Names</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Measure</b>
<b>Place</b>	271367	10506	0.7094	0.2347	0.3528
<b>Person</b>	70922	10691	0.8625	0.4337	0.5772
<b>Organisation</b>	20691	2826	0.8560	0.7343	0.7905
<b>Species</b>	19590	3176	0.7295	0.2548	0.3776
<b>Language</b>	16864	802	0.7475	0.6848	0.7148
<b>EthnicGroup</b>	14872	720	0.8628	0.7276	0.7894
<b>Event</b>	13521	2165	0.8390	0.8386	0.8388

## 6 Conclusions and Further Work

As confirmed for English and Dutch, the described approach to Q&A retrieval improves the precision and reduces the total number of retrieved documents. The reduction in the number of candidates for answer extraction mean that more sophisticated performance intensive algorithms can be used at this stage. Moreover, we confirmed that there are state-of-the-art-quality linguistic resources available for Bulgarian thanks to Kiril Simov, Petya Osenova and other contributors to Bul-treebank project.

We have focused on factoid questions, but term dependencies can also be used for complex procedural Q&A retrieval where answer is several sentences or paragraphs. Scaling question answer type detection to a large ontology is a challenging task, it will be interesting to perform experiments similar to these of Roberts and Hickl [8] with machine learning from large corpus for Bulgarian, and test how it will impact the performance of the system with more diverse and complex questions.

**Acknowledgments.** We are grateful to Kiril Simov and Petya Osenova for providing us with the data from Bultreebank project, and their help with references and suggestion. This research is supported by the SmartBook project, subsidized by the Bulgarian National Science Fund, under Grant D002-111/15.12.2008.

## References

1. Bilotti, M.W., Ogilvie, P., Callan, J., Nyberg, E.: Structured retrieval for question answering. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR 2007, pp. 351–358. ACM, New York (2007)
2. Georgiev, G., Nakov, P., Osenova, P., Simov, K.: Cross-lingual adaptation as a baseline: adapting maximum entropy models to bulgarian. In: Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains, AdaptLRTtoND 2009, pp. 35–38. Association for Computational Linguistics, Stroudsburg (2009)
3. Mihalcea, R., Moldovan, D.I.: Document indexing using named entities. *Studies in Informatics and Control* 10(1), 21–28 (2001)
4. Nakov, P.: Building an inflectional stemmer for Bulgarian. In: Proceedings of the International conference on computer systems and technologies - CompSysTech 2003 (2003)
5. Nivre, J., Hall, J., Nilsson, J.: Maltparser: A data-driven parser-generator for dependency parsing. In: Proc. of LREC 2006, pp. 2216–2219 (2006)
6. Osenova, P., Simov, A., Simov, K.I., Tanev, H., Kouylekov, M.: Bulgarian-english question answering: Adaptation of language resources. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 458–469. Springer, Heidelberg (2005)
7. Osenova, P., Simov, K.: Btb-tr05: Bultreebank stylebook. bultreebank project technical report n 05 (2004)
8. Roberts, K., Hickl, A.: Scaling answer type detection to large hierarchies. In: LREC, European Language Resources Association (2008)
9. Simov, K.I., Osenova, P.: BulQA: Bulgarian–bulgarian question answering at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 517–526. Springer, Heidelberg (2006)
10. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language-model based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis, Tech. rep. (2005)
11. Tiedemann, J.: Improving passage retrieval in question answering using NLP. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 634–646. Springer, Heidelberg (2005)

# Using Semantic Relations for Representing Long-Term User Interests

Muhammad Ali Zeb and Maria Fasli

School of Computer Science and Electronic Engineering, University of Essex  
Colchester, United Kingdom  
email: {mazed,mfasli}@essex.ac.uk

**Abstract.** In this paper we present a new mechanism for representing the long-term interests of a user in a user profile. Semantic relatedness between the profile terms is measured by using the web counting based method. Profile terms are associated through their sets of inductions words, representing highly related words to the terms that are found out through their co-occurrence in the web documents and semantic similarity. The relation between the two profile terms is then calculated using the combination of their corresponding sets of induction words. Although we have used the mechanism for long-term user profiling, applications can be more general. The method is evaluated against some benchmark methods and shows promising results.

**Keywords:** user profile, semantic relatedness, semantic similarity.

## 1 Introduction

In this paper we propose a new approach for measuring semantic relationships between the profile terms aggregated using an RSS aggregator from the activity of the user. The profile terms represent the interests of the user. Semantic relatedness between the profile terms is computed in order to identify the permanent interests of the user. The two step process for calculating the semantic relatedness is performed by first computing the direct relations between the terms, and then in the second step a set of words highly related to each individual term in the profile is created. We call this set of words as “*the set of induction words*” for the profile term. The relatedness between the terms is calculated based on a combination of the terms’ co-occurrence in their respective documents and their *semantic similarity*.

Measure of similarity or relatedness is used in a variety of applications, such as information retrieval, automatic indexing, word sense disambiguation, automatic text correlation. Semantic similarity and semantic relatedness are sometimes used interchangeably in the literature. The terms however, don’t have the same meaning. Semantic relation between the terms or words shows the degree to which they are associated via any type (such as synonymy, meronymy, hyponymy, functional, associative) and other types of semantic relationships. Semantic similarity, on the other hand, is a special case of relatedness and takes into account only hyponymy/hypernamy relations. The relatedness measure may use a combination of

relations existing between words depending on the context or their importance. To illustrate the difference between similarity and relatedness, Resnik [26] provides a widely used example of car and gasoline. The terms are more closely related than the terms car and bicycle, although they are not very similar as compared to the terms car and bicycle, and they have only few features in common. But they are more closely related because cars use gasoline. A number of researcher use distance measure as a measure of opposite similarity.

In this paper we propose a new approach for measuring semantic relatedness between words. Main idea of the approach is to measure semantic relationships using both the direct relation between the profile terms and then using a set of highly related words to a profile term, which we call the set of induction words. We use the co-occurrence of the terms in the documents combining it with semantic similarity measurements of the profile terms to create highly-related set of induction words. Comparison of the experimental results with a benchmark set of human similarity ratings show the effectiveness of the proposed approach.

This paper is organized as follows. Section 2 presents related work. In section 3 the proposed method is explained. The method of evaluating semantic relatedness between the words is explained in section 4 and experimentation results are presented. Conclusions and future work are discussed in the last section.

## 2 Related Work

Measurements of the semantic similarity of words have been widely used in research and applications in natural language processing and related areas, such as the automatic creation of thesauri [10, 19, 16], automatic indexing, text annotation and summarization [18], text classification, word sense disambiguation [15], [16], information extraction and retrieval [4, 28], lexical selection, automatic correction of word errors in text, discovering word senses directly from text [23], and language modelling by grouping similar words into classes [3]. Generally there are two types methods used for computing similarity of two words: edge counting methods and information content methods. There are also some hybrid methods that combine the two types. Edge counting methods, also known as the path-based or dictionary-based methods (using WordNet, Roget's thesaurus or other resources); define the similarity of two words as a function of the length of the path linking the words and on the position of the words in the taxonomy. A short path means high similarity. In WordNet, lexical information is organized according to word meanings. The core unit in WordNet is called a synset. Synsets are sets of words that might have the same meaning, that is, synonyms. A synset represents one concept, to which different word forms refer. For example, the set {car, auto, automobile, machine, motorcar} is a synset in WordNet and forms one basic unit of the WordNet lexicon. Although there are subtle differences in the meanings of synonyms, these are ignored in WordNet. The WordNet::Similarity Software Package<sup>1</sup> implements several WordNet-based similarity measures: Leacock & Chodorow [14], Jiang & Conrath [7], Resnik [26], Lin [19], Hirst &

St-Onge [13], Wu & Palmer [27], extended gloss overlap, Banerjee & Pedersen [1], and context vectors, Patwardhan [24].

If the two words have multiple senses, the similarity between them, out of context, is the maximum similarity between any of the senses of the two words. Three of the above methods are hybrid (Jiang & Conrath [7], Resnik [26], Lin [19]), they use frequency counts for word senses from Semcor, which is a small corpus, annotated with WordNet senses. The work of Rada et al [17] deals with measuring word similarity on the basis of edge counting methods. They compute the semantic relatedness in terms of the number of edges between the words in the taxonomy. In Leacock and Chodorow [14], measurement of semantic similarity takes into account the depth of taxonomy in which the words were found. The Wu and Palmer [27] similarity metric measures the depth of the two given words in the taxonomy, along with the depths of the least common subsumer.

Information content methods, also known as corpus based methods (using statistics), measure the difference in information content of the two words as a function of their probability of occurrence in a corpus. The method was first proposed by Resnik [26]. According to Resnik, the similarity of two words is equal to information content of the least common subsumer. However, because many words may share the same least common subsumer, and therefore might have the same values of similarity, Resnik measure may not be able to obtain fine grained distinctions [26]. Jiang and Conrath [7] and Lin [19] have developed measures that scale the information content of the subsuming concept by the information content of the individual concepts. Lin does this via a ratio, and Jiang and Conrath with a difference. Gloss based methods define the relatedness between two words as a function of gloss overlap [15]. Banerjee and Pedersen [1] have proposed a method that computes the overlap score by extending the glosses of the words under consideration to include the glosses of related words in a hierarchy.

Some researchers define the semantic relatedness between the words using Web. Bollegala et al [2] have proposed a method that exploits the page counts and text snippets returned by a Web search engine to measure semantic similarity between words. An approach to computing semantic relatedness using Wikipedia is proposed in [6]. Strube et al. also investigated the use of Wikipedia for computing semantic relatedness measures [20].

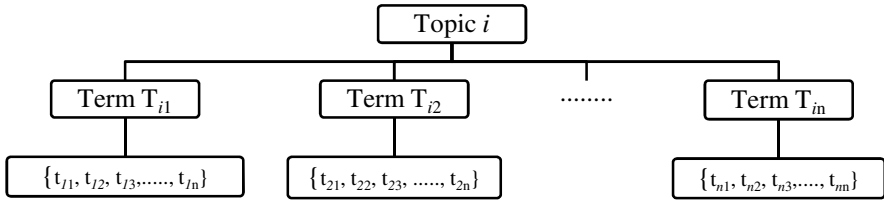
Semantic similarity measurements have also been widely used to create ontology-based user models [8, 21, 29]. Ontology-based user profiling has a number of limitations, when a wide and dynamic domain like the Web is concerned [9]. Although individual profiles are able to manage high number of concepts, these concepts cannot embrace the potentially infinite number of specific user interests. For example, *Yahoo!* Ontology can represent the concept *baseball* inside *sports*, but not going further to represent in a given non-famous *baseball team* or a *player*. Besides failing to capture specific user interests, ontologies impose their organization of concepts to user profiles that are not necessarily in correspondence with user views of such concepts. Moreover, users



can have different perceptions of what the same ontological concept means, leading to inaccurate profile representations. We have tried to address this issue in our approach.

### 3 Methodology

The motivation for our approach is drawn from the need to represent the long term interests of a user based on the semantic relations that exist between the profile terms. The user's profile is represented by a set of  $K$  topics. Each topic in the profile has  $n$  terms associated with it, where  $n$  is variable for each topic. In this paper, we assume that the profile has one topic. Then for a topic  $i$  in the profile,  $T_{i1}, T_{i2}, \dots, T_{in}$  are the terms associated with the topic. We call these terms as the *profile terms* in this paper. We create a set of words that are highly related to each profile term, called the *set of induction words* for that particular term. This set of induction words is created by calculating the frequency of co-occurrence of the profile terms in the corresponding documents from which they were retrieved and their synonyms found using WordNet ontology [22]. Figure 1 shows a topic  $i$  with its associated terms  $\{T_{i1}, T_{i2}, \dots, T_{in}\}$ , and their corresponding sets of induction words.



**Fig. 1** Profile terms associated with a profile topic, and their sets of Induction words

Let  $T_i$  and  $T_j$  be two profile terms associated with a topic in the user profile, for which we want to measure the semantic relation.  $T_i$  is represented by a set of induction words  $S(T_i) = \{t_{i1}, t_{i2}, t_{i3}, \dots, t_{in}\}$  and  $T_j$  is represented by a set of induction words  $S(T_j) = \{t_{j1}, t_{j2}, t_{j3}, \dots, t_{jm}\}$ . Now combining the two sets together, we obtain a common set of words representing the two terms  $T_i$  and  $T_j$ ,  $S(T) = S(T_i) \cup S(T_j)$ :

$$S(T) = \{t_1, t_2, t_3, \dots, t_k\} \quad (1)$$

where  $k$  is equal to or less than  $m+n$ .

Now we measure the relatedness for each word  $t$  in the union set  $S(T)$  with the profile terms  $T_i$  and  $T_j$  using equations (2) and (3) respectively:

$$Rel(t, T_i) = \frac{freq(t, T_i)}{maxfreq_i} \quad (2)$$

$$Rel(t, T_j) = \frac{freq(t, T_j)}{maxfreq_j} \quad (3)$$

Here,  $freq(t, T_i)$  and  $freq(t, T_j)$  show the frequency of the number of web documents in which the word  $t$  and the corresponding term  $T_i$  or  $T_j$  have occurred together.  $maxfreq_i$  and  $maxfreq_j$  represent the maximum number of times the words in the union set  $S(T)$  have occurred together with the corresponding terms  $T_i$  and  $T_j$  respectively, i.e.,

$$maxfreq_i = \max\{Rel(t_1, T_i), Rel(t_2, T_i), \dots, Rel(t_k, T_i)\} \text{ and } maxfreq_j = \max\{Rel(t_1, T_j), Rel(t_2, T_j), \dots, Rel(t_k, T_j)\}.$$

We assume that if an induction word is highly related to a profile term, then the probability of its co-occurrence with the profile term in the web documents is high. In a special case, if a word  $t$  of the induction set synonymous to the profile term  $T_i$  or  $T_j$ ,  $Rel(t, T_i) = 1$  or  $Rel(t, T_j) = 1$ .

Now to calculate the relatedness  $Rel(T_i, T_j)$  between the profile terms  $T_i$  and  $T_j$ , we use equation (4), as follows

$$Rel(T_i, T_j) = \frac{\sum_{t=1}^k \left( \frac{\alpha_t R_t}{1 + R_t} \right) + \beta}{1 + \beta} \quad (4)$$

Here,

$$R_t = \frac{\min\{Rel(t, T_i), Rel(t, T_j)\}}{\max\{Rel(t, T_i), Rel(t, T_j)\}}$$

$\alpha_t$  is the co-occurrence factor, defined as

$$\alpha_t = \begin{cases} 2, & t \text{ occurs in the both the induction sets for terms } T_i \text{ and } T_j \\ 1, & \text{otherwise} \end{cases}$$

$\beta$  is the synonymy factor, defined as

$$\beta = \begin{cases} 1, & \text{Terms } T_i \text{ and } T_j \text{ are synonyms} \\ 0, & \text{otherwise} \end{cases}$$

## 4 Evaluations

To evaluate the semantic relatedness measurement, researchers usually compare the results with several experiments on human judgments. Words and relatedness decided by humans in these experiments have been considered benchmarks for measuring relatedness. To evaluate our proposed mechanism we use the Miller and Charles dataset [22]. They performed the experiment with a group of 38 human subjects using a subset of 30 pairs of nouns. A score of 4.0 was assigned to words considered synonyms and a score of 0.0 was assigned to words considered totally unrelated. The scores of all human judges were averaged and analyzed.

Most researchers have used only 28 pairs of nouns of the Miller and Charles set. We have used the same set of noun pairs. In Table 1 we show a comparison of the proposed mechanism with the Miller and Charles [22], Resnik [26] and Jaccard methods. It can be seen that our proposed method shows significant progress over the other methods.

**Table 1** Pair-wise comparison

Word pair	Miller & Charles	Jaccard	Resnik	Proposed
cord-smile	0.13	0.102	0.1	0.110
rooster-voyage	0.08	0.011	0	0.019
noon-string	0.08	0.126	0	0.005
glass-magician	0.11	0.117	0.1	0.015
monk-slave	0.55	0.181	0.7	0.435
coast-forest	0.42	0.862	0.6	0.201
monk-oracle	1.1	0.016	0.8	0.340
lad-wizard	0.42	0.072	0.7	0.215
forest-graveyard	0.84	0.068	0.6	0.298
food-rooster	0.89	0.012	1.1	0.659
coast-hill	0.87	0.965	0.7	0.473
car-journey	1.16	0.444	0.7	0.450
crane-implement	1.68	0.071	0.3	0.625
brother-lad	1.66	0.189	1.2	0.580
bird-crane	2.97	0.235	2.1	0.450
bird-cock	3.05	0.153	2.2	0.789
food-fruit	3.08	0.753	2.1	0.597
brother-monk	3.82	0.261	2.4	0.745
asylum-madhouse	3.61	0.024	3.6	0.827
furnace-stove	3.11	0.401	2.6	0.578
magician-wizard	3.5	0.295	3.5	0.950
journey-voyage	3.84	0.415	3.5	0.864
coast-shore	3.7	0.786	3.5	0.795
implement-tool	2.95	1	3.4	0.742
boy-lad	3.76	0.186	3.5	0.736
automobile-car	3.92	0.654	3.9	1
midday-noon	3.42	0.106	3.6	0.961
gem-jewel	3.84	0.295	3.5	1

## 5 Conclusion and Future Work

We introduced a novel measure of semantic relatedness for representing the long-term interests of a user in a user profile. Our measure correlates well with the human judgments and can be applied to different domains. Our future work

includes annotation analysis of the profile terms and using them for the classification of web resources.

## References

1. Banerjee, S., Pedersen, T.: An adapted Lesk algorithm for word sense disambiguation using WordNet. In: Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, pp. 136–145 (2002)
2. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring Semantic Similarity between Words Using Web Search Engines. In: Proceedings of the 16th International World Wide Web Conference (WWW2007), Banff, Alberta, Canada, pp. 757–766 (2007)
3. Brown, P.F., DeSouza, P.V., Mercer, R.L., Watson, T.J., Della Pietra, V.J., Lai, J.C.: Class-based n-gram models of natural language. *Computational Linguistics* 18, 467–479 (1992)
4. Buckley, C., Salton, J.A., Singhal, A.: Automatic query expansion using Smart: TREC 3. In: The third Text Retrieval Conference, Gaithersburg, MD (1995)
5. Cilibrasi, R.L., Vitanyi, P.M.B.: The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering* 19, 370–383 (2007)
6. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India, pp. 1606–1611 (2007)
7. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th International Conference on Research in Computational Linguistics, Taipei, Taiwan, pp. 19–33 (1997)
8. Gauch, S., Chafee, J., Pretschner, A.: Ontology-based personalized search and browsing. *Journal of Web Intelligence and Agent Systems* 1(3-4), 219–234 (2003)
9. Godoy, D., Amandi, A.: User profiling in personal information agents: a survey *The Knowledge Engineering Review*, pp. 1–33. Cambridge University Press, Cambridge (2005)
10. Grefenstette, G.: Automatic thesaurus generation from raw text using knowledge-poor techniques. In: Making Sense of Words, 9th Annual Conference of the UW Centre for the New OED and Text Research (1993)
11. Guarino, N., Giaretta, P.: Ontologies and Knowledge bases: towards a terminological clarification. In: Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, pp. 25–32. IOS Press, N. Mars (1995)
12. Guarino, N., Masolo, C., Vetere, G.: OntoSeek: Content-Based Access to the Web. *IEEE Intelligent Systems* 14(3), 70–80 (1999)
13. Hirst, G., St-Onge, D.: Lexical Chains as representations of context for the detection and correction of malapropisms. In: *WordNet An Electronic Database* (1998)
14. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet, *An Electronic Lexical Database*, pp. 265–283. MIT Press, Cambridge (1998)
15. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation, Toronto, pp. 24–26 (1986)

16. Li, H., Abe, N.: Word clustering and disambiguation based on co-occurrence data. In: Proceedings of COLING-ACL, pp. 749–755 (1998)
17. Li, Y., Bandar, Z.A., Mclean, D.: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering* 15, 871–882 (2003)
18. Lin, C.Y., Hovy, E.H.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of HLT-NAACL (2003)
19. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconsin USA, pp. 296–304 (1998)
20. Strube, M., Ponzetto, S.P.: WikiRelate! Computing Semantic Relatedness using Wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Mass., pp. 1419–1424 (2006)
21. Middleton, S.E., Shadbolt, N.R., De Roure, D.C.: Ontological User Profiling in Recommender Systems. *ACM Transactions on Information Systems (TOIS)* 22(1), 54–88 (2004)
22. Miller, G.A.: WordNet: A lexical Database for English. *Comm. ACM* 38, 39–41 (1995)
23. Pantel, P., Lin, D.: Discovering word senses from text. In: Proceedings of ACM SIGKDD, Conference on Knowledge Discovery and Data Mining, pp. 613–619 (2002)
24. Patwardhan, S.: Incorporating dictionary and corpus information into a vector measure of semantic relatedness. MSc Thesis, University of Minnesota (2003)
25. Rada, R., Mili, H., Bichnell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets. *IEEE Trans. Systems, Man, and Cybernetics* 9, 1–30 (1989)
26. Resnik, P.: Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11, 95–130 (1999)
27. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, pp. 133–138 (1994)
28. Xu, J., Croft, B.: Improving the effectiveness of information retrieval. *ACM Transactions on Information Systems* 18(1), 79–112 (2000)
29. Yannis, L., Tim, F.: Yahoo! As An Ontology – Using Yahoo! Categories to Describe Documents. In: Proceedings of the 8th International Conference On Information Knowledge Management (CIKM), pp. 180–187 (1999)

# Integrating Internet Protocol Television and Social Networking

Afshana S. Hussain and Julita Vassileva

University of Saskatchewan  
Department of Computer Science Saskatoon, SK, Canada  
email: afs371@mail.usask.ca, jiv@cs.usask.ca

**Abstract.** Elderly people comprise the highest proportion of television viewers. However, they often struggle with new technologies and often out-rightly reject them due to their complexity. We propose a system to help people, especially the elderly, keep up with new technologies, such as IPTV and social networks with reduced efforts. This system integrates IPTV and social networking website with an interface using mobile phone. Speech to text technology is used as input to reduce the difficulty involved in interaction while viewing television. As speech is a more convenient and natural way of expression than text, we anticipate that people in all age groups would benefit from the system.

## 1 Introduction

In recent years, statistics show that the percentage of elderly people using social networks has been rising in the developed countries, though it is still much lower than that of the younger population. The lower participation of elderly in comparison to younger people is due to their unfamiliarity with modern devices. On the other hand, the elderly population accounts for the highest TV viewing group. New TV technologies can provide an alternative way to connect to social networking sites. With the entry of IPTV, there is an increasing need for the social networking experience to be integrated into the interactive television experience. We propose a system to integrate IPTV to social networking websites with the aid of mobile phone where speech to text technology will be used as input to reduce the difficulty involved in interaction during television viewing activity.

## 2 Literature Survey

### 2.1 Internet Protocol Television

Internet Protocol TV (IPTV) is a system through which Internet television services are delivered [10]. Some of the features of IPTV include live television [5], time-shifted programming [5], and video on demand (VOD) [5]. In future, residential IPTV is expected to grow at a higher pace as broadband was available to more

than 200 million households worldwide in the year 2005[6]. IPTV has been anticipated to grow to 400 million households by the year 2010 [6].

## 2.2 Elderly People Needs and Technologies

As people age they have basic need to be connected without violating their privacy. Many elderly people have difficulty learning new technology and lack the motivation to learn. Elderly are intimidated by the overwhelming complexity they perceive in the technology and have limited economical means.

According to Nielsen study [3], the elderly (aged 65+) account for highest Television viewing group (17.4% of total TV viewers [9]) and approximately spent 4-5 hours watching television every day. The elderly are slowly adopting social networking sites. In 2009, the percentage of elderly holding a profile on social networking site reached 36%, which is more than one-third of the elderly population [4]. Many elderly are adopting Twitter due to their well-developed verbal capabilities to express themselves within the restriction of 140 characters [8]. Though the above studies are restricted to US, they give a clear indication that the elderly are being more and more attracted to social networking. If a social networking facility is provided to the elderly in combination with television viewing then this could be of great benefit to them. The television can provide several services at home and extend their TV viewing activity to accommodate the social activity over the internet. It will fulfill their need to be socially connected [2]. However when it comes to elderly and physically challenged people it is important to consider user-friendly interfaces which is simple to use, natural and intuitive.

The Traditional remote controls have complicated layouts which make it difficult to navigate to a given feature and remote controls aren't apt devices for interaction with the Internet. The devices like mobile phones can come in handy while interacting with interactive television [7]. Nowadays cell phones are accessible to everyone including the elderly due to reduced device and service costs. Pew research center's internet and American life project in August 2010 showed that many elderly aged (65 -75) have cell phones and preferred performing simpler tasks on their cell phone [1]. They were less-likely to do other tedious tasks like texting on mobile phones due to the complexity. Incorporating speech technology in mobile phones can greatly benefit the elderly and physically challenged groups.

Nowadays there are many speech-to-text software products available for deaf people to convert lectures into textual format [11]. Speech to text software such as INTELL and LISTEN are available for language training [12]. There are third party applications on recent smart phones such as Dragon dictation and Vlingo which perform Speech to text conversion [13]. Google text to speech service is also quite good and it has been incorporated massively in Android smart phones.

## 2.3 Problem Statement

Our goal is to provide a convenient way of interaction with IPTV for the elderly and younger users, by incorporating speech technology on mobile phones to

access social networks. The problem is of great importance as speech is the most natural way to express oneself. Therefore everyone can potentially benefit from it.

### 3 Proposed Approach

We propose to integrate the social networking experience for the elderly in the Internet protocol television by using speech as the desired mode of input on a mobile phone. It has 2 main objectives. Firstly, getting the speech input from the user, converting it into text messages and sending it to twitter by means of a mobile phone. Secondly, receiving Twitter messages and displaying them on the IPTV screen.

Our proposed system has four main components, namely the Mobile client (Blackberry Smart phone), Twitter web server, Mediroom web server and Mediroom simulator.

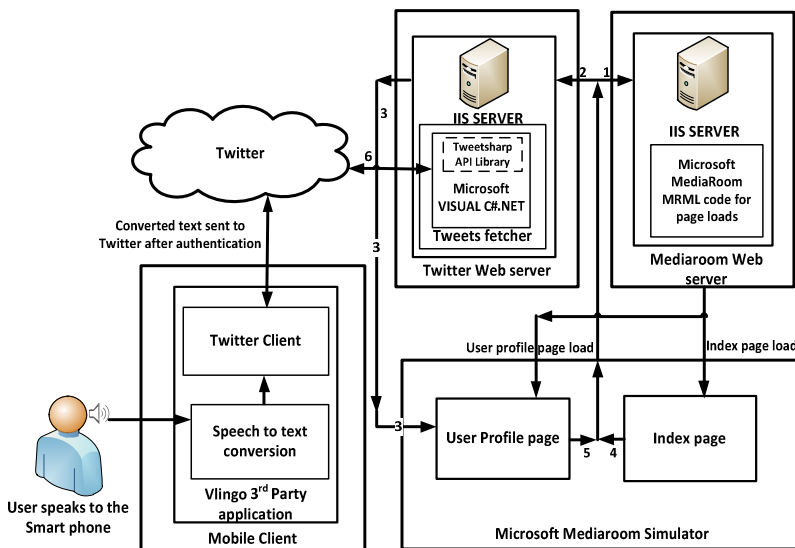


Fig. 1 System Architecture

In Fig.1, 1 is request for the user profile page load (or) Index page load, 2 is request for the user profile page status update information, 3 is actual user profile status updates information, 4 is request for User profile page load, 5 is request for Index page load and 6 is User Status retrieval from twitter account after authentication.

The *Mobile client* is a blackberry smart phone with the Vlingo 3rd party software which does speech to text conversion and sends it to Twitter. The following three steps take place in mobile client. Firstly, the speaker speaks the message intended to be sent to Twitter prefixed by the word "Twitter". Secondly, Vlingo converts the spoken words to text and prompts for approval. Thirdly, if approved



then the message is sent to Twitter. If the message is not approved then the speaker needs to start over re-recording the message.

In *Twitter web server*, the Tweets fetcher application was developed using C# with Tweetsharp API libraries to fetch the status updates from the Twitter server. The Tweets fetcher uses the specific customer key, customer secret, token key and access key associated with registered application to get the authentication and user status information. Currently the user profile selection is restricted to two users only. Since the number of people residing in elderly person residence who will be potential users of the mobile phone is limited. The IIS server has two main responsibilities. Firstly, it receives the name of the user whose tweets are requested from Mediaroom. Secondly, it fetches the tweets of the user requested by Mediaroom using the Tweet fetcher application and sends them to the Mediaroom simulator for display on 5 second interval.

In *Mediaroom Web server*, the IIS Server on Mediaroom web server is responsible for index page loads and user profile page loads based on the requests, using the MRML code. The Mediaroom has two pages, namely an "Index page" and a "User profile page". The Index page is the initial page loaded in the simulator and it has the profile information of the users. The IPTV viewer can choose the user profile whose tweets are to be displayed on the IPTV simulator. Once a user profile is selected on the Index page two steps are carried out. Firstly, the tweet request is sent to the Twitter web server for further processing, to which tweets are received in response and displayed. Secondly, a request is sent to the IIS server on the Mediaroom side to load the user profile page related to the user. There is a valid option on User profile page to return back to the Index page with the help of the IIS Server page load.

## 4 Pilot Study

The System was implemented and a preliminary evaluation of the system was conducted with 7 students from MADMUC lab. The system was tested for responsiveness and accuracy of the speech to text conversion in terms of different accents, sentence length, talking speed and pitch.

One of the performance measures is the turn-around time. It is the time between participant's speech input on the mobile phone and the actual time the twitter message shows on IPTV simulator. The participants were satisfied with 5 seconds turn-around time. The participants preferred transparent and vertical layout. They were tolerant to some conversion errors due to Vlingo's poor handling of different accents and long sentences.

As a part of future evaluation a study with elderly people is being planned. The study will include (i) Usability studies of the interface (ii) Observation studies based on how the IPTV and mobile phones (with speech technologies) are used by them and (iii) Comparative studies targeting usage of social network on the mobile phones, internet and IPTV.

## 5 Discussion and Conclusion

IPTV will likely become a mainstream technology for television viewers in next few years. Elderly people are the highest TV viewing group and they are currently also adopting social networking. To facilitate the adoption of these new technologies, it is imperative to have a simpler interface. Currently existing TV remote control devices are not convenient for navigation due to complexity and Smart phones have the potential to replace them. We saw the recent break-through in speech to text technologies and their roles in recent smart phones.

It can thus be inferred that integrating the IPTV and Social networking with the help of smart phones using speech to text technology will be an important contribution for IPTV field and will help the elderly to be socially well connected. As a part of future work the proposed system will be extended to accommodate other social networking web sites.

## References

1. Berthiaume, D.: MarketingChart, <http://www.marketingcharts.com/direct/85-of-us-adults-have-cell-phone-15958/pew-cell-phone-actions-feb-2011jpg/> (retrieved on February 9, 2011)
2. Cesar, P., Chorianopoulos, K., Jensen, J. (eds.): Proceedings of the 5th European Conference on Interactive Tv: a Shared Experience, p. 74.8. Springer, Heidelberg (2007)
3. Elderly and the Media - Aging and Levels of Viewing, What Do Older Adults Watch? Other Media Use, <http://encyclopedia.jrank.org/articles/pages/6515/Elderly-and-the-Media.html> (retrieved on February 9, 2011)
4. Grove, J.V.: Baby Boomers and Seniors Are Flocking to Facebook (STATS), <http://mashable.com/2010/01/28/baby-boomers-social-media/> (retrieved on February 9, 2011)
5. Hei, X., Liang, C., Liang, J., Liu, Y., Ross, K.W.: A Measurement Study of a Large-Scale P2P IPTV System. *IEEE Transactions on Multimedia* 9(8), 1672–1687 (2007)
6. IPTV, <http://en.wikipedia.org/wiki/IPTV> (retrieved on December 12, 2010)
7. Jacko, J. (ed.): Proceedings of the 12th International Conference on Human-Computer Interaction: Intelligent Multimodal Interaction Environments, p. 979. Springer, Heidelberg (2007)
8. Reuters. Elder like Twitter, Too (February 9, 2011), <http://www.pcmag.com/article2/0,2817,2349535,00.asp>
9. Sditch. Prime time tv viewers by age demographics, <http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/prime-time-tv-viewers-by-age-demog> (retrieved on February 9, 2011)
10. Xiao, Y., Du, X., Zhang, J., Hu, F., Guizani, S.: Internet Protocol Television (IPTV): The Killer Application for the Next-Generation Internet. *IEEE Communications Magazine* 45(11), 126–134 (2007)

11. Stinson, M.S., Elliot, L.B., Kelly, R.R., Liu, Y.: Deaf and Hard-of-hearing Students' Memory of Lectures with speech-to-Text and Interpreting/Note Taking Services. *Journal of Special Education* 43(1), 52–64 (2009)
12. Epp, C.D.: Protutor: A pronunciation tutor that uses historic open learner models. M.Sc. Thesis, University of Saskatchewan, Canada (2010)
13. List of Speech recognition software,  
[http://en.wikipedia.org/wiki/  
List\\_of\\_speech\\_recognition\\_software](http://en.wikipedia.org/wiki/List_of_speech_recognition_software) (retrieved on March 29, 2011)

# Clustering and Visualization of ECG Signals

Anikó Vágner<sup>1</sup>, László Farkas<sup>2</sup>, and István Juhász<sup>1</sup>

<sup>1</sup> University of Debrecen; Faculty of Informatics,  
email: {vagner.aniko, juhasz.istvan}@inf.unideb.hu

<sup>2</sup> Labtech Ltd  
email: medical@labtech.hu

**Abstract.** Holter electrocardiographic (ECG) recordings are ambulatory long-term registers that are used to detect heart diseases. These recordings normally include more than one channel and their durations are up to 24 hours. The principal problem of the cardiologists is the manual inspection of the whole Holter ECG in order to find all those beats which morphologically differ from the normal beats. In this paper we present our method. Firstly, we apply a grid clustering technique. Secondly, we use a special density-based clustering algorithm, named Optics. Then we visualize every heart beat in the record, heartbeats in a cluster, furthermore we represent every cluster with median of heartbeats. We can perform manual. With this method the ECG is easily analyzed and the time of processing is optimized.

**Keywords:** clustering, ECG signals, visualization.

## 1 Introduction

It is part of the job of the cardiologists to evaluate 24 hours ECG recordings. They search for irregular heart beats. Evaluation of long recordings is a lengthy and tedious task. Our program is made for making this work easier.

Figure 1, shows a part of a three channel ECG recording. Normal heart beats are marked with 'N', ventricular heart beats with 'V'. On the figure you can see a vertical line at every heart beat marking the annotation of beat.

The aim of the program is to put seemingly similar heart beats into one group. Thus, cardiologists do not have to examine all the (often more than 100000) heart beat curves. They only have to analyze groups belonging to abnormal beats. On the one hand, the task of the cardiologists is becoming simpler; on the other hand, the possibility of making a mistake is reduced as they discover abnormal beats more easily.

We developed the clustering program in C# 4.0 programming language, in Visual Studio 2010 environment. This program is designed for being used in the Holter system of Labtech Ltd, named Cardiospy.



**Fig. 1** Three Channel ECG recording

In our paper we present the automatic and manual clustering and visualization of ECG signals.

## 2 Processing

We digitalized the analog ECG signals with Cardiospy. The program processes the digitalized, raw ECG signals with a methodology that is similar to the methodology discussed in [2]. The difference between the two methodologies is that we apply wavelet transformation instead of polygonal approximation. Wavelet transformation has been playing an important role in ECG signal processing in the last few years [4,5].

The first step is to locate the specific position of the heart beats performed by the QRS detector. In this step we also get the attributes of every heart beat. We divide the ECG signal into spectral components by wavelet transformation [4]. From the components we create parameters. For clustering we characterize every QRS with a few numerical values.

After QRS detection, we determinate the type of every heart beats (N – normal, S –supraventricular, V – ventricular). After this we perform clustering separately on every type by choosing an ECG channel.

From the well-known types of clustering algorithms we apply the grid-based method. In this method we transform the points into grids, and later we work with these grids only. The main advantage of this method is speed. In our case many points in the set of points have the same coordinate or they are close to each other. With the grid-based method we can radically reduce the number of points and the runtime of our algorithm. In each grid we count how many points are there, and this number is used as a similarity metrics.

Most clustering methods can build only clusters with elliptic shapes. The density-based methods can discover arbitrarily shaped clusters. The basic idea of density-based clustering is that the neighborhood of each point of a cluster with a given radius has to contain at least a minimum number of points. With density-based methods, density can be defined as the number of values in a predefined unit area in the data space. The purpose of this kind of clustering is to group points from each high-density region into a cluster respectively and to ignore the objects in low-density regions. These methods are dynamic methods; we don't need to give the number of clusters. The clustering changes based on parameters, like the radius and the threshold.

DBSCAN is the most frequently used density-based method for ECG signal processing [3]. Optics [1] is also a density-based method, it is less frequently applied than DBSCAN. We use it because it is efficient. The Optics method orders

the points and assigns reachability distance value to each point. We identify the clusters with giving a threshold value for the reachability distance. In our algorithm we use this method but we customized it.

The dispersion of points representing different types of heart beats is not entirely random. They create well-separable sets of points where the sets of points are very dense. There are only a few stand-alone points that are not clusterable.

As a result of the algorithm, clusters appear at heart beat types. We put the not clustered points into a special garbage cluster.

We characterize every cluster by the median of curves. We call it template. We can analyze curves belonging to certain clusters together and separately too by using visualization devices. Heart beats belonging to certain clusters can be examined one by one in their original environment.

### 3 Visualization

The program provides an interactive graphical user interface in which the results of clustering are visualized. So the cardiologist can analyze, manage and can work on them further.

In Figure 2, you can see the full screen of the program. In the right upper part we visualize the templates of certain heart beat clusters. On the left side you can see the heart beats drawn on each other belonging to the template marked with red color. In the right down part you can see the heart beats belonging to the template and their exact position on the ECG recording.

We can visualize the main features of certain heart beat groups by the help of the templates. The templates can help to look through all the heart beat groups. As a main feature, in the left upper corner of the template appears how many heart beats are there in the template and beside it the percentage compared to the total heart beat number. In the upper right corner you can see the type of the heart beats. In the left lower corner the pie chart shows how the heart beats in the group are similar to each other. The more green color it contains, the more resembling the heart beats are. The red number in the centre shows the identifier of the heart beat group. The square in the right lower part helps the cardiologist. He can put a check mark in it if he already analyzed that group.

In the right down part of Figure 2, you can see the whole recording in an enlarged form. You can go through the elements of the cluster marked by red color with the help of the scroll bar in the upper part. The heart beat of the grey column is an element of the heart beat cluster marked by red color.

On the left side of Figure 2, we drew heart beats belonging to certain clusters on each other. The starting points of the drawings on each other are the annotations. We can grab every heart beat at its annotation. After this we cut down areas with a given interval from the left and right side of the recording and we represent certain curves in this way. If there are more heart beats at the given area, its color first becomes darker and after that its color becomes more and more red. The aim is that the more heart beats go to an area, the more powerful the representation should be.

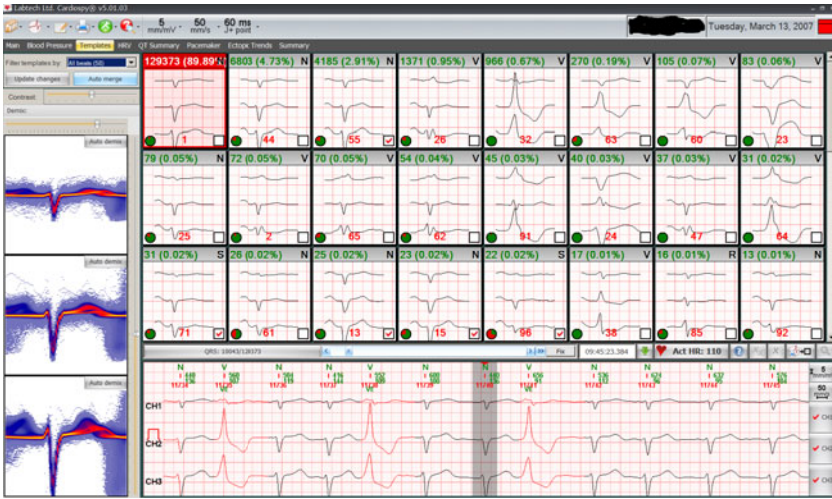


Fig. 2 Full screen

To make its use easier, the figure gets 3 sliders. You can alter contrast with the upper horizontal slider. You can shrink and stretch with the lower horizontal slider in horizontal direction, with the vertical slider in vertical direction. Vertical and horizontal sliders constitute a great help in manual clustering.

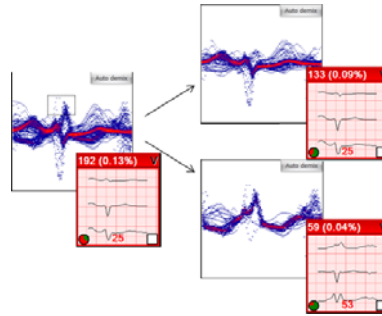


Fig. 3 Manual clustering

## 4 Manual Clustering

You can divide certain clusters into further groups manually. You can select the heart beats in order to be cut by the mouse. By the effect of these, the program divides the cluster into two groups. The selected heart beats constitute a cluster; all other heart beats constitute another one. In the left part of Figure 3, you can see the heart beats drawn on each other and templates belonging to the original

cluster. In the right part of Figure 3, you can see the curves drawn on each other and the templates belonging to the two new clusters. We performed the manual clustering on heart beats of the first channel.

## 5 Discussion and Conclusion

The number of clusters generated by the algorithm could be influenced by the  $r$  (radius) parameter. We applied fix  $r$  value during examinations. We divided the created clusters by Manual and AutoDemix procedures.

We tested our procedure on 20 pieces of 24 hour recordings. Recordings were made about patients who had been examined by Holter ECG due to heart disease. They had no implanted pacemaker.

The amplitude and shape can vary a lot during a recording due to (for example) the position of the body, physical activity, etc. So the parameter created for clustering heart beats can take on varied planar formations.

Our expectation was that the density-based method can put QRS attributes with different planar shapes to one cluster.

On average, it put 97.05% of dominant heart beats into a cluster. In 11 of the 20 recordings, more than 99.0% of dominant heart beats got into a cluster. The algorithm put the QRSs modulated by different noises into the rest of the clusters.

We have built the method into the Labtech Cardiospy Holter system. The system is traded in a number of countries like Japan, Romania, Hungary. The feedback of the users proves our measurements that show that the method written in the article efficiently supports the evaluation of HOLTER ECG.

## References

1. Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J.: Optics: Ordering Points to Identify the Clustering Structure. In: SIGMOD 1999: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM Press, New York (1999)
2. Micó, P., Cuesta, D., Novák, D.: Clustering Improvement for Electrocardiographic Signals. In: Roli, F., Vitulano, S. (eds.) ICIAP 2005. LNCS, vol. 3617, pp. 892–899. Springer, Heidelberg (2005)
3. Zheng, G., Tang, T., Huang, Y.: Automatic Selection of Dynamic Electrocardiogram Waveform by Knowledge Discovery Technologies. In: International Conference on Business and Information, Japan, CD-ROM (2007)
4. Alfaouri, M., Daqrouq, K.: ECG Signal Denoising By Wavelet Transform Thresholding. American Journal of Applied Sciences 5(3), 276–281 (2008)
5. Ciocoiu, I.B.: ECG Signal Compression Using 2D Wavelet Foveation. International Journal of Advanced Science and Technology 13, 15–26 (2009)



# Visualizing and Presenting Story Maps in TM4Book

Lars Johnsen<sup>1</sup>, Darina Dicheva<sup>2</sup>, and Christo Dichev<sup>2</sup>

<sup>1</sup> University of Southern Denmark,  
Institute of Business Communication and Information Science  
Engstien 1, 6000 Kolding, Denmark

<sup>2</sup> Winston-Salem State University, Computer Science Department  
601 S. Martin Luther King J. Drive, Winston Salem, NC 27110, USA  
email: larsjo@sitkom.sdu.dk, {dichevad,dichevc}@wssu.edu

**Abstract.** This paper focuses on the visualization and presentation of *story maps* and suchlike narrative information in TM4Book, a Topic Maps-based semantic annotation tool for creative writing and textual analysis. Drawing on common gestalt principles as well as notions from semiotics and current discourse theory, the paper proposes some general strategies for displaying visual story maps and making them interactive for the user.

## 1 Introduction

TM4Book is a Topic Maps-based semantic annotation tool intended to offer ontological support to writers and readers engaged in planning, structuring and analyzing text, especially narrative prose. Writers and readers will be able to create a semantic index of pertinent information in the text or story and to inspect (subsets of) this information through some kind of visual representation. Visualizing semantic aspects of story-telling (stored in topic maps), however, poses particular challenges. Firstly, dissimilar graphical configurations are needed to convey different perspectives, or views, of the story (such as timelines versus storylines), and secondly, a unifying approach is required to consistently realize diverse meaning types in narrative discourse domains (notably ontological, narrative and rhetorical meaning) as visually cohesive and coherent representations. This paper presents some of these challenges and attempts to point to some possible solutions using storylines as the primary example throughout.

## 2 Fabula Concepts

Ontological support in TM4Book is currently organized around three layers, the item layer, the text layer and the fabula layer [1]. The *item layer* lists entities of relevance in the text. The *text layer* specifies the formal structure of the narrative, parts, chapters, scenes, and so on while the *fabula layer* describes “what happens in the story”. The fabula layer comprises concepts such as event, plot, subplot and storylines. In the current context, *events* are understood as the basic actions that

the characters of a story are involved in. Although a character may perform a range of roles in a specific action (murderer, victim, beneficiary, etc.), it is often enough to indicate who is the active part - “the subject” - of an action and who is the target or “object”. In addition to subjects and objects, descriptions of events may, and often do, carry ancillary information specifying where, when, why and how a certain action is performed or triggered. Events are not only linked in time but are related in a number of ways: they occur because of, or as a result of, other events; they are solutions to specific problems or situations; they have a purpose, and are driven by motivations known or unknown to the reader. Events are not always presented in chronological order but can be rendered as flashbacks or flash-forwards. The *plot* is the set of (central) events in a story. It can be divided into subsets of events having specific narrative purposes such as exposition, rising action, climax, etc. Structurally, a plot may contain *subplots*, narrative threads within the main story often involving minor characters. *Storylines* are events experienced by individual characters, or sets of characters, and thus constitute a certain perspective or view of the plot or parts of the plot. Creative writers can use storylines as a major structuring tool to plan plots and subplots while readers engaged in analysis can employ storylines in describing individual characters, their development and relations to other characters. Storylines can also be used by game designers for creating customizable games. Therefore, being able to present and visualize storylines in a tool like TM4Book becomes a central task.

### 3 Visualization Tools for Topic Maps

Since TM4Book is a Topic Maps-driven tool, the obvious candidates for visualizing the semantic aspects of storylines are techniques and tools suitable for visualizing conceptual structures expressible as topic maps. The early examples of visualizations of web resources include the Hyperbolic Tree [4] for navigating large trees and the Brain<sup>1</sup> for navigating graphs. Another example is Hypergraph<sup>2</sup>, a Java application that provides a hyperbolic layout in 2D allowing interactive repositioning of nodes to provide more magnification to regions of interest. These visualizations focus mainly on syntactic structures, such as link structures. The current generation of tools moves the emphasis to interfaces for manipulating information. For example, systems such as Haystack [3] are emerging that focus on concepts important to the users: documents, messages, properties, annotations, etc.

In the field of Topic Map-based applications, one of the first interactive Topic Map visualization tools was implemented by Le Grand and Sotto [6]. This tool supports visual and navigational forms. However, the presentation is not easily comprehensible and intuitive. TMNav<sup>3</sup> is a combined text and graphical topic map browser that allows users to easily navigate through a topic map. It is based on the TM4J Topic Map library and uses Java Swing, TouchGraph and HyperGraph. The

---

<sup>1</sup> <http://www.thebrain.com>

<sup>2</sup> <http://vtmlgraph.i-scream.org.uk/>

<sup>3</sup> <http://tm4j.org/tmnav.html>

Vizigator<sup>4</sup>, is Ontopia's generic topic map browser, which provides a graphical alternative to the text browsing environment. ThinkGraph<sup>5</sup> is a 2D drawing application specialized for concept maps authoring. It uses standard XML: SVG (Scalable Vector Graphics, a XML language specialized for 2D drawing) for the presentation part (shape and graphical attributes) and XTM for the data part.

The nature of information visualization can be considered as depending on the visual metaphors that it uses to structure information. The process of understanding visualization therefore involves an interaction between these external visual metaphors and the adopted representations. To the best of our knowledge, no tools or conventions for graphically conveying *narrative* meaning structures (based on Topic Maps) exist. In the following we therefore propose a set of principles and strategies for organizing the visual space as a means of communicating aspects of story-telling.

## 4 Visualizing and Presenting Story Maps

A visual rendering of a narrative structure such as a storyline should not only give the viewer a graphic overview of a selected set of events and the characters involved in these events but also, if possible, the basic logical and ontological properties of these events and characters. Also, since a storyline is the perspective of one character, or one set of characters, this perspective ought to be salient in the visual representation. Furthermore, the visualization might somehow indicate the relation between the text layer and the fabula layer, i.e. the correlation between the events of the story, the plot and the formal structure of the text. Last but not least, the user should be able to see the chronological order of events unfolding in the story presented on a timeline.

We are seeking to develop strategies for integrating and visually conveying such text elements in what we call *story maps*, a kind of narrative concept maps [7]. These strategies are centered on:

- Creating and linking visual objects.
- Positioning these objects in a limited two-dimensional space.
- Enabling the user to interact with the visual information through an interface.

As for the generation and grouping of visual objects, our approach takes into account inherent human perception of visual input as formulated in the *gestalt principles* of similarity, proximity, connectedness, common region, good continuation, reinforcement, etc. (see for instance [2]). These principles state that we as humans naturally tend to group visual objects if they are similar in shape, size or color; if they are proximate to one another, connected by lines, framed in some way or form some kind of line. And this perception of visual unity is reinforced if two or more principles apply at the same time, say similarity and proximity.

---

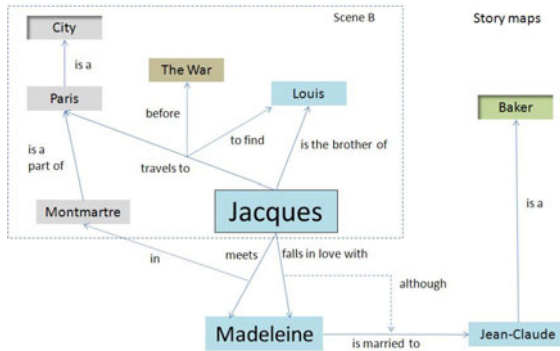
<sup>4</sup> <http://www.idealliance.org/proceedings/xml04/papers/311/311.html>

<sup>5</sup> <http://www.thinkgraph.com/english/index.htm>

To further specify or derive the topology of a story map, i.e. the relative positions of all the visual objects in the map, we employ the notion of *polarized zones* taken from social semiotics [9]. This idea suggests that we, at least in Western culture, tend to assign certain semantic connotations to spatial zones if contrasted. The upper and lower parts of a visual field indicate the *ideal* versus the *real*; the left and the right designate *given* versus *new*, while the semantic contrast of *centrality* and *periphery* can be realized by positions in the middle of the field and in the corners, respectively.

In determining which elements should be allowed to be toggled on and off in a visual representation of a narrative structure, we draw, albeit loosely, upon the *dialogic principle* of Renkema’s connectivity model of discourse [8]. Simply put, the dialogic principle regards discourse and text as a kind of imaginary dialogue between the speaker or writer and the addressee in which new clauses and sentences may be interpreted as a speaker’s or writer’s responses to an addressee’s hypothetical questions or requests. In the same vein, displaying new visual information should arguably answer specific questions by the user about the story and how it unfolds in time.

How these “meta-principles” may be translated into a unifying approach for generating story maps (from topic maps) is exemplified in Figure 1. The example visualizes a small set of ontological facts and events taking place in a simple fictitious story set in pre-war Paris. They constitute the exposition part of the plot and the first couple of events in the storyline of Jacques, the main character.



**Fig. 1** Visual representation of a segment of a story map

We propose the following design strategies or conventions in order to satisfy the meta-principles mentioned above:

1. Subjects and objects of events (e.g. “Jacques” travels to “Paris”) are linked directly, while ancillary information about events is attached indirectly as a kind of satellites (e.g. before “the War”).
2. Items of the same semantic type or items serving the same purpose must be visually similar (e.g. objects designating ontological classes like “city” and “baker” are shown differently from instances like “Paris” and “Jean-Claude”). Likewise, items linked by more significant relations (with

- respect to the storyline logic) are more proximate than less significantly related ones (e.g. “Jacques” is closer to “Madeleine” than “city” )
3. The top/bottom axis indicates hierarchical order whereas the left/right axis indicates narrative order (e.g. “city” is placed on a higher level than “Paris” and the line denoting Jacques’ meeting Madeleine is placed left to the one signifying their subsequent falling in love).
  4. Grouping of related events or characters may be done through applying the gestalt principle of common region (e.g. framing events occurring in scene B). Linking of related events is realized by labeled dotted lines between the lines manifesting the constituent events (e.g. Jacques falls in love with Madeleine “although” she is married to Jean-Claude).
  5. Centrality imposes perspective and relevance, and size indicates salience (e.g. it being Jacques’ storyline he holds the central position of the map and his name is shown in a bigger font-size than Madeleine’s whose name in turn is more salient than Jean-Claude’s).
  6. To instantiate the dialogical principle, the user should be able to progressively display the events as they occur in the story through the manipulation of controls on the interface (not shown in the figure).

## 5 Conclusion

In this paper we proposed a set of design conventions or requirements for visual representations of story-telling. In order to consistently meet these requirements in the development of the TM4Book visualization, we have explored various technological solutions, including existing open source concept map tools and HTML5 with its canvas element and capability to embed SVG. After careful assessment and testing, we decided to build the TM4Book visualization tool based on the open source Visual Understanding Environment (VUE, <http://vue.tufts.edu/>). This approach entails a conversion from TM4Book’s standardized Topic Maps format XTM to VUE’s internal XML format, in effect mapping semantic properties and relations onto a visual configuration. While integrating VUE into TM4Book provides functionality for accessing and manipulating conceptual structures, such as layering, zooming, dimming, searching, etc. “for free”, the complex design of VUE as a very general standalone multi-functional application definitely presents challenges in implementing the additional desired functionality. We believe, however, that it still presents a good basis for realizing the proposed design requirements for visual rendering of story maps as it affords means for creating ‘core’ story maps programmatically as well as personalizing or fine-tuning them manually through an intuitive interface.

## References

1. Dicheva, D., Dichev, C., Damova, M.: Ontological Support for Creative Writing. In: The 24th International FLAIRS Conference, Palm Beach, FL, pp. 576–577 (2011)
2. Horn, R.E.: Visual language: Global Communication for the 21st Century. Macrovue, Inc. (1999)

3. Karger, D., Bakshi, K., Huynh, D., Quan, D., Sinha, V.: Haystack: A Customizable General-Purpose Information Management Tool for End Users of Semistructured Data. CIDR (2003)
4. Lamping, J., Rao, R., Pirolli, P.: A focus+context technique based on hyperbolic geometry for visualising large hierarchies. In: ACM Conf. CHI 1995, pp. 401–408 (1995)
5. Le Grand, B., Soto, M.: Topic Maps Visualization. In: Geroimenko, V., Chen, C. (eds.) Visualizing the Semantic Web. Springer, London (2003)
6. Novak, J.D., Canas, A.J.: The Theory Underlying Concept Maps and How to Construct and Use Them (2008),  
<http://cmap.ihmc.us/publications/researchpapers/theorycmaps/theoryunderlyingconceptmaps.htm>
7. Renkema, J.: The Texture of Discourse. John Benjamins Publishing Company, Amsterdam (2009)
8. Van Leeuwen, T.: Introducing Social Semiotics. Routledge, New York (2004)
9. Ziemkiewicz, C., Kosara, R.: The Shaping of Information by Visual Metaphors. IEEE Trans. Vis. Computational Graph, 1269–1276 (2008)

# A Fuzzy Bat Clustering Method for Ergonomic Screening of Office Workplaces

Koffka Khan<sup>1</sup>, Alexander Nikov<sup>1</sup>, and Ashok Sahai<sup>2</sup>

The University of the West Indies

<sup>1</sup>Department of Computing and Information Technology,

<sup>2</sup>Department of Mathematics and Statistics

St. Augustine, Trinidad and Tobago, W.I.

email: {koffka.khan,alexander.nikov,ashok.sahai}@sta.uwi.edu

**Abstract.** A method for screening of company workplaces with high ergonomic risk is developed. For clustering of company workplaces a fuzzy modification of bat algorithm is proposed. Using data gathered by a checklist from workplaces, information for ergonomic related health risks is extracted. Three clusters of workplaces with low, moderate and high ergonomic risk are determined. Using these clusters, workplaces with moderate and high ergonomic risk levels are screened and relevant solutions are proposed. By a case study this method is illustrated and validated. Important advantages of the method are reduction of computational effort and fast screening of workplaces with major ergonomic problems within a company.

**Keywords:** Clustering, Screening, Fuzzy, Bat algorithm, Ergonomics, Workplaces, Information Extraction, Health Risk.

## 1 Introduction

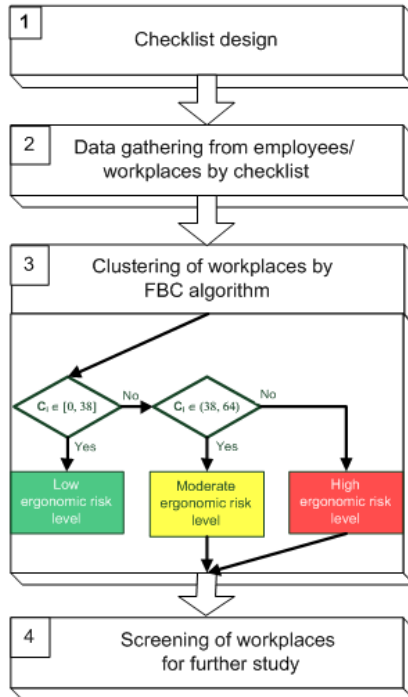
In the last few years, there has been increasing recognition of the importance of ergonomics in office workplace settings [3]. Ergonomic risks at the workplace cause a lot of damage to health. Deterioration in the quality of life of employees results in an economic burden to employers and the economy as a whole [11]. In Europe ergonomic health risks (HS) accounts for a higher proportion of work absences due to illness/injury than any other health condition. Consequently, many office workplaces are poorly designed. For this there is a need to measure the ergonomic risk level for workplaces using a model, which enables clustering of workplaces and screening of highly risky workplaces.

Clusters correspond to the hidden patterns in data (groups/departments of workplaces with similar ergonomic risk). The data clustering has been approached from diverse fields of knowledge like statistics (multivariate analysis) [5], graph theory [16], expectation maximization algorithms [1], artificial neural networks [10] and evolutionary computing [12]. It has been shown in [4] that the clustering problem is NP-hard when the number of clusters exceeds 3.

Clustering of office workplaces according to ergonomic risk level can be well formulated as an optimization problem. Swarm optimization techniques have already been applied to clustering problems [15], [13] and [9]. We propose to combine a clustering algorithm with a swarm optimization algorithm for ergonomic screening of office workplaces. In this paper a method integrating fuzzy c-means clustering algorithm with the bat algorithm for solving this problem is suggested.

## 2 Method Description

For clustering of office workplaces according to their ergonomic risk a method with four steps is proposed (cf. Fig. 1). It includes a checklist (step 1) in which ergonomic dimensions and items/questions are determined. At step 2 data is gathered using this checklist. At step 3, using a heuristic fuzzy swarm model incorporating bat algorithm and fuzzy c-means clustering algorithm, the ergonomic workplace risks are determined. Workplace risk ranges are defined using this method with low, moderate and high ergonomic risk levels. At step 4 workplaces with high ergonomic risk based on clusters are screened for further detailed study. In the following the fuzzy c-means clustering algorithm, the swarm optimization bat algorithm and a fuzzy modification of bat algorithm for ergonomic clustering of office workplaces are presented.



**Fig. 1** Steps of method for ergonomic screening of office workplaces



## 2.1 Fuzzy c-Means (FCM) Clustering Algorithm

The fuzzy c-means (FCM) clustering algorithm [8] generates fuzzy partitions for any set of numerical data, allowing one piece of data to belong to two or more clusters. FCM partitions a set of patterns  $X_i = \{x_1, x_2, \dots, x_n\}$  with  $n$  features [7] into  $c$  ( $1 < c < n$ ) fuzzy clusters with a set of cluster centers  $Z = \{z_1, z_2, \dots, z_c\}$  each being initialized.

$$z_j = \frac{\sum_{i=1}^n \mu_{ij}^m p_i}{\sum_{i=1}^n \mu_{ij}^m} \quad (1)$$

Here, the membership degree  $\mu_{ij} \in [0, 1]$  quantifies the grade of membership of the  $i$ th pattern to  $j$ th cluster. The aim of FCM is to minimize the objective function  $J_{fcm}$  with  $d_{ij}$  being the Euclidean distance [6] measure taken from pattern feature data point  $x_i$  to the cluster center  $z_j$ .  $m$  ( $m > 1$ ) is a scalar which controls the fuzziness of the resulting clusters.

$$J_{fcm} = \sum_{j=1}^c \sum_{i=1}^n \mu_{ij}^m d_{ij} \quad (2)$$

$$d_{ij} = \|x_i - z_j\| \quad (3)$$

The membership degree is  $\mu$ .

$$\mu_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \quad (4)$$

## 2.2 Bat Algorithm (BA)

The bat algorithm [14] uses the echolocation behaviour of bats. These bats emit a very loud sound pulse (echolocation) and listens for the echo that bounces back from the surrounding objects. The  $i$ th bat flies randomly with velocity  $v_i$  at position  $X_i$  with a fixed frequency  $f_{min}$ . The bat varies its wavelength  $\lambda$  and loudness  $A_0$  to search for food. The number of bats is  $n_b$ .

$$f_i = \frac{v_i}{\lambda_i} \quad (5)$$

$$f_i = f_{min} + (f_{max} - f_{min}) \delta \quad (6)$$

$$v_i^t = v_i^{t-1} + (X_i^t - X_{gbest}^t) f_i \quad (7)$$

It is assumed that the loudness varies from a large (positive)  $A_0$  to a minimum constant value  $A_{min}$ . The new solutions of the  $i$ th bat at time step  $t$  are given by  $X_i^t$  and  $v_i^t$ .

$$X_i^t = X_i^{t-1} + v_i^t \quad (8)$$

where  $\delta$   $[0, 1]$  is a random vector drawn from a uniform distribution. For local search procedure (exploitation) each bat takes a random walk creating a new solution for itself based on the best selected current solution.

$$X_{\text{new}} = X_{\text{old}} + \rho A^t \quad (9)$$

where  $\rho \in [-1, 1]$  is a random number,  $A^t$  is the average loudness of all bats at this time step. The loudness decreases as a bat tends closer to its food and pulse emissions rate increases.

$$A_i^{t+1} = \alpha A_i^t \quad (10)$$

$$r_i^{t+1} = r_i^0 [1 - e^{-\gamma t}] \quad (11)$$

where  $\alpha$  and  $\gamma$  are constants.

### 2.3 Fuzzy Bat Clustering (FBC) Algorithm

A modified fuzzy c-means bat algorithm for cluster analysis of office workplaces risk is proposed. The velocity update of the bat is  $v_i$ .

$$v_i^t = v_i^{t-1} + (X_i^t - X_{\text{gbest}}^t) f_i + (X_i^t - X_{\text{lbest}}^t) f_i \quad (12)$$

where  $X_{\text{gbest}}^t$  is the global best of all the bats and  $X_{\text{lbest}}^t$  is the local best of each bat. Any single bat is following the best hunting position found by not only taking all bats into consideration, but also its own preference when searching for food. The reason for this added parameter in the velocity equation is because by choosing its own local hunting area the exploitation of the algorithm will be increased and hence better clustering performance will be achieved by similar bats (workplace vectors) being pulled closer together. The position  $X_i$  and velocities  $v_i$  (cf. (7) and (8)) of bats are redefined to represent the fuzzy relation between them. More fitted cluster partitions are sampled from the search space, that is, those with higher FCM fuzzy values have higher probability of being sampled. The FBC tends to perform better search than FCM. This is because it uses information on the quality of previously assessed partitions to potentially generate better partitions which are not used with FCM.

For evaluating the generalized solutions of the FBC algorithm's fitness function  $f(X)$  the objective function  $J_{\text{fcm}}$  of the FCM algorithm is used:

$$f(X) = \frac{K}{J_{\text{fcm}}} \quad (13)$$

where  $K$  is a constant. The smaller is  $J_{\text{fcm}}$ , the better is the clustering effect and the higher is the individual fitness. FBC algorithm pseudo code is shown on Figure 2.

### FBC algorithm

---

```

initialize the parameters of BA (population size  $n_b$ , frequency  $f$ , pulse rate  $r$ ,
loudness  $A$  and dimensions  $n$ )
initialize the parameters of FCM ( $m > 1$ ,  $\mu_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, c$ )
create a swarm with  $n_b$  bats
for each bat
  initialize  $X_i$  ( $i = 1, 2, \dots, n$ ) and  $v_i$ 
  define pulse frequency  $f_i$  at  $X_i$ 
  initialize pulse rates  $r_i$  and the loudness  $A_i$ 
end for
initialize current-global-best and current-local-best for the swarm
repeat
  calculate the cluster centers for each bat (1)
  generate new solutions by adjusting frequency and updating velocities and
  locations (5), (6), (8) and (12)
  if (rand >  $r_i$ ) then
    select bats with best solutions
    generate a local solution around these best solutions (9)
  end if
  generate a new solution by flying randomly (5), (6), (8) and (12)
  if (rand <  $A_i$  &  $f(X_i) < f(X_{g_{best}})$ ) then
    accept new solutions (8)
    increase  $r_i$  and reduce  $A_i$  (10) and (11)
  end if
  compute Euclidian distance  $d_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, c$ ; for each bat (3)
  update the membership degree  $\mu_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, c$ ; for each bat (4)
  calculate objective function value (13)
  rank the bats and find the current-global-best and current-personal-best locations
until terminating condition is met

```

---

**Fig. 2** FBC fuzzy clustering

FBC algorithm applies FCM to the hoard of bats such that velocities or cluster centres are improved with each iteration/hunting cycle. This leads to improved fitness of clusters. The FCM algorithm is faster than the FBC algorithm because it requires fewer function evaluations, but it can fall into local optima. FCM when used in combination with BA algorithm to form an optimized clustering algorithm FBC will bypass local optima and tend towards global best solutions. The resulting effect is that after each successive iteration loop the resulting partitions provide better prototypes for the FCM, thus reducing the probability of getting stuck in local optima. Thus the FBC uses fuzzy c-means as a local search procedure, which performs the fine-tuning of bat movements obtained by hunting bats, thus speeding up performance. The FCM provides additional information about the spatial distribution of the data contained in the fuzzy partition matrix  $\mu$  and minimizes the variances of the clusters achieved. This yields more compact clusters. The FBC fosters cumulative refinement of fuzzy partitions resulting in reasonable FCM iterations (roughly  $t = 10$ ) for each local search.

Using statistical information theory [2] as a measure of information pooling for the positioning of the individual data points to the three cluster centroids the

coefficient of variation  $CV_i$  was incorporated into the FBC algorithm and calculations for ergonomic risk ranges are defined for this dataset as the point at which most information is captured:

$$P = \frac{CV_1 * z_1 + CV_2 * z_2 + \dots + CV_i * z_i}{CV_1 + CV_2 + \dots + CV_i} \quad (14)$$

where  $CV_i = \sigma_i/\mu_i$  the coefficient of variation of the  $i$ th cluster, where  $\sigma_i$  is the standard deviation of  $i$ th cluster,  $\mu_i$  is the mean of the  $i$ th cluster, and  $z_i$  is the  $i$ th cluster center of workplaces.

### 3 Case Study

For illustration and validation of our proposed method data was collected from a company with computer-aided office workplaces by using an online checklist of 18 questions. Pre-tests of the checklist with employees were done to ensure that the questions were clear, succinct, and unambiguous. Unclear questions were reworded or removed. Data was gathered over a three weeks period. 212 responses were received.

In order to optimize the performance of the FBC, fine tuning has been performed and best values for their parameters are selected. Based on experimental results the FBC algorithm performs best under the following settings:  $A_0 = 1.0$ ,  $A_{\min} = 0$ ,  $r^0_i = 0.01$ ,  $\alpha = \gamma = 0.6$ ,  $f \in [1, 5]$ . The optimization iterations stop when error reaches  $\varepsilon = 10^{-12}$ . Other FBC terminating conditions are the maximum number of iterations 1500, no changes in  $X^t_{\text{gbest}}$  in 500 consecutive iterations, and no  $X^t_{\text{gbest}}$  improvement in 2 consecutive iterations. The weighting exponent  $m=2$ .

We varied the number of clusters  $c$  ( $c=3$ ,  $c=5$ ,  $c=6$ ). Three clusters solution ( $c=3$ ) gave the best result. For these three clusters the coefficient of variation is ( $CV_1=0.04$ ,  $CV_2=0.03$ ,  $CV_3=0.01$ ), the standard deviation is ( $\sigma_1=2.00$ ,  $\sigma_2=1.60$ ,  $\sigma_3=0.58$ ), the mean is ( $\mu_1=22$ ,  $\mu_2=51$ ,  $\mu_3=83$ ), and cluster center of workstations is ( $z_1=15$ ,  $z_2=69$ ,  $z_3=162$ ). Even though the clusters extracted is reasonably small it is appropriate for testing such new metaheuristic algorithms and the results are formally validated using standard statistical techniques and benchmark testing functions.

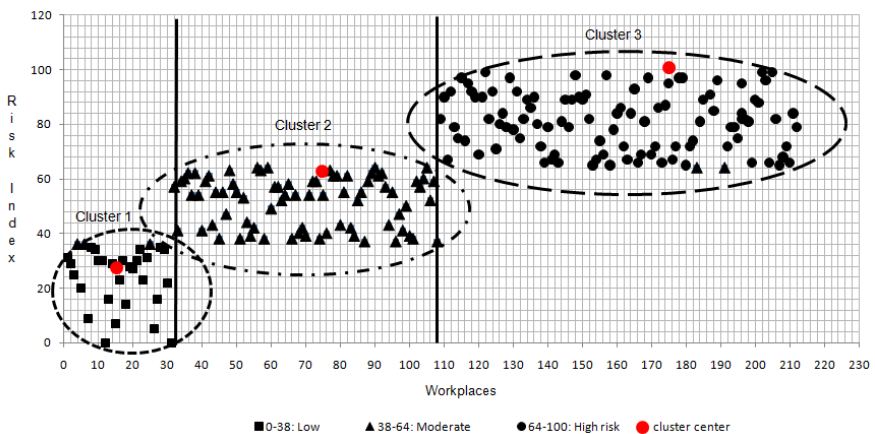
The point with maximum information  $P=51$  and the mean  $\mu=13$ . The ergonomic risk index  $I_r$  range for the first workplaces cluster is  $(0, P - \mu) = (0, 38)$ ; for the second workplaces cluster is  $(P - \mu, P + \mu) = (38, 64)$ ; and for third workplaces cluster is  $(P + \mu, 100) = (64, 100)$  (cf. Table 1).

Three clusters of workplaces are shown (cf. Fig. 3). The first cluster contains 27 workplaces (13%) belonging to the green range with low ergonomic risk indices. Here minor ergonomic improvements are needed. The second cluster contains 83 workplaces (39%) belonging to the yellow range with moderate ergonomic risk indices. The third cluster contains 102 workplaces (48%) belonging to the red range with high ergonomic risk indices. The last two clusters of workplaces need further study for defining improvement measures according to

**Table 1** Ergonomic risk index rating scale

Cluster #	1	2	3
Risk level	Low	Moderate	High
Risk range	[0, 38]	(38, 64)	[64, 100]
Color scale			

step 4 of our method. There is some overlapping of clusters, e.g. some workplaces with moderate and high risk indices are allocated to the first cluster (workplaces with low ergonomic risk); two workplaces with moderate risk indices are allocated to the third cluster (workplaces with high ergonomic risk).

**Fig. 3** Workplaces clusters determined by FBC

## 4 Conclusion

A method for screening of company workplaces with high ergonomic risk is proposed. For clustering of company workplaces a novel fuzzy modification of the bat algorithm is developed. Using data gathered by a checklist from workplaces, ergonomic related health risks are determined. Three clusters of workplaces with low, moderate and high ergonomic risk are defined. Using these clusters, workplaces with moderate and high ergonomic risk levels are screened and relevant solutions are proposed. By a case study this method is illustrated and validated. The suitability of fuzzy bat clustering algorithm for ergonomic risk screening is demonstrated.

Advantages of this method are: 1) fast and effective screening of workplaces with major ergonomic problems within a company; 2) better performance of fuzzy bat clustering algorithm than fuzzy c-means clustering algorithm. Future research

is needed to automatically adapt the FBC parameters, number of clusters and to make comparisons with swarm intelligent and other clustering algorithms.

## References

1. Aci, M., Avci, M.: K nearest neighbor reinforced expectation maximization method. *Expert Systems with Applications* 38(10), 12585–12591 (2011)
2. Arwini, K., Dodson, J.T.C.: *Information Geometry: Near Randomness and Near Independence*. Springer, Heidelberg (2008)
3. Levy, B., et al.: *Occupational and Environmental Health: Recognizing and Preventing Disease*. Oxford University Press, Oxford (2011)
4. Barthelemy, J., Brucker, F.: NP-hard Approximation Problems in Overlapping Clustering. *Journal of classification* 18(21), 59–183 (2001)
5. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21, 768–769 (1965)
6. Gower, J.C.: Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications* 67, 81–97 (1985)
7. Bezdek, J.C.: Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology* 1, 57–71 (1984)
8. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences* 10, 191–203 (1984)
9. Li, H., Li, J., Kang, F.: Risk analysis of dam based on artificial bee colony algorithm with fuzzy c-means clustering. *Canadian Journal of Civil Engineering* 38(5), 483–492 (2011)
10. Lorenzo, J., Mendez, J., Castrillon, M., Hernandez, D.: Short-term wind power forecast based on cluster analysis and artificial neural networks. *Advances in Computational Intelligence* 6691(1), 191–198 (2011)
11. Lundgren, E.R., McMakin, H.A.: *Risk communication: a handbook for communicating environmental, safety, and health risks*, vol. 1, pp. 1–7 (2009)
12. Paterlini, S., Minerva, T.: Evolutionary approaches for cluster analysis. *Soft Computing Applications*, 167–178 (2003)
13. Runkler, A.T.: Ant colony optimization of clustering models. *International Journal of Intelligent Systems* 20(12), 1233–1251 (2005)
14. Yang, X.-S.: A New Metaheuristic Bat-Inspired Algorithm. In: Yang, X.-S. (ed.) *Nature Inspired Cooperative Strategies for Optimization (NICSO)*. *Studies in Computational Intelligence*, vol. 284, pp. 65–74 (2010)
15. Li, X.Y., et al.: Fuzzy clustering algorithm based on improved particle swarm optimization and its application. *Journal of Applied Mechanics and Materials*, 44–47 (2010)
16. Yoshida, T., Okatani, K.: A graph-based projection approach for semi-supervised clustering. In: Kang, B.-H., Richards, D. (eds.) *PKAW 2010*. *LNCS*, vol. 6232, pp. 1–13. Springer, Heidelberg (2010)

# The Influence of Global Constraints on DTW and LCS Similarity Measures for Time-Series Databases

Vladimir Kurbalija<sup>1</sup>, Miloš Radovanović<sup>1</sup>, Zoltan Geler<sup>2</sup>,  
and Mirjana Ivanović<sup>1</sup>

<sup>1</sup> Department of Mathematics and Informatics, Faculty of Science,  
University of Novi Sad, Trg D. Obradovica 4, 21000 Novi Sad, Serbia  
email: {kurba, radacha, mira}@dmi.uns.ac.rs

<sup>2</sup> Faculty of Philosophy,  
University of Novi Sad, Dr Zorana Đinđića 2, 21000 Novi Sad, Serbia  
email: gellerz@gmail.com

**Abstract.** Analysis of time series represents an important tool in many application areas. A vital component in many types of time-series analysis is the choice of an appropriate distance/similarity measure. Numerous measures have been proposed to date, with the most successful ones based on dynamic programming. Being of quadratic time complexity, however, global constraints are often employed to limit the search space in the matrix during the dynamic programming procedure, in order to speed up computation. In this paper, we investigate two representative time-series distance/similarity measures based on dynamic programming, Dynamic Time Warping (DTW) and Longest Common Subsequence (LCS), and the effects of global constraints on them. Through extensive experiments on a large number of time-series data sets, we demonstrate how global constraints can significantly reduce the computation time of DTW and LCS. We also show that, if the constraint parameter is tight enough (less than 10–15% of time-series length), the constrained measure becomes significantly different from its unconstrained counterpart, in the sense of producing qualitatively different 1-nearest neighbour (1NN) graphs. This observation highlights the need for careful tuning of constraint parameters in order to achieve a good trade-off between speed and accuracy.

## 1 Introduction

In many scientific fields, a time series consists of a sequence of values or events obtained over repeated measurements of time [1]. Time-series analysis is comprised of methods that attempt to understand time series, to explain the underlying context of the data points or to make forecasts.

Time-series databases are popular in many applications, such as stock market analysis, economic and sales forecasting, budgetary analysis, process and quality control, observation of natural phenomena, scientific and engineering experiments, medical treatments, etc. As a consequence, the last decade witnessed an increasing interest in querying and mining such data, which resulted in a large amount of work introducing new methodologies for different task types including: indexing, classification, clustering, prediction, segmentation, anomaly detection, etc. [1, 2, 3, 4]

One of the most important aspects of time-series analysis is the choice of appropriate similarity/distance measure – the measure which tells to what extent two time series are similar. However, unlike data types in traditional databases where the similarity/distance definition is straightforward, the distance between time series needs to be carefully defined in order to reflect the underlying (dis)similarity of these specific data, which is usually based on shapes and patterns. As expected, there exists a large number of measures for expressing (dis)similarity of time-series data proposed in the literature, e.g., Euclidean distance (ED) [2], Dynamic Time Warping (DTW) [5], distance based on Longest Common Subsequence (LCS) [6], Edit Distance with Real Penalty (ERP) [7], Edit Distance on Real sequence (EDR) [8], Sequence Weighted Alignment model (Swale) [9].

Many of these similarity measures are based on dynamic programming. It is well known that the computational complexity of dynamic programming algorithms is quadratic, which is often not suitable for larger real-world problems. However, the usage of global constraints such as Sakoe-Chiba band [21] and Itakura parallelogram [22] can significantly speed up the calculation of similarities. Furthermore, it is also reported [10] that the usage of global constraints can improve the accuracy of classification compared to unconstrained similarity measures. The accuracy of classification is commonly used as a qualitative assessment of a similarity measure.

In this paper we will investigate the influence of global constraints on two most representative similarity measures for time series based on dynamic programming: DTW and LCS. We will report the calculation times for different sizes of constraints in order to explore the speed-up gained from these constraints. Also, the change of the 1-nearest neighbour graph will be explored with respect to the change of the constraint size. The proposed research will provide a better understanding of global constraints and offer deeper insight into their advantages and limitations.

All experiments presented in this paper are performed using the system FAP (Framework for Analysis and Prediction) [11]. The data for experiments is provided by the UCR Time Series Repository [12], which includes the majority of all publicly available, labelled time-series data sets in the world.

The rest of the paper is organized as follows. The next section presents the necessary background knowledge about similarity measures and gives an overview of related work. Section 3 briefly describes the FAP system used for performing experiments. The methodology and results of extensive experiments are given in Section 4. Section 5 concludes the paper and presents the directions for further work.

## 2 Background and Related Work

The Euclidean metric is probably the most intuitive metric for time series, and as a consequence very commonly used [2, 13, 14, 15, 16]. In addition, it is also very fast – its computation complexity is linear. The distance between two time series is calculated as a sum of distances between corresponding points of two time series. However, it became evident that this measure is very brittle and sensitive to small translations across the time axis [10, 17].

Dynamic Time Warping (DTW) can be considered as a generalization of Euclidian distance where it is not necessary that the  $i$ -th point of one time series must be aligned to the  $i$ -th point of the other time series [10, 17, 18, 19]. This



method allows elastic shifting of the time axis where in some points time “warps”. The DTW algorithm computes the distance by finding an optimal path in matrix of distances between points of two time series. The Euclidian distance can be seen as special case of DTW where only the elements on the main diagonal of the matrix are taken into account.

Longest Common Subsequence (LCS) applies a different methodology. According to LCS, the similarity between two time series is expressed as the length of the longest common subsequence of both time series [20].

Both DTW and LCS are based on dynamic programming – the algorithms seek the optimal path in the search matrix. The types of matrices are different but the approach is the same. DTW examines the matrix of distances between points, while LCS examines the matrix of longest common subsequences of different-length subseries. As a consequence, both algorithms are quadratic. However, the introduction of global constraints can significantly improve the performance of these algorithms. Global constraints narrow the search path in the matrix, which results in a significant decrease in the number of performed calculations. The most frequently used global constraints are the Sakoe-Chiba band [21] and the Itakura parallelogram [22]. These constraints were introduced to prevent some bad alignments, where a relatively small part of one time series maps onto a large section of another time series.

The quality of similarity measures is usually evaluated indirectly, e.g. by assessment of classifier accuracy. The simple method combining the 1NN classifier and some form of DTW distance was shown to be one of the best-performing time-series classification techniques [4, 17, 18, 23]. In addition, the accuracy of 1NN directly reflects the quality of a similarity measure. Therefore, in this paper we report the calculation times for unconstrained and constrained DTW and LCS, and we focus on the 1NN graph and its change with regard to the change of constraints. The influence of global constraints is not investigated well in the literature, and the results presented in this paper will provide a better understanding of theirs essence.

### 3 The FAP System

There are three important concepts which need to be considered when dealing with time series: pre-processing transformation, time-series representation and similarity measure. The task of pre-processing transformations is to remove different kinds of distortions in raw time series. The task of time-series representation is to reduce the usually very high dimensionality of time series while preserving their important properties. Finally, the task of a similarity measure is to reflect the essential similarity of time series, which are usually based on shapes and patterns.

All these concepts, when introduced, are usually separately implemented and presented in different publications. Every newly-introduced representation method or distance measure has claimed a particular superiority [4]. However, this was usually based on comparison with only a few other representatives of the proposed concept. On the other hand, to the best of our knowledge there is no freely available system for time-series analysis and mining which supports all mentioned concepts, with the exception of the work proposed in [4]. Being motivated by these observations, we

have designed a multipurpose, multifunctional system FAP – Framework for Analysis and Prediction [11]. FAP supports all mentioned concepts: representations, similarity measures and pre-processing tasks; with the possibility to easily change some existing or to add new concrete implementation of any concept.

At this stage of development, all main similarity measures ( $L_p$ , DTW, CDTW (Constrained DTW), LCS, CLCS, ERP, CERP, EDR, CEDR and Swale) are implemented, and the modelling and implementation of representation techniques is in progress. All constrained measures employ the Sakoe-Chiba band. Furthermore, several classifiers and statistical tests are also implemented.

## 4 Experimental Evaluation

In this section we will investigate the influence of global constraints on two most illustrative similarity measures based on dynamic programming: DTW and LCS. Furthermore, two aspects of applying global constraints are considered: efficiency and effectiveness of the 1NN classifier for different values of constraints. For both similarity measures, the experiments are performed with the unconstrained measure and a measure with the following constraints: 75%, 50%, 25%, 20%, 15%, 10%, 5%, 1% and 0% of the size of the time series. This distribution was chosen because it is expected that measures with larger constraints behave similarly to the unconstrained measure, while smaller constraints exhibit more interesting behaviour [10, 18].

A comprehensive set of experiments was conducted on 38 data sets from [12], which includes the majority of all publicly available, labelled time-series data sets currently available for research purposes. The length of time series varies from 24 to 1882 depending on the data set. The number of time series per data set varies from 60 to 9236.

### 4.1 Computational Times

In the first experimental phase we wanted to investigate the influence of global constraints on the efficiency of calculating the distance matrix. The distance matrix for one data set is the matrix where element  $(i,j)$  contains the distance between  $i$ -th and  $j$ -th time series from the data set. The calculation of the distance matrix is a time-consuming operation, which makes it suitable for measuring the efficiency of global constraints.

In Table 1, the calculation times of DTW in milliseconds are given for some datasets and for different values of constraints. Table 2 contains the same data for the LCS measure. The complete tables are available in extended version of the paper at Computing Research Repository – CoRR (<http://arxiv.org/corr/home>). All experiments are performed on AMD Phenom II X4 945 with 3GB RAM.

It is evident that the introduction of global constraints in both measures significantly speeds up the process of distance matrix computation, which is the direct consequence of a faster similarity measure. The difference of computation times between an unconstrained measure and a measure with a small constraint is two and somewhere three orders of magnitude. Furthermore, it is known for DTW that

**Table 1** Calculation times of distance matrix for DTW

Name of dataset	DTW									
	unconstrained	75%	50%	25%	20%	15%	10%	5%	1%	0%
Car	79844	73391	58656	34500	28562	22047	15141	8016	2016	672
CBF	258375	242703	198969	124719	105375	86031	62672	41766	23047	17203
cinc_ecg_torso	88609875	79638047	63711094	36991468	30533875	23531718	16107266	8290062	1814203	146672
fish	434297	392656	317093	185672	154672	119906	83672	45390	12031	3969
Haptics	4257391	3789922	3052828	1774359	1468031	1135609	781547	404907	88234	10844
Inlineskate	25407250	22014203	17571907	10359203	8563921	6618313	4534843	2341546	491640	32109
Lighting2	101641	90781	72750	42828	35171	27391	19890	9875	2594	719
Mallat	97847485	88454641	70498062	41180453	34403500	26572297	18149719	9492531	2189141	284188
OSULeaf	592328	536562	431062	254844	210047	164515	113000	61312	17672	5906

**Table 2** Calculation times of distance matrix for LCS

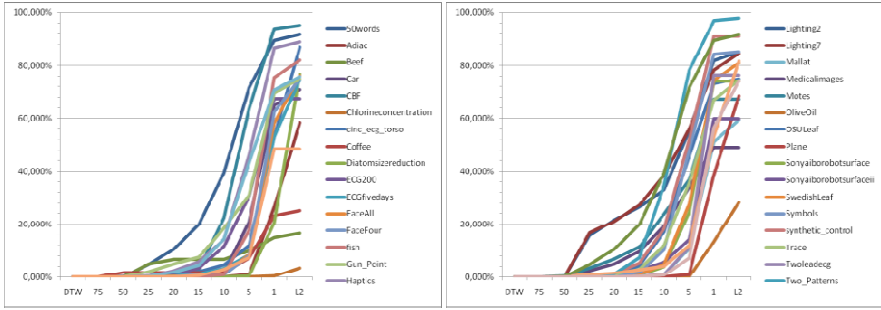
Name of dataset	DTW									
	unconstrained	75%	50%	25%	20%	15%	10%	5%	1%	0%
Car	52282	48844	39187	23078	18984	14875	10281	5516	1531	609
CBF	170016	161250	134297	85671	72359	60656	45797	32406	20204	17547
cinc_ecg_torso	62996109	59056579	46699297	27158750	22369187	17262734	11882062	6155704	1402765	161766
fish	285922	268282	214875	127484	105765	82329	58000	31906	9390	4156
Haptics	2793547	2593672	2074313	1218812	1008203	777266	539672	280437	64437	12359
Inlineskate	16524687	15507859	12291281	7163172	5895296	4556593	3139500	1633141	351640	37016
Lighting2	71000	66578	53329	31531	26422	20484	14188	7547	2250	750
Mallat	67029406	62582938	50095359	29302062	24642968	18588750	12998219	6910859	1670969	307656
OSULeaf	388375	371156	295813	174734	144938	114547	79140	43750	13984	6375

smaller values of constraints can tend to more accurate classification [10]. The authors also reported that the average constraint size, which gives the best accuracy, for all datasets is 4% of the time-series length. On the other hand, the influence of global constraints on the LCS measure is still not well investigated. However, it is evident that the usage of global constraints contributes to the efficiency of both measures, and, at least for DTW, improves classification accuracy.

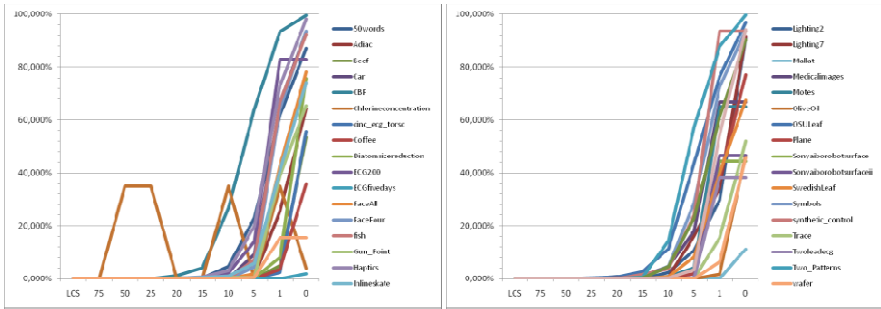
## 4.2 The Change of 1NN Graph

In the next experimental phase we wanted to investigate the influence of global constraints on the NN graph of each dataset. This decision was mainly motivated by the fact that the 1NN classifier is among the best classifiers for time series [18].

The nearest neighbour graph is a directed graph where each time series is connected with its nearest neighbour. We calculated this graph for unconstrained measures (DTW and LCS) and for measures with the following constraints: 75%, 50%, 25%, 20%, 15%, 10%, 5%, 1% and 0% of the length of time series. After that, we focused on the change of the 1NN graph for different constraints compared to the graph of the unconstrained measure. The change of nearest-neighbour graphs is tracked as the percentage of time series (nodes in the graph) that changed their nearest neighbour compared to the nearest neighbour according to the unconstrained measure. The graphical representation of results can be seen in Figure 1 and Figure 2 for DTW and LCS, respectively. Each figure is represented by two charts showing one half of the data sets for the sake of readability. The numerical results are available at CoRR.



**Fig. 1** Change of INN graph for DTW



**Fig. 2** Change of INN graph for LCS

The presented results clearly show that both measures behave in a similar manner when the constraint is narrowed. The INN graph of the DTW measure remains the same until the size of the constraint is narrowed to approximately 50%, and after that the graph starts to change significantly. The situation with LCS is more pronounced: the LCS INN graph remains the same to approximately 10-15%, while for smaller constraints it changes even more drastically.

Only one data set does not follow this rule for LCS measure: *Chlorineconcentration*. For some values of the constraint (75%, 20%, 15% and 5%) the graph is the same as the unconstrained, while for other values of the constraint the difference of graphs is 34.87%. Additionally, we investigated the structure of this dataset and found that the time series are periodical, where all time series have approximately the same period. Since the LCS measure searches for the longest common subsequence, it turns out that for some constraint values the LCS algorithm finds the same sequence as the unconstrained LCS. Other values of the constraint break that sequence, which is then no more longest, and as a consequence some other time series is found as a nearest neighbour. This behaviour is caused by the strict periodicity of this data set.

All other datasets (for both measures) reach high percentages of difference (over 50%) for small constraint sizes (5-10%). This means that when the constraint size is narrowed to 10% of the length of time series, then more than

50% of time series in the data set change their first neighbour with regard to the unconstrained measure. This fact strongly suggests that constrained measures represent qualitatively different measures than the unconstrained ones.

## 5 Conclusion and Future Work

Although the Euclidian measure is simple and very intuitive for time-series data, it has a known weakness of sensitivity to distortion in the time axis. Many elastic measures (DTW, LCS, ERP, EDR, etc.) were proposed in order to overcome this weakness. However, they are all based on dynamic programming and have quadratic computation complexity. Global constraints are introduced in dynamic programming algorithms to narrow the search path in the matrix and to decrease computation time.

In this paper, we examined the influence of global constraints on two most representative elastic measures for time series: DTW and LCS. Through an extensive set of experiments, we showed that the usage of global constraints can significantly reduce the computation time of these measures, which is their main weakness. In addition, we demonstrated that the constrained measures are qualitatively different than their unconstrained counterparts. For DTW it is known that the constrained measures are more accurate than the unconstrained, while for LCS this issue is still open.

In future work we plan to investigate the accuracy of the constrained LCS measure for different values of constraints. It would also be interesting to explore the influence of global constraints on the computation time and INN graphs of other elastic measures like ERP, EDR, Swale, etc. Finally, the constrained variants of these elastic measures should also be tested with respect to classification accuracy.

**Acknowledgments.** The authors acknowledge the support of this work by the Serbian Ministry of Education and Science through project "Intelligent Techniques and Their Integration into Wide-Spectrum Decision Support" no. OI174023.

## References

1. Jiawei, H., Micheline, K.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, CA (2005)
2. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases. In: SIGMOD Conference, pp. 419–429 (1994)
3. Keogh, E.J.: A Decade of Progress in Indexing and Mining Large Time Series Databases. In: VLDB, pp. 1268–1268 (2006)
4. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. In: VLDB 2008, Auckland, New Zealand, pp. 1542–1552 (2008)
5. Keogh, E.J., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* 7(3), 358–386 (2005)
6. Vlachos, M., Gunopulos, D., Kollios, G.: Discovering similar multidimensional trajectories. In: ICDE 2002, pp. 673–684 (2002)

7. Chen, L., Ng, R.T.: On the marriage of lp-norms and edit distance. In: VLDB 2004, pp. 792–803 (2004)
8. Chen, L., Ozsu, M.T., Oria, V.: Robust and fast similarity search for moving object trajectories. In: SIGMOD Conference, pp. 491–502 (2005)
9. Morse, M.D., Patel, J.M.: An efficient and accurate method for evaluating time series similarity. In: SIGMOD Conference, pp. 569–580 (2007)
10. Ratanamahatana, C., Keogh, E.: Three Myths about Dynamic Time Warping. In: proc. of SIAM Inter. Conf. on Data Mining, Newport Beach, CA, April 2005, pp. 506–510 (2005)
11. Kurbalija, V., Radovanović, M., Geler, Z., Ivanović, M.: A framework for time-series analysis. In: Dicheva, D., Dochev, D. (eds.) AIMS 2010. LNCS, vol. 6304, pp. 42–51. Springer, Heidelberg (2010)
12. Keogh, E., Xi, X., Wei, L., Ratanamahatana, C.: The UCR Time Series Classification/Clustering Page (2006), [http://www.cs.ucr.edu/~eamonn/time\\_series\\_data/](http://www.cs.ucr.edu/~eamonn/time_series_data/)
13. Agrawal, R., Lin, K.I., Sawhney, H.S., Shim, K.: Fast similarity search in the presence of noise, scaling, and translation in times-series databases. In: Proceedings of the 21st International Conference on Very Large Databases, pp. 490–501 (1995)
14. Chan, K.P., Fu, A., Yu, C.: Haar wavelets for efficient similarity search of time-series: with and without time warping. IEEE Trans. Knowl. Data Eng. 15(3), 686–705 (2003)
15. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. J. Knowl. Inf. Syst. 3(3), 263–286 (2000)
16. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of ACM SIGMOD Conference on Management of Data, May 2001, pp. 151–162 (2001)
17. Keogh, E.: Exact indexing of dynamic time warping. In: 28th International Conference on Very Large Data Bases, Hong Kong, pp. 406–417 (2002)
18. Xi, X., Keogh, E., Shelton, C., Wei, L., Ratanamahatana, C.A.: Fast time series classification using numerosity reduction. In: Proceedings of the 23rd International Conference on Machine Learning (ICML 2006), pp. 1033–1040. ACM, New York (2006)
19. Berndt, D., Clifford, J.: Using dynamic time warping to find patterns in time series. In: AAAI-1994 Workshop on Knowledge Discovery in Databases, pp. 229–248 (1994)
20. Vlachos, M., Kollios, G., Gunopulos, D.: Discovering Similar Multidimensional Trajectories. In: Proc. of 18th Inter. Conf. on Data Engineering (ICDE), San Jose, CA, pp. 673–684 (2002)
21. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics Speech Signal Process ASSP 26, 43–49 (1978)
22. Itakura, F.: Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoustics Speech Signal Process ASSP 23, 52–72 (1975)
23. Radovanović, M., Nanopoulos, A., Ivanović, M.: Time-series classification in many intrinsic dimensions. In: Proceedings of SDM, 10th SIAM International Conference on Data Mining, Columbus, Ohio, USA, pp. 677–688 (2010)

# Social Network Analysis in Professional e-Recruitment

Kostadinka Toteva and Elissaveta Gourova

Sofia University, 125 Tzarigradsko shosse Blvd. bl. 2., 1113 Sofia, Bulgaria  
email: kossy.toteva@gmail.com, elis@fmi.uni-sofia.bg

**Abstract.** The present paper considers the importance of social networks for the success of the recruitment process in the knowledge society. It provides a short theoretical background on social network analysis (SNA) and the most common recruitment practices. The paper provides some results from a survey on the social networks usage by Bulgarian recruiters. On bases of the analysis made, a short specification of a tool supporting professional recruitment is provided.

**Keywords:** social networks, social network analysis, recruitment, human resource management.

## 1 Introduction

The development of information technologies (IT) in the last few years, and in particular, the appearance of Web 2.0, has resulted in deep changes in work and life. Nowadays, Web 2.0, and social networks in particular, reflect business and social communications of individuals, and become essential tools for transfer of knowledge and information. In the knowledge-based society, where knowledge and skills are acknowledged as an important resource for growth and competitiveness, human resource (HR) management becomes a business process with strategic importance, both for design and implementation of corporate strategy, motivation, recruitment and preserving of highly-skilled personnel. [13].

Nowadays, Web 2.0 changes the selection process of employees, providing many new opportunities. It provides opportunities for relationship-based approaches and proactive recruiters [6]. Social networks, in particular, provide a link between candidates for the position and the recruiters [8]. In addition, in order to facilitate the work of recruiters, tools that facilitate the selection process by taking advantage of the benefits of Web 2.0 have been developed [7].

Taking into account these rapid changes, it is interesting to find out how new technologies and the opportunities associated with them have affected the daily practices of Bulgarian recruiters. Some of the questions this paper tries to highlight include: How they are using social networks? Which social networks do recruiters prefer? What information is relevant to them? Do they apply SNA to find the right candidate? Do they use specialized tools for social networking? The paper initially provides an insight into social network analysis and the recruitment process. Subsequently, it presents the methodology and the results of a study

carried out among Bulgarian recruiters in order to investigate their current recruitment practices and the usage of SNA in this process. On this base, a short specification of an appropriate tool for recruiters in Bulgaria is proposed.

## 2 Social Networks and Recruitment Processes

The term social network was first coined by Professor J. A. Barnes in the 1950s as “an association of people drawn together by family, work or hobby” [4]. Nowadays, the use of the term in society is highly related social network sites (SNSs) - an IT tool for support of social networks, which provides a communication platform and specific tools for organizing events, knowledge sharing, easy messaging, etc.

In recent years, an increasing penetration of social networks on the web can be noticed [9]. Many studies are devoted on social networks and provide various classifications: according to the functionality they offer (e.g. searching and browsing capabilities [12], privacy protection scale [10] etc.) or the characteristics of consumers who use them (e.g. relationship classification [11]).

A well-known technique for extracting information from SNSs is the social network analysis (SNA) - a structural approach, which studies interaction among social actors [2]. It is based on the assumption that there are patterns in relations and these patterns are based on live individuals' relationships [1]. There are four important features of SNA [2]: (1) it is a structural approach, which studies the structure of the network; (2) it is grounded in systematic empirical data, especially relational, or network data; (3) it draws on graphic imagery; (4) it uses mathematical and computational models. SNA uses many concepts from the graph theory and the network analysis, interpreted from social theory point of view. For example, high degree centrality for a node indicates a high popularity or activity for the actor, which is represented by that node.

Recruitment is a part of human resource management (HRM) which refers to the process of attracting, screening, and selecting qualified people for a particular position. The recruitment process according to Armstrong [5] contains four steps: defining requirements, planning recruitment campaigns, attracting people and selecting people. SNA could bring many benefits in the phase of attracting and selecting people. For example, in the case of subnet of good specialist, a high value of degree centrality of a node means that the person, represented by this node has many connections to the high-level specialists. By communicating with them the individual most probably exchanges knowledge and information, and it might be concluded that he also possesses good expert knowledge and skills. If the subnet represents professional group pages and the edges represent belongings of a person to a page, a high value of degree centrality indicates interests and hobbies of a person. If the definition of edge is changed to ‘an edge connects two nodes if the person has published information in the professional group page’, the degree centrality would measure competence. Both methods could be extended by associating weights on the edges in order to obtain more precise results. They could be used for comparison of people and could contribute to choosing the right person in the recruitment process. They are based on a common algorithm:



1. Define a subnet, which includes defining nodes, edges, weight of edges, and an interpretation of the edges.
2. Calculate degree centrality.
3. Analyze the value of degree centrality.

The homogeneity of these steps allows defining a module with such functionality. Similar algorithms could be designed for other characteristics from the graph theory.

### **3 SNSs in Bulgarian Professional Recruitment**

An initial survey was conducted with the objective to understand the usage of SNSs in HR practices and to extract the common patterns in the recruitment processes on Bulgarian labor market. The target group includes Bulgarian HR specialists who are involved in recruitment processes. The survey methodology contains a questionnaire and interviews with respondents in order to deepen the understanding on SNSs usage. The questionnaire includes closed and open-ended questions aimed to collect information about the most used SNSs. Generally, the survey aims to answer the following questions: Are recruiters in Bulgaria using SNSs in their practice? Which SNSs they use and how? Are they using SNA for determining competency? What information they gather? Are they using any tool to support their work with SNSs and if so, what they like and dislike in it?

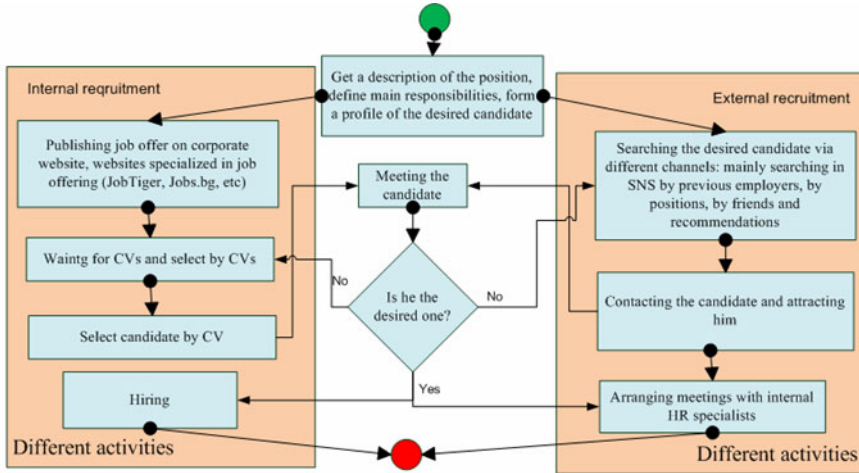
In addition, a method for determining the competency based on SNA was described and the respondents were asked if they use it. A question about the relevance of a SNSs IT tool was asked. In the interviews the recruiters describe step by step their way to analyze the information published in SNSs.

A group of 14 professional recruiters from different organizations were asked to fill the survey form. The recruiters work on Bulgarian labor market. They are selected from different organizations in order to cover the majority of practices used by Bulgarian recruiters. Most respondents recruit mainly IT specialists. They are divided into two subgroups according to the type of recruitment they practice. The first subgroup comprises specialists, who work in HR departments and their main responsibility is to recruit specialists for the whole organization. Therefore, they are called internal recruiters. The specialists of the second group work in recruiting companies (e.g. Adecco Bulgaria Ltd, MiNDS People & Solutions, PFG Bulgaria Ltd, etc.). They are called external recruiters and often are referred to as 'head-hunters' because they attract good professionals and make them change their employer.

The two groups use different methodologies and tools for recruitment (Fig. 1). The main method for attracting new specialist used by internal recruiters is publishing job offers on the website of the company or posting it on specialized sites for job finding (JobTiger, Jobs.bg, etc.). It is a passive way to find the right person, driven by the interests of job seekers, not by the recruiter. They are interested in the candidates' previous experience, previous employer, education, etc. This subgroup of recruiters does not use SNSs for reaching the candidates or for building the complete profile of the candidate. For this purpose latter occasionally Facebook is used.

The recruitment process for external HR specialists is driven by the recruiter and includes searching SNSs for a person with accurate profile, contacting and

attracting him/her. As the survey results show, the external recruiters are turning to social networks for almost every position they are working on. They rely mainly on LinkedIn for evaluating professional skills, followed by Facebook for determining personal characteristics. Other SNSs, like SkillsPages and Xing, are rarely used.



**Fig. 1** Processes used by internal and external recruiters

It could be noted that internal recruiters are interested mainly from education and experience, as well as available certificates of the applicants. The external recruiters do not put such high weight on education and certificates, but rather on previous employers, the profession and experience of the applicants. It is interesting to point out that in the evaluation process relatively high value is given to 'Friends, who are good specialists in the area'. This factor is as significant as education and certificates for external recruiters while internal ones find it for less important.

One of the issues found out during the survey is that all of the respondents do not use an IT tool supporting their activities for searching SNSs. They use only built-in functionalities of SNSs, and lose time in searching information. The respondents who were interviewed claimed that they spend over 60% of their time on searching relations in SNSs. Therefore, an SNSs IT tool would be very useful.

## 4 Conclusion

Although some companies still prefer the traditional job offering sites, many professional recruiters are going beyond them and are entering the SNSs in order to find and attract the most talented candidates for the job. Currently, recruiters are searching the social network manually and their success mainly depends on their intuition. Nowadays, when SNSs are increasingly used in business processes, a tool

for analyzing social networks is not yet available on Bulgarian market. According to the survey, it is obvious that for internal HR specialists such tool is meaningless, however external ones need it. In addition, the application of SNA could facilitate the recruitment process and could lead to more effective work of recruiters. SNA as a structural approach could help recruiters to find complex patterns and derive links between such patterns and individual characteristics of objects in the social network.

A short specification of an adequate IT tool supporting the recruiting activities is defined by listing the minimum needed features. The tool must:

- be integrated with social network sites (at least LinkedIn);
- support information for education, profession, previous experience, current and ex employers and owned certificates;
- support defining relations or use relations from a social network site;
- provide functionality for searching (at least search functionality of LinkedIn);
- provide a mechanism for evaluating people based on relations to other people;
- provide functionality for searching people by their relationship index.

The described survey was made as initial study of the problem of the recruitment methodologies and SNSs usage. It will serve as foundation for deeper research on the ways recruiters operate with social network sites. The need of an IT tool, supporting work with social network sites was identified and it should be analyzed in more details in further studies, which will lead to detailed software specification.

## References

1. Freeman, L.C.: What is Social Network Analysis?, <http://www.insna.org/sna/what.html>
2. Freeman, L.C.: The Development of Social Network Analysis: A Study in The Sociology of Science. Empirical Press Vancouver, Canada (2004)
3. Zhang, M.: Social Network Analysis: History, Concepts and Research, Handbook of Social Network Technologies and Applications, pp. 3–21. Springer, Heidelberg (2010)
4. Barnes, J.A.: Class and Committees in Norwegian Island Parish (1954)
5. Armstrong, M.: A handbook of Human Resource Management Practice. Kogan Page Ltd. (2006)
6. Girard, A., Fallery, B., Italy, M.: E-recruitment: new practices, new issues. In: An Exploratory Study 3rd International Workshop on Human Resource Information Systems HRIS in conjunction with ICEIS, pp. 6–7. INSTICC Press (2009) ISBN 9789898111975
7. Buttinger, C., Pröll, B., Palkoska, J., Retschitzegger, W., Schauer, M., Immler, R.: JobOlize - Headhunting by Information Extraction in the era of Web 2.0. In: Proceedings of the 7th International Workshop on Web-Oriented Software Technologies, IWOST 2008 (2008)
8. Han, J.: Employer Branding Leveraging Web 2.0, Master thesis (2010)
9. Boyd, D.M., Ellison, N.B.: Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* 13(1), article 11 (2007), <http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html>

10. Goettke, R., Christiana, J.: Privacy and Online Social Networking Websites (2007)
11. Tang, S., Yuan, J., Mao, X., Li, X., Chen, W., Dai, G.: Relationship Classification in Large Scale Online Social Networks and Its Impact on Information Propagation. In: IEEE INFOCOM (2011)
12. O'Murchu, I., Breslin, J., Decker, S.: Social Networking Portals: An Overview and Evaluation. In: Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web, FOAF Galway (2004)
13. Gourova, E., Toteva, K.: Electronic tools for support of strategic human resource management. *W. Trans. on Comp.* 9(10), 1133–1148 (2010)

# A System Integrating Agent Services with JMS for Shared Battlefield Situation Awareness

Albert Esterline, William Wright, and Srinivas Banda

Dept. Of Computer Science, North Carolina A&T State University,  
Greensboro, NC, USA  
email: {esterlin,wtwright,srbanda}@ncat.edu

**Abstract.** This paper reports on work in progress on a prototype information dissemination and management system for a hierarchical military command structure. Each unit is provided with a suite of services that integrates three paradigms for collaboration, communicates with its neighbors, and directs its immediate subordinates. This paper emphasizes concepts largely ignored but critical for shared situation awareness. As a prerequisite for coordination, we focus on common knowledge. Something is common knowledge in a group if not only does everyone in the group know it but also everyone knows that everyone knows it. Constraints on inter-unit communication motivated using, in one unit's services, proxy agents for its subordinate units. This use of proxies amounts to a version of the psychological concept of theory of mind (the ability to attribute to others mental states often conflicting with our own), seen as a mechanism for achieving common knowledge.

**Keywords:** multiagent systems, shared situation awareness, common knowledge, workflows, JMS.

## 1 Introduction

We report on work on a battlefield command and control system that maintains shared situation awareness among units. The goal of the US Army's Tactical Information Technologies for Assured Net Operations (TITAN) Program [1] is to show how emerging information technologies can improve tactical operations. A key area is Information Dissemination and Management (ID&M), using an XML-based common information exchange data model for command and control information. The key software here is a collection of agent-based software services that collaborate during tactical mission planning and execution. A TBS (TITAN Battle Command Support) is a suite of services associated with a unit in a command hierarchy. Since there is a hierarchy of commanders, there is a hierarchically integrated set of TBSs. The TBS for a unit must communicate with its peers in its echelon and with its subordinates—and with its parent in receiving command and control documents and updates to them and in sending feedback and warnings. Figure 1 illustrates information dissemination.

The next section outlines prototype TITAN services. Three paradigms for collaboration and information dissemination are used: multiagent systems, Web

services, and distributed event-based systems (here JMS, Java Messaging Service). The third section indicates how these paradigms are integrated and the division of labor. The main goal of TITAN is coordination, a prerequisite for which is common knowledge. Roughly, something is common knowledge in a group if not only does everyone in the group know it but also everyone knows that everyone knows it (with deeper nestings of “everyone knows” as needed by the analysis). Common knowledge and the importance of self reference is the topic of the fourth section. Because of communication constraints, a TBS maintains proxies of its neighbors and subordinates that are kept up to date by light-weight messaging and help the TBS determine what messages to send. These proxies are agents communicating with each other (and other agents in the given TBS) in full-blown agent-based ways. The fifth section points out that this is a version of the psychological concept of theory of mind (the ability to attribute to others mental states that possibly conflict with one’s own) and enables common knowledge. The penultimate section points out that the literature on shared situation awareness generally ignores the simple but critical self-referential nature of a system.

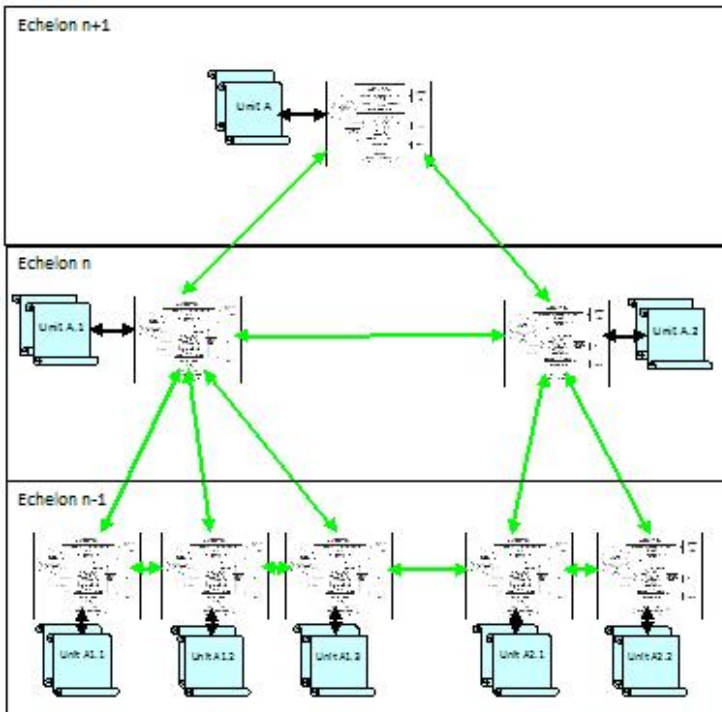


Fig. 1 Information dissemination in a hierarchy of TBSs [1]

Common knowledge and related notions are addressed in terms of collaboration of software systems, but the commanders also share in the common knowledge. The way communication is managed in modern warfare in fact gives insight into the prerequisites of human collaboration, justifying our perspective.

## 2 Prototype TITAN Services

We have implemented prototypes of three of the services planned to be developed under the ID&M portion of TITAN and have focused on the development of prototypes of some of the envisioned agents [2]. The three implemented services are the Alert and Warnings Service (AWS), the OPORD Support Service (OPS), and the Workflow Orchestration Support (WOS). The AWS provides predictive assessments and generates warnings and alerts. The OPS disseminates Operational Orders (OPORDs) and Operational Plans (OPLANs) in the form of XML documents. An OPORD is a directive and the basis for synchronizing operations, while an OPLAN is a proposal for executing a command. Finally, the WOS generates workflows from tasks and executes those workflows (interacting with other services).

These services support both a Web Service interface and a JMS interface. Web Services provide synchronous information retrieval by remote clients and are widely used by the military. JMS provides message-oriented APIs for a publish/subscribe messaging model, a push-based model where messages are automatically broadcast to consumers. The Java Agent DEvelopment (JADE) framework [3] is used; agents do the work in a TITAN service and collaborate most effectively in JADE's native Agent Communication Language (ACL). To support WOS workflows, we use the JADE-based WADE (Workflows and Agents Development Environment) software platform, which adds the ability to define agent tasks according to the workflow metaphor [4].

The prototype system shows the feasibility of integrating three paradigms for collaboration and information dissemination: multiagent systems, Web services, and distributed event-based systems (here JMS). The JADE distribution site provides gateway agents to integrate JADE agents, Web services, and JMS. Agent collaboration via ACL messages occurs only within a TBS although agents are ultimately responsible for integration of TBSs. JMS is the communication medium among TBSs. The push paradigm is particularly appropriate where monitored information must be passed along posthaste or requests for immediate action are posted. Direct JMS messaging among TBSs is restricted to communication between parent and child and among siblings; communication between more distant echelons is achieved by passing the message along. Such handling of messages approximates a military chain of command. Since the services are provided by teams of agents, the TBSs are autonomous, and information dissemination results in coordination, not control, of units.

## 3 Common Knowledge

In a hierarchy of TBSs, an OPORD amounts to a coordination script. The OPS service receives an OPORD from the parent TBS and fills in the detail needed to direct and coordinate its child units, for which the TBS in turn provides OPORDs. Detail is thus filled out as the script is disseminated down the hierarchy.

The salient concept is that of common knowledge, a necessary condition for coordinated action. To explain this [5], let  $G$  be a group of  $n$  agents, each agent denoted by a distinct ordinal, so that they are named  $1, 2, \dots, n$ ; then  $G = \{1, 2, \dots, n\}$ .

We introduce  $n$  modal operators  $K_i$ ,  $1 \leq i \leq n$ , where  $K_i \phi$  is read “agent  $i$  knows that  $\phi$ ”.  $E_G \phi$ , read as “everyone (in  $G$ ) knows that  $\phi$ ,” is defined as  $K_1 \phi \wedge K_2 \phi \wedge \dots \wedge K_n \phi$ . Let  $E_G^k$  be  $E_G$  iterated  $k$  times. Then “it is common knowledge in group  $G$  that  $\phi$ ,” in symbols,  $C_G \phi$ , is defined with the infinite conjunction  $E_G^1 \phi \wedge E_G^2 \phi \wedge \dots \wedge E_G^i \phi \wedge \dots$ . For example, traffic lights would not work unless it were common knowledge that green means go, red means stop, and lights for opposite directions have different colors. If not, we would not confidently drive through a green light. In an epistemic logic (i.e., logic of knowledge) augmented with these operators, it is easy to show that, if everyone in  $G$  agrees that  $\psi$ , then the agreement is common knowledge [5]. It can also be shown that coordination implies common knowledge. Some authors speak of mutual (or common) belief, which is the same as common knowledge as they involve the same self reference and recursive unfolding even though “ $A$  knows that  $\phi$ ” implies  $\phi$  while “ $A$  believes that  $\phi$ ” does not.

Besides characterizing common knowledge via an infinite conjunction, Barwise [6] identifies two other approaches. In the fixed-point approach, we view  $C_G \phi$  as a fixed-point of the function [6]  $f(x) = E_G(\phi \wedge x)$ . The third approach (here following [7]) is the shared situation approach. Where  $A$  and  $B$  are rational, we may infer common knowledge among  $A$  and  $B$  that  $\phi$  if

1.  $A$  and  $B$  know that some situation  $\sigma$  holds.
2.  $\sigma$  indicates to both  $A$  and  $B$  that both  $A$  and  $B$  know that  $\sigma$  holds.
3.  $\sigma$  indicates to both  $A$  and  $B$  that  $\phi$ .

Barwise concludes that the fixed-point approach (implied by the shared situation approach) is the correct analysis. Some conclusions in epistemic logic [5] are paradoxes if we regard common knowledge as a disposition, but, as Barwise notes, common knowledge is not properly knowledge: knowing that  $\phi$  is stronger than carrying the information that  $\phi$  since it relates to the ability to act. He concludes that common knowledge is a necessary but not sufficient condition for action and is useful only in a shared situation that “provides a stage for maintaining common knowledge.”

Barwise addresses situations, but common “knowledge” is more embracing. H. H. Clark and Carlson [7] identified three “co-presence heuristics” giving rise to different kinds of shared “situations.” Two, physical co-presence (e.g., shared attention) and linguistic co-presence (as in conversation), properly relate to situations, but the third, community membership (presupposed by the others), is not temporally or spatially restricted. It is essentially the social part of what Andy Clark [8] called *scaffolding*: a world of physical and social structures on which the coherence and analytic power of human activity depends. Let “common state knowledge”, abbreviated “CSK”, refer to what is established by the two non-scaffolding heuristics. Common knowledge thus is either scaffolding or a self-referential feature of the situation.

Discussing conversation, H. H. Clark addresses “... shared information or common ground—that is, mutual knowledge, mutual beliefs, and mutual assumptions ...” [9]. Stalnaker [10], in a thorough account of common ground, discusses what it is for a speaker to presuppose a proposition  $\phi$  in a conversation. It is, apparently, to believe that  $\phi$  is common ground, identified as common belief. But what is presupposed may



diverge from what is mutually believed—e.g., what is assumed may temporarily be part of the common ground. So he defines common ground in terms of a notion of *acceptance* broader than that of belief: “It is common ground that  $\phi$  in a group if all members accept (for the purpose of the conversation) that  $\phi$ , and all *believe* that all accept that  $\phi$ , and all *believe* that all *believe* that all accept that  $\phi$ , etc.” This again involves self reference and an infinite conjunction (the critical points for us), and speaker presuppositions are still the speaker’s beliefs about the common ground.

Common ground has found application in shared situation awareness, but the simple self-referential aspect is generally lost in the detail, and common ground is characterized as a shared perspective [11] or (simply) shared knowledge, beliefs, and assumptions [12], or the negotiation is explicated in terms of state transitions [13] with no focus on self reference. In TITAN, the protocols and ontologies are designed into the TBSs as scaffolding, and the XML documents communicated are essentially coordination scripts. TITAN CSK is similar to common ground in conversation, but communication is scripted and has larger scope in that message syntax and content are understood as intended when correct with respect to the protocol and schema.

The notion of common knowledge has proved useful in a wide array of disciplines. It was introduced by the philosopher Lewis and has been used in the analysis of protocols (scaffolding) in distributed systems [5] with an eye to CSK as the distributed state evolves. Chew [14] demonstrated that rituals are rational at a meta-level since they establish common knowledge enabling coordinated actions with significant payoffs. In game theory, games of complete information assume that players’ strategies and payoffs as scaffolding, while games of perfect information take players’ moves as CSK. Rituals convey messages by brute-force stylization, and games of perfect information assume infallible communication. TITAN achieves CSK in more flexible ways that coordinate complex activities over extended periods.

In designing a system (scaffolding) to maintain CSK, one should identify what is included in the CSK and how situations are self-referential. Usually, however, these issues remain tacit, and sometimes an analysis in terms of CSK is hardly called for, as in designing protocols for distributed systems, where experience sets the stage. But when the participants are autonomous, content relates to the real world, and the tasks are novel for the analyst, it is good to consider these issues.

## 4 Theory of Mind

Coordination of child TBSs by a parent TBS with the paradigm used within a TBS—a multiagent system—is infeasible because of the communication burden. So the WOS maintains proxy agents to model the children. The WADE workflow contains agents that themselves implement workflows that mirror the salient aspects of the workflows of the children. The mission is partitioned into stages identified in the OPORD, and the sub-flows synchronize on stage transitions by receiving JMS reports sent to the TBS from the child TBSs and passed to the WOS. The proxies also respond to alerts and to updates to the OPLAN. And the WOS must have enough information to simulate the siblings of the TBS; the proxies for these are not part of the workflow but are coordinated with it. Finally, the TBS must have information on its parent’s script to form a proxy for it, which also is not part of the workflow.

This use of proxies implements a scaled-back version of what is called theory of mind (ToM) in developmental psychology [15], the ability to attribute mental states (beliefs, desires, intentions, etc.) to oneself and others and to understand that others have mental states different from one’s own. Early work in ToM focused on the improvement in children between ages three and five on false-belief (FB) and Level 2 visual perspective-taking (PT) tasks. As an example of an FB task, a child watches as a puppet sees a cookie put in one of two boxes and leaves, someone moves the cookie to the other box, and when the puppet returns, the older child (having a notion of false belief), but not the younger, says the puppet will look in the original box. As a PT task, the older child, but not the younger, understands that a picture oriented correctly for them looks upside down to a person seated opposite. ToM provides a scaffold for language development: when a child hears an adult speak a word, they recognize that the word refers to what the adult is looking at. And it is suggested [16] that the connection of pretend play (involving role assignments) with false belief understanding is in the representation of the other’s beliefs and goals, often conflicting with one’s own.

ToM emphasizes differences, and in cooperation it is critical that these differences be common knowledge. Generally, the roles assigned to participants and their intentions are common knowledge that involves recognizing differences. Such recognition is supported in the WOS version of ToM, which is scaled back principally in being restricted to intentions and beliefs—but intentions are critical. Tomasello [17, p. xiii] uses “shared intentionality” for the underlying psychological processes that support our species-unique forms of cooperation. Shared intention lets us, in cooperative endeavors, “create with others joint intention and joint commitment,” which “are structured by processes of joint attention and mutual knowledge.”

We have shown [18] how the physical co-presence heuristic may be used for groups of artificial agents to attain common knowledge by perceptual means. To focus on the episodic nature of co-presence evidence, we introduced into epistemic logic a modal operator  $S_a^t \phi$  for agent  $a$  seeing at time  $t$  that  $\phi$ . Given time parameters for other operators, we have the axioms  $S_a^t \phi \Rightarrow K_a^t \phi$  and (for simplicity)  $K_a^t \phi \Rightarrow K_a^u \phi$  for all  $u > t$  and focus on cases where, e.g., both  $S_a^t S_b^t \phi$  and  $S_b^t S_a^t \phi$  hold. The formalism exposes reasons to hold that, to attain CSK, agents must model each other’s perceptual abilities, requiring common knowledge of shared abilities. The linguistic co-presence heuristic can be handled somewhat similarly. With the appropriate scaffolding, ToM abilities provide a mechanism for achieving CSK. Indeed, ToM would be pointless if it did not do so and provide access to the scaffolding.

## 5 Common State Knowledge and Shared Situation Awareness

The key feature that is established by TITAN and enables coordination is often identified as shared situation awareness (SA). This is typically characterized as, e.g., “a reflection of how similarly team members view a given situation” [19]. Similarly, as we have seen, common ground (roughly what is shared in shared SA) is characterized as a shared perspective or shared knowledge, beliefs, and assumptions. As noted, what is ignored is the simple but critical self-referential nature of the system.

Our account emphasizes the participants, their coordination, and their communication; perception is de-emphasized. Although mental models for individuals are seen as especially useful in SA [20] for understanding how information is integrated and the future is projected, the subject of shared SA does not typically connect one mental model with another. As we have seen, however, ToM suggests that our understanding of the beliefs and intentions of others is quite natural, and we suggest that this understanding is quite accurate if we share a common social scaffolding.

We accept the emphasis on projecting into the future [20] and agree that one should address how well a design “supports the operator’s ability to get the needed information under dynamic operational constraints” [21]. For TITAN, the threat is not so much being overwhelmed by information about the physical environment as being overwhelmed by messages from collaborators. A military hierarchy is part of the solution, and scripting is another since it passes only relevant information through the hierarchy to keep the proxies consistent and synchronized.

## 6 Conclusion

We described prototypes of several services of the TITAN program for disseminating command information. A TBS is a suite of TITAN services associated with a unit in a command hierarchy mirrored by a hierarchically integrated set of TBSs. The most important service here is the WOS, which generates workflows from XML-encoded operational orders (OPORDs). Three paradigms for collaboration are used: multi-agent systems, Web services, and JMS. JADE is the agent framework and provides bridges to JMS and to Web services. Agent collaboration via agent messages occurs only within a TBS, and JMS is the means of communication among TBSs.

TITAN’s main goal is for TBSs to coordinate, and a prerequisite for coordination is common knowledge, which can be defined in several ways. In all cases a nested self reference within the group is critical. Common knowledge in a situation (CSK) can arise under certain physical or linguistic co-presence conditions. There is also common knowledge by virtue of a mutual sharing of social scaffolding. In TITAN, the protocols designed into the TBSs provide scaffolding. Given an OPOrd as a coordination script, JMS messaging maintains CSK. Because of communication constraints, the workflow that directs the child TBSs uses proxy agents to represent the children. This implements a version of the concept of theory of mind (ToM). With the appropriate scaffolding, ToM provides a mechanism for achieving CSK.

What the literature on shared situation awareness often ignores is the simple but critical self-referential nature. For TITAN; the conditions for CSK and the appropriate communication patterns are critical. The threat of being overwhelmed with messages is met by a hierarchical organization and scripting that passes only relevant messages. We have characterized the common knowledge as shared by TBSs, but the commanders using the TBSs share in it as well. The analysis could focus on the TGBs because of the way communication is managed in modern warfare.

**Acknowledgements.** The research reported here was performed with support of Viecore Subcontract BAA8095022 under direction of Israel Mayk of the US Army C2D.

## References

1. US Army, A Guide for Developing Agent-Based Services. Revision 14.0 (2008)
2. Krishnamurthy, K., Esterline, A.: A System That Integrates Agent Services with Web Services and a Distributed Event-Based System for Processing XML Information. In: IEEE Southeast Conference, pp. 367–370. IEEE Press, New York (2010)
3. Bellifemine, F., Caire, G., Greenwood, D.: Developing Multi-Agent Systems with JADE. Wiley, Hoboken (2007)
4. WADE Workflow Environment for JADE, <http://jade.tilab.com/wade/>
5. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, Y.: Reasoning About Knowledge. The MIT Press, Cambridge (1995)
6. Barwise, J.: On the Model Theory of Common Knowledge. In: Barwise, J. (ed.) The Situation in Logic, Univ. of Chicago Press, Chicago (1989)
7. Clark, H.H., Carlson, T.B.: Speech Acts and Hearers' Beliefs. In: Smith, N.V. (ed.) Mutual Knowledge, pp. 1–37. Academic Press, New York (1992)
8. Clark, A.: Being There: Putting Brain, Body and World Together Again. The MIT Press, Cambridge (1997)
9. Clark, H.H., Brennan, S.E.: Grounding in Communication. In: Resnick, L.B., Levine, J.M., Teasley, J.S.D. (eds.) Perspectives on Socially Shared Cognition, pp. 127–149. American Psychological Association, Washington, D.C (1991)
10. Stalnaker, R.: Common Ground. *Linguistics and Philosophy* 25, 701–721 (2002)
11. Introduction. In: Bilger, R., Guest, S., Tauber, M.J. (eds.) Conference on Human Factors in Computing Systems, CHI 1996, Vancouver, BC (1996), <http://www.sigchi.org/chi96/proceedings/>
12. Nofi, A.A.: Defining and Measuring Shared Situational Awareness. Center for Naval Analyses, Alexandria, VA (2002), <http://www.cna.org/documents/D0002895.A1.pdf>
13. Pérez-Quiñones, M.A., Sibert, J.L.: A Collaborative Model of Feedback in Human-Computer Interaction. In: Electronic Conf. on Human Factors in Computing Systems (CHI 1996), Vancouver, BC (1996), <http://www.sigchi.org/chi96/proceedings/>
14. Chwe, M.S.-Y.: Rational Ritual: Culture, Co-ordination and Common Knowledge. Princeton University Press, Princeton (2001)
15. Flavell, J.H.: Theory-of-Mind Development: Retrospect and Prospect. *Merrill-Palmer Quarterly* 50(3), 274–290 (2004)
16. Astington, J.A., Jenkins, J.M.: Theory of Mind Development and Social Understanding. *Cognition and Emotion* 9, 151–165 (1995)
17. Tomasello, M.: Why We Cooperate. The MIT Press, Cambridge (2009)
18. Wiriyaconkasem, S., Esterline, A.: Heuristics for Inferring Common Knowledge via Agents' Perceptions in Multiagent Systems. In: 5th World Multi-conference on Systemics, Cybernetics, and Informatics, Orlando, FL (2001)
19. Bolstad, C.A., Cuevas, H.M., Gonzalez, C., Schneider, M.: Modeling Shared Situation Awareness. In: 14th Conference on Behavior Representation in Modeling and Simulation (BRIMS), Simulation Interoperability Standards Organization, Los Angeles (2005)
20. Endsley, M.R.: Design and Evaluation for Situation Awareness Evaluation. In: Human Factors Society 32nd Annual Meeting, Santa Monica, CA, pp. 97–101 (1988)
21. Endsley, M.R.: Theoretical Underpinnings of Situation Awareness: A Critical Review. In: Endsley, M.R., Garland, D.J. (eds.) Situation Awareness Analysis and Measurement, pp. 3–18. Lawrence Erlbaum Associates, Mahwah (2000)

# A Model for Fuzzy Logic Assessment of Real Estate Investment Risks

Plamena Zlateva<sup>1</sup>, Dimiter Velev<sup>2</sup>, and George Zabunov<sup>2</sup>

<sup>1</sup>Institute of System Engineering and Robotics, Bulgarian Academy of Sciences  
Acad. G. Bonchev Str., Bl. 2, P.O.Box 79, 1113 Sofia, Bulgaria  
email: plamzlateva@abv.bg

<sup>2</sup>University of National and World Economy  
UNSS, Studentski grad, 1700 Sofia, Bulgaria  
email: dvelev@unwe.acad.bg, zabunoff@gmail.com

**Abstract.** Assessing the real estate investment risks is a major issue for the responsible management and the sustainable regional development. The paper proposes a fuzzy logic model for complex estimation of the real estate investment risks, based on the available information sources and the expert knowledge. The fuzzy logic model is designed as a hierarchical system that includes several variables. This model is intended to be implemented as a Web service in a cloud computing environment as a next natural step for increasing the span and efficiency of real estate manager activities.

**Keywords:** real estate management, risk assessment, fuzzy logic model, cloud computing.

## 1 Introduction

Innovation and regional development in Bulgaria expands over a vast variety of business areas and real estate management is one of them. The real estate market bubble, which was observed two years ago in Bulgaria, showed the real estate manager's central role on that market. Despite the regulations and legal constraints, driven by personal interests real estate managers succeeded to impose their goals on the society. The state control system in the fields of construction and real estate was inefficient and incapable of working in current conditions. These laws were created years earlier, when Bulgarian economy was centralized and there were no private property real estate managers on the market. Few months later it turned out that the laws of the market were stronger than state laws. New self-made real estate managers were supposed to learn how to operate in free market conditions [4].

Risk assessment of construction projects was one of the most complex real estate managers' problems, still waiting for its decision. Due to their characteristics residential real estate projects (investment alternatives) are of particular interest. They have to be assessed in terms of potential risks, risk scale and degree of influence.

There are many qualitative and quantitative methods for the complex risk assessment of investment alternatives. However, it is necessary to point out, that the risk factor assessment is done under the subjective and uncertain conditions. The intelligent methods are an appropriate tool for real estate investment risk assessment. These methods, using the fuzzy logic theory, provide adequate processing the expert knowledge and uncertain quantitative data [1].

The aim of this paper is to propose a fuzzy logic model for real estate investment risk assessment using the available information and the expert knowledge. The fuzzy logic model is designed as a hierarchical system that includes several variables.

The proposed fuzzy logic model will be included as a Web service in a cloud computing environment as a next natural step for increasing the span and efficiency of real estate manager activities.

## 2 Fuzzy Logic Assessment of the Real Estate Investment Risk

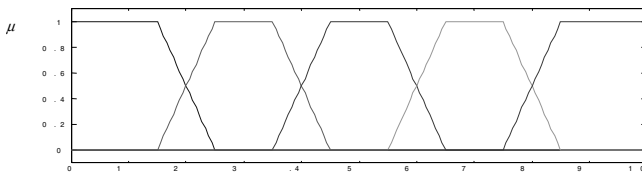
The idea is to design a fuzzy logic model that describes efficiently the subjectivity in the complex assessments of different experts regarding the size of the real estate investment risk for examined investment alternatives with respect to the various risk criteria and factors with different weights.

In the current paper the fuzzy logic model for complex real estate investment risk  $R$  is established on the basis of determined real estate risk criteria  $R_k, k = 1, \dots, m$  and corresponding risk factors  $F_{ki}, i = 1, \dots, n_k, n_k = n_1, \dots, n_m$ .

Five values for the linguistic variables  $F_{ki}, i = 1, \dots, n_k, n_k = n_1, \dots, n_m$  are introduced to reflect five levels for each of the factor types.

The proposed five levels of damages are set with five fuzzy subsets, correspondingly: *Very small*, *Small*, *Medium*, *Large* and *Very large*.

All linguistic variables vary in the  $[0, 10]$  interval and they are set with a trapezoid member functions (see Figure 1).



**Fig. 1** Membership functions of the linguistic variables

A risk node point vector  $r = (r_1, r_2, r_3, r_4, r_5)$  is introduced, which in the particular case has the following form:  $\alpha = (1, 3, 5, 7, 9)$ .

For the linguistic variable – complex assessment of the real estate investment risk  $R$ , five levels are introduced as well, as shown in the Table 1.

Each variable  $F_{ki}, i = 1, \dots, n_k, n_k = 1, \dots, m$  has a corresponding membership function  $\mu_{ij}, j = 1, \dots, 5$  to the five fuzzy subsets.

The membership functions  $\mu_{ij}$  are defined with the following formulae:

$$\mu_{i1} = \begin{cases} 1, 0 \leq F_{ki} < 1.5 \\ 10(2.5 - F_{ki}), 1.5 \leq F_{ki} < 2.5 \\ 0, 2.5 \leq F_{ki} \leq 10 \end{cases}; \quad \mu_{i2} = \begin{cases} 0, 0 \leq F_{ki} < 1.5 \\ 10(F_{ki} - 2.5), 1.5 \leq F_{ki} < 2.5 \\ 1, 2.5 \leq F_{ki} < 3.5 \\ 10(4.5 - F_{ki}), 3.5 \leq F_{ki} < 4.5 \\ 0, 4.5 \leq F_{ki} \leq 10 \end{cases};$$

$$\mu_{i3} = \begin{cases} 0, 0 \leq F_{ki} < 3.5 \\ 10(F_{ki} - 3.5), 3.5 \leq F_{ki} < 4.5 \\ 1, 4.5 \leq F_{ki} < 5.5 \\ 10(6.5 - F_{ki}), 5.5 \leq F_{ki} < 6.5 \\ 0, 6.5 \leq F_{ki} \leq 1 \end{cases}; \quad (1)$$

$$\mu_{i4} = \begin{cases} 0, 0 \leq F_{ki} < 5.5 \\ 10(F_{ki} - 5.5), 5.5 \leq F_{ki} < 6.5 \\ 1, 6.5 \leq F_{ki} < 7.5 \\ 10(8.5 - F_{ki}), 7.5 \leq F_{ki} < 8.5 \\ 0, 8.5 \leq F_{ki} \leq 1 \end{cases}; \quad \mu_{i5} = \begin{cases} 0, 0 \leq F_{ki} < 7.5 \\ 10(F_{ki} - 7.5), 7.5 \leq F_{ki} < 8.5 \\ 1, 8.5 \leq F_{ki} \leq 1 \end{cases}$$

**Table 1** Levels of the real estate investment risk.

<i>R</i> intervals	Levels of the real estate risk
$8 < R \leq 10$	“Very large real estate risk”.
$6 < R \leq 8$	“Large real estate risk”
$4 < R \leq 6$	“Medium real estate risk”
$2 < R \leq 4$	“Small real estate risk”
$0 < R \leq 2$	“Very small real estate risk”

The complex assessment of the real estate risk on the basis of the proposed fuzzy logic model is calculated as follows:

$$R = \sum_{j=1}^5 r_j \sum_{k=1}^m w_k \mu_{kj}(R_k), \quad (2)$$

$$R_k = \sum_{j=1}^5 r_j \sum_{i=1}^{n_k} w_{ki} \mu_{ij}(F_{ki}), \quad (3)$$

$$\sum_{i=1}^{n_k} w_{ki} = 1, \quad n_k = 1, \dots, m \quad (4)$$

The obtained value for *R* shows the level of the real estate investment risk in the examined investment alternatives. The higher value of the *R* variable shows a higher investment risk. A table with example values of risk criteria and factors and the corresponding number values for two alternatives are given in Table 2.

**Table 2** Example values of risk criteria and factors for two investment alternatives.

Risks (Criteria)	Factors (Sub-criteria)	Factors' weights	Alternative projects	
			A	B
$R_1$ : Political risks	Government policy on construction business	0.3	4	4
	Government policy on housing	0.3	1	3
	Municipal policy on housing	0.4	1	3
$R_2$ : Social risks	City planning	0.6	3	3
	Social processes	0.4	4	2
$R_3$ : Economical risks	General state of the economy	0.2	6	6
	State of supply	0.2	5	2
	State of demand	0.3	7	7
	Access to credit (for business)	0.3	6	6
$R_4$ : Contractual risks	Management of the relations with suppliers	0.5	4	5
	Customer relationship management	0.5	2	1

### 3 Real Estate Investment Risk Assessment by Cloud Computing

Cloud computing has recently emerged as an alternative to in-house IT investment. This approach encompasses pay-per-use services that extend IT's existing capabilities. Cloud computing is viewed as an efficient way to increase capacity or add capabilities on the fly – without investing in new infrastructure, hiring or training new personnel or licensing new software. Most cloud services are scalable which means that users with a sudden need for greater capacity can simply increase the level of their cloud service instead of investing in more hardware and software and expanding in-house data centers. Wikipedia defines cloud computing as computing in which dynamically scalable and often virtualized resources are delivered as a service over the Web [5], [6], [7]. Online solutions are used instead of downloaded software to run and manage one's data and business. The term “cloud” is used as a metaphor for the Internet, based on the cloud drawing used in the past to represent the telephone network, and later to depict the Internet in computer network diagrams.

The cloud is an ideal environment for companies and organizations to build and deliver an inventory of business services targeted to individual market segments and specific customers. Business process management (BPM) is the basis for offering business services over the cloud [2]. By using it companies can assemble an appropriate bundle of business services needed to best serve certain kinds of customers. BPM allows real estate companies to break their business processes into collections of interconnected tasks. This is an important step in enabling them to extend their operations beyond their own company boundaries to embrace services provided over the cloud. In this way, they can outsource certain processes and tasks so that they can concentrate on their core value-added processes, continue to improve them, and invent new ones.

The lifecycle of a property is measured in decades while IT is often measured in much more shorter terms. By exporting and basing all email, applications and real estate software in the Cloud can bring a number of advantages, such as [2], [3]:



- Free-up internal resources since IT focuses on software applications and the associated innovations that drive business in comparison with permanent infrastructure troubleshooting.
- Online hosted real estate software costs less.
- Data used by the software is stored online eliminating the need to manage the logistics of data storage or backup locally.
- Real estate software and data are both available from anywhere at any time without the need for complex networking facilities. In such a way Cloud solutions enable employees to work remotely using the same familiar desktop interface, thus drastically reducing commute time and improving employee efficiency.
- Improved Risk Management – risk increases with more IT investments. Cloud services reduce the organization’s dependence on onsite systems by assuming the costs and risks of the entire IT lifecycle, including hardware, backups, security and support.
- Greater financial visibility - cloud-based managed services help accurately forecast the costs of adding new users or locations.

Real estate companies can use the Cloud to reduce their overheads and become more flexible in the way they work – this has implications for the amount of real estate they use and for how investors will assess the property.

## 4 Conclusion

A fuzzy logic model for complex assessment of the real estate investment risks is proposed. The fuzzy logic model is designed as a hierarchical system including different risk factors and their weights. The results can support the real estate managers to take more informed decisions for the sustainable regional development and established investment strategies. The designed fuzzy system is envisaged part of the Web Integrated Information System implemented by Cloud computing.

## References

1. Chen, Z., Khumpaisal, S.: An Analytic Network Process for Risks Assessment in Commercial Real Estate Development. *Journal of Property Investment & Finance* 27(3), 238–258 (2009)
2. Hugos, M., Hultzky, D.: *Business in the Cloud*. John Wiley, Chichester (2011)
3. Marks, E., Lozano, B.: *Executive’s Guide to Cloud Computing*. John Wiley, Chichester (2010)
4. Pelov, T., Jovkova, J., Zabunov, G., Stamenova, V., Tagarev, S., Galabov, M., Bojadzhiev, D., Belev, D., Ignatova, N.: *Economics and Management of Real Property (in Bulgarian)*. Stopanstvo Publishers, Sofia (2011)
5. Reese, G.: *Cloud Application Architectures: Building Applications and Infrastructure in the Cloud*. O’Reilly, Sebastopol (2009)
6. Rittinghouse, J.W., Ransome, J.F.: *Cloud Computing: Implementation, Management and Security*. CRC Press, N.Y (2009)
7. Williams, M.I.: *A quick start guide to cloud computing: moving your business into the cloud*. Kogan Page Limited, London (2010)

# Towards the Foundation for Read-Write Governance of Civilizations

Alois Paulin

A. Paulin D.O.O., Ročevnica 59, 4290 Tržič, Slovenia  
email: alois@apaulin.com

**Abstract.** The research presented in this paper aims towards defining a novel - governance oriented, layer of nodes for interaction between governmental data and data consumers, which aims to satisfy the need for a flexible and participative infrastructure for governing a modern society based on the rule-of-law and existing democratic conventions. We describe herein the Secure SQL Server - a system composed of a data format for describing complex rules of access to data stored in relational databases and a middleware server that mediates read/write interaction between clients and the database management system based on mentioned rules. Rather than designing a brand new protocol for this citizen-to-government interaction, we make use of existing well-supported standards, namely SQL, XML, XAdES and HTTP to model a system that fully complies with legal requirements of modern governments and which allows both read and write access to governmental data based on complex rules applied at run-time.

## 1 Introduction

In the past 10 years many great governmental e-services have emerged that enable citizens to be more productive: in Slovenia for example – besides many other online services, laws and court rulings are published on the web, cadastral information can be retrieved online, interaction with the land registry is conducted entirely in electronic form, tax forms can be submitted via web services, relevant information about legal subjects is available online in the electronic business register and even companies can be registered through the Web.

Architecturally, each of these governmental web services is an n-tier application, where the presentation tier consists of a web server serving either a *graphic user interface* (GUI; defined typically in HTML, CSS and JavaScript) and/or an *application programming interface* (API; typically SOAP). The business logic of such applications is defined in the middle tier and contains rules that have been hardcoded at design-time. Consequently, these e-services serve fixed, predefined use-cases and it can be said that governmental web applications behave *digital*, as overall, only a discrete number of immutable services are available for the subject to access.

In contrast to the static, *digital* offerings, which are able to sufficiently satisfy only predefined requests, the relations among subjects and between subjects and

the government – in a society based on the rule-of-law, behave *analogue* and are highly dynamic.

The protocols of accesses to governmental electronic services or data, as well as their mere availability, are usually not regulated by primary law (law issued by the parliament). Instead, either secondary law (legal acts defined by non-legislative bodies such as ministries and public agencies) defines crucial characteristics of the particular services, or existing bureaucratic rules are arbitrarily translated to the electronic dimension. Consequently the legal boundary, within which electronic services perform, is not always clear.

The introduction of a new law or a change of the existing can render e-services obsolete or even illegal, and the same applies for court decisions or rulings of inspectorates. Even a simple change in the organizational or political structure of the public body in charge of the e-service can result in a major re-development of the particular system. But unplanned change of legal boundaries is not the only challenge of modern governmental e-services.

Governmental e-services are planned, developed and maintained by individual governmental organizations. Effectively this means that either the governmental organization contracts a private company to build the information system plus web frontend, or that a state-owned company conducts this work. Either way, the resulting e-service application comes as a “black box” with hardcoded rules and a web frontend targeting a limited number of user agents (UA; e.g. web browsers).

Unlike the private sector, which fosters strict and public standards for data transfer between tiers of information systems (cf. e.g. EBICS<sup>1</sup>), public sector e-services lack technical regulation. Law falsely seems to perceive e-services as monolithic, one-tier applications and correspondingly regulates only the human interaction with the presentation layer, while leaving technical issues to arbitrariness. An example of inadequately regulated governmental e-services is the Slovenian “e-Justice” system: In 08/2010, Slovenia’s Ministry of Justice passed and published an ordinance that regulates electronic transactions in the field of e-justice<sup>2</sup> It would be expectable that this ordinance would rigorously define the format for data exchange and the technical protocol, in a way technical standards do it, but instead the ordinance talks about indefinable concepts such as a “portal e-justice” (no URL defined – where should we find it?), and “electronic applications” (no format defined – is it XML + XML-DSig/XAdES? If so – where is the schema? Is it a Pkcs-7 signed plaintext message?). In order to submit the e-application, the ordinance instructs the citizen to “choose the corresponding ‘e-task’ on the portal and enter data into the required fields of the provided form”. This ordinance reads like a user-manual for a particular software product and does in no way define or mention any technically relevant characteristics. A similar style of regulating technical procedures has been chosen for the 2011 amendment

---

<sup>1</sup> The *Electronic Banking Internet Standard* is obligatory in Germany and is used for straight-trough processing of SWIFT orders in electronic banking.

<sup>2</sup> Rules on electronic operations in civil procedures / Pravilnik o poslovanju v civilnih sodnih postopkih.

of Slovenia's Land Register Act<sup>3</sup>, which defines the electronic land register as software backend including several "modules" and a public web portal as the frontend.

This kind of regulation does not allow a technically clear implementation; neither does it regulate the interaction between the server and the client. Several important issues are undefined, such as: Where is the URL? How should HTTP requests be formulated? What data format does the web server accept and respond? How does the client authenticate to the server? Who guarantees that the technical protocol will not arbitrarily change over time? Who is the responsible legislative body that can be called for accountability?

These questions are not only of importance from a technical perspective; they are important legal questions as well. The exchange of requests and responses between the web server and the UA is essentially a series of interactions between the citizen and the government in which the citizen's HTTP request is a formal application towards the government, which must evaluate the application and respond lawfully. Additionally, the rule-of-law principles require that the rules, against which applications are evaluated, are transparent and published in advance.

The interaction between citizens and the government over the Internet is a novel experience for both the legislative, executive and the judicative branch, which all severely lack the technical knowledge required to cope with the challenge of structured data exchange. While ICT-literacy of the population rises, new situations will have to be resolved, such as: Has the government permission to prevent citizens from interacting with governmental services in an automated manner – e.g. through bots (cf. [1])? Or: Is it legal to force citizens to use only certain (though undefined) user agents and system configurations to interact with e-services?

Arbitrariness of the design and non-existence of technical regulation for governmental e-services could potentially represent a breach of the Human rights<sup>4</sup>, esp. art. 21/II, which states: "Everyone has the right of equal access to public service in his country."

The third important issue with modern governmental e-services is their efficiency. While the first two described problems are to our knowledge not yet present on the international research agenda, the issue of is a major topic within the *open data*, *linked data* and *open government* communities (hereinafter OGD). OGD is a field of research that is concerned with the transparency of governmental data. The basic idea behind OGD is that governments and public organizations/bodies should make their data available online for the public to consume and to draw added value out of it.

The OGD movement took shape with the rise of several OGD project "in countries around the world from the United States, Australia and New Zealand to The Netherlands, Sweden, Spain, Austria and Denmark, not to mention and increasing number of city- and local-authority-based initiatives from Vancouver to London" [2]. However, despite their positive vision, OGD portals have become dumping-yards for governmental analyses and high-level statistical data with little

---

<sup>3</sup> Land Registry Act / Zakon o zemljiški knjigi.

added value. Authorities have published barely relevant statistics about child-seat safety, the jail population and the population count for wild horses and burros [3]. According to nonpartisan organizations, US federal agencies which by Obama's decree had to publish at least three of their high-value sets of statistics or other information in a downloadable format, "went for the low-hanging fruit for things that are already out there and not terribly controversial" (ibid.).

Furthermore, governmental OGD initiatives rarely offer their data in a coherent, structured, machine-readable format, but instead focus on providing various dedicated web sites with relatively little added value. Robinson & al. [4] demand that the government should shift its focus away from designing "sites that meet each end-user need" towards "creating a simple, reliable and publicly accessible infrastructure that 'exposes' the underlying data. Private actors, either non-profit or commercial, are better suited to deliver government information to citizens and can constantly create and reshape the tools individuals use to find and leverage public data" (ibid.).

O'Reilly [5] shares Robinson's concerns regarding governmental involvement in developing web pages and calls for a *government as a platform* (GaaP). The GaaP idea envisions the hegemony as a provider of infrastructure on which subjects can conduct their exchange of goods and services in a transactional manner. According to this idea, "Government 2.0 is not a new kind of government; it is government stripped down to its core, rediscovered and reimagined as if for the first time." (ibid.) O'Reilly's *platform* is an analogy to modern computer platforms, like iPhone or Android, hence the vision encompasses a two-tier architecture of e-governance.

Although we reject the API approach as proposed by O'Reilly, we follow his call to *reinvent governing*. Consequently our research presented in this paper first presents a short insight into the fundamentals of governing as elaborated in the fields of political philosophy and jurisprudence. Based on these theoretical findings we describe a novel technical solution that responds to the problems outlined hereinbefore.

## 2 Rights – Structured Pieces of Information Stored in a Database

Rights are the fundamental legal relations between subjects and the sovereign within a governed society. Social contract theory (cf. [6, 7]) tells us that each society is grounded in its *social contract*, an implicit mutual agreement between the society's members about their rules of conduct. At the constitution of civil society each member of the community surrenders his natural liberty and all the resources at his command - including the goods he possesses, to the community (the State), which in return gives him civil liberty and *proprietorship* over all his possessions [7]. Therefore "the State, in relation to its members, is master of all their goods by the social contract, which, within the State, is *the basis of all rights*" (ibid. book 3/1). Contrary to natural liberty, which can be exercised within natural borders (walls, rivers, gravity) and social liberty, which is limited with social borders (morals, habits, conventions), rights represent artificial liberty, which must be granted by the sovereign in order to exist.

By granting a right, the sovereign creates a virtual space of legal liberty, and promises the grantee not to interfere with the subject's execution of the right; cf. [8]. Eventually, the sovereign may promise to defend the right, which it does by establishing a defensive system of subjects who have the right to use repression in order to secure the given right (e.g. a police force, a judicial system, state attorneys, etc.).

Rights are of various flavors and names – the right to exploit natural materials is granted through a concession, the right to teach at a school or university is obtained through habilitation or tenure, the right to lead a mission is called an appointment, rights in political issues are called mandates, etc. But as soon as we disregard the naming, the requirements and the procedure in obtaining it, the right at its basic level is information about the expressed decision of the person (or jointly of a committee) in charge, stored in a database. (A more elaborate argumentation is available in our previous work on this topic: [9]; for a juridical theory on rights see also [10].)

Every expressed right can be defined as a set of discrete electronic data that can be stored in a relational electronic database. Many rights are already today stored exclusively electronic – the land registry in Slovenia for example is from 2011 on kept electronic only and the information about rights is stored in a relational database.

### 3 The Secure SQL Server

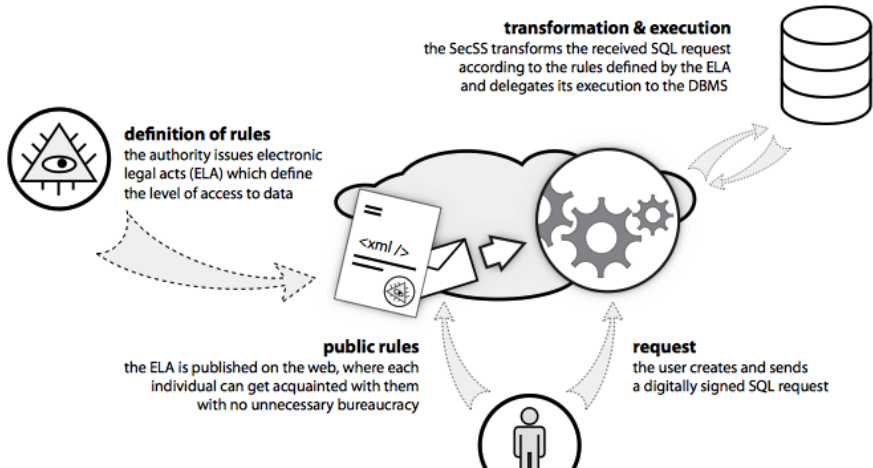
The *Secure SQL Server* (SecSS; Fig. 1) is a novel electronic interface, which allows the public as well as known users to fully transactional – without human intervention, read and write data in remote relational databases using digitally signed standardized SQL statements.

SecSS allows users to send digitally signed SQL queries of any kind to a publicly known URI on the server. The semantics of the query are not important – it can be a simple *freedom of information request* for public data in form of a SELECT statement, a *registration in the Land Register* expressed through an UPDATE statement, request for *matriculation to an university*, *application for a governmental job* or even just a *bidding at a public auction* expressed through INSERT statements.

After SecSS receives the SQL request, which is treated as a formal application, it first validates the digital signature and the signer's certificate. Based on the identity stored in the certificate, SecSS can apply personalized rules, if rules for the identity are explicitly defined. After the identity-check, SecSS dynamically applies public and personal rules to the original SQL request in form of SQL sub-queries, which limit the range of data to which the applicant has access-permissions.

The rules, which SecSS applies to the original request, are stored in an XML Infoset, which has been signed by an official with a legislative mandate. This Infoset is a legally binding set of rules, which are subject to usual legal principles and is called the *electronic legal act* (ELA). Because the ELA is public, anybody can view and evaluate the rules contained. In case that somebody suspects that the ELA violates her rights or that it is in any form unlawful, the disputed validity of the ELA can be brought to an inspectorate's attention, or even evaluated by a court.

Each rule defined in the ELA is an explicit SQL query bound to a field in case the specified statement type – e.g. *update*, *insert*, or *select*, is requested. The rule may include SQL variables, to which values from the original request are assigned. Each rule has also access to the identity of the applicant, which is crucial for strictly personal applications, e.g. when a change of ownership at the applicant’s real estate is requested. This SQL statement is later applied as a filter to the applicant’s request.



**Fig. 1** SecSS enables dynamic, fully transactional access to governmental data in accordance to the principles of the rule-of-law

In order to apply the correct rules, SecSS utilizes a SQL language parser, which analyses the incoming request and extracts the mentioned fields. The applicant can only access fields for which an explicit regulation exists. If the applicant tries to access fields that are not regulated by the ELA, the application is not processed.

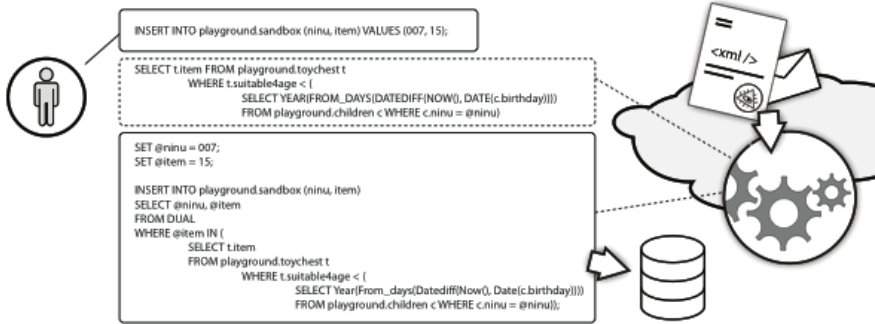
SecSS provides only an electronic interface that can be accessed over the Internet – HTTP may be preferred, but also other protocols, like SMTP should not be discriminated. Developers on the free market should provide higher-level applications that e.g. allow users to interact via a graphical user interface.

Furthermore, every request that is received by SecSS must be stored in its original form, as well as the corresponding response. This assures that each formal application is appropriately archived so that in case of a future dispute the non-repudiable request of the applicant can be evaluated again.

### **Proof of Concept: The Sandbox**

In order to prove and demonstrate the working of SecSS, we have published a prototype server and client application. The server has access to a MySQL database management system. The database hosts the testbed scheme “playground”, which represents a fictional playground on which children can play with toys in a sandbox. For each child we store the following personal information: The national

identification number (ninu), name, surname, and the date of birth (birthday). The toys are kept in the toychest and for each toy we know its unique ID (item), the name of the toy, the image and information about the suitable age (suitable4age). In the sandbox we store information about which child (ninu) is playing with which toy (item) and where the child is geographically located within the sandbox (posx and posy).



**Fig. 2** The SecSS applies rules in form of SQL sub-queries to the original request in order to allow access only to the permitted sub-quantity of data

To the playground scenario several simple and complex requirements apply:

1. The public may *read* any data except the children's birthdays, which are protected personal information by law.
2. A child may play only with toys for which it is old enough.
3. If a child plays with a toy, it must not be given to another child.
4. Anybody may at any time add a new toy into the toy chest or put a new child into the playground.

The given conditions are complex and cannot be handled by individual rules applied to either the children or toys. Instead, the rules must be generic and applicable to all requirements.

Requirements #4 and #1 are simple; they can be realized by applying appropriate read or write permissions for the particular fields. However, requirements #2 and #3 are complex and require the definition of filters in form of sub-queries. Fig. 2 shows the sub-query for requirement #2 and demonstrates how an INSERT-statement is transformed by the corresponding rule before it is sent into the MySQL server.

The *playground* prototype application, which is available online at <http://sex.apaulin.com>, is a proof-of-concept that SecSS is capable to handle complex real-life scenarios.

Let us take for example that an ordinance would be passed, which would make it illegal to play with toys made in Azerbaijan: in that case, the authority in charge of the relational database would have to add a new field to the table *toys* - *countryOfOrigin* and add the appropriate SQL sub-query to the ELA. Besides those two simple changes, no additional modifications would be needed, neither server- nor client-side.



## 4 Conclusion

In this article we argued that rights could be effectively represented as structured data within relational databases. Consequently a society can be governed in a similar manner as a player of computer games governs his virtual community – by making actions that effectively trigger the change of values of digital objects in the virtual sphere of the computer game.

Following the finding that a change of data in a governmental relational database can have real legal impact on the rights within the real world, we developed a prototype system – the Secure SQL Server (SecSS) which allows fully transactional access to databases of rights according to legal, technical and political maxims of modern rule-of-law based states.

We believe that SecSS can revolutionize the way modern societies are governed and that our system can make a significant contribution to development towards fully transactional, self-service governance with practically no need for bureaucrats.

## References

1. Lundblad, N.: e-Exclusion and Bot Rights. *First Monday* 12 (2007), <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/251> (accessed June 23, 2011)
2. Sheridan, J., Tennison, J.: Linking UK Government Data. In: *Proceedings of the Linked Data on the Web Workshop LDOW 2010*, vol. 628 (2010), <http://CEUR-WS.org/Vol-628/> (accessed June 23, 2011)
3. O’Keefe, E.: Info released under Obama transparency order is of little value, critics say. *The Washington Post* (2010), <http://www.washingtonpost.com/wp-dyn/content/article/2010/01/27/AR2010012704589.html> (accessed June 23, 2011)
4. Robinson, et al.: Government data and the invisible hand. *Yale Journal of Law & Technology* 11 (2009), <http://www.yjolt.org/files/robinson-11-YJOLT-.pdf> (accessed June 23, 2011)
5. O’Reilly, T.: *Government as a Platform*. O’Reilly Media, inc. (2010), <http://ofps.oreilly.com/titles/9780596804350/> (accessed June 23, 2011)
6. Hobbes, T.: *The Elements of Law Natural and Politic*. Published online by: University of Virginia Library (1650) (2002), <http://etext.lib.virginia.edu/toc/modeng/public/Hob2Ele.html> (accessed June 23, 2011)
7. Rousseau, J.J.: *The Social Contract*. Published online by: Thomas S (1762) (2010), Adelaide, [http://ebooks.adelaide.edu.au/r/rousseau/jean\\_jacques/r864s/](http://ebooks.adelaide.edu.au/r/rousseau/jean_jacques/r864s/) (accessed June 23, 2011)
8. Jellinek, G.: *System der subjektiven öffentlichen Rechte*. J.C.B. Mohr, Tübingen (1905)
9. Paulin, A.: *The Role of Rights in the Transactional Civilization*. In: Parycek, P. (ed.) *Proceedings of the International Conference for E-Democracy and Open Government*, Danube University Krems, Austria (2011), <http://apaulin.com/research/2011/CeDEM11-Rights>
10. Hohfeld, W.N.: *Fundamental Legal Conceptions as Applied in Judicial Reasoning and Other Legal Essays*. Yale University Press, New Haven (1920)

# An Analysis and Forecast of Software and Services Research in Bulgaria – Comparing the Situation Two Years Later

Irena Pavlova<sup>1</sup>, Desislava Petrova-Antonova<sup>1</sup>, Avram Eskenazi<sup>2</sup>, Sylvia Ilieva<sup>1</sup>, and Aleksandar Dimov<sup>1</sup>

<sup>1</sup>Faculty of Mathematics and Informatics, Sofia University “St. K. Ohridski”, Sofia 1164, J. Bouchier str., No. 5  
email: {irena\_pavlova, d.petrova, sylvia, aldi}@fmi.uni-sofia.bg

<sup>2</sup>Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia 1113, Acad G. Bonchev str., bl.8  
email: eskenazi@math.bas.bg

**Abstract.** In [1], we have published the results of our analysis of software and services research in Bulgaria during the 2004-2008 time period. This analysis was part of the activities performed within the EU FP7 SISTER project. By using the same methodology, which is expanded now with some new features, we started analyzing a new period – 2009-2010. In this paper we compare the results obtained for both periods (overcoming their different duration through a simple data transformation). The first results obtained show an intensification of the research work in the majority of the subject areas in view with a few new appearing now. The analysis is planned to be finalized in the next few months.

**Keywords:** ACM Computing Classification System, software and services, software engineering, research, priorities, methodology.

## 1 Introduction

This paper presents an updated view of the future in the area of Software and Services (S&S), primarily focused on the Faculty of Mathematics and Informatics (FMI) at Sofia University. It is based to the one, presented in [1], which was derived from a systematic analysis of the current state and future trends in Bulgaria and other countries worldwide covering a period of 4 years – from 2004 until 2008. The current analysis covers the next two years period – 2009-2010 – and in addition an initial comparison of the results from both views is presented.

This investigation was performed within the FP7 SISTER project, which aims at supporting the strengthening of the research capacity of FMI in Software and Services. The main purpose of project’s tasks is to enable the FMI research group in S&S to capitalize on existing research capacities, while providing a strategy for harmonization of research focus. It has to identify topics of S&S research that (i) constitute research urgencies at European levels – from NESSI technology platform initiative roadmap, as well as other emerging working groups roadmaps

and FP7 priorities (ii) map on local and regional interests and capacities, and (iii) are in the research focus of FMI research group. Another goal is to propose and apply a research approach that should increase the benefits to the research group.

Within this general task we intend to refine the methodology for identification of research priorities, as described in [1], into several directions. First, we should take into account the views of Bulgarian industry on the research priorities. The real exploitation of research results in industry is identified as an important factor. Second issue is to resolve the problem with publications and other sources, which cover more than one area, according to the accepted classification scheme. For example research related to security was investigated in different contexts. Last issue is to break down the broad area of “Online information services” into sub-areas because of the very intensive research for the last years in the area. Also the refined methodology is to be applied on the data from the last two years. The obtained results are compared with the results for the period of 2004–2008 and predictions for the most promising research areas in S&S are made, which should be the main focus in Sofia University.

The rest of the paper is organised as follows: Section 2 “Methodology” introduces shortly the methodology applied for identification of the appropriate research topics and the sources of information used; Section 3 named “Results” summarises the data collected from conferences and journals, defended and current PhD thesis, research projects; Section 4 concludes the paper by pointing out the open issues and the possible ways to improve the methodology for identification of research priorities in Software and Services field.

## 2 Methodology

### 2.1 General

The methodology we use is described in [1]. Here we will only briefly mark the main points.

First decision to be taken was to decide about the most appropriate classification frame for the needs of the task. After analyzing ACM CCS 1998 [2], SWEBOK[3] and relevant standards, we decided that ACM CCS would be more appropriate mainly because of the degree of its granularity, as well as its relatively higher popularity.

We believe that the main determining factors are:

- The current state and trends of the research in the world,
- The capabilities, traditions and trends in Bulgaria in S&S research,
- The needs of the employers in Bulgaria.

In order to take all these factors into account we modified the basic idea of [4]. Its theoretical basis was provided by elements of the theory of non-equilibrium thermodynamics in open systems [5], whose main points were projected on the fields of information science and science of science.

Consequently, first we have to collect the appropriate information. For this purpose, we determined the following criteria:

- Scientific publications,
- PhD theses completed and defended,
- PhD theses under development,
- Research projects,
- The opinion of representatives of the Bulgarian software industry.

Few other criteria were identified but rejected due to problems with the collection of relevant information.

During the investigation in the first period (2004-2008) we took into account the opinion of foreign partners. We also collected the understanding of experts about the importance of each criterion selected through a questionnaire and following a formalized procedure [6] determined weights and ranking [1].

The number and importance of citations would also be a valuable source of information, but our efforts proved it to be quite difficult and expensive to obtain full and reliable data.

Because of the fact that one of our goals is to compare the 2004-2008 and 2009-2010 periods, we had to decide on how to compensate the difference in their duration. A logical decision is to apply the following formula (taking into account that the duration of the first period is 5 years and the second one is 2 years):

$$(N-O*0.4)/(O*0.4)*100. \quad (1)$$

where  $O$  is the number of objects for a given parameter in the first period,  $N$  is the number of objects for a given parameter in the second period.

## 2.2 Sources

To cover the first criterion – scientific publications, we reviewed the following sources.

A. Journals for the 2009 and 2010 years:

- Comptes rendus de l'Académie bulgare des Sciences,
- Cybernetics and Information Technologies,
- Serdica Journal of Computing,
- Information Technologies and Control,

B. Scientific events – International conferences in Bulgaria for the 2009-2010:

- Information Technologies,
- Computer System and Technologies,
- Software, Services and Semantics Technologies (S3T)
- Challenges in Higher Education and Research in the 21st Century
- Automation and Informatics,
- Automation and Informatics - School for young scientists.

At this stage we don't include events held outside Bulgaria because we are still not in a position to encompass all relevant data.

For the second criterion – PhD theses defended - we used the database [7] of the former Higher Attestation Commission (HAC), as well as the protocols (2009-2010) of the Scientific Council on Informatics of HAC.

For the third criterion - PhD theses under development – we contacted the respective universities and research institutes, authorized to tutor PhD students.

For the fourth criterion – research projects in the area - we investigated the CORDIS data base [8] for Framework Programme (FP) 7 projects and National Science Fund data base [9] for national projects.

For the last criterion - the opinion of representatives of the Bulgarian software industry and trade - we created and distributed questionnaires. We are still in the process of collecting the answers.

### 3 Results

This section of the paper presents the analytical results of our study, with respect to each of the measures identified above.

#### 3.1 Publications

The data collected, is summarized in Table 1. In Figure 1 we can observe in a diagram form the relative change of publication activity for both periods.

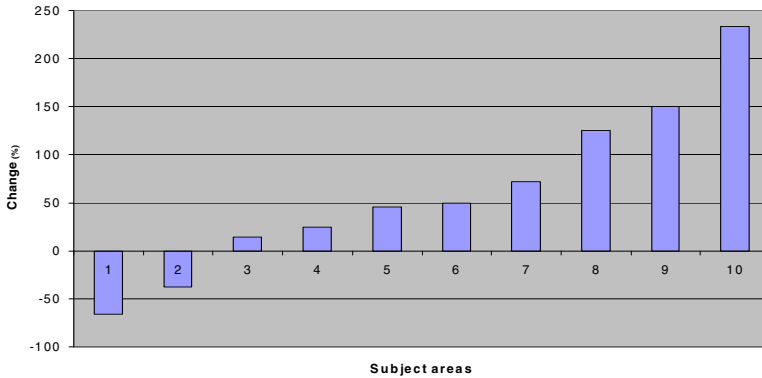
**Table 1** Distribution of research papers according to the ACM classification index

No	Classifier	O 2004-2008	N 2009-2010	$(N-O*0.4)/(O*0.4)*100$
1	H.2.4 Systems	22	3	-66
2	D.2.13 Reusable Software	4	1	-38
3	H.3.5 Online Information Services	46	21	14
4	D.2.5 Testing and Debugging	6	3	25
5	D.4.6 Security and Protection	12	7	46
6	D.2.2 Design Tools and Techniques	30	18	50
7	D.2.11 Software Architectures	16	11	72
8	D.2.9 Management	10	9	125
9	D.3.4 (Software) Processors	1	1	150
10	D.1.3 Concurrent Programming	3	4	133
11	D.2.4 Software/Program Verification	1	8	1900

For the other half of the 22 subject areas there was no activity in either of the periods. Therefore, no change in percentage can be calculated. But such calculations are not vital to our research. We can conclude that these areas are of marginal interest for Bulgarian authors. Moreover, the number of publications for each of them is most often 1 or 2. The only exceptions are H.3.3 “Information search and retrieval” with 6 and D.2.12 “Interoperability” with 4. An explanation might be that both areas are of a particular importance for the implementation of software services. Because of the fact that this is all data from the second period, we can interpret it as a recently born interest.

The half of the subject areas presented in Table 1 reflects those areas where a publication activity has been observed for both periods. In 9 out of 11 of them there is

a positive change – a relative increase of the number of publications. In most cases this growth is “normal” – between 14 and 72%. In three cases – D.2.2 “Management”, D.3.4 “(Software)Processors” and D.1.3 “Concurrent Programming” the change is between 125% and 233%, however the data about D.3.4 has to be neglected, since it is “artificial” – 1 publication for each of the two periods. The last case – D.2.4 “Software/Program Verification” shows a tremendous growth of 1900%. Obviously, it should be regarded in the same way as H.3.3 and D.2.12.



**Fig. 1** Change in % of the number of research papers (subject areas 1-10 of Table 1)

### 3.2 PhD Theses

We broke down the data accumulated into two groups

- PhD (and DSc) theses completed and successfully defended
- PhD (and DSc) theses still ongoing.

Table 2 shows the status of the theses completed for both investigated periods. As we already explained, we compensate the interval duration difference by multiplying with 0.4 the data concerning the first period.

The rows marked in dark grey do not belong to the area of Software and Services, but to the larger area of Computer Science (Informatics). We include them to show the correlation between the two. The most obvious conclusion is that the number of theses completed and defended has almost doubled during the second period. The fact that the number of theses in S&S is still more than three times lower that the total number could be commented in various ways, but we refrain from doing this at this stage of our research. Unfortunately the numbers in the various areas is so low, that it is not appropriate to make general conclusions about the development trends relying just on this criterion. As far as the ongoing theses are concerned, we succeeded in collecting the relevant information from four of the main institutions (universities and research institutes) tutoring PhD students, but still need to cover the remaining ones. However, as preliminary information we can convey that there is a very substantial increase of the number of PhD students in S&S. They are spread over a lot of subject areas with 1 to 4 PhD students in each of them.

**Table 2** Distribution of PhD theses according to the ACM classification index

No	Classifiers	2004-2008	2009-2010
1	C.2.1 Network Architecture and Design		1
2	D.1.3 Concurrent Programming		1
3	D.2.2 Design Tools and Techniques	2	1
4	D.2.5 Testing and Debugging	1	
5	D.2.9 Management	1	
6	D.2.11 Software Architectures	1	
7	D.2.12 Interoperability		1
8	D.4.6 Security and Protection	2	2*
9	G.2.2 Graph Theory		1
10	H.2.1 Logical Design (DB)		1
11	H.2.7 Database Administration		1
12	H.3.5 Online Information Services	2	2
13	H.5.1 Multimedia Information Systems		2
14	I.1.1 Expressions and Their Representation		1
15	I.2.7 Natural Language Processing		3*
16	I.2.11 Distributed Artificial Intelligence		1
17	I.3.3 Picture/Image Generation		1
18	I.5.4 Pattern recognition applications		1
19	I.6.3 Simulation and modelling applications		1
20	I.6.8 Types of simulation (distributed)		1
21	J.1 Administrative data application (education)		1
	Total S&S (non-marked rows)	<b>9</b>	<b>7</b>
	Relative growth S&S (%) :	<b>94</b>	
	Total Computer Science (2009-2010)		<b>23</b>
	* - 1 of them is a DSc		

### 3.3 Projects

During the last two years (2009-2010) funding for research at national level has drastically been decreased. The National Innovation Fund had no calls for proposals in 2009 and 2010. Since 2010 it has been operating under EUROSTARS European programs, oriented towards SMEs. Unfortunately, there are still no projects with Bulgarian participation. The National Science Fund is the main source of research funding. While in 2008, 24 projects in the area of S&S were funded, in 2009 and 2010 the numbers decreased to 17 and 13 accordingly. Table 3 presents the research areas identified as most topical with respect to the number of national projects.

The participation of Bulgarian organizations (both from academia and industry) in EU research programs and projects is still low. The figures presented here are based on the results of the six FP7 ICT Calls with focus on Calls 3, 4, 5 and 6 (2009-2010).

The number of proposals submitted in FP7 Call 5 and 6 is lower than the number of proposals in FP7 Call 1. Still, according to CORDIS [8], S&S is the area in which Bulgaria has shown best performance with 11 new funded projects. It should be noted that for compliancy with FP7 ICT classification, in this paper S&S are considered in a broader sense, including communication networks and other application areas. On Table 4 a summary of projects distribution per ACM topic is presented.

**Table 3** Distribution of S&S national research projects

No	Classifier	National Scientific Fund 2008	National Scientific Fund 2009	National Scientific Fund 2010
1	C.1.3 Other Architecture Styles	2		2
2	C.1.4 Parallel Architectures		1	1
3	C.2.0 Computer Communication Networks	4	2	
4	C.3.3 Real-time and embedded systems	1	2	1
5	D.2.0 Software engineering (K.5.1)	1	1	
6	D.2.5 Testing and Debugging		1	
7	D.2.11 Software Architectures	2		2
8	D.2.13 Reusable Software	1		
9	D.2.m Miscellaneous	3		
10	H.2.7 Database Administration			1
11	H.3.5 Online Information Services	2	2	
12	H.4.0 Information Systems Applications	3	1	
13	H.5.1 Multimedia Information Systems		2	3
14	J.7.0 Computers in Other Systems		1	
15	J.2.0 Physical Sciences and Engineering	4		
16	J.3.0 Life and Medical Sciences		2	2
17	K.4.2 Social Issues	1	2	1
	<b>Total</b>	<b>24</b>	<b>17</b>	<b>13</b>

**Table 4.** Distribution of Software and Services projects with Bulgarian participation in FP7

No	Classifier	FP7 2007-2008	FP7 2009-2010
1	C.2.0 Computer Communication Networks	1	2
2	D.2.11 Software Architectures	3	
3	H.3.5 Online Information Services	2	2
4	H.4.0 Information Systems Applications		1
5	H.5.1 Multimedia Information Systems	1	1
6	J.7.0 Computers in Other Systems		1
7	J.2.0 Physical Sciences and Engineering	1	
8	J.3.0 Life and Medical Sciences	2	2
9	K.4.2 Social Issues	1	2
	<b>Total</b>	<b>11</b>	<b>11</b>

A general conclusion can be made that more appealing are research projects, in which the research interest is shifting from core technological software engineering areas to more practical and close to industry and society ones.

## 4 Conclusion

The goal of this paper was to refine the approach, applied in [1] to determine the research priorities in S&S field for Sofia University and extend the data collected for analysis. We are reporting early results from our investigation and still there are collected data which we are analyzing. Comparing the periods 2004-2008 and



2009-2010 some positive trends can be noticed from the data collected and systemized – increase of scientific activity in selected areas of S&S. This is particularly valid for the first criterion – scientific publications.

For topical research directions, like *Online Information Services, Testing and Debugging, Security and Protection, Design Tools and Techniques* and *Software Architectures*, there is a relatively small change in the research activity. The positive change reveals that these topics are still of importance for researchers. Bigger figures for other directions like *Software/Program Verification* could be a result of either increased activity or an accidental event - a longer research period is needed to refine this in recent years.

In the following few months we plan to concentrate on the following topics in order to complete our research:

- Finalizing the data about ongoing PhD theses
- Systematizing the answers of the representatives of the Bulgarian software industry;
- Trying to properly resolve the classification problems with some objects (PhD theses, publications) and the subject area of *Online information services*.

The current investigation and the obtained results could be a base for extension in a broader context - not only for Sofia University but all research organizations in S&S area in Bulgaria.

**Acknowledgments.** The work reported in this paper is partly supported by the SISTER project, funded by the European Commission in FP7-SP4 Capacities under agreement no. 205030.

## References

1. Ilieva, S., Eskenazi, A., Dimov, A., Pavlova, I.: An Analysis and Forecast of Software and Services Research in Bulgaria. *Serdica Journal of Computing* 4(2), 133–158 (2010)
2. <http://www.acm.org/about/class/1998/>
3. <http://www.swebok.org/>
4. Pavlovska, E.: Early Identification of Development Trends in Science. In: II International Forum on Information and Documentation, vol. 16(1), pp. 28–33 (1991)
5. Glensdorf, P., Prigogine, I.: *Thermodynamic Theory of Structure, Stability and Fluctuations*, NY (1971)
6. Saaty, T.L.: *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill, New York (1980) ISBN 0-07-054371-2
7. <http://phd-center.bvu-bg.eu/index.php?Cmd=bazidanni>
8. CORDIS, [http://cordis.europa.eu/fp7/projects\\_en.html](http://cordis.europa.eu/fp7/projects_en.html)
9. Bulgarian National Science Fund, <http://www.nsfb.net>

# A Fuzzy System Model for Task Implementation in Extreme Programming Process

Ensar Gul

Marmara University, Department of Computer Engineering,  
Göztepe Kampusu,34722, Istanbul, Turkey  
email: ensar.gul@marmara.edu.tr

**Abstract.** Software simulation enables us to see how a software process is working and gives indications of some of the parameters of the process. In this study the implementation of the tasks in extreme programming software process was modeled using a fuzzy system. The inputs of the system are defined as communication between pair programmers, the writing of unit tests and the coding rules, while the output is the implemented task. The defuzzified output of this fuzzy system provides quantitative results that can be used to determine to how good a task has been implemented.

**Keywords:** software process models, agile software development, extreme programming, fuzzy systems.

## 1 Introduction

There has been a considerable effort in the field of software process simulation to understand the dynamics of software development before actually implementing the code, thus avoiding any pitfalls that may occur during the development. Software process simulation is also used to explain why software processes perform as they do [1][2].

Wernick and Hall [2] define software process simulation as:

“...a software process simulation is a simplified abstracted model, enactable on a computer, of a real or proposed software development or evolution process, usually producing results reflecting real situations or the expected results for proposed process changes”

In software engineering it is easy to propose hypotheses; however, it is very difficult to test them [3]. Raffo [4] provides us with a list of the benefits of software process simulation; however, in this study we will only concentrate on the task implementation of the extreme programming (XP) software process. The task implementation is the core activity of extreme programming, and the model of task implementation provides important clues and suggestions for software development managers who use XP in their projects.

The software development process involves a great deal of input and uncertain variables; therefore, finding a mathematical model to simulate a development process is rather difficult. Fuzzy systems enable us to model systems where a mathematical model does not exist or does not work properly in practice. In this paper, a fuzzy system model is proposed to simulate the implementation of tasks in extreme programming practices.

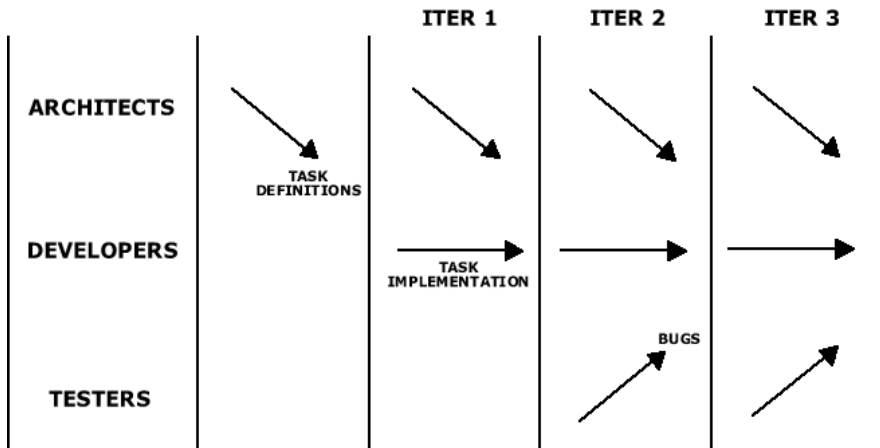
In section 2, how extreme programming and fuzzy systems are related to one another will be explained, while the fuzzy model for task implementation is described in section 3. Finally, the conclusions that can be drawn are given in section 4.

## 2 Extreme Programming and Fuzzy Systems

Extreme programming is a lightweight software process. It reduces software development risks, adapts itself to changing requirements, increases productivity, and reduces overall development cost [5][6][7][9]. The most important aspects of XP that differ from other software development methodologies may be listed as: being lightweight, agile, test-oriented and pair-programming [6][10]. The practice of implementing tasks by pair programmers is modeled as a fuzzy system which takes into account the XP rules that affect the development of a code by a pair of programmers. The task implementation is the core activity in XP, and it is for this reason that it was selected to be examined here.

In a typical control system it is assumed that the system model exists and that it can be modelled by mathematical equations, such as state space or differential equations [11]. On the other hand, in practice it is not always possible to make a complete model of a system. To be able to control such systems, fuzzy logic is more appropriate [12][13]. The fuzzy systems were pioneered by L. Zadeh [14] and have gained widespread applications in control systems. Fuzzy logic provides an unorthodox approach to control problems. This method focuses on what the system should do, rather than trying to understand how it works. One can thus concentrate on solving the problem rather than trying to model the system mathematically [15]. Large scale software is much more complex than a typical control system that is controlled by a fuzzy logic controller, but both systems approach the desired values as corrective actions to be taken at every increment. Levary and Lin [16] developed a software tool which includes two expert systems that use fuzzy engines to simulate software development process. This study does not attempt to address all aspects of the XP process, but rather only simulates the task implementation.

In XP, user stories (tasks) are defined by the customer and software architects. A task is assigned to a pair of programmers. This pair is responsible for coding, writing unit tests and producing a working piece of program in a certain period of time. It is assumed that everybody on the team will function as software architect, developer and tester. However, in practice, these roles are often carried out by different team members. Therefore, even if there are no distinct groups of architects, developers or testers, it can be assumed that there are three virtual teams: i) software architects, ii) developers and iii) testers. Information flow between these teams is shown in Fig. 1.



**Fig. 1** Information flow between virtual teams

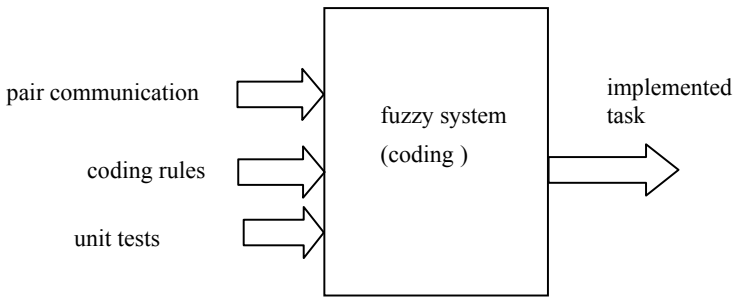
When a pair of programmers is implementing a task, one uses the keyboard while the other watches the code and provides feedback. If the pairs communicate well, they then can smoothly resolve the issues related to implementation. If this is not the case, if they have communication problems, then some issues will be overlooked, while other issues will be left unresolved. Consequently, the resulting code quality will be lower than expected. We can consider pair communication to be an input to the fuzzy system.

Writing unit tests during the code development is essential for detecting bugs in the software. In practice, some pairs write all the necessary unit tests before coding, but in some cases the writing of unit tests is ignored or only partially done. This also affects the resulting code quality.

XP advocates common code ownership. Every member of the team is thought to be responsible for the code and able to modify any part of the code. Coding rules must be observed by the programmers. However, modifying a piece of code in a way that is not compliant with coding rules may not be well understood by the pair that has been assigned with this task.

The output of the system is the code which implements the assigned task. Let's assume that the quality of the code is measured by the number of bugs in the code. For any particular task, pair communication, unit tests and coding rules affect the number of bugs in the code.

Task implementation can be modelled as a dynamical system, as shown in Fig. 2.



**Fig. 2** Task implementation as a fuzzy system

### 3 Designing a Fuzzy Model for Task Implementation

Fuzzy logic maps an input space to an output space, using the rules expressed as *IF variable IS set THEN action*

Fuzzy logic uses linguistic variables such as: poor, fair, good and excellent. For example, in pair programming we can say that communication between pairs is good, but how good is it? Is it possible to measure “pair communication” level in real terms? Rather than saying the level of communication is “five”, we prefer to express communication level using the terms mentioned above such as fair, good or excellent.

In the pair programming model there are three inputs; pair communication, unit tests and coding rules. The fuzzy subsets of these three inputs are:

pair communication: poor, fair, good, excellent

unit tests: poorly done, some missing, well done

coding rules: poorly applied, some applied, most applied, all applied

Triangular membership functions were selected as the membership function of pair communication, unit tests and coding rules. Other membership functions, such as a gaussian or trapezoidal curve, may also be used.

The output set, task implementation, is defined as:

task implementation: poorly done, fairly done, satisfactory, well done and excellent

This type of system, in which the output set is a fuzzy set, is known as Mamdani type fuzzy systems [17][18][19]. Matlab[20] was used to do fuzzy simulations. Now we can define fuzzy rules to relate the output set, the implemented task, to the input variables; pair communication, unit tests and coding rules. The rules reflect the experience of the author [21][22], therefore these rules may vary from one project to another. The software managers who want to use this model should construct the fuzzy model as described in this paper and modify the rules given here according to their project needs.

**Rules:**

1. if pair communication is **poor** and unit tests are **poorly done** and coding rules are poorly applied, then task implementation is **poorly done**.
2. if pair communication is **fair** some unit tests are **missing** and **some** coding rules are applied, then task implementation is **fairly done**.
3. if pair communication is **good** and **some** unit tests are missing and **all** coding rules are applied, then task implementation is **satisfactory**.
4. if pair communication is **excellent** and **some** unit tests are missing and **all** coding rules are applied, then task implementation is **well done**.
5. if pair communication is **excellent** and unit tests are **poorly** done and **some** coding rules are applied, then task implementation is **fairly** done.
6. if pair communication is **excellent** and unit tests are **well** done and **all** coding rules are applied, then task implementation is **excellent**.
7. if pair communication is **good** and unit tests are **poorly** done and **some** coding rules are applied, then task implementation is **fairly** done.
8. if pair communication is **good** and unit tests are **poorly** done and **most** coding rules are applied, then task implementation is **fairly** done.
9. if pair communication is **fair** and unit tests are **well** done and **most** coding rules are applied, then task implementation is **satisfactory**.
10. if pair communication is **good** and unit tests are **well** done and **all** coding rules are applied, then task implementation is **well** done.

In the task implementation, the fuzzy AND operator was used to get the fuzzy output variable T. The membership functions were chosen as triangular functions. The fuzzy rules reflect the experiences of XP practitioners. More rules can be added to the system, or some rules may be removed to fine-tune the system.

To get a numeric value, the fuzzy output set T is defuzzified using the centroid operator. This numeric value shows us how good the task has been implemented. For the rules given above, the defuzzified numerical output was found to be 2.6/10.0.

As a second alternative, Gaussian membership function was used in all input and output variables, while maintaining the same rules and defuzzification method. The numerical output was 3.17/10.0. This numerical output is less than half the maximum value, which was set as 10.0. By using a fuzzy system model, a quantitative value that demonstrates the quality of the implemented task can be obtained.

In an iteration there may be several tasks that need to be implemented. For each task the fuzzy model designed gives an indication of how well a task has been implemented. Assume that there are n tasks in the iteration; it is then possible to calculate a qualitative measure by averaging the outputs of the fuzzy systems as:

$$\text{iteration quality} = \text{sum of } n \text{ tasks } / N$$

Here it is assumed that every task in an iteration has equal weight.

## 4 Conclusion

Software development is a complex process. Therefore, it is not easy to find a mathematical model for a software development process. Similarly, in system theory there are complex systems that cannot be described by mathematical equations. Fuzzy logic is used to describe these types of systems. In this study, fuzzy logic was used to model the task implementation in extreme programming process in order to give valuable input to software project managers.

Project managers can observe the communication level of the pairs; in addition, they can monitor how a pair does the unit tests and if they follow the coding rules. The project manager can then define the rules and get a numeric value that will indicate the “quality” of a particular task. Taking the overall average of these values a numerical quality indication could be found for an iteration.

Fuzzy system rules and linguistic input and output variables may differ from one project to another. However, in all cases the fuzzy system model provides insight into the task implementation in the Extreme Programming process.

The model presented here can be improved by taking into account all extreme programming rules. It would also be interesting to investigate the interaction between the applied rules.

## References

1. Chatters, B.W., Lehman, M.M., Ramil, J.F., Wernick, P.: Modelling A Software Evolution Process, *Software Process. Software Process: Improvement and Practice* 5(2-3), 91–102 (2000)
2. Wernick, P., Hall, T.: Getting the Best of Software Process Simulation. In: *Second Int. Workshop on Realizing Evidence Based Software Engineering, REBSE 2007* (2007)
3. Setamanit, S.-o., Wakeland, W., Raffo, D.M.: Exploring the Impact of Task Allocation Strategies for Global Software Development Using Simulation. In: Wang, Q., Pfahl, D., Raffo, D.M., Wernick, P. (eds.) *SPW 2006 and ProSim 2006*. LNCS, vol. 3966, pp. 274–285. Springer, Heidelberg (2006)
4. Raffo, D.M.: Getting the Benefits from Software Process Simulation. In: *International Conference on Software Engineering and Knowledge Engineering (SEKE 1999)*, Kaiserslautern, Germany (June 1999)
5. Beck, K.: Embracing change with Extreme Programming. *IEEE Computer* 32(10), 70–77 (1999)
6. Beck, K.: *Extreme Programming Explained: Embrace Change*. Addison Wesley, Reading (2000) ISBN: 0201616416
7. Abrahamson, P., Salo, Q., Ronkainen, J.: *Agile Software Development Methods*, 4th edn. VTT Publications (2002) ISBN 951-38-6009-4
8. Mauer, F., Melnik, G.: Agile Methods: Moving Towards the Mainstream of the Software Industry. In: *ICSE 2006*, Shanghai, China, May 20-26 (2006)

9. Jeffries, R., Anderson, A., Hendrickson, C.: *Extreme Programming Installed*. Addison Wesley, Reading (2001) ISBN: 0201708426
10. Fowler, M.: *The new Methodology*, *Extreme Programming FAQ*, <http://www.jera.com/techinfo/xpfaq.htm>
11. Benjamin, K., Golnoraghi, F.: *Automatic Control Systems*, 8th edn. Wiley, Chichester (2002)
12. Kosko, B.: Fuzzy Systems as Universal Approximators. *IEEE Transactions on Computers* 43(11), 1329–1333 (1994)
13. Kosko, B.: The Probability Monopoly. *IEEE Transactions on Fuzzy Systems* 2(1), 32–33 (1994)
14. Zadeh, L.: Fuzzy sets. *Information Control* 8, 338–353 (1965)
15. Steven, D., [http://www.seattlerobotics.org/encoder/mar98/fuz/fl\\_part6.html](http://www.seattlerobotics.org/encoder/mar98/fuz/fl_part6.html)
16. Levary, R.R., Lin, C.Y.: Modelling the Software Development Process Using an Expert Simulation System Having Fuzzy Logic. *Software Practice and Experience* 21(2), 133–148 (1991)
17. Mamdani, E.H., Assilian, S.: An Experiment with in Linguistic Synthesis with a Fuzzy Logic Controller. *Int. Journal of Man-Machine Studies* 7, 1–13 (1975)
18. Mamdani, E.H.: Twenty Years of Fuzzy Control: Experiences Gained and Lessons Learnt. In: *Second IEEE International Conference on Fuzzy Systems*, vol. 1, pp. 339–344 (1993)
19. Ross, T.J.: *Fuzzy Logic with Engineering Applications*. McGraw Hill, Inc., New York (1995)
20. <http://www.mathworks.com/>
21. Gul, E., Sekerci, T., Yuceturk, A.C., Yildirim, U.: Using XP in Telecommunication Software Development. In: *ICSEA, The Third International Conference on Software Engineering Advances*, Malta, October 26-31, pp. 258–263 (2008)
22. Gul, E., Yildirim, U.: Servis Tabanlı Kurumsal Yazılım Geliştirilmesinde XP Kullanımı ve Eleştirisi (in Turkish) 2. Ulusal Yazılım Mühendisliği Sempozyumu, Middle East Technical University, Ankara (2005)



# Integration of Natural Language Processing Chains in Content Management Systems

Diman Karagiozov

Tetacom Interactive Solutions Ltd,  
18 Prof. N. Genchev, 1700 Sofia, Bulgaria  
email: diman@tetacom.com

**Abstract.** Modern web application hype revolves around a rich user interface experience. A lesser-known aspect of modern applications is the use of techniques that enable the intelligent processing of information and add value that can't be delivered by other means. This article presents a scalable, maintainable and inter-operable approach for combining content management functionalities with natural language processing (NLP) tools. The software, based on this architecture, is open for chaining various NLP tools and integration of languages in a standardized manner. As a demonstration of the concept, we have developed two web sites using a content management system, featuring the English NLP based language processing chain. Language processing chains for Bulgarian, Croatian, German, German, Greek, Polish and Romanian languages are in a process of development and integration.

**Keywords:** multilingual content management, text mining, software architecture, natural language processing, linguistic tools, UIMA.

## 1 Introduction

Content management systems (CMS) are used to organize and facilitate the collaboration content creation and presentation. Most of them provide functionality and user interface to manually manage and inter-link the content. The increasing amount of information nowadays requires adequate and almost real-time reaction in order to provide up-to-date information to the readers. The ability of content managers to interpret and respond to these events is increasingly constrained by the volume, variety and velocity of the information. The professionals are obliged to collect and analyze vast amounts of data, and to identify correlations between disparate pieces of information. The conventional CMS fail to streamline the process of interlinking of the content and turning the mass information into useful knowledge. This article outlines innovative methods for automatic and semi-automatic semantic analysts of textual content managed by the CMS. The practical application of such technologies and their integration in an existing CMS are described in details.

## 2 Related Work

Modern web application hype revolves around a rich user interface experience. A lesser-known aspect of modern applications is the use of techniques that enable

the intelligent processing of information and add value that can't be delivered by other means [1]. This section of the article outlines the techniques and algorithms that are required in order to build an intelligent web application using a CMS.

*Indexing and full text searching* – a modern CMS allows the information designers to structure the content and the relations between the content items dynamically, and later to run full-text search queries in the pool of content items. The most widely used full-text search engine library that is integrated in CMS is Apache Lucene or tools based on Lucene, such as Apache Solr [2].

*Identification of important “cue” words and phrases* - Nouns (and noun phrases) are traditionally defined as "persons, places, things, and ideas." [3]. Amazon first defines the term “Statistically improbable phrases” as “*the most distinctive phrases in the text of a particular books ... relative to all books (in a collection)*”<sup>1</sup>. The main added value to a CMS is the presentation to the user of the main concepts and ideas of a content item.

*Identification of named entities* – The named entities are noun phrases which are further disambiguated and categorized by their meaning and function in the text. The extracted named entities are used for answering the 5W1H questions (who, what, why, where, when and how) and for finding similar content. Popular services providing NE extraction are OpenCalais<sup>2</sup>, Stanford CoreNLP<sup>3</sup> and OpenNLP<sup>4</sup>.

*Clustering similar content items* – Filtering, reviewing and maintaining the relations between the content items is time and effort consuming task for the information designers and content providers. Thus, a CMS needs tools which provide functionalities like “more like this”, “recommended reading”, and “see also”. According to the cluster hypothesis (“*Documents in the same cluster behave similarly with respect to relevance to information needs.*” [4]) the most significant features of a content item are almost the same in similar content items form one and the same cluster.

*Automatic assignment of tags to the content items* – tagging the content (assigning keywords) facilitate its searching and finding, however this process require a lot of manual efforts. Taxonomy building and tags assignments are two techniques that can be performed semi-automatically by the computers and reviewed and corrected manually [4].

*Computer aided translation for multilingual web applications* – being a thriving research field, the machine translation (MT) is a new functionality, poorly integrated in the process of content management. On the other hand, the demand for multilingual web sites is rapidly increasing. The MT engines assist the content providers with the initial translation of textual materials; they also help the web application users to cross the language barriers. Existing services providing MT are Moses [5], Google Translate<sup>5</sup>, Bing Translator<sup>6</sup>.

<sup>1</sup> Amazon SIPs, <http://www.amazon.com/gp/search-inside/sipshelp.html>, 2011

<sup>2</sup> OpenCalais, <http://www.opencalais.com>, 2011

<sup>3</sup> Stanford CorNLP, <http://nlp.stanford.edu/software/corenlp.shtml>, 2011

<sup>4</sup> OpenNLP, <http://incubator.apache.org/opennlp>, 2011

<sup>5</sup> Google Translate, <http://translate.google.com>, 2011

<sup>6</sup> Bing translator, <http://www.microsofttranslator.com>, 2011

## 2.1 Language Processing Chains

Textual information is generally unstructured, however humans are able to process it and find the most important pieces of it. Computers, on the contrary, cannot perform such analysis – they are programmed to execute a sequence of tasks in order to revile the main concepts and interrelations in the text. The sequential tasks, called a language processing chain (LPC), consist of atomic NLP tools which add low-level annotations in the text and thus make it structured. We use the low-level annotations to extract important words and phrases, and named entities on a later stage of the processing. Furthermore, we apply statistical algorithms to the low-level annotations in order to find the most significant features of the analyzed text.

A sample LPC consist of the following atomic NLP tools [6]: *Tokenizer* (splits the raw text into tokens) → *Paragraph splitter* (splits the text in paragraphs) → *Sentence splitter* (splits the paragraphs in sentences) → *POS tagger* (marks up each token with its particular part of speech tag) → *Lemmatizer* (determines the lemma for each token) → *Word sense disambiguation* (disambiguates each token and assigns an unique sense to it) → *NP Extractor* (marks up the noun phrase in the text) → *NE Extractor* (marks up named entities in the text).

In order to achieve optimal precision of the LPC, we combine statistics-based NLP tools with language specific linguistic rules. The output of an LPC run over a raw text is stored in a data store. The stored records are later used by language independent higher-level NLP tasks.

## 2.2 Higher-Level NLP Tasks

The low-level annotations from the LPC output cannot be directly used mainly because of their volume and quality. Thus, we clean these annotations using different statistical and rule-based methods so that only the most significant features of the input text are presented to the end users.

*Important words.* A common technique to find the most significant tokens of a content item is to use a ranking functions such as TF (term frequency), TF-IDF [4], or Okapi BM25 (a variant of TF-IDF) [7];

*Important phrases.* We also apply the ranking functions on the level of the noun phrases (NP). Prior applying the ranking functions, the NPs are “normalized” so that that they can be compared. We have identified two comparison methods for NPs – evaluation of the lemmatized version of the NPs (quite complex language-dependent task) or evaluation the sequences of tokens in the NPs;

*Similar content.* The important words identified in the previous steps are used for the weighting the similarity between two content items. Special “more-like-this” Lucene [2] queries are used for finding all content items that have the same (or similar) important words as the first content item;

*Automatic categorization.* Various classification and clustering techniques, algorithms and combinations between them are used for achieving a good quality of the automatic categorization functionality [1]. Such technique uses the tokens of the content items, creates a vector space, applies feature reduction, builds multi-label, multi-class model and then uses the model for getting predictions of the classes a content item belongs to;

*Summarization.* The goal of a résumé is to shorten an initial text and make it more understandable to the user. The extractive and abstractive summaries are the two main streams in the résumé creation. They both identifies sentences which provide the most significant information in a document using TextRank, LexRank, Grasshopper [8] or other similar algorithms working with the paragraphs, sentences, tokens and anaphora resolver and co-reference annotations. The sentences in the abstractive summary, however, are straight simple sentences, generated by a computer program.

### 3 Software Architecture

Each CMS have a unique architecture which best serves the main design and functional requirements of the specific system. The aim of the currently described CMS is to enable the integration of various linguistic tools in the process of content management. The major non-functional requirements for the CMS are:

- Responsiveness – the classical request-response scenario should be performed as fast as possible;
- scalability – the CMS should scale horizontally and vertically in order to achieve maximum performance;
- Maintainability – a CMS is rich of major and minor functionalities which are often overlapping and/or complimentary. The maintenance of a CMS is not a trivial task, thus the architecture should support this process as much as possible;
- Inter-operability – the interface between a CMS and other systems should be as standard as possible. This will allow future extensions and integration of external functionalities.

The LPC modules must not compromise any of these four major requirements.

*Responsiveness.* Usually, the NLP tasks are *slow*. Their overall performance depends on performance of the atomic NLP tools and the size of the input text. This is the reason why a LPC cannot be instantiated in the classical request-response chain because response time cannot be predicted. Thus, we are using an asynchronous communication channel between the CMS and LPC components.

The CMS asynchronously sends a message, identifying the document and providing its content, to the LPC engine and informs the user that the request is being processed. The appropriate status of the task is shown to the user while the message is being processed by the LPC engine. The results of the task become available in the CMS once the message is eventually processed.

*OSGi LPC engine.* The OSGi framework is a module system and service platform for Java that implements a complete and dynamic component model. Applications or components can be remotely started, stopped, and updated without requiring a reboot [9]. Equinox, an OSGi framework implementation, has been chosen as a backbone of suggested LPC engine architecture. Our architecture consists of three main components. It can be easily extended with more entity points because of the flexible OSGi framework.

1. *Message queue*. Java Messaging Service (JMS) API is a message oriented middleware for sending messages between two or more clients. It allows the communication between different components of a distributed application to be loosely coupled, reliable, and asynchronous. We have based the implementation of the transport messaging agent, between the CMS and the different LPC components on the Apache ActiveMQ.

2. *Atomic annotator*. The atomic annotator is responsible for the initial set of annotations needed by the higher NLP tasks. The annotator checks-out a message from the queue and delegates the processing to:

- *Pre-processor*. The component identifies the mime-type of the message content, extracts the text if needed, detects the language of the text and sends an internal message to the NLP processor;
- *NLP processor*. The component provides the basic annotations in the message text. Similar to the OSGi for Java UIMA (Unstructured Information Management Applications) allows the complex NLP applications to be decomposed into components. Each atomic NLP tool is wrapped into UIMA primitive engine; the primitive engines are sequenced by an aggregate engine. UIMA is not OSGi compliant, thus we wrapped the UIMA aggregate engine in an OSGi component (NLP processor), making it available to the rest of the components in the installation;
- *Post-processor*. The component is invoked when the annotations are ready, saves the annotations, provides performance report, informs the CMS that the annotations are available, and invokes the higher-level *categorizer* and *summarizer* components.

3. *Categorizer and Summarizer*. The *categorizer* and the *summarizer* have one and the same internal architecture, thus, only the *summarizer* is described in details. The component checks-out a message from a queue, loads the needed sentences and tokens for the requested document, instantiates a summarization engine (LexRank [8] implementation or OpenText Summarizer external tool [10]), creates a summary of the document and sends the summary to a queue to be further processing (saved in a data store).

**Scalability.** The usage of message queue in the architecture of the LPC engine enables a trivial horizontal scalability by simply installing new instances of the LPC engines. Usually, the documents in English are a lot more than the documents in other languages, so it makes sense to deploy several English LPCs working in parallel. In this way we are able to minimize the time prior a message has been processed.

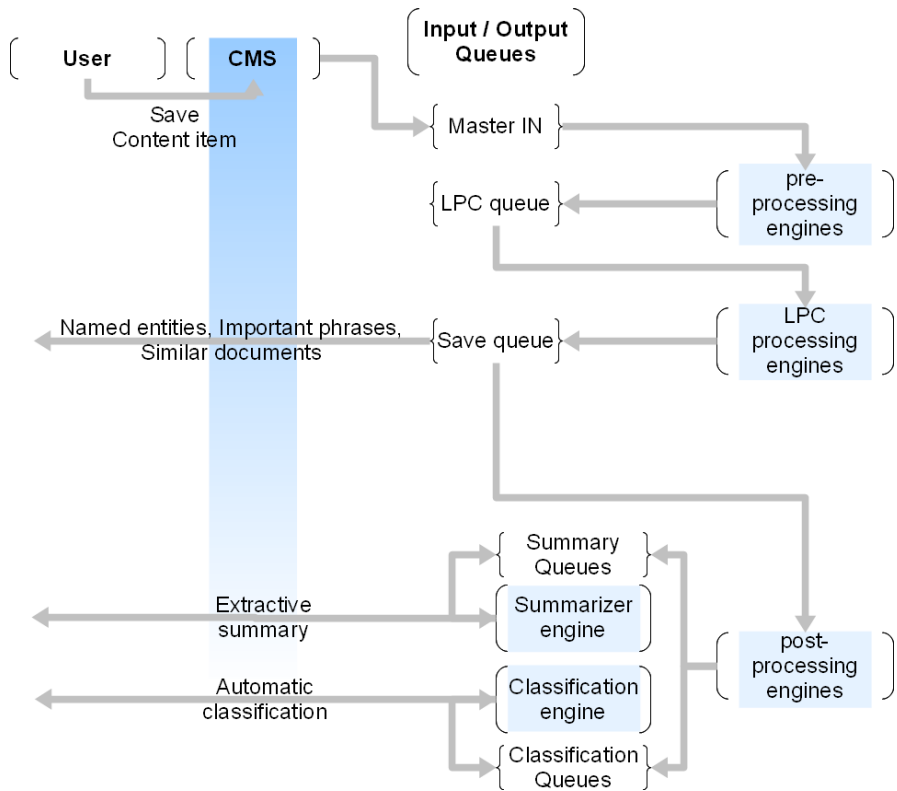
### 3.1 Schema of the Architecture

We have implemented the above-described architecture as part of the project “ATLAS – Applied Technologies for Language-aided CMS”<sup>7</sup>. Details in the

---

<sup>7</sup> ATLAS is still ongoing CIP-ICT-PSP.2009.5.3 EU-funded project.  
<http://www.atlasproject.eu/>

architecture and their implementation may change or will be extended by the end of the project. The diagram below (Fig. 1) depicts the major architectural components and communication channels between them.



**Fig. 1** Major architectural components and communication channels between them

*User.* The users trigger the language processing chains by manipulating (adding or updating) the content on a web site.

*Content management system.* The CMS “communicates” with the LPC engines via message queue through a well-defined API. Currently, only OSGi-based API is available.

*Input & Output queues.* The asynchronous communication between the components is empowered by a JMS implementation. A message is sent to an input queue; a component checks-out the message, transforms it and sends it to another queue. The LPC component and the CMS implement the message router, message translator, messaging gateway, event-driven consumer and competing consumers enterprise design patterns [11].

*Pre-processing engines.* The component provides mime-type detection, text extraction, language identification and text cleanup.

*LPC processing engines.* The component wraps a LPC for a given language.

*Post-processing engines.* The components store the annotations in a data store.

*Summarization and Categorization engines.* These components provide a summary and list of categories applicable to a document. The architecture of the engines allows the integration of multiple summarization algorithms and categorization tools.

## 4 Scenarios

We have developed two web sites in order to illustrate how the linguistic building blocks are interconnected and integrated into a useful web application, configured entirely within ATLAS CMS. Both sites can operate in a multilingual setting; the implemented functionality currently is offered for English and Bulgarian. The language processing chains for other languages (Croatian, German, Greek, Polish and Romanian) will be available by December 2011.

***i-Librarian: a personal library.*** i-Librarian service<sup>8</sup> is a demonstration of ATLAS functionalities in the form of a digital library web site. The library addresses the needs of authors, students and researchers by providing them an easy way to create, organize and publish documents and then search for similar documents in different languages, and to locate the most essential texts from large collections of unfamiliar documents. The library uses language technology to extract important phrases and named entities from indexed documents; similar items are then displayed on demand, abstract translated and document summaries are produced. 4.4K documents (165M tokens) from Project Gutenberg have been uploaded in order to evaluate the performance and scalability of the site even before users started uploading their papers.

***EUDocLib.*** Another demonstration of ATLAS functionality is the EUDocLib<sup>9</sup> web site. The library offers easy access to law documents of the European Union with automatic categorization, extraction of important phrases, named entities, and similar items. Currently the web site covers 140K documents (182M tokens) in English.

## 5 Conclusion and Further Work

The described architecture and its implementation in the ATLAS project gives a prospect to standardized multilingual online processing of language resources within a CMS and offers localized demonstration tools built on top of the linguistic modules. The framework is ready for integration of new types of tools and new languages to provide wider online coverage of linguistic services in a standardized manner.

An obvious direction to be followed is to provide implementation of the “LPC integration API” for a wider range of platforms and programming languages such

---

<sup>8</sup> i-Librarian, <http://www.i-librarian.eu>, 2011

<sup>9</sup> EUDocLib, <http://eudoclib.atlasproject.eu/>, 2011

as PHP and .Net. In this way the most widely spread CMS can benefit from language processing chains software. We will provide a LPC engine web service in order to enable the integration with CMSs in other languages, such as Python, Ruby, and Perl.

The limits of the conventional relational databases are easily reached with the amount of data that is stored as a result of the LPC engines. The data store should be easily replaceable with another one that provides transparent horizontal scalability.

Language processing chains for Bulgarian, Croatian, German, German, Greek, Polish and Romanian languages are in a process of development and integration in the ATLAS project.

## References

1. Marmanis, B.: Algorithms of the Intelligent Web. Manning (2009)
2. McCandless, M., Hatcher, E., Gospodnetić, O.: Lucene in Action, 2nd edn. Manning (2010)
3. Huddleston, R.: Introduction to the Grammar of English. Cambridge University Press, Cambridge (1984)
4. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press, Cambridge (2008)
5. Koehn, P., et al.: Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics, ACL (2007)
6. Cristea, D., Pistol, I.C.: Managing Language Resources and Tools using a Hierarchy of Annotation Schemas. In: Proceedings of Workshop 'Sustainability of Language Resources and Tools for Natural Language Processing', LREC (2008)
7. Jones, K.S., Walker, S., Robertson, S.E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. Information Processing and Management 36(6), 779–840 (2000)
8. Chen, S., Huang, M., Lu, Z.: Summarizing Documents by Measuring the Importance of a Subset of Vertices within a Graph. In: Proc. IEEE WIC ACM Int. Conf. Web Intell. Agent Technol. (2009)
9. OSGi Service Platform, Core Specification, Release 4, Version 4.2, OSGi Alliance (2009)
10. Yatsko, V., Vishnyakov, T.: A method for evaluating modern systems of automatic text summarization. Allerton Press (2007)
11. Hohpe, G., Woolf, B.: Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions. Addison-Wesley Professional, Reading (2003)



# Supporting Interactive IPTV Apps with an Enterprise Resource Bus

Sunny Sharma and Ralph Deters

Department of Computer Science, University of Saskatchewan  
email: sps885@mail.usask.ca, deters@cs.usask.ca

**Abstract.** Delivering TV services via internet protocols over high-speed connections is commonly referred to as IPTV (Internet Protocol TeleVision). A particularly interesting aspect of IPTV is the augmentation of the subscriber's TV experience with interactive applications (apps). Due to the service-oriented infrastructure of many IPTV platforms it is fairly easy to integrate 3<sup>rd</sup> party services and expose them in form of IPTV apps. However, due to the very competitive market in which IPTV providers operate, they are forced to minimize their costs and thus avoid costly customer complaints. This in turn introduces the need for highly dependable IPTV apps/web services with minimal downtimes. This paper focuses on the development of dependable web services for IPTV. Using Brewer's CAP theorem the paper investigates the role of state-management within web services and presents the idea of the Enterprise Resource Bus (ERB) and an evaluation of its Erlang implementation.

**Keywords:** IPTV, Dependable Web Services, Erlang, REST, State.

## 1 Introduction

IPTV (Internet Protocol Television) is the delivery of TV services via internet based protocols over packet-switched networks. IPTV differs from internet-based multimedia platforms (e.g. Netflix, YouTube, iTunes, etc.) in terms of content, delivery and costs. IPTV offers its subscribers, in addition to the video-on-demand streaming and/or downloading services of multimedia platforms, *live* content. To ensure that content-providers grant access to premium content, IPTV platforms offer very dependable (secure, safe, reliable and available) service delivery. Recently, IPTV providers have begun the move towards *interactive* TV experiences via apps to further differentiate their offerings from internet-based multimedia platform. IPTV apps blur the lines between classical TV and computers and allow for more engaging user experiences. Microsoft's Mediaroom [1] is a leading IPTV platform that enables IPTV providers to embed applications into the video-stream by injecting XML encoded documents. Mediaroom is a mature service-oriented IPTV platform that allows easy integration of 3<sup>rd</sup> party apps and services. Within Mediaroom, IPTV apps are XML documents (figure 1) that follow a proprietary format called MRML (Media Room Markup Language). Upon receiving a MRML document, the client (e.g. set-top box) renders the XML

document and blends it into the video-stream . A main feature of the Mediaroom platform is its web-centric design. This web-compliant model allows 3<sup>rd</sup> party developers to provide services/apps that can be used by Mediaroom subscribers. Given the very competitive market in which IPTV providers operate, they are forced to ensure minimal customer complaints by ensuring that only dependable and scalable apps are allowed to enter their service ecosystem. This paper focuses on how to IPTV apps using an Enterprise Resource Bus. The rest of the paper is structured as follows: Section 2 focuses on state management within web services and the futility of enforcing a consistent state across a distributed system. Section 3 presents the Enterprise Service Bus (ERB) as a pattern/architecture to overcome the challenges identified in the CAP theorem and an overview of the implementation and evaluation results. The paper concludes in chapter 4 with a summary and an outlook.

## 2 State and Web Services

In 1996 Gartner [2] introduced services as a new integration and development paradigm. Within web services, two competing architectural styles namely the Service Oriented Architecture (SOA) [3,4] and the Representational State Transfer (REST) [6] exist. When Gartner [2] introduced their new Service-Oriented Architecture (SOA), it was at first assumed that services should be stateless to ensure better scalability. Since a stateless service is expected not to maintain a state altered by processing a request it is possible to scale-out the service and thus increase scalability. However, the stateless services rely on state-servers that add additional overhead. To minimize interactions with a state-server, stateful services maintain state between serving requests. Obviously, by allowing services to maintain state, the performance is improved at the expense of scalability. The Representational State-Transfer (REST) approach, which was developed by Roy Fielding [6] offers a lightweight and very well grounded alternative to SOA. A Restful service is expected to maintain state and to offer operations with clear operational semantics (read/write) that modify its state. The interaction in a REST system typically follows a request/response pattern in which each side assumes that all information is contained in the request and response. Leonard Robinson [7] identifies within his 4-level web maturity model two patterns that have become popular within REST. The first is the CRUD [8] (Create Read Update Delete) approach that follows a basic data-centric style. CRUD has gained significant interest due to the easy mapping on HTTP verbs. Create, read, update and delete are achieved via POST, GET, PUT/PATCH and DELETE. The second pattern, that is less widespread, is the use of hypermedia controls. Unlike the data-centric CRUD pattern, the hypermedia control pattern focuses on the use of embedding links into the responses of request. By offering links, the server provides the client with possible next steps and ways to obtain further information. Consequently, this approach allows the service provider to *push the application state* to the client. While SOA and REST both require stateless communication, between requester and service provider, they differ in respect to the management of application and web service state. While REST is a state-centric approach, SOA

tends to marginalize state since it considers it a provider/domain specific aspect. Consequently a key difference in the deployment of SOA and REST services is the ability to cache the response to a request. Because SOA has no clear read/write semantic for operations on service it is impossible to cache; unless additional protocols/information is introduced. REST however has clear read/write semantics for the operations on its services (aka resources) which in turn allows for the caching of responses. In fact caching is so natural within REST that it is considered a standard performance enhancing technique. While the handling of state seemed at first a minor issue with the web service community, it transformed into a major challenge in 2000. Up till 2000, web service dependability and scalability were considered largely SLA/QoS issues that could be addressed by monitoring providers, provisioning resources to providers and scheduling requests [5]. However in 2000 Brewer [12] introduced the CAP theorem, which states that for any physically distributed system it is impossible to ensure consistency (nodes share same state at same time), availability (in case of node failure, surviving nodes can continue to operate), and partition-tolerance (system handles arbitrary loss of messages) all at the same time. The impact of the CAP theorem on web services was studied by Gilbert and Lynch [10]. Their work provided a formal proof of the CAP theorem and showed that it is necessary to relax the consistency constrains across web services if high availability and partition tolerance has to be achieved. Due to the absence of any common state concept in SOA, it is impossible to provide generalizable mechanisms for dealing with soft-state. In REST however, there is a clear notion of state which allows to reason over state-changes and their consequences. As a result, there is no shortage of approaches for dealing with soft-state scenarios. The simplest approach for dealing with soft-state is to use caching and to require services to include cache-control headers in their responses. To deal with network loss or temporarily provider failures, it seems reasonable to use stale cache data and to require clients of services to register compensation handlers that get called once more accurate data becomes available. The caching approach can be improved by allowing client to register callback functions with the service providers (e.g. Functional Observer REST pattern (FOREST [11]) and by requiring the providers to use the callback to notify former consumers of state-changes. A notified consumer would then have the ability to resubmit the request and thus obtaining the more current request.

### **3 Enterprise Resource Bus (ERB) Implementation**

The Enterprise Resource Bus (ERB) is a pattern that builds on the model of the Enterprise Service Bus (ESB) [3]. In addition to the functionalities provided by ESB, the ERB is designed to provide better support for Resource-Oriented Systems. The most important differences between an ESB and ERB are the use of caching, views, and request routing. Due to the clear focus on states caching and database like views are natural additions that boost performance (cache), add access security (views) and convenience (views). Request routing, a common feature of ESBs, is different due to the clear operational semantic of the request which allows to spawn concurrent read activities across hosts. A key issue in the

development and deployment of the ERB and Restful services for the IPTV domain was minimizing downtime and thus improving the perceived dependability. Erlang [7] was chosen as the programming language for implementing the ERB and the Restful services due to its ability to handle large number of concurrent activities and scale-out. It uses message-passing, supports fault-tolerance and allows for declarative programming. Within Erlang, patterns for designing servers are defined in the OTP modules and it is expected that developers use them to implement their code. In the ERB and the Restful services, the `gen_server` template/behavior is used extensively since it allows the definition of supervisors that automatically monitor and if necessary restart the servers. To test the Restful services in a controlled manner with realistic IPTV services, the ERB and services were placed on an Amazon EC2 cloud instance (“m1.xlarge”, 1 Xeon 2.27GHz 4 core CPU, 15GB RAM, 64Bit Windows Server 2008 Datacenter). The load generator was placed on more powerful Amazon EC2 instance with 8 Xeon cores (“c1.xlarge”, 2 x Xeon 2.27GHz 4 core CPU, 7GB RAM, 64Bit Windows Server 2008 Datacenter). ApacheBench[13] (AB) was used as the tool to create the loads for the Restful services. To evaluate the impact of various clients engaging the services, AB generated 1000 requests with the concurrency settings 1,10,20,30,40,50,60,70,80,90,95 and 100. The concurrency setting determines the number of concurrent request e.g. setting 10 simulates 10 concurrent clients.

**Table 1** AB results for various concurrency settings executing write operations

Concurrency	Min	Mean	Median	Max	50%	75%	90%	95%
1	0	5	0	62	0	0	16	16
10	0	24	32	31	31	31	31	31
20	0	48	47	62	47	47	47	62
30	0	73	78	140	78	78	78	78
40	16	95	94	109	94	94	109	109
50	0	118	125	125	125	125	125	125
60	16	141	140	156	140	156	156	156
70	16	166	172	187	172	172	172	172
80	16	187	187	203	187	203	203	203
90	16	210	218	234	218	218	218	234
95	16	219	234	390	234	234	234	234
100	16	2933	3510	6568	3510	3541	3541	3557

Table 1 show the results of the 12 runs in ms. The columns 50% - 95% show the maximum round-trip times for 50, 75, 90 and 95 percent of all requests. The results were obtained by disabling caching and represent direct calls to resources. As can be seen, the system performs very well up to 95 concurrent calls. 96 concurrent calls already resulted in some response times over 2 seconds. Please note that handling 95 concurrent clients is a very good result and that by scaling-out it significantly larger number of clients can be supported.

## 4 Conclusion and Outlook

The paper presents the Enterprise Resource Bus, an extension of the widely used Enterprise Service Bus, as a pattern to allow for easy integration of message- and event-oriented patterns. The evaluation of the Erlang implementation showed that the ERB pattern in combination with Erlang-based Restful web services provides good scalability and stable performance even when faced with large numbers of concurrent clients. Future work will focus on experimenting with different arrival rates, investigating the impact of response size and by conducting a more controlled injection of faults. We will also want to investigate different event push techniques to better understand the costs and benefits of our current callback approach. Finally we are also interested in investigating the use of P2P approaches as a means for propagating state-change events across distributed systems. Connecting the various caches and resources via P2P protocols seems to offer a very robust and efficient mechanism for avoiding stale caches.

## References

1. Microsoft Mediaroom, <http://www.microsoft.com/mediaroom/>
2. Schulte, R., Natis, Y.: Service Oriented Architecture, Gartner, April 12 (1996)
3. Chatarji, J.: Introduction to Service Oriented Architecture (SOA), 5 pages (2004), <http://www.devshed.com/c/a/WebServices/Introduction-to-Service-Oriented-Architecture-SOA>
4. Box, D.: Four Tenets of Service Orientation, <http://msdn.microsoft.com/msdnmag/issues/04/01/Indigo/default.aspx>
5. Dyachuk, D., Deters, R.: Transparent Scheduling of Web Services. In: IEEE DEST (2007)
6. Fielding, R.: Architectural Styles and the Design of Network-based Software Architectures, Dissertation University of Irvine (2000), <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>
7. Robinson, L.: Richardson Maturity Model, <http://martinfowler.com/articles/richardsonMaturityModel.html>
8. CRUD: Create Read, Update and Delete, [http://en.wikipedia.org/wiki/Create,\\_read,\\_update\\_and\\_delete](http://en.wikipedia.org/wiki/Create,_read,_update_and_delete)
9. Erlang, <http://www.erlang.org>
10. Gilbert, S., Kynch, N.: Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. ACM SIGACT News, 33.2 (2002)
11. FOREST, <http://duncan-cragg.org/blog/post/forest-functional-observer-rest/>
12. Brewer, E.: Towards Robust Distributed Systems, <http://www.eecs.berkeley.edu/~brewer/cs262b-2004/PODC-keynote.pdf>
13. ApacheBench, <http://httpd.apache.org/docs/2.0/programs/ab.html>

# Service Oriented Architecture Applied to Provide Inter-operability between Macedonian MoD and NATO

Zoran Gacovski and Goko Vancov

FON University, Bul. Vojvodina, bb, Skopje, Macedonia  
email: zoran.gacovski@fon.edu.mk, gokov@morm.gov.mk

**Abstract.** In this paper we present SOA (Service oriented architecture) that we are developing to enable interoperability between Macedonian Ministry of Defence (MoD) and NATO. First we give a brief overview of Service-Oriented Architecture (SOA) and an explanation of how it can be used for the Ministry of Defense purposes. MoD uses IEEE 1472 definition of an architecture description to define a standard approach to describing, presenting, and integrating a defense architecture that can be used with a service oriented approach to capability based planning. The principal objective of our work is to ensure that architecture descriptions can be compared and related across organizational boundaries, including NATO and multi-national boundaries. SOA can provide easier identifying of required capabilities, the ways (operational activities), the means (human or system services) and the conditions (under which capability is required).

## 1 Introduction

Technologically advanced nations are increasingly digitizing their military forces. Operations are no longer conducted by a single service, but are Joint and more likely to involve either a coalition of willing countries, or existing alliances such as NATO. There has been a growing awareness that the traditional exchange of information that has been limited to data exchange and the use of common message text formats. The increasing use of commercially supported open standards pushes the information technology (IT) infrastructure from proprietary military solutions towards web-enabled Service-Oriented Architectures (SOA). SOA is an architectural approach that enables flexible connectivity of applications or resources implemented as services. Such services have well-defined, platform independent interfaces that hide the underlying technical complexity of the environment (encapsulation), they are self-contained (loosely coupled), and reusable [5].

The SOA's greatest advantage is that it provides seamless information exchange based on different policies and loose coupling of its components. In a military domain it enables to make the military information resources available in the form of services, which can be discovered and used by all mission participants that do not need to be aware of these services in advance. The most mature implementation

of SOA, recommended by NATO and widely applied in the commercial sector are Web Services (WS). WSs are described by a wide range of standards that deal with different aspects of WS realization, transport, orchestration, semantics, etc. They provide the means to build a very flexible environment that is able to dynamically link different system components to each other. These standards are based on the eXtensible Markup Language (XML) and have been designed to operate in high bandwidth links. XML gained wide acceptance and became very popular because it solves many interoperability problems and it facilitates the development of frameworks for software integration, independent of the hardware platform [1].

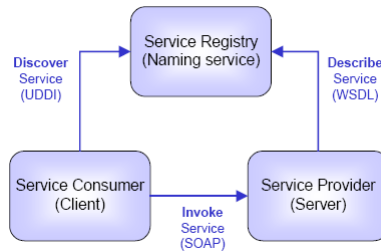
In this paper we present SOA (Service oriented architecture) that we are developing to enable interoperability between Macedonian Ministry of Defence (MoD) and NATO. This paper is organized as follows: first we give an overview of SOA architecture and its components. Then we describe the Service Oriented Architecture that is currently developing in the Macedonian MoD. In the forth Chapter the topic of inter-operability between national MoD and NATO is covered. Finally, we give some conclusion remarks.

## 2 SOA Description

The current software paradigm to cope with the challenges of net-centric operations is to apply services within a service-oriented architectures (SOA). An SOA is a collection of composable services. A service is a software component that is well defined, both from the standpoint of software and operational functionality. In addition, a service is independent, i.e. it doesn't depend on the context or state of any application that calls it. Currently, these services are typically implemented as web services. The advantage of using web standards in an SOA is that the services can more easily handle distributed applications in heterogeneous infrastructures. Nothing in particular has to be done programmatically to the service, except to enable it to receive requests and transfer results using web-based messaging and transportation standards. In many cases, web services are straightforward and existing software can easily be “web enabled” to create new services usable within an SOA. Web Services are a set of operations, modular and independent applications that can be published, discovered, and invoked by using industry standard protocols – Extensible Mark-up Language (XML), Simple Object Access Protocol (SOAP), Web Service Description Language (WSDL), and Universal Distribution Discovery and Interoperability (UDDI). It is a distributed computing model that represents the interaction between program and program, instead of the interaction between program and user. Web services can also be defined as discrete Web-based applications that interact dynamically with other web services.

How do web services work? Web services send and receive data described in XML. XML is a platform, programming language, and operating-system independent way to structure data and describe these data using tags. SOAP is used to send and receive data packages described in XML. Web services describe their data, operations, bindings, protocols, and all other relevant information in a standardized way, WSDL. This WSDL package is send to a UDDI repository. If a user needs a service, he looks through the WSDLs in a UDDI repository. If he

finds what he needs, he prepares the data the service needs as input and uses SOAP to send these data to the service. The service delivers the output via SOAP back to the user. Figure 1 shows how these standards interplay. In summary, web services are procedures with descriptions of data and operations in a common syntax to be found in a known repository. To invoke the service, a simple protocol is used for a general form of remote procedure call.



**Fig. 1** Web Service Standards

With XML the IT community agreed on a powerful standard to promote general data exchange. The application of XML enabled a new level of interoperability for heterogeneous IT systems by enabling separation of data definition and data content. Second, SOAP is an easily applicable and easily implemented protocol available on many platforms from PCs to handheld systems. These two concepts have been agreed upon by many vendors and IT providers and are supported by many applications. Many tools provide XML migration for legacy systems, such as database applications or client-server oriented structures. The step from distributed systems to web service based systems is relatively easy; the integration of web service is a solved problem.

### 3 SOA Developed for Macedonian Ministry of Defense (MOD)

In order to participate in NATO exercises, missions and other activities, Macedonian MoD has to develop its own SOA structure (more as consumer, but in some cases - as a provider of services). In brief, we have to implement following:

- **Dynamic Service Discovery:** The services are accessible on the Network, platforms don't need prior knowledge of their location except the Service registry location.
- **Publish-Subscribe Service:** This method allows a service consumer to subscribe to a delivery information service that has been published in the service registry.
- **Request-response Service:** This method allows a platform to request a service and then to receive the response. The Sensor Request Service uses this method.
- **Services Registry:** A central Services Registry is used for sharing information about services and their publishers.



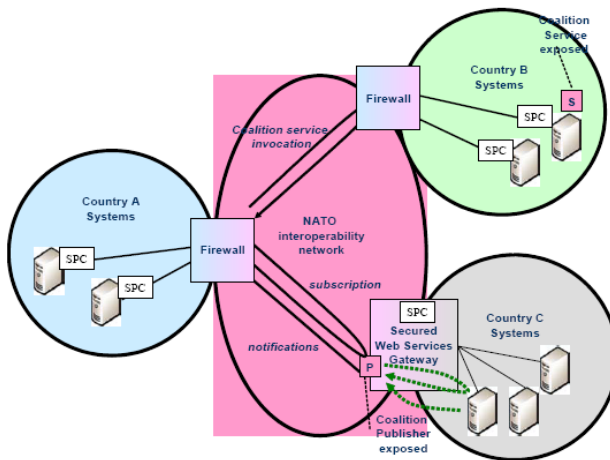
- **Security Certificates Directories Replication:** Platforms can be synchronised and Certificates can be exchanged.
- **Secured Exchanges:** All exchanges between platforms are performed in a secure way using signature, labeling and ciphering.
- **Data Exchange Format:** All exchanged Messages are XML-based.

To implement the Publish/Subscribe pattern different protocols and standards can be used. We have to rely on the specifications provided by OASIS, denoted WS-Notification [6]. This is actually a collection of the three specifications WS-BaseNotification, WS-BrokeredNotification and WS-Topics. For the purpose of interoperability - we have to utilise the WS-BaseNotification and the WS-Topics specifications. Using WS-Notification terminology, a service that publishes data (publisher) at a specified Topic is called a *NotificationProducer*. Topics are a way to group together, represent and categorize items of interest.

The data format of each topic is defined by an XML schema. A client, called a *NotificationConsumer* (subscriber), first creates a subscription to the service. The client will subsequently receive notifications as they are produced by the NotificationProducer. Since WS-Notification is a Web Services specification, all messages are exchanged using SOAP.

#### 4 Inter-operability with NATO Structures

Modern coalition operations are conducted in a dynamic environment, usually with unanticipated partners and irregular adversaries. In order to act successfully they need technical support that gives modularity and flexibility in connecting heterogeneous systems of cooperating allies. To support such cooperation in NATO community, SOA is recommended as crucial Network Enabled Capability (NEC) enabler. Figure 2 shows possible communication scenario during NATO exercise.



**Fig. 2** Communication scenario during NATO exercise

Participant platforms have to connect each other according a three-step process in order to get capabilities of browsing the Service registry and then Service invocation:

- 1) **Planning:** Exchange of necessary security Certificates for directory replication and exchange of addresses of platform gateways through trusted files;
- 2) **Assembling:** Platforms replicate directories of security Certificates, hence all necessary Certificates, for enabling service invocation through the network, have been retrieved; and
- 3) **In operation:** Platforms can publish services that they are willing to expose in the NATO Services Registry. This Registry is hosted on the one national Platform. Platforms can browse the NATO Services Registry and subscribe to selected services. They will receive updates of information from the services to which they have subscribed.

In summary, each Nation (participant in NATO activities) should have the following features [7]:

- Adding/Removing of active services in operation.
- Adding/Removing of Systems providing/consuming services in operations when appropriate adaptation connectors are ready.
- Insertion of legacy systems or infrastructures.

## 5 Conclusion

In this paper we have presented SOA (Service oriented architecture) that we are developing to enable interoperability between Macedonian Ministry of Defence (MoD) and NATO. First we gave a brief overview of Service-Oriented Architecture (SOA) and an explanation of how it can be used for the Ministry of Defense purposes. The principal objective of our work was to ensure that architecture descriptions can be compared and related across organizational boundaries, including NATO and multi-national boundaries.

The benefits of using SOA to support NEC may be summarized as follows:

- Military resources are made available as services over a communication network;
- Efficient discovery of and subscription to as well as downloading of relevant information;
- Faster deployment of new technology and functionality;
- Dynamic reconfiguration of functionality within a relatively short timeframe;
- Integration of functionality over different networks and heterogeneous technologies;
- Minimal pre-planning required – loose coupling of systems.

## References

1. Sliwa, J., Gleba, T.K., Amanowitz, M.: Adaptation Framework for Web Services Provision in Tactical Environment. Concepts and Implementations for Innovative Military Communication, 52–67 (2009)
2. Pullen, J.M., Brunton, R., Brutzman, D., Drake, D., Hieb, M.R., Morse, K.L., Tolk, A.: Using Web Services to Integrate Heterogeneous Simulations in a Grid Environment. Journal on Future Generation Computer Systems (FGCS) 21, 97–106 (2005)
3. Tolk, A., Gaskins, R.C.: Challenges and Potential of Service-Oriented Architectures for Net-Centric Operations. In: Virtual Media for Military Applications, pp. 1–14 (2006)
4. Winters, L.S., Tolk, A.: The Integration of Modeling and Simulation with Joint Command and Control on the Global Information Grid. In: Proceedings IEEE Spring Simulation Interoperability Workshop. IEEE CS Press, Los Alamitos (2005)
5. Ang, H.W., Dandashi, F., McFarren, M.: Tailoring DODAF for Service-Oriented Architectures, Mitre Corporation Report, McLean, Virginia (2008)
6. OASIS Web Services Notification (WSN) Technical Committee (2006), [http://www.oasis-open.org/committees/tc\\_home.php?wg\\_abbrev=wsn](http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsn)
7. NATO RTO Technical Report IST-061, Secure Service Oriented Architectures supporting NEC (2009)

# Empirical Analysis of Software Reliability Models Predictive Ability

Aleksandar Dimov

Sofia University “St. Kl. Ohridski”,  
Department of Software Engineering, FMI, 1113, Sofia Bulgaria  
email: aldi@fmi.uni-sofia.bg

**Abstract.** Each software system needs an evaluation of its non-functional characteristics. Reliability is one such important non-functional characteristic and currently there exist a lot of models that asses it from different perspectives. In this paper we make an empirical comparison of several such models and analyze their ability to correctly estimate software reliability.

## 1 Introduction

Non-functional characteristics and quality of software systems are taking increasing attention from both researchers and practitioners in the area of software engineering. One significant quality parameter is dependability [1], which is defined as the ability of a computing system to deliver services that can justifiably be trusted. Dependability is represented by several attributes, such as reliability, availability, safety, confidentiality, integrity and maintainability. One important attribute of dependability is reliability. It is defined as the continuity of correct service, i.e. the belief that a software system will behave as per specification over a given period of time and is usually modeled as a stochastic value. It may have different measures like: probability of failure; mean time between system failures or failure rate.

Reliability is crucial to be considered in all phases of software development, especially in the context of embedded and safety-critical software systems [2]. From one side it is necessary in order to assess when enough testing has been performed on a given software unit. On the other side it allows to select a best candidate component to be integrated into a software system from architectural viewpoint. In order to calculate reliability in the first case, the so-called black box models (also called reliability-growth models) are applied, which regard the software as a monolithic whole and rely on testing data [3], [4]. In the second case models that take into account the internal structure of the software (like architecture) should be applied [5], [6].

In this paper we are focusing on the models from the first of the aforesaid groups – black box models. Despite their large number, in the literature there does not exist enough research in terms of comparison and analysis of their capability to predict the reliability of software components.

The goal of the paper is to compare different black-box reliability models with respect to their ability to predict the time to next failure of the software system. This will enable quality assurance engineers select the best model to fit their needs and to have more utensils to determine when it is appropriate to stop testing of software.

The remainder of the paper is organized as follows: Section 2 describes in more detail the motivation for our research and reviews the related work; Section 3 describes the research experiment and results of models comparison; Section 4 discusses the results and finally Section 6 concludes the paper.

## 2 Motivation and Background

Black-box software reliability models are statistical models that should predict software system's failure rate, given some failure history of the system. They are usually applied on raw data from system testing and this way they identify the failure density of software<sup>1</sup>. They are used when past failure information about the software system is available and assume extensive testing and observation of failures. When in state of failure, the system produces some kind of unexpected behavior in terms of wrong result, late reaction, etc. Failures are usually provoked by faults (also known as bugs), which reside somewhere in system code.

Usually input data for black-box models are the number of failures and the time that have passed between two subsequent failures. Most models assume that when a failure is detected, the fault that caused it is removed and the process continues with the assumption that the correction did not introduce new faults into the code.

The closest relations to our work mainly concern classification, survey and selection methods for reliability models. Several very sound model reviews already exist in the literature with some of them done in the seventies and eighties of the 20<sup>th</sup> century. For instance, the work presented in [8] classifies models in four big groups (Times between failures models, Failure count models, Fault seeding models and Input domain based models) and further discuss model applicability with respect to their assumptions and limitations. Another interesting model review [7] presents a thorough classification and theoretical comparison of the models, dividing them in two classes: time and data domain approaches. However, none of these reviews provide empirical analysis of the models with real-world failure data. A newer survey is published by Tandem Computers [4]. It evaluates models with respect to predicted number of faults that they estimate about the software system, as suggested in [12].

Another direction of related research is concerned about how to select a particular reliability model that will best match the examined system [9], [10]. However these works do not focus directly on model analysis and comparison is made just to validate the approach described. It should be noted that the focus there is also the ability of the model to predict total number of failures in the software.

In nowadays practice it is of highest importance to predict not the number of failures that reside in the system, but the time to next failure. Indeed complexity of modern software systems increased so much, that long time is needed in order to be sure that enough testing has been completed, for the reliability model results to converge and to come to realistic estimate about total number of failures. In this context, it would also be practical to evaluate model applicability with respect on

---

<sup>1</sup> Have in mind that the notions of black-box software reliability models and black-box testing have different meanings.

its ability to correctly predict the time to next failure of software system. Actually, this comparison parameter is closer to the definition of reliability, as given in the introduction of the paper.

In the next section we continue with the comparison of the models. The results are discussed and analyzed in Section 4.

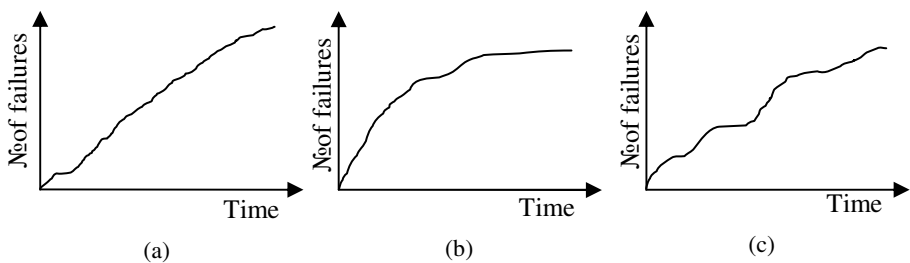
### 3 Comparison of Software Reliability Models

This section first describes the comparison framework and the fundamentals of the experiment conducted and next it presents the results obtained.

#### 3.1 Description of the Experiment

The Data & Analysis Center for Software (DACS) [13] has provided a number of publicly available datasets, containing failure data about software systems in different domains – real time and control, commercial, military, operating systems, etc. Datasets provided contain the number of failures, experienced during testing, times between two consecutive failures and also the total testing time. This information may be used as an input to the CASRE (Computer Aided Software Reliability Engineering) tool [14]. It implements a set of models that can be executed over results from any stage of system testing - unit testing integration testing, acceptance test and also during system operation.

In order to compare different models, we apply some of the significant of them over the testing results of more than one system (i.e. dataset), given by DACS. We have selected the datasets, according to encompass a broad range of system failure behaviours. Under the term failure behavior, here we mean a certain distribution of the number of failures over time. For example, it may follow linear (Fig. 1a.), exponential (fig. 1b.) or multi-exponential (fig. 1c.) curve. These are the most-widespread examples and in this paper we do not focus on other types of failure behaviours.



**Fig. 1** Different type of software system failure behaviours

The following DACS datasets were considered in this paper:

- A Real-time command & control system (System #1), with 21 700 lines of code and 136 failures detected over about 89 000 seconds of system testing.
- A Commercial subsystem (System #6), with 5 700 lines of code and 73 failures detected over about 5 100 seconds of system testing.

- A Military system (System #40), with 180 000 lines of code and 101 failures detected over about 19 million seconds of system testing.
- An operating system (System #SS1B), with hundreds of thousands lines of code and 375 failures detected over about 50 million seconds of system testing.

For the analysis, we have selected a number of models, belonging to different model groups and implemented in CASRE, namely: Geometric (G) model [18], Jelinski-Moranda (J-M) model [15], Littlewood-Veral (L-V) model [16], Musa-Okumoto (M-O) model [17] and Non-homogeneous Poisson Process (NHPP) model [19].

Description of these models is outside the scope of this paper and for more details, the reader is referred to the respective literature. For all models and systems we choose the maximum likelihood method for parameter estimation, as available in CASRE options. There is also a possibility to select a range at the start of the failure data over which initial parameter estimates are made. After that range, CASRE makes parameter estimates for each observation in the failure data set. For all experiments, this initial data window needed for the estimations to converge is set to half of the total number of failures (as by default).

Outputs of the models are compared towards raw failure data of the system in terms of actually experienced Time Between Failures (TBF). Visualization of results is made in next subsection by plotting the relative model prediction error  $\bar{e}$ , given by equation (1).

$$\bar{e} = \frac{TBF_r - TBF_m}{MTBF_i}, \quad (1)$$

where  $TBF_r$  is the actually experienced (real) time between failures, as taken from the raw dataset,  $TBF_m$  is the model prediction and  $MTBF_i$  is the mean time between failures, experienced at the  $i^{th}$  failure observation. In this case negative value means that the model overestimates the potential of software and positive value means that it underestimates it. It should be noted that the second case is preferred in practice, because if we are sure that the model always underestimates the software than we may take its results as an input for a worst-case reliability analysis. Hereinafter, underestimation of the system is referred as *the ability of reliability model to estimate potential of the software*. Assessment of model ability (*MA*) to correctly estimate the potential of software is done by equation (2)

$$MA = \frac{N_p}{N}, \quad (2)$$

where  $N_p$  is the number of failure observations for which  $\bar{e} \geq 0$  and  $N$  is the total number of failure observations.

### 3.2 Results

The results from our experiment for system #1 are shown on Fig. 2. The results for Systems #6, #40 and #SS1B are shown on Fig. 3, Fig. 4 and Fig. 5 respectively. Table 1 shows the *MA* parameter for each model, i.e. how many times each model has underestimated the respective system.

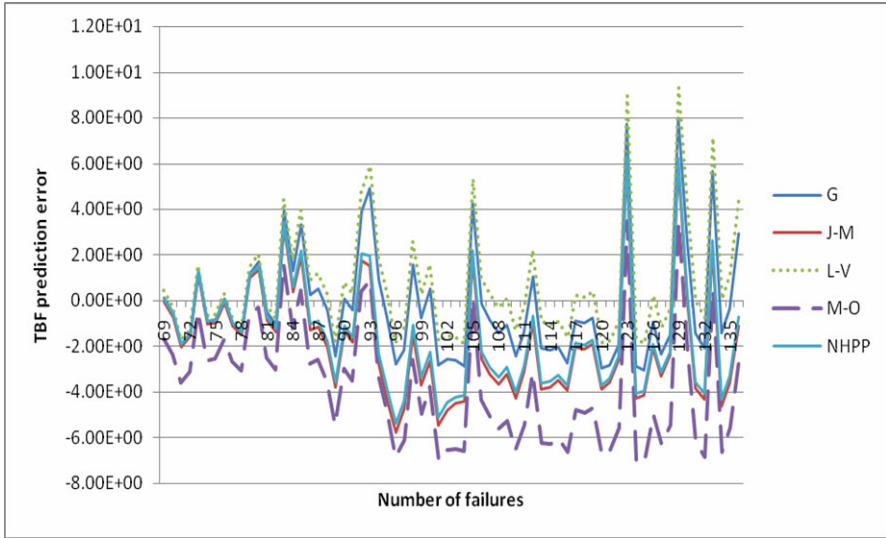


Fig. 2 Model results for System #1

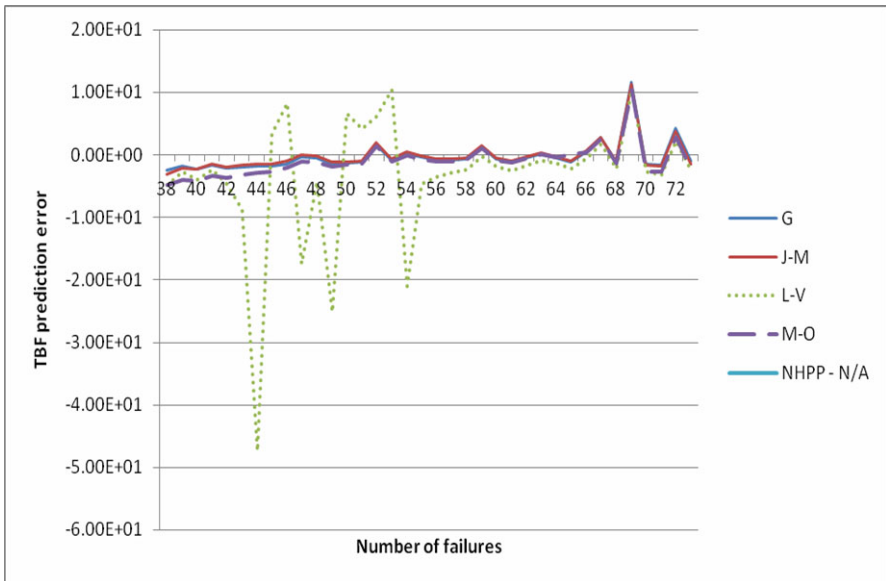


Fig. 3 Model results for System #6



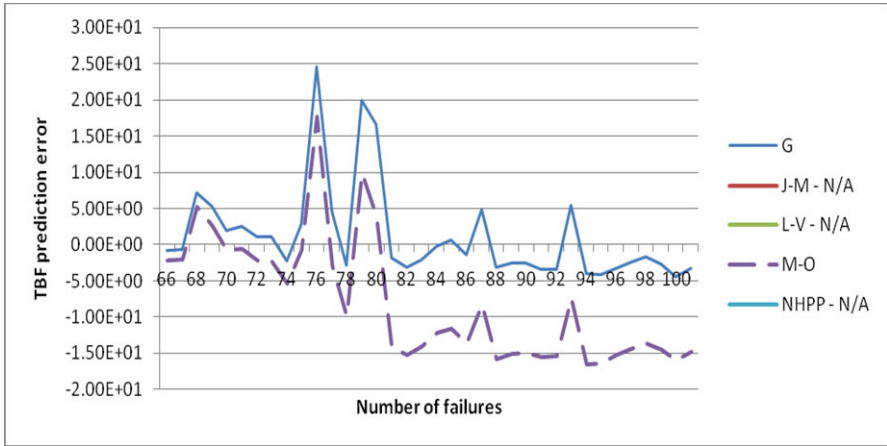


Fig. 4 Model results for System #40

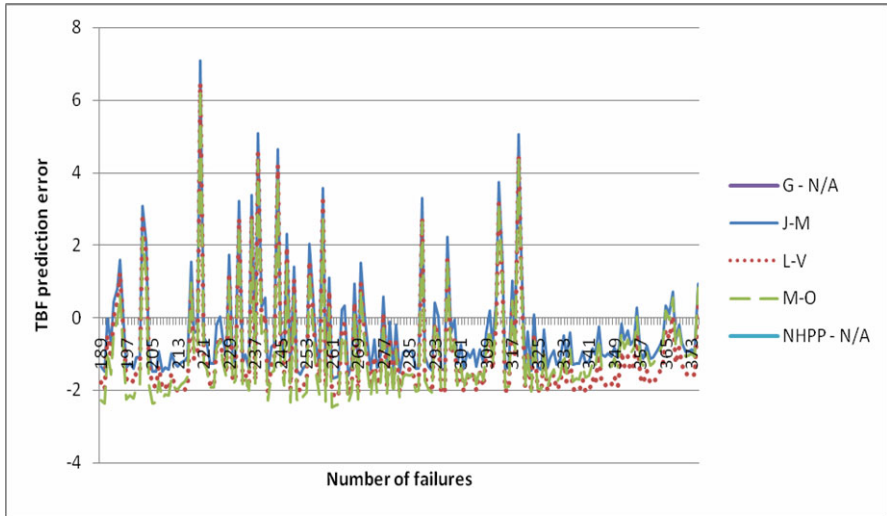


Fig. 5 Model results for System #SS1B

Table 1 Model ability to correctly asses the potential of software system

	System #1	System #6	System #40	System #SS1B
G	0,338	0,222	0,389	n/a
J-M	0,191	0,222	n/a	0,235
L-V	0,529	0,25	n/a	0,166
M-O	0,118	0,222	0,139	0,150
NHPP	0,221	n/a	n/a	n/a

## 4 Discussion

A general look at the results shows that in most cases the models show similar behaviour with respect to a particular system. The only exclusion of this rule is System #6, where L-V model show very large deviations of the predicted time to failure. In the contrary it shows lower deviations from all models in the case of System #SS1B. J-M and G model give best results in terms of deviation of actually experienced TBF in the case of System #6. Similar considerations infer that G model is best applicable to System #1, and System #40. In this sense we may conclude that different software systems may have different models as best fits to estimate their reliability.

The only model that is able to give predictions for all 4 systems is M-O model, however, as seen from table ,1 it has relatively low ability to assess system potential. All models give similar results when applied to System #6 with respect to their ability to assess system potential. It is important to note that only one model (L-V) is able to underestimate (the better from reliability viewpoint case) the software with more than 50% and this is the case only for System #1.

Results shown and discussed above, show that although this is an area with several decades of research work, there still exist a need to address some challenges in it. In particular new software reliability models that are able to better predict next time to failure and have higher ability to estimate the potential of software are needed.

## 5 Conclusion

Reliability is one important non-functional characteristic (as an attribute of the general notion of dependability) which should be regarded both during development and usage of software systems. Currently there exist a lot of software reliability models that assess reliability from different perspectives. A big group of models, called black-box or reliability-growth models make statistical processing of software failure history in order to estimate reliability. In this paper we analyze five of these models and apply them to failure datasets obtained by testing of four different software systems. Results show that there still exist work to be done in improving models predictive ability at least in two directions – ability to predict the potential of software and ability to predict correctly the time to next system failure.

Directions for future research include improvement the existing models or development of a new reliability model, which should have better predictive ability.

**Acknowledgements.** The work presented in this paper was partially supported by grant from the Bulgarian National Science Fund, under the PD01-0106 ARECS Project / Д002-298.

## References

1. Avižienis, A., Laprie, J.-C., Randell, B.: Basic concepts and Taxonomy of dependable and secure computing. IEEE Trans. on Dependable and Secure computing 1(1) (January-March 2004)

2. Dimov, A.: Reliability Assessment of Safety-Critical Embedded Software Systems. In: Proceedings of the International Conference on Information Technologies (InfoTech), Bulgaria (to appear, September 2011)
3. Farr, W.: Software reliability modeling survey. In: Lyu, M.R. (ed.) Handbook of Software Reliability Engineering, pp. 71–117. McGraw-Hill, New York (1996)
4. Wood, A.: Software Reliability Growth Models. Tandem Technical Report 96.1 (September 1996)
5. Goseva-Popstojanova, K., Trivedi, K.S.: Architecture Based Approach to Reliability Assessment of Software Systems. Performance Evaluation 45(2-3) (June 2001)
6. Dimov, A., Punnekkat, S.: On the Estimation of Software Reliability of Component-Based Dependable Distributed Systems. In: Reussner, R., Mayer, J., Stafford, J.A., Overhage, S., Becker, S., Schroeder, P.J. (eds.) QoSA 2005 and SOQUA 2005. LNCS, vol. 3712, pp. 171–187. Springer, Heidelberg (2005)
7. George, S., Wolverton, R.: An Analysis of Competing Software Reliability Models. IEEE Transactions on Software Engineering SE-4(2), 104–120 (1978)
8. Goel, A.: Software Reliability Models: Assumptions, Limitations and Applicability. IEEE Transactions on Software Engineering SE-11(12), 269–281 (1985)
9. Stringfellow, C., Andres, A.A.: An empirical method for selecting software reliability growth models. Empirical Software Engineering 7(4), 319–343 (2002)
10. Andersson, C.: A replicated empirical study of a selection method for software reliability growth models. Empirical Software Engineering 12(2), 161–182 (2007)
11. Wood, A.: Software reliability growth models: assumptions vs. reality. In: Proceedings of the Eighth International Symposium on Software Reliability Engineering, pp. 136–141 (1997)
12. Iannino, A., Musa, J., Okumoto, K., Littlewood, B.: Criteria for Software Reliability Model Comparisons. IEEE Trans. on Software Engineering SE-10(6), 687–691 (1984)
13. DACS Software Reliability datasets,  
<https://www.thedacs.com/databases/sled/swrel.php>  
(last visited on May 16, 2011)
14. Nikora, A.: Computer Aided Software Reliability Estimation User's Guide (CASRE), Version 3.0 (2002)
15. Jelinski, Z., Moranda, P.: Software reliability research. In: Freiberger, W. (ed.) Statistical Computer Performance Evaluation, pp. 465–484. Academic Press, New York (1972); IEEE Trans. Reliability R-34, 216–218
16. Littlewood, B.: The Littlewood-Verrall model for software reliability compared with some rivals. Journal of Systems and Software (1), 251–258 (1980)
17. Musa, J.D., Okumoto, K.: A logarithmic poisson execution time model for software reliability measurement. In: Proc. of 7th Int. Conf. on Software Engineering, pp. 230–238 (1984)
18. Moranda, P.: Prediction of software reliability during debugging. In: Proceedings of the Annual Reliability and Maintainability Symposium, Washington DC, pp. 327–332 (1975)
19. Goel, A., Okumoto, K.: A Time Dependent Error Detection Model for Software Reliability and Other Performance Measures. IEEE Transactions on Reliability R-28(3), 206–211 (1979)

# Living Labs in E-Learning, E-Learning in Living Labs and Living Labs for E-Learning

Albena Antonova and Korneliya Todorova

Faculty of Mathematics and Informatics, University of Sofia

Tzarigradsko Shose 125 bl.2 fl.3 Sofia 1113, Bulgaria

email: a\_antonova@fmi.uni-sofia.bg, cornelia@fmi.uni-sofia.bg

**Abstract.** Within knowledge-based society companies are the main engines of technological innovations, boosted by severe global competition and shorter product life-cycle. In parallel learning systems slowly adapt to increasing needs of learners, society and companies. Learning often becomes isolated practice, concentrated in educational institutions and not reflecting regional, company or practical context. Therefore, the present research proposes a new approach for discussion and comparison of emerging living labs environments with e-learning systems as technology-mediated social participation systems. Matching e-learning with living labs can increase value both for education and for innovation processes. That is why there will be discussed three main scenarios for implementation – e-learning in the context of living labs, living labs in the context of e-learning and implementation of living lab for e-learning.

**Keywords:** Living labs, e-learning, open innovations, technology-mediated social participation systems.

## 1 Introduction

The capacity for innovation and fast knowledge realization become the main company advantage. Increased competition on global markets fosters companies to lead technological innovation growth and to exploit new sophisticated solutions, complex services and advanced business models. It results to more knowledge-intensive products and services, with shorter life-cycle, and increased demand for support and customization. In order to retain their leading positions, companies largely depend from educational institutions. On one side companies need educated employees and customers that are able to use and further develop and support enhanced products and services. On the other side, universities and research centers are expected to deliver high-quality scientific outputs, accelerating further technological innovations and R&D production. However, educational institutions are often in the position only to follow the fast changing technological trends of the streamline and slowly transform and adapt its learning processes to the increasing needs for life-long learning. Thus both companies and educational institutions have to improve collaboration and cooperation.

Recently emerged, Living labs (LL) propose an innovative infrastructure enhancing end-user involvement in the context of complex products development. In the same time e-learning technologies have already entered in the mature phase, merging various learning contexts: formal learning, life-long-learning and ubiquitous learning context. Can both approaches facilitate knowledge sharing and collaboration between industry and academia?

The present research aims to discuss the phenomena of living labs and e-learning and the possibility to integrate both in order to improve educational and innovation process. In the first place it is presented the concept of living lab, its features and mode of operation, followed by description of e-learning characteristics. The second part of the paper proposes an innovative approach to investigate living labs and e-learning as technology-mediated social participation systems. There are identified main features of both approaches, considering LL as social system and e-learning as social media. The third part provides a broad discussion of different integration mechanisms as LL in the context of e-learning, e-learning in the context of LL, and there are discussed practical approaches for LL for e-learning. Finally, there is proposed a complex model integrating the three approaches, analyzing the limits and benefits of LL and e-learning use for better industry-education-research cooperation, regional development and knowledge sharing.

## **2 Theoretical Review - Living Labs, E-Learning and TEL**

Living labs (LL) is evolving concept, firstly emerged in USA, but fast spreading around Europe (EnoLL). This is a form of user-driven open innovation ecosystem, based on a partnership which enables users to take an active part in the research, development and innovation process. It can be defined as “an environments for innovation and development where users are exposed to new solutions in (semi)realistic contexts, as part of medium- or long-term studies targeting evaluation of new solutions and discovery of innovation opportunities”[10]. LL represent a research methodology for sensing, prototyping, validating, and refining complex solutions in multiple and evolving real life contexts. The main concept behind is that LL bring users early into the creative process in order to better discover user patterns; LL bridge the innovation gap between technology development and the uptake of new products and services; LL allow early assessment of the socio-economic implications of new technological solutions by demonstrating the validity of innovative services and business models [12].

Living labs become main test place for development of innovation, as it combines simultaneously open innovation approach, end-users active involvement and distributed value co-creation. Living labs are organized on regional principle, enhancing local knowledge sharing in specific industry areas. LL is not a network of infrastructure and services but much more a network of real people with rich experiences [6].

Therefore Living labs can be identified by its participants and complex services and innovation offerings. The main participants of Living labs are: researchers, end-users and developers [3]. Developers organize and manage the innovation

experiences and users' involvement, so thus competent management of the innovation process depends on their efforts. Developers have to fulfill the end-users needs, but also search for their own market and business opportunities. In the context of LL, researchers explore case studies, collect primary data from experiments and actively observe and collaborate with end-users and developers. Finally end-users are interested to search better solutions to specific needs and are involved in many use-case experiences and situational exercises. End-users can be individual, like consumers, or group as workteams.

The main processes of Living labs usually follow the technology innovation phases – ideation process, design and prototyping, production, testing and validation phase and usability testing. LL is human-centric, opposed to technology-centric approach [3]. Typical offerings include R&D projects, pre-studies, end-users events, workshops, need finding activities, and different kinds of product-service or market evaluations, formalized meetings, special interest groups, counselling groups, clusters and others [3].

E-learning is defined as complementary channel of communication allowing computers and computer networks to connect learners with learning media, with other people (fellow learners, sources, facilitators), with data (about learning, about media, about people) and with processing power [5]. E-learning technologies can perform customized, cheaper, flexible and learner-oriented training, reflecting personal attitudes and allowing new type of learning process, fostering significant improvements in accessibility and opportunity to learn [9]. It couples innovations in technology to eliminate barriers of time, distance and socio-economic status, creating a whole new dimension of learning. E-learning is closely linked with the concept of technology-enhanced learning (TEL), and ICT implementation in education and learning. Information technologies and e-learning largely influence life-long learning of all generations, due to new possibilities to access and share information, new roles and pedagogical paradigms [7]. Therefore, in the context of the present paper, we will understand e-learning as broad concept of technology-enhanced learning modules both in formal (primary, secondary, tertiary level) and informal learning experiences (life-long-learning and self-learning).

However, although the popularity of e-learning today, practitioners report lack of interactivity, low contextualization, lack of simulations and in fact e-learning serves traditional learning methodologies mainly as mediums for dissemination of learning materials [1]. Even the most important aspects of e-learning (reusability and learner personalization) are not realized, as organizations developing e-learning resources follow their own cycle and do not adapt tools and learning methodologies to the needs of the learners [9]. The main value of e-learning consists of the meta-information about the content, when it is useful and how to reuse it. Therefore, the heart of successful e-learning system is the methodology of the system to define and manage knowledge content. Thus e-learning systems have to change the perspective from content-oriented approach to knowledge-synthesis approach. As further Teo&Gay [9] continue, the two main tasks of the e-learning systems should be to facilitate the creation and synthesis of new knowledge and to manage the way people share and apply it in various context.

E-learning systems value should deliver context-rich content shared and evolved within social community.

Finally, according to [6], e-learning and TEL recently evolved to ubiquitous learning. Thus e-learning technologies provide continuous and context-based educational material to learners anytime, anywhere, and from any device.

### 3 Living Labs and E-Learning as TMSP

Living labs and e-learning are both social participation systems. The both e-learning and living labs approaches should be implemented within some technology-mediated environment. Therefore we will use the methodology and analysis of Chi et al. [2], to investigate technology-mediated social participation (TMSP) systems, in order to describe main characteristics of the both e-learning and Living labs systems.

The emergence of social media transformed the logic of the users involvement within technological, organizational and economic aspects, enhancing learning, working and collaboration. Social systems on the other hand create new supportive environment for social creativity and collaborative design. While e-learning is mainly referring to social media, living labs can be classified successfully as social ecosystem. The main approach will be to compare the both – e-learning and living labs, according to main considerations of TMSP systems. According to Chi et al. [2], the design for social participation rely on the concept of usability (ability of all users to contribute), sociability (users skills for networking and participation), social capital (different positions in social networks), and collective intelligence (evolution of collective ideas). The principle concern for designing TMSP systems is to provide large-scale participation, and ensuring that participants both give and get something back from the system.

In the context of TMSP, three main factors for consideration include knowledgeware, toolware, peopleware. The knowledgeware refers to the understanding of domains and contexts of impact, social experiences, community life stage and individual differences. Toolware refer to IT system components that enable effective social participation. Finally peopleware describe how people interact in social cognitive systems, both as individuals and as social agents. How to improve social interactions, conflict management, system governance and control?

**Table 1** Comparing LL and e-learning systems according to three criteria of TMSP.

TMSP	Living labs-social ecosystem	E-learning – social media
Knowledgeware	Focus on specified domain and region	Focus on learning content
Toolware	Participation and communication, collection of data and use-cases	Focus on Personalization, participation, knowledge sharing
Peopleware	Focus on End-users experiences	Focus on learners

### 3.1 Living Labs in the E-Learning Context

Living labs can be broadly described as complex social ecosystems, facilitating participation of different heterogeneous agents in the process of open innovation. Thus in the context of e-learning, LL propose valuable opportunity to involve different social actors in the learning process, in order to make it specific, context-based and problem-oriented. Living labs can provide unique problem-based learning experiences for learners, while solving real innovation problems of companies and regions. Lecturers often discuss examples and case studies that are not relevant to the learners cultural, economical, educational, or geographical context. During their studies, learners are regularly involved in research and practical exercises, course works, individual and group assignments. Serving as end-users, learners can contribute to all phases of the innovation process, from idea generation to end-product testing. Collaboration between LL and e-learning can enlarge learners' involvement in local environment and economy, providing real-life examples and solutions for local community.

As illustration, two case studies reported by [4] and [8] are discussed. Exploring Living labs concept, Luoju & Vilky [4] developed research-oriented, problem-based educational approach, linking instruction to company-based production process in cooperation with local businesses in the field. The reported results indicate that learners succeeded to develop new complex skills to: flexible use and master diverse development tools and models, get used to problem-solving, gathering and structuring knowledge in genuine development contexts, apply user-driven production process and research method.

Reichel & Schelhowe [8] present the learners' experience in the LL of smart textiles. Authors reported that learners are an interesting target group for companies, because of their capacity to generate innovative ideas. After the experience, Reichel et al. concluded that learners can invent complex innovative products quite similar to real designers. Thus Living labs involvement in various learning situations can be beneficial both for learners and for industry partners from living labs.

### 3.2 E-Learning Concepts in the Context of Living Labs

In this section we will focus on Living labs benefits from implementation of modern e-learning. E-learning is discussed from two main perspectives: first as social media, delivering specific learning content in the context of technology-mediated social participation systems, and secondly as method for developing social skills and competences.

With adoption of social media, e-learning becomes creative, collaborative experience. The personal assessment of learner prevails, informal learning spaces are used for group collaboration, content is created via participation and collective intelligence, personalized learning is enabled [11]. The new participatory applications allow learners to improve time planning, deeper study of colleagues works and result in deeper knowledge acquisition. Having in mind that learners have to be active members (participant) in the process of design and development of e-learning resources, the e-learning integration in the context of Living labs could present additional benefits. Collaborative tools and social software seems to be very appropriate for creation of shared e-learning content. Using Web 2.0 tools



learners could set requirements and criteria for efficient education using different media as blogs, wikis and other tools for opinion sharing, commenting and voting for new ideas, products designs and others.

In parallel to deliver learning content in the context of LL, participatory e-learning system contribute for development of specific skills and competences, increasingly important for the further users of Living labs. Participatory and communication skills allow learners to express better their identity in online systems and to create rich social media content through participation. Moreover, e-learning social media can develop skills for active civic participation, improving social and cultural expression, and developing better competences for involvement in knowledge-intensive activities.

Therefore in the context of living labs modern e-learning media can propose not only participatory environment for sharing relevant knowledge and information. E-learning develops skills and competences, enhancing access to knowledge and information and contributes to complex communication and collaboration models. Learners are actively involved in many participatory scenarios and this contribute for further Living lab success.

### **3.3 Development of Living Labs for E-Learning**

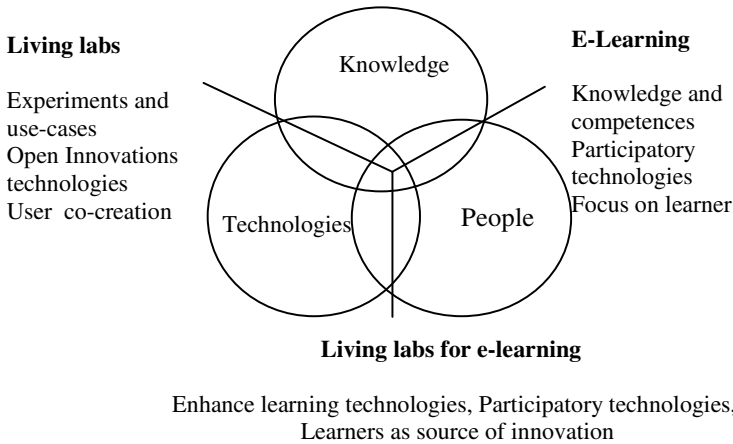
The third perspective in the analysis includes practical considerations about implementation of Living labs for E-learning applications. As Living labs are recognized as successful practices for software technologies improvement, they could be successfully implemented in the context of developing modern and sophisticated e-learning solutions, adapted to the end-users needs and concerns. Thus implementing LL for e-learning, proposed by Serra [1] can significantly improve the quality and attractiveness of e-learning systems. Designing better collaborative tools can enhance learners to co-create learning content and to expand learning resources repository. On the other hand involvement of users in the process of collaborative tools design and development will allow the e-learning experts and developers to create more customized and adapted to the specific needs of student tools for communication and collaboration. Several examples include: learning approach and methodology (LL for Serious games), learning technologies (LL for e-learning platforms), user orientation (LL for elderly and Life-long-learning, e-learning for students, etc.).

As admitted by Leven& Holmstrom [3], LL can attract a heterogeneous set of actors with heterogeneous set of needs. Thus on one side Living labs can improve design and development of new e-learning solutions, relevant to different regional contexts of life-long learning, and on the other side, living labs can propose research and learning base for educators, learners, researchers and companies – technology developers.

In the first place e-learning LL can propose social environment where users can be involved in the early stage of design, development and testing of e-learning solutions. The end-users of e-learning experiences can be learners, but as well lecturers (who design and deliver learning content), company administrators, HR experts, KM experts and others. Developers or companies delivering e-learning platforms have a leading role and should explicitly design learning experiences

and simulations, measuring on one side outcomes of the learning and on the other side evaluating learning process.

Finally, researchers can contribute with innovative learning approaches and methodologies for better e-learning solutions development, searching synergy between social and technological settings, intercultural and interdisciplinary approaches for TEL. Researchers can get additional information in the context of end-users feedback, collecting data about different use-cases, propose improvements and others.



**Fig. 1** E-learning and Living labs in the context of TMSP

Developing a framework for integration of Living labs and e-learning can enhance both innovation and learning processes. As summarized on Fig.1, both e-learning and LL take part in technology-mediated social participation systems, improving social interactions, knowledge sharing and various learning opportunities within specific location. TMSP theory allow better settings for understanding complex socio-technological consideration of LL and e-learning.

Using Living Lab methodology can enable learners to get information about real life problems and existing challenges in different field of business delivered by other participant in Living Labs (organization, firms, regional organizations). This way they will be able to solve real problems and will be more competitive specialist in their professional experience.

## 4 Conclusion

Living labs and e-learning take part in participatory social systems, influencing large impact on many social domains as education and research. While Living labs represent a social eco-system, contributing for open innovation processes, e-learning provide social media, increasing digital competences and delivering learning content. We discussed the multiple opportunities to involve living labs in learning process, improving academia-industry partnership and collaboration of

learners with wider regional community of researchers, companies and users. On another side, the recent trends in e-learning contributed for development of new set of technologies and competences, contributing to active learning, participatory learning and learning experiences. Finally various considerations about Living lab for e-learning purposes are discussed. Following our research, we would like to conclude that living labs and e-learning are interconnected and provide multiple benefits for new enhanced experiences.

Our future research plans will be dedicated on implementation of Living Labs methodology in high level education and vocational training in order to verify the efficiency of proposed solution, to revise the ideas for improvement in order to deliver and support high quality learning. We intend to develop an e-Learning course representing basic concepts of Living Labs and to involve learners with main principles, concepts, methods and methodologies of this relative new filed.

**Acknowledgments.** The work on this paper has been sponsored by the SISTER Project funded by the EC 7th FP.

## References

1. Antonova, A., Todorova, K.: Serious Games and Virtual Worlds for high-level learning experiences. In: Proceedings of S3T Conference, Varna, September 12-13 (2010)
2. Chi, E.H., Munson, S., Fischer, G., Vieweg, S., Parr, S.: Advancing the design of Technology-Mediated Social Participation Systems. *Computer* 43(11) (2010) ISSN: 0018-9162
3. Leven, P., Holmstrom, J.: Consumer co-creation and the ecology of innovation: A living lab approach. In: Proceedings of IRIS31, Are, Sweden (August 10-13, 2008), <http://umu.diva-portal.org/smash/get/diva2:311009/FULLTEXT01>
4. LuoJus, S., Vilky, O.: Development of User-Driven Research Methods as the Starting Point for Living Lab Activities. In: Proceedings of NordCHI, Lund, Sweden, October 18-22 (2008)
5. Morrison, D.: E-learning strategies. Wiley & Sons, Chichester (2003)
6. Muntean, C.H., Muntean, G.M.: Open corpus architecture for personalised ubiquitous e-learning. *Personal Ubiquitous Computing* 13, 197–205 (2009)
7. Kikis IADIS, What we need to know if we want to assess the impact of ICT in education (2009)
8. Reichel, M., Schelhowe, H.: Living labs: driving innovation through civic involvement. In: The Proceedings of IDC 2008, Chicago, USA, July 11-13 (2008)
9. Teo, C.B., Gay, R.K.L.: A Knowledge-Driven Model to Personalize E-Learning. *ACM Journal of Educational Resources in Computing* 6(1), Article 3 (March 2006)
10. Foelstad, A.: Towards a Living Lab For Development of Online Community Services. *Electronic Journal of Virtual Organisations* 10, 47–58 (2008)
11. Adzgauskiene, D., Butrime, E., Zuzeviciute, V.: Impact of Web 2.0 on e-learning as a socio-cultural system. In: Proceedings of IADIS, pp. 249–252 (2009)
12. EC, Living labs for User-driven innovation, report Information society and media (2009)
13. Serra, A.: Open innovation systems, living labs and SMCs: The case study of Citilab, Cornella de Llobregat, Spain, <http://www.slideshare.net/arturserra/the-case-of-citilab-truefinalv2>

# An Ontology-Based Model for Driving the Building of Software Product Lines in an ITS Context

Alan Pedro da Silva, Ig Ibert Bittencourt, Walker Ataide, Olavo Holanda, Evandro Costa, Thyago Tenório, and Patrick H.S. Brito

GrOW - Group of Optimization of the Web, Federal University of Alagoas  
email: {alanpedro, ig.ibert, walker, olavoholanda, evandro, patrick}@ic.ufal.br

**Abstract.** The Intelligent Tutoring Systems (ITS) have been used daily to support education in various domains. For this reason, fast and easy construction of ITSs are fundamental requirements. In this sense, a Software Product Line has been used accordingly for building Intelligent Tutoring System families. However, the construction of such family systems is still a hard and complex task which involves the representation and manipulation of different knowledge sources with distinct artifacts. To alleviate these issues, this paper proposes an ontology-based model for driving the building of software product lines in an ITS Context. It also provides a case study describing the construction of an ITS in the programming domain. In addition, an evaluation is presented aiming to show the feasibility of the proposed model. The main conclusion is that this model reduces the effort and the complexity in the construction of such systems.

## 1 Introduction

Intelligent Tutoring Systems (ITS) have been used daily to support education in several educational contexts and domains. For instance, ASSISTment [8], a system that assists students in solving mathematical problems has been used as a large-scale system. Although it is used to solve problems only in the field of mathematics, it has been used by thousands of secondary school students as a tool that assists the teaching process by diagnosing students knowledge in the math field. As a matter of fact, it offers support in other knowledge fields as well (Physics and Chemistry), supporting further more than 300,000 students. Carnegie Learning is actively involved in the development of Intelligent Tutoring Systems [1], especially in mathematics. The acceptance of their products (they are used by more than 500,000 students [5]) has shown that they are effective in improving the education of students. On the other hand, although the needs for ITSs is increasing, there are few approaches for building ITS in various domains and contexts. In this sense, Ontology-Based Software Product Lines have been used accordingly for building families of Intelligent Tutoring Systems [10]. This kind of software families has some advantages: (i) It provides a mechanism for reducing the effort of building complex software families; (ii) It can integrate the

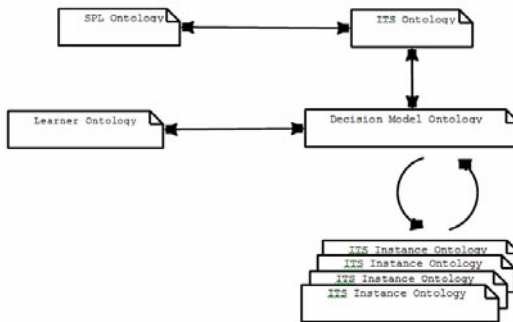
knowledge embedded in the Intelligent Tutoring Systems with several kinds of artifacts; (iii) It provides a distinct evolution for all the family. Existing approaches that focus on SPL based on Ontology produced some progress [9] such as large-scale, fast construction and support to different domains. However, there is not a model for ontology-based SPL in the context of ITS.

The rest of this paper is organized as follows. Section 2 presents the proposed solution. Section 3 presents a case study which validates our proposal. Section 4 presents related works. Section 5 contains a discussion and concludes.

## 2 The Proposed Solution

The goal of this section is to describe the proposed model for building Intelligent Tutoring Systems, using a Software Product Line approach based on the use of ontologies. The use of ontologies is motivated by the need of providing a semantic and consistent description of the knowledge present in intelligent tutoring systems. In addition, it is important to provide a general way to instantiate families of intelligent tutoring systems in different domains with different requirements. Moreover, it provides a way to divide the knowledge of the software implementation, providing the software product line artifacts separated of the ITS domain.

For this reason, it is important to have ontologies to represent: i) the software product line artifacts; ii) the intelligent tutoring system requirements; iii) the specification (decisions) of intelligent tutoring systems; and iv) the instances of each ITS. As a result, four ontologies were combined in order to provide the semantic description of the software product lines in the context of Intelligent tutoring Systems, as presented in Figure 1.



**Fig. 1** Models for Product Instantiation

Firstly, the SPL Ontology was designed to be used as a meta-model for a semantic specification of the Feature Oriented Domain Analysis (FODA)[4]. With this model is possible to create ontologies of any application domain, regarding FODA specifications. Then, the SPL-ITS Ontology was built with the definitions of the ITS Feature Model presented in Section 3, allowing that changes can be done independently of the ITS knowledge domains.

In the ITS context, there is a focus on the student and on the individualized teaching. So, the Learner Ontology was adapted from [2], which provides a mechanism to represent the learners, also relating for each student a set of features. Finally, the Decision Model Ontology represents the instantiation of an e-learning system. With this ontology, it is possible to properly set up the characteristics of each product.

### 3 Case Study

The aim of this section is to describe the features of the proposed model through the development of an intelligent tutoring system with different features. The main idea is to present each step in the construction of such an ITS. The model is composed by six steps of which the first step is related to the representation of the features of any intelligent tutoring system according to the ontology instantiation and the other steps are related to the use of an authoring tool based on the proposed architecture.

**Step 1 - SPL Feature Model:** A feature model provides a global view of the Software Product Line regarding the variable and common aspects of it. Thus, it becomes possible to realize the variable points of the system architecture. With this, from the SPL general diagram, a SPL for programming tutors was produced. The feature model of this SPL is presented below.

It can be seen that the created ITS has all the mandatory features as predicted. Besides, some optional functionalities/features were provided, such as:

- Pedagogical Strategy - Cognitive: This feature ensures the proper sequencing of educational resources for each student profile.
- Form: This aims to obtain an initial knowledge about each student;
- True or False Problem: This type of problems is important for evaluating specific questions about a certain curriculum;
- Hint: This functionality is responsible for providing hints to the student when necessary;
- Domain Learning Report: This functionality produces reports that can help the teachers to diagnose possible problems with each student.

It is noteworthy that at an implementation level, each feature was related to a software component. Through this it can dynamically add new features, without any additional price of implementation in other components.

**Step 2 - SPL Selection:** In this step the system designer must select the SPL and the name of the product (ITS) that will be generated.

**Step 3 - Product Customization:** In this step the designer of the system must customize the intelligent tutor from the feature model of the SPL chosen in the previous step (SPL Selection). The features can be “mandatory”, which must be mandatorily selected, “optional”, and “alternative”. If one alternative feature is selected, the alternative features related to it can not be selected.

**Step 4 - Feature Validation:** At this step the designer should click on the button “Validate Product ” for the authoring tool to verify if errors exist in the selection

of the features. If errors exist, the tool alerts the designer to correct them. Otherwise, the tool alerts a message that the product is correct and allows the designer to generate the product (Product Generation).

**Step 5 - Product Generation:** For realizing this step, the designer should have validated the features (as shown in the previous step), received a message that the product is correct, and should click on the button “Generate Product.” The authoring tool will generate the configured ITS. In case of ITS that utilizes Java Web technologies, the tool will generate a war file with the feature modules selected.

**Step 6 - Product Deployment:** At this step, the ITS is deployed on the web server. The designer should copy the URL provided in the message issued in the Product Generation phase and paste it in the web browser, so that he or she can access the generated and deployed tutor. This completes the process of authoring of the ITS.

The steps used to instantiate an ITS through the authoring tool are closely linked to the establishment, persistence and change in the ontology that describes it in terms of its features. Below the steps of the ontology instantiation of ITS are described and related to the steps of the authoring process:

1. **Ontology creation:** the ontology of ITS is created at the step “SPL Selection”. It generates an OWL file that describes the ontology that represents the tutor to be generated.
2. **Ontology persistence:** after the creation of ontology it is persisted in an ontology repository, such as SESAME.
3. **Ontology change:** this happens at the step “Product Customization”, when the designer indicates (the state of the feature is changed to “SELECTED”) or clears (the state is changed to “ELIMINATED”).

The checkbox of the feature model, an ontology of ITS, reflects these changes both in the ontology OWL file (used for verification) and in the ontology repository.

## 4 Related Works

There are several works related to building of Intelligent Tutoring Systems, however none of them is focused on the use of Software Product Lines for building Intelligent Tutoring Systems. For this reason, this section presents only the three main related works. The first two [6, 3] provide an architecture for building an ITS with its main characteristics. Although they enable facilities to create ITS, they do not provide mass customization (no use of SPL) and support to different ITS specification (different purpose on the use of ontologies).

In [7] the authors propose a methodology for the development of service oriented software product lines based on a multi-phase specialization of features model. Even so, this approach is not adapted for the building of cognitive systems, such as Intelligent Tutoring Systems. This approach is not intended to model cognitive knowledge, due to the reason that the artifacts related to the nonfunctional requirements are expressed on a not expandable formalism.

## 5 Conclusion and Future Work

In this paper a model for developing intelligent tutoring system based on the use of software product lines and ontologies is described. The introduced approach promotes an easy and efficient way for building intelligent tutoring systems. On one hand, the use of software product lines offers low development costs, mass customization, and systematic reuse. On the other hand, the use of ontologies offers knowledge sharing, knowledge reuse, and specification of different ITS domains.

The case study showed positive results concerning the facilities and effectiveness for building ITS applications. Our future plans include the development of new case studies and also the use of different metrics for evaluating of the proposed model.

## References

1. Aleven, V., McLaren, B., Sewall, J., Koedinger, K.: A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education* 19(2), 105–154 (2010)
2. Bittencourt, I., Nunes, C., Bezerra, C., de Barros Costa, E., de Oliveira, R.N., Costa, M., Tadeu, M., da Silva, A.P.: Ontologia para construção de ambientes interativos de aprendizagem. *Simpósio Brasileiro de Informática na Educação*, 559–568 (2006)
3. Cruz, A.C., Valderrama, R.P., Cano, O.F., Humberto Sossa, A.: Architecture for development of wbes based on components and agents. In: *15th Int. Conference on Computing, CIC 2006*, pp. 223–228 (2006)
4. Kang, K.C., Cohen, S.G., Hess, J.A., Novak, W.E., Peterson, A.S.: Feature-oriented domain analysis (foda) feasibility study. Technical report, Carnegie-Mellon University Software Engineering Institute (November 1990)
5. Koedinger, K.R., Corbett, A.T.: Cognitive tutors: Technology bringing learning science to the classroom. *The Cambridge Handbook of the Learning Sciences*. Cambridge University Press, Cambridge (2006)
6. Lin, F., Holt, P., Leung, S., Li, Q.: A multiagent and service-oriented architecture for developing adaptive e-learning systems. *Int. J. Cont. Engineering Education and Lifelong Learning* 16(1/2), 77–91 (2006)
7. Mohabbati, B., Kaviani, N., Gasevic, D.: Semantic variability modeling for multi-staged service composition. In: *Proceedings of the 3rd International Workshop on Service-Oriented Architectures and Software Product Lines* (2009)
8. Razzaq, L., Patvarczki, J., Almeida, S.F., Vartak, M., Feng, M., Heffernan, N.T., Koedinger, K.R.: The assistment builder: Supporting the life cycle of tutoring system content creation. *IEEE Transactions on Learning Technologies* 2, 157–166 (2009)
9. Rusk, J.J., Gasevic, D.: Semantic web services-based reasoning in the design of software product lines. In: *SPLC*, vol. 2, pp. 123–130 (2008)
10. Silva, A., Costa, E., Bittencourt II., Brito, P., Holanda, O., Dermeval, D.: Semantic web-based software product line for building intelligent tutoring systems. In: *Second International Conference of Software, Services and Semantic Technologies, S3T 2010*, pp. 127–136 (2010)



# A Proposal for Storysculpturing

Pavel Boytchev

KIT, Faculty of Mathematics and Informatics, Sofia University, Sofia, Bulgaria  
email: boytchev@fmi.uni-sofia.bg

**Abstract.** Storytelling is a core element of the design of Serious Games. Traditionally it is implemented by chaining narrative blocks. This paper presents a proposal for a new approach – stories are built via removal of blocks. The internal structure is briefly discussed along with the main advantage of using this approach – the possibility to shape a story depending on the goals. The didactical modeling and the representation of competences are also described, as well as the possibility for automatic design of stories and automatic assessment of gained competencies.

**Keywords:** Storytelling, serious games, subtractive approach.

*“Io vedo l'Angelo intrappolato nella roccia, e scavo per liberarlo”*

*Michelangelo*

## 1 Storytelling Challenges

The traditional storyline writing is based on chaining card-block or narrative entities. Although this provides a systematic and consecutive approach to storyline building, it has intrinsic disadvantages, like a limited degree of freedom and exponentially growing complexity. These disadvantages are dormant in a traditional application of a storyline (e.g. in a book or a film), but emerge as a notable restriction in virtual game environments.

The fuzziness of the business world could be expressed with sufficient level of freedom. Unfortunately, the design of a serious game could be chaotic process [2] and the provision of additional freedom comes at the cost of losing control over the scenario [3]. Thus, some storytelling authoring tools rely on emergent narrative approach as *Façade* [4] and *Scenejo* [5]. Others, like [6], define manually chained scenes enriched by sets of freely combinable scenes. Often scenes are represented as atomic story units [3] or scripted scenario building blocks [7].

The rest of this paper discusses a proposal of how to carve a story instead of building it by stacking narrative building blocks – a story could be created not by adding entities, but by taking the whole graph of narrative building blocks and eliminating unwanted areas.

## 2 Story Sculpturing and Didactical Modeling

A serious game can be represented as *a game layer* – a graph which *nodes* are game states and *links* are state transitions. Once we have a complete game layer

we can remove surplus nodes and links. This removal process is what we call *sculpturing a story* – Fig 1 (left).

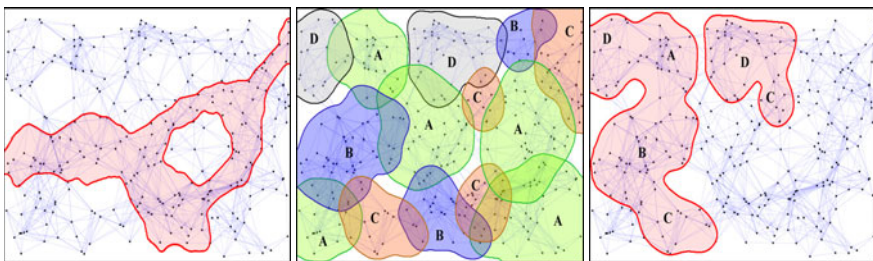
To support the business perspective of the game, the current model is enriched by a *competence layer*. It is a graph with its own *nodes* (i.e. *competences*) and *links* (i.e. *relations* between competences). Nodes of both layers are mapped in a many-to-many relationship – Fig. 1 (middle). The relationship guarantees that one game node could be related to several competence nodes. Additionally, one competence could be related to different nodes from the game layer.

The creation of a story by adding narrative building blocks is a tedious process. The story writer should carefully chain blocks balancing business complexities and story believability. Story ornaments should be done explicitly, thus adding more load to the story writer.

The sculpturing of a story is based on the opposite point of view. Instead of adding, the writer only disables areas that must not be a part of the story. It is possible to quickly disable large clusters of games nodes to shape the general story-line. Then, disabling individual nodes defines details and specific elements of the story similar to the DPE framework describe in [2].

Carving can be automated to the extent that stories are completely designed by software, provided there is an adequate description of the pedagogical target of the story and the *competence profile* of the player (this profile lists competences that the user has gained or need to gain). By selecting required competencies we can automatically select areas from the competence layer. When these areas are mapped down to the game layer we get a fully shaped story.

It is possible to create totally different stories focusing on the same set of initial and target competences – Fig. 1 (right). The *cognitive load* of each story can be automatically calculated by the difference of the competences in the story and the competences in the player's profile. A larger difference indicates bigger cognitive load.



**Fig. 1** A story sculptured from the full graph of the virtual environment (left) and a two-layered graph of states and competences A, B, C and D (middle). Two stories targeting a competence development path from competence C to D, personalized for players with different competence profiles (right).

### 3 Diversity, Perspectives and Assessment

One of the main benefits of the sculpturing approach is the variety of story shapes that can be carved. The shape of a story is intrinsically bound to the freedom of

choices in the game. Yet, it is possible to carve any shape between both extremes: from a linear story having a complete plot control to an unrestricted game environment of total freedom. The rest of this section describes just a few of the game shapes.

Fig. 2 (left) represents *the museum shape*, named after the way people visit museums – they walk through galleries in any way, but exit through a single gate. The museum shape gives the player a broad freedom of choices in the middle of the game. However, to complete it successfully, the player must reach a very specific preselected exit situation. If the story writer wants to force a specific action in the middle of the game, then *the hourglass shape* can be used – Fig. 2 (middle).

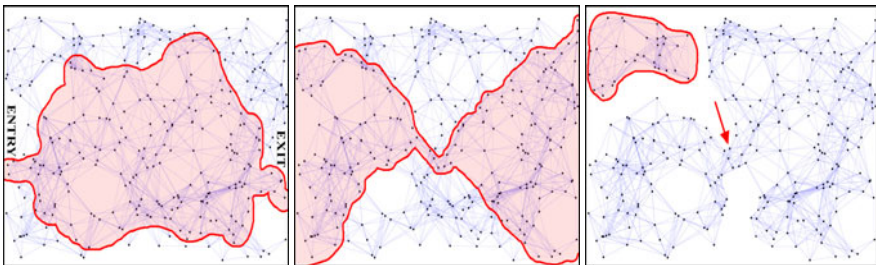
Serious games have two main aspects – being games and being serious. Both perspectives are equally important [1], [2]. The gaming perspective of a carved story is supported by several features which are intrinsic to the dual layer (game/competence) representation. Some of the most important features are:

- *Multidirectional* – any game node of a story could be initial or final.
- *Multiplayer* – a story can be co-played by several interacting players.
- *Dynamic players* – a player can “fall in” the middle of a story. This can be used in case of emergency, as a gateway for mentors to join the game.
- *Dynamic states* – the structure of a story as emergent graph can be changed during an active game, by adding or disabling nodes and links.

The business perspective is the second important component of a serious game. Actually, it is the only component that distinguishes a serious game from other games. A serious game must represent business processes adequately. If these processes are represented as a graph, then this graph could be converted into a game layer, where one cluster of nodes represent negotiating, another – contracting, and so on.

The main features of the business perspective are:

- *Closed business worlds* – it is conceptually acceptable if there are specialized game “islands”. When a player is placed in such an island, he can only play inside it – Fig. 2 (right).



**Fig. 2** A museum-like shape providing a specific exit point (left); a story shaped like an hourglass focusing on an internal situation (middle); and a story spanning over an isolated island of the game, unrelated to the rest of the world (right). The arrow shows a new bridge connecting two halves of the “mainland”

- *Opened business worlds* – initially closed worlds could be opened by building bridges to other worlds – see the arrow in Fig. 2 (right).
- *Role playing* – a player may take different roles (e.g. clerk or manager) which “come” with different sets of available resources and connections.

The goal of a business component of a game is to provide an attractive and unobtrusive way of gaining competencies. A crucial element of the successfulness of the playing of a serious game is the final assessment.

The possibility to visualize the graph of the story along with the actual path of the player (Fig. 3, left) can be projected back on to the competence layer. Thus we will have two mapping paths – the initial path of player’s experiences within the story; and the derived path of possibly gained competencies – Fig. 3 (middle).

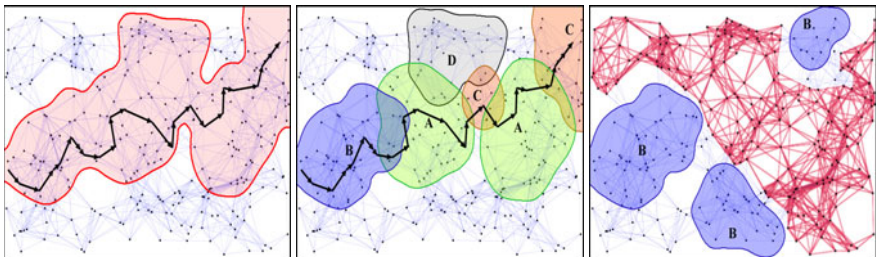
When a person plays several games, the collection of his game paths may indicate the competencies that are gained or avoided – see Fig. 3 (right). Most likely avoided competences represent *threshold concepts* for the player. The possibility to identify such concepts is important for the successful story writing and business education.

## 4 Implementation Challenges

The story sculpturing presented in this paper is a proposal for a new storytelling approach. There is no any functioning implementation yet, however, there are some initial considerations, which address the most evident implementation challenges:

- the volume of data,
- the low-level representation of nodes,
- automatic story definition.

The full graph of game nodes could easily become too big to manage in real time. Fortunately, the actual game play does not need the complete graph; it only needs to know the current nodes and the transitional rules to other nodes. However, story authoring and experience assessment tools may want to visualize maps of the whole graph. These tools are typically run off-line, so they do not need the computational power to process large data in real time.



**Fig. 3** A path mapped onto the story (left) and onto the competences layer (middle). Although competence D is present in the story, the player bypassed it. On the right: an aggregated map of player’s paths for a large set of played stories. Situations requiring competence B have been routinely avoided

The low level representation of the graph is crucial to the functionality of the system. A simple map of nodes and links might not be sufficient to support the game mechanics. Most likely nodes and links must be enriched with attributes. The decision of moving from one node to another depends on various factors, which must be determined and described. Some of these factors are related to the objective state of the game, others are related to the current emotional mood of the players.

Although the game layer is emergent from the nodes and the transition rules, the competence layer and its mapping to the game layer are done manually. Once created, the mapping can be used by the serious game tools to make automatic decisions, like the ones described in the didactical modeling and the cognitive (over)load analysis. Such automatic decisions do not require the invention of new algorithms, because these tasks can be reduced to standard algorithms for graph analysis and management.

In a game universe where nodes have a high degree of connectivity, local modifications may cascade globally. A typical example of this case is when the story writer wants to eliminate areas of the game layer by imposing resource restrictions. However, these restrictions may affect other areas and disrupt the intended shape of the story. Further research is needed to identify how resource restrictions affect the automatic storytelling.

## References

1. Zeller, D.: *Serious Games, Seminar on Computer Graphics*. Institute for Computer Graphics and Algorithms, Vienna Technical University (2007)
2. Winn, B.: *The Design, Play, and Experience Framework*. In: Ferdig, R.E. (ed.) *Handbook of Research on Effective Electronic Gaming in Education*, vol. III, pp. 1010–1024. Information Science Reference, Hershey (2009)
3. Göbel, S., Mehm, F., Radke, S., Steinmetz, R.: *80Days: Adaptive Digital Storytelling for Digital Educational Games*. In: Cao, Y., Hannemann, A., Manjón, B., Göbel, S., Hockemeyer, C., Stefanakis, E. (eds.) *Proceedings of the 2nd International Workshop on Story-Telling and educational Games*. CEUR Workshop Proceedings, vol. 498 (2009)
4. Mateas, M., Stern, A.: *Structuring content in the Façade interactive drama architecture*. In: Young, M., Laird, J. (eds.) *Proceedings of the Conference on Artificial Intelligence and Interactive Digital Entertainment*. AAAI, Menlo Park (2005)
5. Spierling, U., Weiß, S., Müller, W.: *Towards Accessible Authoring Tools for Interactive Storytelling*. In: Göbel, S., Malkewitz, R., Iurgel, I. (eds.) *TIDSE 2006*. LNCS, vol. 4326, pp. 169–180. Springer, Heidelberg (2006)
6. Mehm, F., Göbel, S., Radke, S., Steinmetz, R.: *Authoring Environment for Story-based Digital Educational Games*. In: *Proceedings of the 1st International Open Workshop on Intelligent Personalization and Adaptation in Digital Educational Games*, vol. 1, pp. 113–124 (2009)
7. Est, C., Poelman, R., Bidarra, R.: *High-level Scenario Editing for Serious Games*. In: Richard, P., Braz, J. (eds.) *Proceedings of the 6th International Conference on Computer Graphics Theory and Applications*. INSTICC Press, Algarve (2011)

# Ad-Hoc Radial Visualization of SCORM Objects with Extended Search and Keywords Handling

Ivan Angelov<sup>1</sup> and Ivan Ganchev<sup>2</sup>

<sup>1</sup> Faculty of Mathematics and Informatics,  
University of Plovdiv, Plovdiv, Bulgaria  
email: iangelov@uni-plovdiv.bg

<sup>2</sup> Department of Electronic and Computer Engineering,  
University of Limerick, Limerick, Ireland  
email: Ivan.Ganchev@ul.ie

**Abstract.** This paper presents a novel, mash-up web service amending the Sharable Content Object Reference Model (SCORM) objects' presentation. It is designed to be used on its own or as complementary visualization in Learning Management Systems (LMS) or other SCORM runtimes. As a proof-of-concept, a SpicyNodes-gearred radial visualization of the extracted semantic-, keyword-, and index structures is built by parsing the objects' internal content and activities organizations, Learning Object Metadata (LOM) notations, HTML's content, structuring tags, hyperlinks and metadata.

**Keywords:** e-learning, semantic, SCORM, LOM, LMS, LCMS, radial visualization, searching, mash-up service.

## 1 Introduction

Access to knowledge and information seems quite unlimited nowadays, and - although it might look like a positive thing at a glimpse- when it comes to knowledge and learning objects, it also comes with a good amount of attached challenges. One such challenge, for example is the loss of curiosity [1] and the need to learn new things due to the widely spread (though mostly wrong) opinion that everything worth knowing is already present somewhere in the Internet. This in turn leads to the wrong conclusion that knowing is downscaled to nothing more than simply having the location or whereabouts of a given piece of information or knowledge.

Add that in the computer era every piece of information, course, learning object, etc. needs an appropriate method for visualizing its structure and content digitally. And that this content and knowledge base is constantly growing and increasing exponentially. Initially, a simple hierarchical list, e.g. a table of contents as shown on the left-hand side in Fig.1, was employed to display this just fine for a limited number and levels of topics and subtopics. However, with the hierarchy of knowledge complexity expanding, that basic visualization proved more and more inappropriate [2].

At the same time, nowadays we witness the blooming variety of Learning Management Systems (LMS) and Learning Content Management Systems (LCMS)

along with a rapid improvement and development of their quality and growing number of provided services. Also we see an improved support for widely adopted standards [3], such as Sharable Content Object Reference Model (SCORM) [4] and Learning Object Metadata (LOM) [5], and faster implementation of their most recent additions and advancements, raising their adjacent version numbers.

During our ongoing research on the current state-of-the-art of such systems and their expansion modules, we reached an impression that too few or none such modules target the need for alternative methods of presentation and visualization of content. The particular needs for different kinds of learners [6] seem also left quite unaddressed. Attempts for alternative and better visualization would optimize the learners' interaction with such systems and improve their adoption, adaptation and accommodation, and most importantly, enrich the understanding of the internal connections within the content. They would also allow for a choice of the way used to present knowledge and course content, often delivered in a shape of a SCORM package.

Present research shows that there are such methods and means for increasing the usefulness and utilization of the visualization and interaction process of semantically structured knowledge and information representation compared to the widely adopted, in the current LMS realm, mostly static HTML content [7] (with attached "table of contents" in a form of simple, expandable linear tree). The gain is experienced especially stronger when such alternative methods are interchangeable and can be selected by the users themselves, e.g. when they are able to try looking at the content in different contexts. Besides the level of understanding, such methods are also able to better match the learners' preferences and their learning style, and to significantly improve the presentation of knowledge, thus raising the learners' interest and concentration, increasing their engagement in the learning process, as expected in [8], and fostering their creativity [9].

Based on conducted research and undertaken analysis on the matter, an advisable description, architecture, and implementation of one such alternative are presented here for further study, extension and evaluation. The result of these efforts is a generalized, value-added visualization service, which could add to the existing options available in the current releases of the widely adopted LMSs and LCMSs.

The rest of the paper is organized as follows. Section 2 briefly describes the developed visualization service prototype. Section 3 presents the selected third-party visualization method and the corresponding web-based service providing it. Section 4 outlines and discusses the implementation of the prototype. Finally, section 5 concludes the paper by summarizing the results and outcomes of using the prototype and by drawing up a roadmap for future research, improvements and extensions applicable to the described service.

## 2 Visualization Service Prototype

As an example on which to demonstrate our service, we use a freely distributable, sample SCORM package, available from the SCORM repository of the "SELF" project [10], namely "The not so short guide to Latex" [11]. Fig.1 depicts this SCORM object imported and presented in the ATutor 2.0.2 LCMS [12].

My Network  
My Contacts  
Network Profile  
Gadgets  
Network Groups  
Settings

Search People

intent Navigation  
course Home  
1. Thank you!  
2. Preface  
2.1. Contents  
2.2. List of Figures  
2.3. List of Tables  
3. Things You Need to Know  
4. Typesetting Text  
5. Typesetting Mathematical F...  
6. Specialities  
7. Indexing  
8. Fancy Headers  
9. The Verbatim Package  
10. Installing Extra Packages  
11. Working with pdfLaTeX  
12. Creating Presentations  
13. Producing Mathematical Gr...

## 1. Thank you!

Much of the material used in this introduction comes from an Austrian introduction to LaTeX 2.09 written in German by: Hubert Partl [partl@mail.boku.ac.at](mailto:partl@mail.boku.ac.at)

Zentraler Informatikdienst der Universität für Bodenkultur Wien

Irene Hyna [Irene.Hyna@bmf.ac.at](mailto:Irene.Hyna@bmf.ac.at)

Bundesministerium für Wissenschaft und Forschung Wien

Elisabeth Schlegl [es\\_email](mailto:es_email)

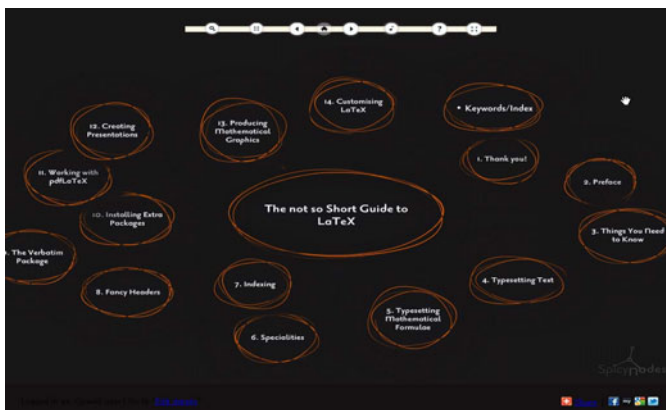
in Graz

If you are interested in the German document, you can find a version updated for LaTeX2<sub>ε</sub> by Jörg Knappen at [CTAN/tex-archive/info/short/german](http://CTAN.tex-archive.info/short/german)

The following individuals helped with corrections, suggestions and material to improve this paper. They put in a big effort to help me get this document into its present shape. I would like to sincerely thank all of them. Naturally, all the mistakes you'll find in this book are mine. TF

**Fig. 1** A sample SCORM object imported and presented in ATutor

The developed visualization service is flexible enough to be utilized as an extension module in LMSs and LCMSs (or as an external service co-existing in the browser window), or even as an independent, self-contained application (or layer) that simply uses and amends presented content and its associated metadata in SCORM packages. The service processes the HTML and textual content inside the package and generates a relevant, radial mapping played in a third-party web environment providing interactive browsing, clear and convenient notion of the keywords, and a handy option for searching the content, right in the learners' web browser as illustrated in Fig. 2.



**Fig. 2** An interactive radial representation of the sample SCORM package in SpicyNodes (using different styles)

The example can be experienced live in [13], while Fig. 3 demonstrates a detailed view of the content of the focus node skinned in yet another visual style.





Fig. 3 Basic content proved useful for visualization when the node is in focus

### 3 SpicyNodes

Our visualization service prototype employs the visual environment provided by the web-based service of SpicyNodes [14].

#### 3.1 Description

SpicyNodes is an innovative interface, built around the concept of radial maps, for interactively browsing and finding information, used mainly in problem domains like knowledge representation, site mapping, etc. Aiming at organizing online data by illustrating both the concepts and the relationships between them, it helps users intuitively search and understand complex semantic and organizational structures of information, presenting a virtual scene of the visual representation of that structures and their content. This promotes the exploration of larger context within which that information resides.

#### 3.2 Features and Tools

The most notable additional features and tools of SpicyNodes, which proved of great benefit to our solution, are described below:

- *Interaction* – allowing users to zoom, move or rotate the radial map involving intuitive, organic animation that employs scaling, panning and rotation based on physical models of motion.
- *Maintaining orientation and sense of history* – supported by the users' ability to know where they are, where they can go next, and which pages are related. As the user browses and explores the nodemap, the navigation system also keeps and displays intuitive notions of the browsing history and the current route to the home node. This, essentially, forms a logical connection of related contexts – from the most general context at the home node level to the currently narrower local context of information.

- *Indexing support and search function* – presented in-place by the visualization service as an interactive transparent overlay illustrated in Fig. 4. This nifty, hands-on addition is provided completely “under-the-hood” by the engine’s player.
- *Animation effects and interaction abilities* – allowing for defining various actions in reaction to the user-generated events like mouse clicks or changing the node in focus.
- *Exemptions to allow cross-nodes links* inside the content’s structure, where applicable and necessary, breaking the formal rules of the tree structure and shifting it towards graphs and networks. Despite how powerful this feature may look like, it should be exploited with caution. It could be extremely useful and beneficial in some cases, but when used unconscionable it could not only neglect the positive effect but also bring-in negative influence, causing syndromes like “lost in cyberspace”, loss of focus or loss of the notion of context.
- *Ability to alter the view and style of the radial presentation* chosen from a gallery of different available “skins” by the author or the publisher of the radial map. Such skins are provided by the SpicyNodes’s environment or by other users who decided to share their own settings and arts.
- *Ability to further fine-tune* the way and details of visual presentation and interaction, and to apply and utilize various sound and graphic attributes on the structured content to be presented.

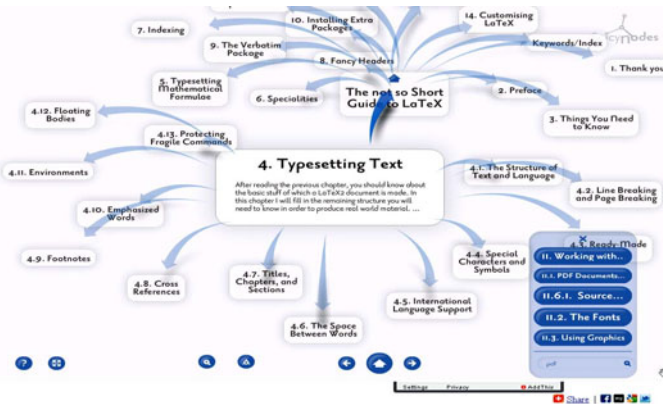


Fig. 4 An example of the overlay search results found for “pdf”

We also decided to append an additional “special” node composed by our service, which contains all the keywords found in the package, its manifest and the content itself. It is organized to enlist all the keywords encountered, acting as an interactive index of content (Fig. 5). If the user clicks on any of its sub-nodes, it executes a new web search for the attached keyword in Google, Wikipedia or other relevant search provider.

## 4 Implementation

The implementation of our own web service is based on the SpicyNodes, which was made available as a free web service. Not only this makes it just as much accessible to the extension modules' developers as well as to the web content authors and integrators, but it also offers opportunity to build an additional, highly usable and human-considering User Interface (UI) [15], which to provide an user-friendly description of the input document as an uploaded file or a Uniform Resource Locator (URL) to its web residence. Such wrapper UI would also offer visualization of the output XML document along with means for browsing, navigation and searching in it as an embedded SpicyNodes' nodemap player.

At the bottom of our visualization service lies an extended algorithmic transformation of input data, mostly defined as a combination of eXtensible Markup Language (XML), HTML, graphics and text content, packed accordingly to the SCORM standard in a single Zip-compressed file. The resultant document describes the noted structure of the input content and information by applying the particular syntax and validation rules of a SpicyNodes XML document, also known as the nodemap XML [16].



Fig. 5 The specially generated “Keywords” branch

Having in mind the fact that most LMSs and LCMSs are web-based and, even more, that all SCORM runtimes are required to execute on a web infrastructure by the standard itself, gives us the option to mash-up the result's visualization -or even overlay it over the window of the eLearning system presenting the SCORM object simultaneously- by employing only the standard HTML5 features and techniques. Of course, this algorithmic transformation could be easily configured or ported for off-line execution as well. This way the content authors would have full control over the intended use and presentation of the output by enabling its further alteration or conversion.

The base working structure gets parsed, extracted and built out of the `imsmanifest.xml` file, which is obligatory for each SCORM package. In particular, from the

defined inside organizations of resources (sometimes also referred to as the Activity trees), while a major part of the keywords' list is built from LOM metadata (when presented), the resources themselves and their content are extracted from the package and processed at a later stage. Once recognized as valid HTML- or text-based ones, they get further indexing and processing in an attempt to find and extract amending details or sub-structures of the content. The content itself is copied and reformatted at this stage, and keywords are assigned to it to get enlisted in the general list.

## 5 Conclusion and Future Work

This paper has presented a service prototype that could be used for alternative, radial visualization of learning content in eLearning (and other) systems. The developed service also allows further enriching and amending of the tutoring and learning content, whenever the content author finds it useful or appropriate to allow so, without all the usually attached overhead and demanding excess efforts (at least not making these obligatory).

Further research directions for improving and enhancing the presented service prototype are listed below:

- Better and more generalized parsing of *imsmanifest.xml*, and especially the HTML content, due to its not strongly standardized structure.
- Improved support for embedded images, audio notes and Flash objects.
- Wire up the navigation by using cross-scripting, "javascript:" calls or "http redirects" (i.e. link the keyword or index entry to the resource playing in the LMS browser window).
- Add support for sequencing rules as described in the SCORM 2004 standard [4].
- Further analysis and processing of keywords and indexing stats that may lead to observations and generation of cross-tree references and content links.

Although some research has already been conducted as regards the pros and cons of employing radial visualization in different applications and content topics [17, 18], a premeditated evaluation of the improvements this approach contributes to the users' learning experience and to the learnability of the knowledge content would greatly benefit the reasoning behind further promoting of the solution to the general public. Being an alternative tool of choice, we don't expect any drawbacks of using it, because this presentation peacefully co-exists with the tools already provided by a LMS or its authoring tool.

**Acknowledgements.** This work was partially supported by the Plovdiv University under the Research Projects Ref. No. NI11-FMI-004/2011

## References

1. Dreyfus, H.: Anonymity versus commitment: The dangers of education on the Internet. *Educational Philosophy and Theory* 34, 369–378 (2009)
2. Turetken, O., Sharda, R.: Visualization of web spaces: state of the art and future directions. In: *SIGMIS Database*, vol. 3, July 2007, pp. 51–81 (2007)

3. Ganchev, M., O'Droma, Andreev, R.: Functionality and SCORM-compliance Evaluation of eLearning Tools. In: Proc. 7th IEEE Int. Conference on Advanced Learning Technologies (ICALT 2007), June 2007, pp. 467–469 (2007) ISBN: 0-7695-2916-X
4. Advanced Distributed Learning, SCORM's standardization body's page, <http://www.adlnet.gov/Technologies/scorm/> (accessed: March 17, 2010)
5. IEEE, Learning Object Metadata Standard (IEEE LTSC 1484.12.1), <http://ltsc.ieee.org/wg12/par1484-12-1.html> (accessed: March 17, 2010)
6. Sabry, K., Baldwin, L.: Web-based learning interaction and learning styles. *British Journal of Educational Technology* 34, 443–454 (2003)
7. Draper, G., Livnat, Y., Riesenfeld, R.: A Survey of Radial Methods for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 15, 759–776 (2009)
8. Douma, M., Ligierko, G., Angelov, I.: Promoting Creative Engagement with Spicy Nodes. In: Proc. AMEE 2008, vol. 1067, pp. 196–208 (2008) ISBN: 978-0-7354-0598-1
9. Hewett, T.: Informing the design of computer-based environments to support creativity. *Int. J. Hum.-Comput. Stud.* 63, 383–409 (2005)
10. Project “SELF”, Short description of the SELF project, <http://www.selfproject.eu/en/project/> (accessed: March 17, 2010)
11. Project “SELF”, URL of the downloadable, sample SCORM package, [http://www.selfproject.eu/files/scorm\\_packages/Office\\_tools/The\\_not\\_so\\_Short\\_Guide\\_to\\_LaTeX\\_ims.zip](http://www.selfproject.eu/files/scorm_packages/Office_tools/The_not_so_Short_Guide_to_LaTeX_ims.zip) (accessed: March 17, 2010)
12. ATutor, Learning Management System, <http://atutor.ca/atutor/> (accessed: March 17, 2010)
13. Live Sample, Hosted resultant document, <http://www.spicynodes.org/a/920b50540394b1ada9f54b5d37d2fd28> (accessed: March 17, 2010)
14. Spicy Nodes, Main service page, <http://www.spicynodes.org/> (accessed: March 17, 2010)
15. Rahnev, S.M.: Principles and technologies for building user interfaces for web and desktop applications. In: Proc. Education in the Information Society (ARIO 2010), May 2010, pp. 308–316 (2010) ISSN: 1314-0752
16. Douma, M., Ligierko, G., Ancuta, O., Gritsai, P., Liu, S.: Spicy Nodes: Radial Layout Authoring for the General Public. *IEEE Transactions on Visualization and Computer Graphics* 15, 1089–1096 (2009)
17. Angelov, I., Menon, S., Douma, M.: Finding Information: Factors that Improve Online Experiences. In: Yuen, S., Yang, H. (eds.) *Handbook of Research on Practices and Outcomes in E-Learning: Issues and Trends*, pp. pp. 493–506. IGI Publishing, ISBN: 978-1-6056-6788-1
18. Angelov, I.: An evaluation of radial mapping's usability for navigation and visualization of structured information and knowledge. In: Proc. Education in the Information Society (ARIO 2010), pp. pp. 298–307 (May 2010) ISSN: 1314-0752

# Applied Cognitive Task Analysis in the Context of Serious Games Development

Albena Antonova and Krassen Stefanov

Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

email: a\_antonova@fmi.uni-sofia.bg, krassen@fmi.uni-sofia.bg

**Abstract.** Sophisticated Technology Enhanced Learning (TEL) instruments as training simulations and Serious Games (SG) do not provide knowledge content in an explicit form, but propose interactive solutions for learners to build their own skills and competences in close to real situations. Thus, expert knowledge should be implicitly embedded into design of SG and training simulations. The present paper discusses the problem of expert knowledge elicitation for building SG, and provides an analysis of the use of cognitive task analysis methods for SG. The use of Applied Cognitive Task Analysis (ACTA) as effective and time-efficient tool for codification of complex expert skills in the context of serious games and complex simulations is discussed, on the base of experience gained in the TARGET project.

**Keywords:** Serious games, Knowledge elicitation, expert knowledge codification.

## 1 Introduction

While increasingly popular in recent years, serious games and training simulations still represent a challenging learning environment. In traditional TEL systems expert knowledge is usually presented in explicit form and subject matter experts (SME) prepare and provide their own learning materials as text, audio or video content. Thus designers of TEL environments do not need to codify specific expert knowledge, as TEL role is to facilitate delivery and building of rich knowledge ecosystems around that learning content. This approach is not possible in serious games and training simulations.

The logic of serious games and training simulations is to develop complex scenarios, where learners can build skills coping with number of challenging situations. In serious games the Kolb's cycle for knowledge acquisition [1] is adopted, where learning is developed through number of trial-and-error situations. Building successful SG include synchronization of multiple elements (game mechanics, appealing graphic environment, engaging scenarios), and therefore achieving good mix of learning elements is very difficult. Moreover, expert knowledge should be incorporated in game scenarios and learning path. So expert knowledge is crucial to make learning simulations useful and meaningful for learners, and to put them in situation where they can substantially build new skills.

Using expert knowledge in the design and development of TEL in general, as well as in the case of design of SG and training simulations, is very important. As noticed in [2-3], learners who receive explanations from experts perform better on knowledge transfer tasks than learners who received explanations from non-experts. Different research evidences prove that accurate identification of experts' cognitive processes can be adapted into training materials that are substantially more effective than those developed through other means [2-3].

Two main problems with expert knowledge elicitation for TEL and serious games development can be observed. The first problem consists of proper identification of expertise and complex cognitive processes of experts, because subject matter experts cannot easily externalize automated tacit knowledge. This is due to complex cognitive models for decision making developed with experience and practice.

The second problem refers to identification of suitable models for capturing and codification of expert knowledge in a way, appropriate for further use in SG design. The expert knowledge has to be transmitted via different game elements, as critical situations, game scenario, game tasks, communication with non-playing characters (NPCs) and others.

The present research aims to review the ACTA model for expert knowledge elicitation, and to discuss its application for design of serious games and training simulations. It provides a short review of serious games, skills, and general cognitive task analysis methods. Then ACTA methodology is presented as suitable tool for capturing and presenting expert knowledge for serious games design.

In the second part of the paper we present the case of TARGET project for building serious games. We discuss how ACTA methodology was applied on practice and how enhanced the process of creating game mechanics and game scenarios in the case of Sustainable manufacturing game scenario. Advantages and limitation of applying ACTA in the context of serious games are outlined. The last section provides some general conclusions and directions for further work.

## **2 Theoretical Background**

### **2.1 Serious Games for Learning**

Serious games often overlap and extend the terms e-learning, edutainment (education and entertainment) and game-based learning [4]. Although slight variances among different authors, serious games are commonly described as (digital) games used for purposes other than mere entertainment or fun [5]. Serious games usually refer to games used for training, simulation, or education that are designed to run on personal computers or video game consoles. Thus, serious games transfer positive experiences of building skills and competences while entertaining and playing on computer games to apply it in more complex context and purpose-oriented learning.

The common elements of SG include: back story (plot/ story line), game mechanics (physical functions/actions), rules (constraints), immersive environment (including 2D/3D, animations), interactivity (impact of player's actions), and challenge/competition (against the game or against other players). In

SG players have to perform a set of actions and take different decisions, following preliminary defined rules and constraints. Usually players receive instructions and feedback on their performance and are virtually assisted with additional learning materials.

## **2.2 Cognitive Skills**

The overview of skills provided in [6], classify several approaches for conceptualizing skills, from manual or motor skills to cognitive skills, perceptual skills, response selection skills, and problem-solving skills [7]. Most schemes for categorizing skills are hierarchical, starting with the simpler form of skill and ending with the most complex. Welford [8] defines skills as “combination of factors resulting in competent, expert, rapid and accurate performance, equally applicable to manual operations and mental activities”. Proctor and Dutta [7] define skill as “goal-directed, well-organized behavior that is acquired through practice and performed with economy of effort”. Seamster et al. [9] propose a framework for hierarchy of skills, ranking on first place strategic skills, then decision-making skills, representational skills, procedural skills, and automated skills. Automated skills are subconscious and are characterized with rapid execution and economy of effort.

A “cognitive skill” is a skill that is predominantly cognitive in nature. All skills have a perceptual, motor, and cognitive component [7], but cognitive skills form the basis for training because they can be trained in relatively short period of time. The recognition that some skills have a predominant cognitive component, and the use of cognitive methods to analyze these skills, allows the application of meaningful Cognitive Task Analysis (CTA) methods [9].

## **2.3 Cognitive Task Analysis and ACTA Methodology**

Task analysis represents a methodology for describing the physical tasks and cognitive plans required of a user to accomplish a particular goal. Traditional task analysis segments a job into distinct behavior tasks and their component activities. All forms of task analysis rely on the idea that human action can be decomposed, and that the decomposition can be used to reason about what people should do and know to complete a task.

The CTA is a set of methods designed to elicit information about the knowledge thought processes, and goal structures that underlie observable performance. (CTA is used to elicit and represent knowledge and information about thought processes in a systematic way [10]. CTA describes and represents cognitive elements underlying goal generation, decision making, judgments and others. In CTA skills are analyzed in substantially more details based on their cognitive components. CTA uses a variety of interview and observation strategies to capture a description of the knowledge that experts use to perform complex tasks. Complex tasks are defined as those for which their performance requires the integrated use of both controlled (conscious, conceptual) and automated (unconscious, procedural, or strategic) knowledge to perform tasks. It is a valuable



approach when advanced experts are available who reliably achieve a desired performance standard on target tasks [2].

Researchers have identified over 100 types of CTA methods currently in use [11]. The number and variety of CTA methods are due primarily to the diverse paths that the development of CTA has taken, including behavioral task analysis, computer system interfaces, and military applications. Applications of CTA include system design, training design, human-computer interface design, accident investigation and the development of tests to assess competence [10], [12]. CTA methods have been applied within a wide range of domains including aviation, nuclear power plant operation, consumer behavior, air traffic control and military operations, and consumer research [12]. One of the more extensive reviews of CTA [11] identified three broad families of techniques: (1) observation and interviews, (2) process tracing, and (3) conceptual techniques. Observations and interviews involve watching experts and talking with them. Process tracing techniques typically capture an expert's performance of a specific task via either a think-aloud protocol or subsequent recall. In contrast, conceptual techniques produce structured, interrelated representations of relevant concepts within a domain.

The CTA method called Applied Cognitive Task Analysis (ACTA) can be selected as appropriate technique for knowledge elicitation and codification in SG [13]. Compared to traditional CTA techniques, ACTA methodology requires considerably less training for application, less time and resources [13]. ACTA is easy to use, flexible method that don't require interviewers to be experts in the knowledge domain. Moreover, ACTA methodology is suitable for job domains where observational data are difficult to obtain, and can be used for identification and codification of complex skills on the workplace.

ACTA provides three interview protocols: the task diagram interview, the knowledge audit, and the simulation interview. The task diagram interview elicits information about the task structure within a particular task domain (e.g. the main tasks and sub-tasks), and helps to identify which of these task components are typically experienced as challenging or difficult. The knowledge audit and simulation interviews generally focus on the more difficult/challenging components and elicit more detailed information about the underlying knowledge, thought processes and goal structures. The main output of the ACTA method is the Cognitive Demands Table. This framework includes information about why each element is often found to be difficult, identifies common pitfalls/errors incurred by novices, and identifies cues and strategies that experts use to overcome the difficulties [13].

### **3 Application of ACTA Methodology in TARGET**

#### **3.1 Approaches for SG Design in TARGET**

The main aim of the TARGET Project is to research, analyze, and develop a new genre of Technology Enhanced Learning (TEL) environment that supports rapid competence development of individuals, namely knowledge workers within the

complex domains of project management, innovations and sustainable global manufacturing. In TARGET, the learner is presented with complex situations in the form of game scenarios, and via interacting with the game results into enriched experiences that are gradually leading to knowledge and complex skills acquisition.

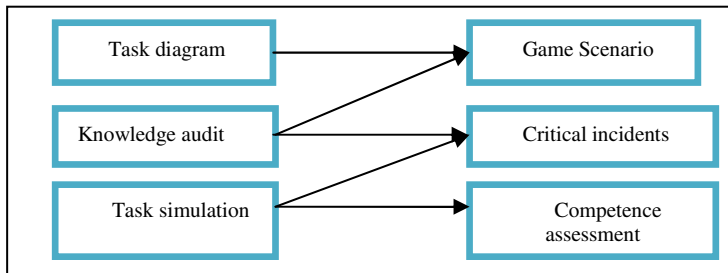
Developing serious games and training simulations require general understanding of the knowledge domain. Moreover, knowledge domain need to be further specified concerning skills that will be trained, general objectives for game application and target users. While in traditional game development, the whole phase of game design is described as creative process, in TARGET approach, game design is segmented into three stages: knowledge elicitation, knowledge representation and game design [14].

The knowledge elicitation phase objectives are to clarify cognitive models of experts, leading to high performance in tasks execution. Thus interviews with subject-matter experts establish the broad high-level structure for decision model competences, concerning specific knowledge domains. The knowledge representation phase includes codification of data in a way appropriate for incorporating in game design. The game design phase consists of scenario building, identification of critical incidents and learning situations, and description of non-playing characters and their role in the process. The game design phase is the most complex phase as it involves as well conception, design and development of software environment and other game elements.

### 3.2 Methodology for Application of ACTA for SG

The use of the ACTA method produced valuable data for the initial phase of SG and the goal is to learn about the task, the cognitive challenges associated with task performance.

The outputs of the first component of the ACTA interview is a task diagram interview, that provides useful insights of most challenging cognitive tasks and subtasks. The Knowledge Audit phase aims to identify how the expertise is used in the application domain and provides cognitive difficult elements, why difficult, potential errors and cues& strategies. They can provide useful information about critical incidents and learning situation in game dynamics.



**Fig. 1** Application of ACTA methodology to serious games design

The third step is a task simulation, where the SMEs are asked to imagine particular organizational role and to describe how they would think and act in this situation. They will identify the key events, the possible actions, the models to assess situation, and potential errors. Very important is analysis of critical cues of the event and potential errors, that novice can make. This information is summarized in a table that can be used in SG design, skills performance levels and potential errors identification.

### **3.3 Application of ACTA and Discussion of the Results in the Sustainable Global Manufacturing Knowledge Domain**

In order to identify specific competence set in the field of sustainable global manufacturing there were organized several semi-structured interviews with SME experts. The interviews included sections for critical incidents technique and job analysis. In total there were performed 14 interviews with SMEs from business and academia in 5 countries – Germany, Italy, Poland, Bulgaria and Slovenia. The outcomes were analyzed and allowed project partners to identify general level of understanding of the domain. There was realized that subject matter experts are mainly focused on their specific experiences and context and rarely provided summarized information that can be directly used in game development. Moreover, the information obtained from SMEs about critical incidents was hardly comparable in scope and importance. Therefore, much of this information remained unused on practice.

On a second stage, it was used the ACTA methodology to elicit expert knowledge for development of scenario for serious game in the field of Sustainable global manufacturing. Therefore 5 narrow-domain experts were identified in 4 of the countries, and interviews were performed using structured ACTA templates and tables. As result, there was quickly produced a general framework, allowing project partners to structure the game process and the game flow, to identify difficult cognitive elements and potential errors, and finally – to identify possible game paths. Proper identification of tasks contributed for better structuring of the game scenario. An emphasis is made on cognitive difficult elements and potential errors. As final output all collected structured data is accumulated and stored in unified tables, allowing project partners to have access in later stages of game development.

The application of ACTA facilitated TARGET team to [14]:

- Generate a scoped task model of the domain.
- Identify task elements that novice learners often find particularly challenging.
- Generate information about the knowledge, thought process and goal structures that underlie observable task performance in the domain.

Therefore, the application of ACTA method for knowledge elicitation in TARGET was successful, as it enabled SG designers to approach SME without deep knowledge in subject domain. It was easy to use, fast for application and not specific training was needed to apply it on practice. Moreover, the obtained results are comparable, storable and can be easily transferred to SG design specifics.

### 3.4 Limitations for Application of ACTA Methodology in SG Design

ACTA can be easily used for knowledge elicitation in the case of SG and TEL, because it enables SG designers to approach SME without deep knowledge in subject domain. However, when applying ACTA on practice, SG designers should take in consideration several limitations. ACTA methodology prioritize knowledge gained through first-hand experience and there should be identified SMEs with practical experience. Thus SMEs with theoretical knowledge (for example Lecturers) could not provide useful results as they lack practical insights for task execution. Another limitation of the model is that it requires decomposition of expertise on structured task processes. This could be difficult in complex and broad areas as sustainable global manufacturing and innovations. Thus before applying ACTA to scope knowledge domain, there should be identified basic task structure and working processes. This can be used in defining the game scenario further.

## 4 Conclusion and Future Work

ACTA methodology can be successfully used for design of SG and training simulations. The benefits of ACTA is based on compelling evidence that experts are not fully aware of about 70% of their own decisions and mental analysis of tasks and so are unable to explain them fully even when they intend to support the design of training, assessment, job aids, or work [2]. ACTA methods attempt to overcome this problem by specifying interview strategies that permit SG designers to capture more accurate and complete descriptions of how experts succeed at complex tasks. It enables SG designers to identify critical phases of task execution and to capture SGM assessment of cognitive demanding sub-tasks, where will be concentrated most of the learning processes and challenges in Game scenario. Further work can identify more strategies for identification of scenarios for SG design. The ACTA methodology will be further applied in other serious games development and further analysis and observations will be made.

**Acknowledgments.** The present work is supported by the TARGET project, funded by the European Community under the Seventh Framework Programme (Grant Agreement N° 231717) <http://www.reachyourtarget.org/main/>.

## References

1. Antonova, A., Todorova, K.: Serious Games and Virtual Worlds for high-level learning experiences. In: Second S3T Conference, pp. 250–254 (2010), ISBN 978-954-9526-71-4
2. Clark, R.E., Feldon, D., van Merriënboer, J., Yates, K., Early, S.: Cognitive task analysis. In: Spector, J.M., Merrill, M.D., van Merriënboer, J.J.G., Driscoll, M.P. (eds.) Handbook of research on educational communications and technology, 3rd edn., pp. 577–593. Lawrence Erlbaum Associates, Mahwah (2007)

3. Hinds, P.J., Patterson, M., Pfeffer, J.: Bothered by abstraction: the effect of expertise on knowledge transfer and subsequent novice performance. *J. Appl. Psychol.* 86, 1232–1243 (2001)
4. Antonova, A., Martinov, M.: Serious Games and Virtual Worlds in education and business. In: SAI Conference, Sofia, Bulgaria (October 2010)
5. De Freitas, S.: Serious Virtual Worlds. A scoping study prepared for the JISC e-Learning Programme (2008)
6. Winterton, Delamare, Stringfellow: Typology of knowledge skills and competences: clarification of concept and prototype. Research report, final draft, CEDEFOP Project, No RP/B/BS/Credit Transfer/005/04 (2005)
7. Proctor, R.W., Dutta, A.: Acquisition and transfer of response-selection skill. In: Healy, A.F., Bourne Jr., L.E. (eds.) *Learning and memory of knowledge and skills: Durability and specificity*, pp. 300–319. Sage, Thousand Oaks (1995)
8. Welford, A.T.: *Fundamentals of Skill*. Methuen, London (1968)
9. Seamster, T.L., Redding, R.E., Kaempf, G.L.: *Applied Cognitive Task Analysis in Aviation*. Ashgate Publishing Limited, Aldershot (1997)
10. Schraagen, J.M.C., Chipman, S.E., Shalin, V.L. (eds.): *Cognitive task analysis*, ch. 1. Lawrence Erlbaum Associates, Mahwah (2000)
11. Cooke, N.J.: Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies* 41(6), 801–849 (1994)
12. Baber, C., et al.: *Cognitive Task Analysis: Current Use and Practice in the UK Armed Forces and Elsewhere*. Research report HFIDTC/WP2.3.1/1 (2005)
13. Militello, L.G., Hutton, R.J.B.: Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. *J. Ergonomics* 41(11), 1618–1641 (1998)
14. Seager, W., Ruskov, M., Sasse, A., Fradinho, M.: Eliciting and Modelling Expertise for Serious Game Design. In: Derby Workshop on SG (2010)

# Using Robotics to Achieve Meaningfully Engaged Learning

Rebecca Caldwell and Elva Jones

Department of Computer Science, Winston-Salem State University, Winston-Salem, NC  
email: {caldwellr, jonese}@wssu.edu

**Abstract.** There has been a significant decline in the number of college students choosing majors in computer science or technology related fields. Additionally, within the United States, there is an achievement gap between under-represented minority students and majority students at a time when underrepresented groups are becoming an increasing proportion of the national labor force. This reluctance to study Science, Technology, Engineering, and Mathematics (STEM) disciplines must be confronted and changed if the United States is to maintain a competitive position within the global market. Effective use of learning technologies is vital to solving many of our current STEM learning challenges. We use robotics as a technology tool to captivate and engage students in learning computer science concepts. Robotics engages multiple modes of learning, including: sensory, perceptual, and cognitive information processing.

## 1 Introduction

Robotics is being used as a tool to revitalize interest in the Computer Science major. Robotics, as a motivational tool, is also a growing research area in Computer Science Education. Used in the computer science classroom to both teach hardware and software concepts, robots are also being used to attract students to the computer science discipline. Why use robotics? Robotics systems are powerful, engaging and affordable. The use of robotics in an educational setting offers the student multiple modes of learning. Not all students are able to understand and retain information through the traditional lecture style. Robotics tasks incorporate a range of learning, including: cognitive (knowledge), affective (attitude), and psychomotor (skills). A benefit of using robotics as an instructional tool is the development of effective learning strategies such as time management, motivation, concentration, positive attitude, and comprehension, information processing and self-testing. Robotics enhanced instruction is presently used across the control to successfully teach computing concepts in higher education. This paper discusses how we use robotics as a teaching tool to captivate and engage students in learning computer science concepts. We define engaged learning as students actively participating in their learning. Oblinger [9] proposes that several characteristics of today's university students to consider in designing new learning spaces for them including: a penchant for highly active and participatory experiences both face-to-face and digitally and often at the simultaneously; technological adeptness and ubiquity, using mobile

phones, digital cameras, MP3 players, and wireless Internet to browse, download, and message. These students often have multiple priorities, including school, work, sports, and volunteer activities.

## **2 Learning Strategies**

Learning strategies are methods employed by students to assist them in understanding material and problem solving. Successful implementation of learning strategies may involve changes in mode of instruction and course management. Weinstein and Mayer (1986) defined learning strategies as behaviors and thoughts that a learner engages in during learning intended to influence how the learner processes information. With effective learning strategies, students can learn faster and easier. Active and cooperative strategies require students to actively participate in the learning process. Research shows that these are the learning styles that are most beneficial to students from underrepresented groups. These learning strategies have proven to increase student retention, enhance student learning, and improve student academic achievement. Active and cooperative learning strategies provide mutually beneficial learning environments for students with different learning styles. [5, 6]

Effective use of learning technologies is vital to solving many of our current STEM learning challenges. Robotics is a growing research area in computer science education. We use robotics as a technology tool to captivate and engage students in learning computer science concepts. Robotics engages multiple modes of learning, including: sensory, perceptual, and cognitive information processing.[2, 3, 4]

## **3 Integrating Robotics in a Computer Software Systems Course**

### **3.1 Course Implementation**

Assessment of the sophomore level Introduction to Computer Software System course revealed that students who had difficulty mastering the course had a difficult time learning to adapt to the abstract nature of Assembly Language programming. Concepts such as structuring code, reusability, easy of maintenance, meaningful naming and user interface design need to be considered for all but the simplest programs. Visual learners often struggle with the abstract concept of Assembly Language programming. In an attempt to address the retention of students at the sophomore level, educational robotics was implemented as a unit in the Introduction to Computer Software System Course during the fall semester 2006. Study participants were all African American students, with a majority being male.

### **3.2 Sumobot Construction and Laboratory Assignments**

Students are placed in groups of 4 students per team to build the robots. Two Parallax sumobots are assigned to each team. The Sumobot-Mini Sumo Robots Assembly Documentation and Programming Manual provides step-by-step instructions to build the sumobot. It normally takes the students two lab sessions

to fully assemble the sumobots. Most students who enroll in this course have never built a computer system so therefore they are intrigued from the start. Students are excited when they are able to view their finished sumobots. We observed that at first the female students were hesitant to assist in building the sumobots. Generally, after the first lab session, they tend to become actively involved.

After assembling the sumobot, motion must be controlled. The sumobot motion is controlled by using two continuous rotation servo motors. Motion control introduces systems programming and mathematical computation of voltage. Students see a relationship between computer science, mathematics, and physics.

After downloading motion control software, students test the sumobot for essential motion control. The students cheer loudly when they see the sumobots move for the first time - an indication of the captivation associated with robot programming. If any of the students experience difficulty, it is not the instructor that assist them, but the other students. After the students get the motors aligned and sumobot moving they are given a group programming assignment. The students are required to view the code and listen while it is explained by the instructor. After the explanation of the code students are required to perform the following tasks as a group assignment. Students are strongly encouraged to only work with their three other partners. The code is modified to do the following:

- 1) Move straight at low and how speeds by changing the motor speed constants
- 2) Move the sumobot to turn 30 degrees, 45 degrees, and 90 degrees by finding the proper loop control
- 3) Move the sumobot in three geometric patterns – square, triangle, figure-8

Students who experienced difficulty programming in the CS1 and CS2 courses are amazed when they are able to complete the first programming assignment. They realize that they did internalize some of the concepts presented in these courses.

### **3.3 Final Project—Blending the Psychomotor, Affective, and Cognitive**

The psychomotor domain includes actions that are neuromuscular in nature and demand certain levels of physical dexterity. The affective domain is hierarchical with higher levels being more complex and depending upon mastery of the lower levels. Factors such as motivation, attitudes, perception, and values are included in the affective domain. As tasks progress to greater complexity, students become more involved, committed, and self-reliant. Cognitive domain progresses through six levels of complexity: knowledge, comprehension, application, analysis, synthesis, and evaluation. The higher the level, presumably the more complex mental operation is required.

The final project for the robotic unit requires the students to prepare for the sumobot competition. This competition requires students to use psychomotor skills to modify the robot and connect the line sensors and download the program to test and evaluate the QTI sensors prior to the competition. Affective development is motivated by the challenge of the completion, attitudes of impending success in the completion, perceptions of team capability and individual confidence, and the values of adhering to the rules of the competition.



Cognitive development is exhibited through knowledge of the robot construct, increasingly improved and creative programming skills, and complex problem solving skills in determining robot opponent avoidance algorithms. Students quickly realize the importance of efficient code in detecting the borders of the ring. The sumobot is required to stay within the borders and detect the opponent, and push the opponent out of the ring to successfully win the competition.

Usually the sumobot competition is the last assignment for the sumobot unit. However, Fall 2010, one of the female students suggested an additional project be included called “Dancing with the Sumobots”. The instructor agreed and the teams started programming their sumobots and producing their videos. The students worked hard and challenged their peers. The students’ results for “Dancing with the Robots” were excellent. All groups successfully completed the assignment within the short time limit. The assignments were blindly judged and the average score was 92/100.

## 4 Course Observations and Reflections

The students tackle the assembly language without moaning and groaning about its difficulty. The students understand the problem solving and programming process much better after the sumobot unit. The students evaluations for the class are always in the range of very high (4.6 or above out of 5). All the students reflected that they enjoy building the robots. The end of course grades are mostly above average (B) to excellent (A). The attendance for the class is good. Most students only miss class because of other required university activities. Some comments from the students are listed below:

- *“I like these sumobots and I learned more about programming.”*
- *“I feel better about my major now. Are they any more classes like this?”*
- *“Programming actually made sense. I saw my program in action “*
- *“I enjoyed working with my team. We were able to explore with the sumobot”*
- *“I loved the hands on”*
- *“I had a great team. They helped me understand!”*
- *“We need more courses like this one in the Computer Science department. I learn better with hands on.”*

### 4.1 Informal Analysis of Robotics Learning Strategies

After the final programming project of the semester, students were given a survey to complete. In this section, we present the questions and a summary of the students’ responses. The questions and a summary of responses appear in Table 1. The results of the survey were consistent with the student reflections. Both the survey and the reflections show that students were positive about programming.

The positive results of this unit in a course that often triggered the departure of students from the major, convinced the faculty that using engaged and cooperative learning strategies was beneficial. Approval was granted to develop an introductory level course in robotics.

**Table 1** Summary of Student Responses to Survey

QUESTION	SA 4	A 3	DA 2	SDA 1	AVG
Working with the robot has changed my perception of programming from unfavorable to favorable?	20 51%	18 46%	1 3%	0	3.4
I enjoyed this class more than the other computer science classes that I have taken.	39 100%	0	0	0	4.0
The robotics unit helped me to see the relevance of mathematics to computer science.	39 100%	0	0	0	4.0
I felt better working with a team than individually.	25 64%	12 31%	2 5%	0	3.5
I enjoyed building the robot.	39 100%	0	0	0	4.0
I prefer courses which include hands-on-experience	20 51%	19 49%	0	0	3.5
I would enroll in a robotics course if one was offered.	38 97%	0	1 3%	0	3.9

Legend: SA-Strongly Agree, A-Agree, DA-Disagree, SDA-Strongly Disagree, AVG-Average

The Introduction to Robotics course offers a hands-on introduction to robotics, relying on the use of the iRobot Create. Topics covered in the course include a C++ review, the basic concepts in robotics, such as sensors, actuators, and describes the most important approaches to robot control. The textbook for the course is *The Robotics Primer* by Maja J Matarić. The laboratory projects include: Introduction to Player Stage, Explore Create Tutorial, Maneuvering a Maze, Motors – Driving Base, Sensing – Touch, Light, Sound, Ultrasonics, and Introduction to Tekkotsu.

## 5 Conclusion

The use of the Parallax Sumobots in the Introduction to Computer Software Systems course changed the atmosphere of the course from being boring assembly language programming to programming in action. Working with the robots requires the students to utilize problem solving, programming, and critical thinking skills. Building and programming the sumobots enhances their knowledge of hardware and software concepts. Using robots in the computer science classroom reinforced student involvement and participation in their course of study. Integrating robotics into the computer science curriculum introduces the students to a hands-on experience which they are not likely to forget. The hands-on programming experience changed some students' outlook on programming and motivated some of them who were considering changing their majors to continue in the computer science area.

The use of robotics in the undergraduate curriculum has proven to be instrumental in meaningfully engaging students and motivating them to achieve. In order to continue to attract under-represented students to the STEM disciplines there must be an atmosphere which consider their learning styles and nurture them. STEM educators need to continue to explore strategies that best fit the needs of this increasingly diverse student population. The authors have observed in working with diverse populations that students from under-represented groups majoring in computer tend to learn and retain information more when they are actively engaged. Future work will included a controlled two-sample study of the effectiveness of engaged learning, and the effects of engaged leaning on developing specific computing concepts such as recursion, objects, and polymorphism.

## References

1. Bogdan, D., Sambit, B.: Programming robots with state diagrams. *J. Computing Sciences in Colleges* (2009)
2. Boonthum, C., Touretzky, D., Jones, E., Humphries, T., Caldwell, R.: The ARTSI Alliance: Using Robotics and AI to Recruit African-Americans to Computer Science Research. In: *Proceedings of FLAIRS-24*, Palm Beach, FL (2011)
3. Boonthum, C.: Robotics Introductory with NXT at Hampton University. In: *The Proceedings of ADMI 2009*, Baltimore, MD (2009)
4. Burhans, D.: A versatile tool for student projects; an ASM programming language for the Lego Mindstorm. In: *2007 AAAI Spring Symposium* (2007)
5. Caldwell, E.R.: Beyond Wrestling: Using Sumobots in the Computer Science Classroom at Clemson University. In: *Proceedings of ADMI 2011*, Clemson, SC (2011)
6. Ford, A., Obiakor, F., Patton, M.: *Effective Education of African American Exceptional Learners- New Perspectives*. pRO-ED, Inc., Austin (1995)
7. Grossman, H.: *Teaching In A Diverse Society*. Needham Heights, MA Allyn & Bacon; A Simon and Schuster Company (1995)
8. Klassner, F.: A case study of LEGO Mindstorms suitability for artificial intelligence and robotics courses at the college level. In: *Proceedings of SIGCSE 2002*, pp. 8–12 (2002)
9. Oblinger, D.G.: *Learning spaces* (2006), <http://www.educause.edu/ir/library/pdf/PUB7102a.pdf> (retrieved May 4, 2011)
10. Touretzky, D.: Preparing computer science students for the robotics revolution. *Communications of the ACM* 53(8), 27–29 (2010)
11. The Sumobot-Mini Sumo Robots Assembly Documentation and Programming Manual, <http://www.parallax.com/dl/docs/prod/robo/sumobotman.pdf>
12. Schmeck, R.R.: *Learning Styles and Learning Strategies*. Plenum, NY (1986)
13. Weinstein, C.E., Goetz, E.T., Alexander, P.A.: *Learning and Study Strategies*. Academic Press, NY (1986)

# Using a Fuzzy Emotion Model in Computer Assisted Speech Therapy

Ovidiu A. Schipor, Stefan G. Pentiuc, and Maria D. Schipor

“Stefan cel Mare” University of Suceava, Universitatii Street, 13, 720229, Romania  
email: schipor@eed.usv.ro, pentiuc@eed.usv.ro,  
vmdoina@yahoo.com

**Abstract.** Affective computing – machine’s ability to recognize and simulate human affects – has become a main research field for Human Computer Interaction. This paper deal with emotion recognition within a CBST (Computer Based Speech Therapy System) for preschoolers and young schoolchildren. Identifying the emotions of children with speech disorders during the assisted therapy sessions requires an adaptation of classical recognition techniques. That is why, in our article we focus on finding and testing the best emotion representation model to be used in this narrow field. An experiment that validates our proposed approach and indicates the probabilistic coefficient matrix is also presented. The proposed emotion recognition framework can be seen as a future extension of our CBST – Logomon.

**Keywords:** computer assisted speech therapy, emotion recognition, fuzzy emotion representation.

## 1 Introduction

One of the main differences between machines and humans refers to emotional intelligence (i.e. the ability to perceiving, using, understanding and managing emotions). Since the intelligent behavior is associated with the adaptation to the world around, the skills of emotional intelligence reflects the adaptability to affective stimuli. The artificial systems that integrate these capabilities are likely to be perceived as more natural, efficient and trustworthy by the human users [1].

From the standpoint of educational community (parents, teachers, and SLTs – Speech and Language Therapists), the most important limitations of CBSTs is the lack of adaptability and empathy [2]. Despite the remarkable progress in terms of providing real time feedback, highlighting the evolution of the pronunciation and establishing a personalized therapeutic program, CBSTs are still seen as a “secondary assistant” of the SLT.

That is why, since 2005, our team has begun the development of Logomon – a “next generation” CBST for Romanian language. We started implementing three classical modules: the main program installed on SLT’s PC, the child monitor program installed on a mobile device such as a PDA, and an interactive, animated 3D Model of the phono-articulatory system [3]. Then we extended this architecture

with a fuzzy expert system whose role is to determine the optimal exercises set for each child [4]. Next, in order to reduce the gap between the classical and the computer assisted therapy, we want to develop an affective computing framework. So our next challenge is to adapt general emotion recognition techniques to children's assisted speech therapy.

Therefore, in our paper we present an experiment whose purpose is to identify the specific emotional patterns (probabilistic coefficients matrix) for each therapy sequence. Since the emotions recognition was made by human experts (SLTs and psychologists), the experiment also reveals us the benchmark (i.e. the reference point) that we will consider to determine the performance of automatic recognition.

What are the emotions that could occur in the assisted therapy of speech? Which is the most appropriate emotions representation model? What is the probability associated with a specific emotion and a specific stage of therapy? These questions have been around since the start of our project but will become more pressing as the necessity of clever assisted therapy becomes more of a reality.

## 2 The Methods

### 2.1 Basic Models of Emotions

Many theorists have tried to define a list of basic emotions and, consequently, a wide range of research on identification of these states could be referred. In their article [5], Ortony and Turner have collated 14 different linear models. In addition, different hierarchical models have been proposed. For example, Parrot [6] offers a model on three levels: primary (6), secondary (25) and tertiary emotions (over 130).

There are two basic models of emotions:

- 1) *Labeling approaches* – this is a model that relies on basic emotions. The human experts (e.g. psychologist, SLT) choose a specific emotion from a predefined list (e.g. anger, joy, pleasure, sadness). The major advantage of this approach is the simplicity in terms of automatic recognition due to a small number of basic emotions that form the recognition universe. On the other hand, it cannot be used in confusing situations, when emotional state would be described by several words, each of them eventually associated with a certain weight [7];
- 2) *Dimensional approaches* – in this model, the emotion classification is performed according to specific dimensions such as: valence, arousal, intensity etc [8]. This method involves identifying affective state as a location on several continuous scales (e.g. pleasant – unpleasant, calm – arousal, etc). Two – dimensional (valence and arousal) or three – dimensional (valence, arousal, and stance) models are currently used. For example, joy has a positive valence, a high arousal level and, in addition, reflects an open stance.

## 2.2 Fuzzy Emotion Representation

In this paper we use a combination of the above models. The resulted approach - *Weighted labeling approach or Fuzzy model* – is flexible and appropriate for our system. In addition, it can be integrated with the fuzzy expert system that we have already implemented [4]. The simplified and natural integration of information taken from multiples channels (i.e. speech, video, physiological sensors) is also an advantage [9].

Instead consider a single emotional state associated with an activity, more emotions are accepted, each of them being related with a specific weight [10]. Weighting can be performed taking into consideration how long the emotion was manifested in the activity and the intensity level. So, for each emotion that forms the recognition universe, a subunit coefficient must be found [11], [12].

Let consider  $E$  – a finite set of  $n$  basic emotions and  $FM$  – an infinite set of fuzzy membership functions.

$$E = \{e_1, e_2, \dots, e_n\}, FM = \{\mu_j : E \rightarrow [0,1], j = 1, 2, \dots\}. \quad (1)$$

If we denote by  $ES$  the set of all emotional states ( $es_j$ ) that could be described by this model, the relations will be as follows:

$$ES = \{es_j, j = 1, 2, \dots\}, es_j = \{(e_i, \mu_j(e_i)) \mid e_i \in E\}. \quad (2)$$

So each emotional state can be seen as a fuzzy set with  $n$  elements. Each element is an ordered pair formed by the emotion ( $e_i$ ) and the value of membership function ( $\mu_j(e_i)$ ). The number of all emotional states that can be represented is, theoretically, infinite. Practically, however, this number depends on the precision of membership function representation. If we denote the representation precision by  $\varepsilon$ , then the cardinal of  $ES$  set can be calculated as follows:

$$|ES| = (1/\varepsilon)^{|E|}. \quad (3)$$

Each emotional state can be graphically represented as a point inside a hypercube into an  $n$ -dimensional space, where  $n$  is the cardinal of  $E$  set. In this representation, each axis corresponds to one basic emotion ( $e_i$ ) and each emotional state ( $es_j$ ) is an  $n$ -dimensional vector.

$$es_j = (\mu_j(e_1), \mu_j(e_2), \dots, \mu_j(e_n)). \quad (4)$$

In a previous article [13] we identify five basic emotions that could occur in an assisted therapy session: happiness, contentment, neutral, tenseness, and nervousness. The point described by the vector  $(0.2, 0.8, 0.3, 0.1, 0.0)$  represents a positive state and the corresponding fuzzy set is:  $\{(happiness, 0.2), (contentment, 0.9), (neutral, 0.3), (tenseness, 0.1), (nervousness, 0.0)\}$ .

This approach for emotion representation has several advantages:

- very good ability to handle blended emotions;
- quasi-continuous emotion representation and recognition;
- the fuzzy representation of emotions is familiar and natural to humans.

### 3 The Experiment

#### 3.1 The Methodology

The aim of this research is to use the fuzzy emotion representation in order to obtain a probabilistic pattern indicating the affective states that are likely to occur in a specific therapeutic sequence.

*The subjects* (N= 41) were children with moderate speech disorders, 4 to 9 years old, selected from Regional Speech Therapy Center – Suceava, Romania.

*Procedure:* The children were observed by three independent human experts during assisted speech therapy and an observation sheet was filled out by each of them (reliability analysis test Alpha-Cronbach  $\alpha = .83$ ). The observation was performed in three distinct sequences of speech therapy: 1. Speech evaluation by recording with feedback; 2. Exercises for phonematic hearing development; 3. The pronunciation of affected sound using 3D model. For each child and each stage of therapy five scores (in a Likert type scale ranging from 0 – “absence” to 5 – “maxim intensity”) were obtained.

#### 3.2 Results and Discussions

In order to identify emotional patterns (i.e. fuzzy sets  $eS_{seq1}$ ,  $eS_{seq2}$ ,  $eS_{seq3}$ ) associated with each therapy sequence the average scores were calculated and converted into probabilistic coefficients (Figure 1).

	0.00	0.09	0.44	0.73	0.33	1st sequence
	0.02	0.08	0.24	0.63	0.37	2nd sequence
	0.04	0.07	0.34	0.73	0.46	3rd sequence
nervousness						
tenseness						
neutral						
contentment						
happiness						

**Fig. 1** The probabilistic coefficient matrix indicates what emotional states are likely to occur in a specific therapy sequence

The Paired Sample t Test has shown us that despite the existence of the absolute differences between the therapy sequences, not all these differences are significant. So, we identified distinct emotional patterns for neutral (for each of the sequences), contentment (for 1-2 and 2-3 sequences), and happiness (for 1-3 and 2-3). For the others two emotional states (nervousness and tenseness) we identified but a global pattern, common to all therapeutically sequences. This probabilistic pattern will help the emotion recognition framework to deal with ambiguous situations (different channels indicate different emotional states).

## 4 Conclusion

The purpose of this study was to use fuzzy representation of emotions in order to obtain probabilistic emotional patterns for three stages of assisted speech therapy. The experiment involved SLTs, psychologists, and children with speech disorders and the results indicated the opportunity of fuzzy approach utilization. These outcomes encourage us to extend the Logomon CBST with an automatic emotion recognition framework.

**Acknowledgments.** This paper was supported by the project "Progress and development through post-doctoral research and innovation in engineering and applied sciences– PRiDE - Contract no. POSDRU/89/1.5/S/57083", project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

## References

1. Cassell, J., Bickmore, T.: External manifestations of trust-worthiness in the interface. *Commun. ACM* 12(43), 50–56 (2000)
2. Schipor, M.D., Pentiu, S.G., Schipor, O.A.: End-User Recommendations on LOGOMON - a Computer Based Speech Therapy System for Romanian Language. *J. AECE* 4(10), 57–60 (2010)
3. Pentiu, S.G., Schipor, O.A., Danubianu, M., Schipor, M.D.: Therapy Of Dyslalia Affecting Pre-Scholars. In: ECUMICT, Gent, Belgium, pp. 317–326 (2008)
4. Schipor, O.A., Pentiu, S.G., Schipor, M.D.: Computer Based Speech Therapy Using a Fuzzy Expert System. *Computing and Informatics* 2(29), 303–318 (2010)
5. Ortony, A., Turner, T.J.: What's basic about basic emotions? *Psychological Review* 97, 315–331 (1990)
6. Parrot, W.: *Emotions in Social Psychology*. Psychology, Philadelphia (2001)
7. Kim, J.: *Bimodal Emotion Recognition using Speech and Physiological Changes*. I-Tech Education and Publishing, Vienna (2007)
8. Mower, E., Metallinou, A., Lee, C.C., Kazemzadeh, A., Busso, C., Lee, S., Narayanan, S.: Interpreting Ambiguous Emotional Expressions. In: *International Conference on Affective Computing (ACII)*, Amsterdam, The Netherlands (2009)
9. Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., Metze, F.: Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features. In: *13th INTERSPEECH*, Brighton, UK, pp. 340–343 (2009)
10. Kusstatscher, V., Cooper, C.L.: *Managing Emotions in Mergers and Acquisitions*. Edward Elgar, Cheltenham, UK (2005)
11. Esau, N., Wetzel, E., Kleinjohann, L., Kleinjohann, B.: Real-Time Facial Expression Recognition Using a Fuzzy Emotion Model. In: *IEEE International Conference on Fuzzy Systems*, London, UK, pp. 23–26 (2007)
12. Wollmer, M., Eyben, F., Reiter, S., Achuller, B., Cox, C., Douglas-Cowie, E., Cowie, E.: Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In: *12th INTERSPEECH*, Brisbane, Australia, pp. 597–600 (2008)
13. Schipor, O.A., Schipor, M.D., Crismariu, E., Pentiu, S.G.: Finding key emotional states to be recognized in a computer speech therapy system. In: *WCPCG*, Antalya, Turkey (2011)



# Towards Better Discoverability and Use of Open Content

Christo Dichev, Bandana Bhattacharai, Curan Clonch, and Darina Dicheva

Winston-Salem State University, 601 S. Martin Luther King J. Drive,  
Winston Salem, NC 27110, USA  
email: {bbhattacharai109,dichevad,dichevc,cclonch109}@wssu.edu

**Abstract.** Availability and open access to resources is an important factor in educational development, but not a sufficient solution by itself. Open educational practices need to be fostered by an appropriate supportive environment including discoverable and shareable content, combined with tools for adaptation and redistribution. Although a number of open content resources have been created, the accessibility for end users remains low because resources are spread across different OER repositories. This paper is an attempt to respond to OER challenges by proposing an innovative architecture extending the set of features typical for portal OER repositories with resource discoverability, social networking features and (re)publishing functionality.

## 1 Introduction

Open Education movement is aimed at improving education access and quality by enabling educators to develop, use, re-use, and share learning resources. Open educational resources (OER) include instructional materials, tools, and media used for teaching and learning that are free from copyright restrictions or are publicly licensed for anyone to use, adapt, and redistribute. The vision of educational material, openly accessible on the Web, has attracted substantial attention [1]. It builds on the idea of open source software that communities of common purpose can achieve more by contributing ideas and efforts in an environment freed from the constraints of copyright and monetary exchange. The OER movement has greatly benefited from the Creative Commons licensing scheme that enables creators to give away their material for use while having the option to retain certain rights such as identification of the original author and barring the selling of open content or its derivatives.

From a technical perspective, OER repositories are a natural choice to host and locate appropriate open educational content. Some repositories are created by institutions to host their own resources, while others are held outside institutions. Some repositories are general in nature (e.g. JorumOpen [2], Merlot [3]), others are specialized in a particular theme or discipline. There can be made a distinction between repositories that store their content locally, and such that provide metadata with links to OERs housed at other sites. Yet, there are repositories that

provide both content and links to external content. *Content OER repositories* store content on a site following a centralized model. These repositories range from the widely known MIT OpenCourseware [4] site with more than 2000 courses, to the SOFIA (Sharing of Free Intellectual Assets) website [5] that hosts only eight courses for community college level students. WikiEducator [6] is an ambitious project aiming to become a central repository for storing large numbers of OERs. *Portal OER repositories* are hubs that host links to educational content provided by others. Many of them do house some content locally, but they are primarily aggregators of links. The most well known such a portal is MERLOT, which links to a wide variety of pedagogical content. Others, like CITIDEL [7], provide links to resources, ranging from research papers, notes, and lesson plans to multimedia applications and videos. *Hybrid OER repositories* function as both a content host and a portal. Several sites such as the multilingual ARIADNE [8] and the Commonwealth of Learning's OER repositories are examples of this type. EDNA [9] in Australia is one of the most comprehensive repositories enabling access to a range of hosted and linked educational content.

The number of OERs available in a repository is highly variable. It ranges from more than a million objects to just a few. The larger ones generally consist of components and the smaller ones of full courses or specialized multimedia applications.

The majority of repositories are open to anonymous users but require users to sign-in when submitting or reviewing materials; the latter is required as a form of quality control [10]. For example, Merlot requires reviewers to register.

Open access to resources is an important element in educational innovation, but not the ultimate solution. Repositories can offer the advantage of advanced searching facilities and can attract their own audience of resource users and contributors. The key enabler is an appropriate supportive environment, including easily accessible and shareable content, tools, and services. Such an environment should facilitate finding relevant materials to be directly used or repurposed for specific learning contexts. Although a number of OER has been created by several initiatives and projects, the accessibility for end users remains low because useful resources are distributed across many different OER repositories. Identifying and searching these repositories is a big barrier for the large-scale uptake and use of OER.

This paper is an attempt to respond to these limitations by addressing two key aspects of OER – finding relevant content and reusing it. It first outlines an experiment intended to demonstrate how easy it is to find relevant OER materials and then describes our OER Portal highlighting the support for resource finding and reuse.

## **2 How Easy Is It to Find OER Relevant to the Task in Hand?**

The open education movement has been successful in publishing a large amount of educational resources. At present, there are more than 13,000 courses published by 150 universities. This success however creates new challenges; for example,

finding relevant educational content on the Web is a recognized problem. We did a small scale experiment to check if the OER movement has improved this situation. Our experiment indicated that searching for good quality and clearly licensed resources can still be a frustrating experience.

The goal of the experiment was to collect some evidence on how easy (difficult) is to find open content on selected topics that can be used as (a source for) lecture slides, notes or other instructional materials. We selected general Computer Science topics assuming a lecture level as a granularity criterion for the results. Our experiment started with topics selected from a typical Operating Systems course covering Process Synchronization. We made an assumption that the instructor wants to illustrate the algorithms using Java. For maximizing the coverage and accuracy of our search for relevant OER material, we employed several searching strategies. First, we used Google's advanced search to narrow down the results to materials licensed under a Creative Commons Attribution ("Free to use, share or modify, even commercially"). The top three hits were Wikipedia pages. The next was a Java example implementing barrier. The fifth result was irrelevant, followed by the first relevant resource on Synchronization and CPU Scheduling provided by Connexions [11]. Neither of the next results returned by Google (from 7 to 50) provided any relevant page.

For comparison with an open web search we repeated the search experiment with twelve OER sites providing Computer Science materials, namely: Connexions, MIT OpenCourseWare, CITIDEL, The Open University, OpenLearn, *OpenCourseWare Consortium*, OER Commons, Merlot, NSDL, Wikibooks, SOFIA, Textbook Revolution, and Bookboon. We used again Google's advanced search but this time to localize the search within the corresponding domain, which resulted in twelve search sessions triggered by queries such as "operating systems synchronization java site:merlot.org" (for searching within Merlot site). In total, the thirteen search sessions found only one truly relevant material, "Synchronization, CPU Scheduling" (<http://cnx.org/content/m28019/1.1/>) stored in Connexions' repository. Most of the "somewhat relevant" results were from MIT OCW and Wikibooks on Java Threads.

The initial experiment gave us some clues about the distribution and discoverability of university level Computer Science related teaching materials. For more solid evidence, we extended the experiment to a set of introductory and upper-level courses following the last ACM Curriculum Recommendation (<http://www.acm.org/education/curricula/ComputerScience2008.pdf>), namely: Analysis of Algorithms, Artificial Intelligence, Computer Architecture, Computer Networks, Database Management, Introduction to Programming, Operating Systems, Programming Languages, and Software Engineering. For each of these introductory or core courses we selected topics for searching resources that can be used as lectures plans or notes (see Table 1). In addition, we included an elective type courses and more advanced subjects (see the last seven rows in Table 1) in order to weigh the corresponding OER material availability against the more conventional subjects.

The results showed that the coverage of the core and introductory computer science courses is quite low - around 2.5 (average) resources per subject. The coverage of elective computer science courses and advanced subjects is even lower - 0.4 (average) resources per subject. The latter results indicate a disturbing imbalance:

almost no OER materials were found on subjects, where instructors need even stronger support since up-to-date information is typically not covered in textbooks.

Another purpose of this experiment was to identify certain factors that need to be addressed for improving the growth and discoverability of OER. With a few exceptions, the twelve OER sites are institutional repositories that do not provide support for resubmitting derivative work. Typically there is no software support for social/community recognition (e.g. rating) of resources. Most OER repositories do not collect user reviews/comments on the quality of the resources hosted in their sites. They are general repositories providing learning resources from a wide variety of subject areas. Most of the metadata in the repositories is about the content itself, and not about its use/application. The OER resources returned as results of the experiment varied greatly in their size and target audience. If the site location is not explicated (e.g. as url), OER resources are not easy to find, even with the Google advanced search. Moreover, locating OER repositories themselves is a challenge. This observation is in line with the opinion expressed in [12]. Another problem of the current open access educational repositories is that despite their philosophy of sharing, they see teachers and learners as consumers of content who primarily want to download useful material.

**Table 1** Results of Google and OER search for resources on selected topics.

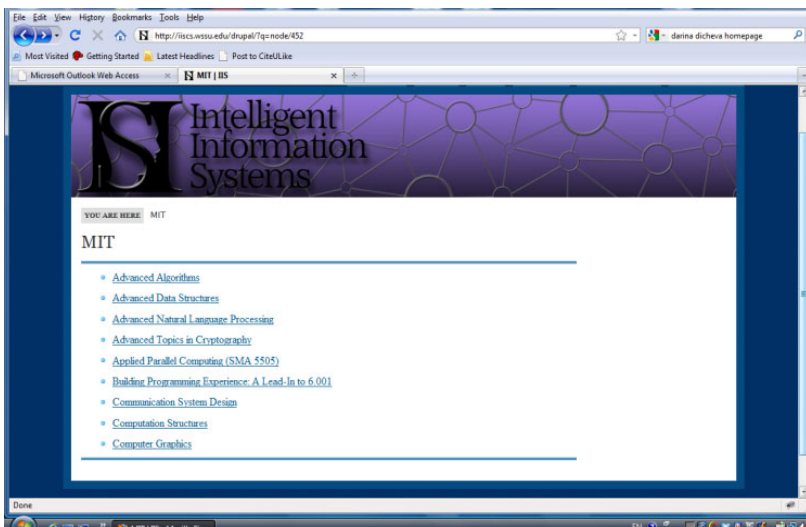
No	Search Keywords	Found Relevant	
		Google	OER
1	analysis algorithms spanning trees	0	0
2	artificial intelligence reasoning systems	0	0
3	computer architecture pipelining	1	3
4	computer networks congestion control	1	6
5	database management normalization	1	9
6	data structures binary trees	2	3
7	introduction programming java inheritance	1	3
8	operating systems synchronization java	1	1
9	programming languages data types	1	0
10	software engineering agile development	1	4
11	artificial intelligence definite clause grammars	0	0
12	e-commerce web services	0	0
13	hardware media security blue-ray prm	0	0
14	information system security access matrix model	0	0
15	information system security public key infrastructure	0	0
16	web programming php arrays	3	2
17	web programming xml xpath	0	0

### 3 The OpenCS Portal

We propose a new type of OER repository, intended to respond to the challenges involved in making specialized course materials openly available and easy to find and (re)use. As a test bed we are implementing OpenCS - a portal for Computer

Science instructors and students that provide resource links and metadata and is aimed to support finding, using, adapting, re-publishing and sharing resources in an open way. The open resources are collected and disseminated based on the metadata provided by the users themselves. At present it contains more than 1200 links to CS open content learning resources, harvested across seven major OER sites (out of 49) offering Computer Science related material (Figure 1).

The novelty of OpenCS is that it provides a set of features typical for portal OER repositories combined with social networking features and extended with (re)publishing functionality, where the focus is on finding relevant resources and providing motivational factors for contributors. Despite the fact that OpenCS is a portal type OER repository, it is designed to support all stages of OER: from search and identification of resources, to re-use, to adaptation, re-publishing and sharing. However, in contrast to content portals such as Connections, we used the term republishing to mean that users adapt and publish the resources on their own sites and then publish and annotate the links to the adapted versions in OpenCS.



**Fig. 1** Screenshot from the OpenCS interface

The top-level criteria impacting the design strategy of OpenCS was not to build “yet another repository” but to try some innovative solutions aimed at addressing known obstacles. Among the biggest challenges facing the open education movement, according to Brown [13] are to help people find open educational resources and to help people (re)use open educational resources. As we have worked to design a repository that would address those challenges, our strategies coalesced around two ideas: provide help to users by finding resources matching the context and task at hand (resources for a user’s current job); show contributors how their resources are used, liked and shared. Most users (including the authors of this article) search for the resources/content they need not in specialized

repositories, but rather by general Web searches. Our explanation for this phenomenon is that the repositories are designed for a small set of “common jobs”, ignoring the long tail of less conventional “jobs”. Therefore, one of the principles, on which the OpenCS design is grounded, is the Christensen’s “Jobs-to-Do Marketing” theory<sup>1</sup>.

So far OER repositories are following the standard paradigm of dividing the resources into categories (topical disciplines) and resource types (like “courses”, “videos”, “simulations”) and also segmenting their users (e.g. “high school” vs. “college” instructors). Using such categories and hierarchies does provide value toward matching content to educators. For jobs such as “I am looking for PowerPoint slides on CPU scheduling” the standard categorization is acceptable for certain set of tasks. However there might be some users with more specific tasks (jobs) such as “Need PowerPoint slides on CPU scheduling focusing on algorithmic details”, “..with examples in Java”, “...with emphasis on thread scheduling”, “...skipping complex issues”, etc. Besides, there are task requirements not expressible in terms of keywords or proper categories. Creating repositories satisfying all possible tasks is unachievable goal. More realistic objective is to allow users with minimal efforts to find the “best fit” solution for their requirements. This strategy is in line with our “job-to-do” approach – for any job, make the resources that are requiring less repairing efforts easily findable. This assumes providing options for selective exploration of the repository. One such option under development is “find more like this” service, where OpenCS users can direct and narrow their search towards the best fit. It allows control of the similarity requirements by controlling specific attributes. As a result users are provided a way to “find more like this” form the same author, “more like this” but more recent, “more like this” but Java based, etc. Pragmatically this approach is analogical to a customer in a shoe store who asks “show me pair of shoes like this one but with heels and proper leather”. By taking the “jobs-to-do” perspective we aim at a new approach that can make the OER repositories responsive for a wider variety of users.

In the following we outline the OpenCS functionality and the combination of strategies resulting in a system that is more than the sum of its components.

**Browsing and searching.** The OpenCS portal contains an organized list of links to resources harvested from known OER sites that provide Computer Science related material. It is organized under the taxonomy: *Institution/Organization* > *Courses* > *Topics*, with an additional possibility for browsing the topics alphabetically, by instructional methods such as notes, assignments, exams and others, or by tags.

In addition to “find more like this” search, we are exploring the possibility of utilizing customized web search, by tuning up a custom search engine to a specific context, based on varying the OER sites to be searched, specific metadata, recency, etc. For example, by providing a sitemap, one can search specific sites and pages as well as control the coverage and freshness of the search results; by providing adequate keywords, weighted labels, and scores or with promoting results one can control ranking of the results; by using *synonyms*, which are

---

<sup>1</sup> <http://hbswk.hbs.edu/item/6496.html?wknews=02142011>

variants of a search term one can expand search queries. Further, one can categorize sites by topics and take into account the most popular queries for extending the search options. The various options for customization are intended to expand the alternatives and cover the possible “job-to-do” needs.

Another tool that is part of the integrated browsing and searching support is “find similar” for recommending content similar to given resource in the collection.

**Sharing and dissemination.** The site offers RSS feed intended for selective dissemination of the available open educational resources. Instructors and learners are able to subscribe to course-specific RSS feeds, so that they can stay up to date as new resources/updates are added. This way they can receive the latest derivative works, the latest annotations and ratings for resources under the selected courses. The updates passed to the users reflect the corresponding updates occurring within the hosting OER sites, since the RSS feeds technology is used also to syndicate content *as it is published* from the selected OER sites. More specifically, RSS feeds are utilized to create two types of information flows from the viewpoint of OpenCS: incoming and outgoing. The incoming flow provides updated information from the hosting sites, while the outgoing flow delivers the updated information from OpenCS to the subscribed users customized to their needs. Thus we utilize the dual feature of the RSS feeds - as a mechanism for content syndication and content distribution. The incoming information is used also to automate the process of harvesting new content: the new resource URLs forming the incoming flow are converted into static links.

The sharing of information about useful resources by signed-in users is facilitated by including a “share” button, with options of emailing the resource details to other users or sharing them on Facebook or Twitter, among others.

**Support for (re)use.** Another aspect addressed in the repository design is the support for resource contribution and reuse. The corresponding functionality is intended to encourage an active user engagement and leverage the repository uptake. The idea was to employ easy-to-use social software tools that will extend the role of the repository from a storage system to an open platform where users can participate and contribute. Our approach exploits the concept of *social and technological affordances* [14,15]. Social affordances refer to the properties of an artifact that encourage users to generate social interaction. Technological affordances refer to an artifact’s usability. The support for resource contributors includes the following functionality:

- Submitting (a link to) a new open content resource;
- Submitting (a link to) an adapted version of a given open content resource;
- Acknowledging the use of a given resource;
- Social networking functionality such as tagging, liking, sharing, syndication and commenting.

Among the affordances that we consider of particular importance is tracking the use of and rating (liking) resources. We have added a “Used it” button next to

each resource to allow users to indicate the fact that they have used the resource and thus recognize its value. Recognition is a key factor for promoting participation in a community. The community of open source software developers has demonstrated how a gift economy (a social system in which status is given by how much one shares or gives to their community) works and how a reputation-based gratification system rewards the work of volunteers [16]. Another indication of the value of this type of recognition is the new feature publicized by Google. For recommending a particular link Google is adding a new +1 button next to each result. We plan to further exploit the “Used it” feature to enable resources with high “Used it” values float to the top of the repository search. In addition to improving the search results, user reputation rating could be used as a rewarding mechanism for contributions. Credit should be given to contributors for sharing quality content. On the other hand, such reputation systems make it possible to relax the prerequisites for entry to the repository. Widely used material typically is of high quality.

**Implementation.** Drupal, an open source content management platform, was chosen as a platform for developing OpenCS, as we wanted to rapidly prototype and test the portal. Additional factors in favor of its selection were that Drupal has already a variety of modules offering Web 2.0 functionality; it has interoperability functionality for many popular services like, Creative Commons, Facebook, Flickr, Google API, YouTube, etc. Our work aims to make the OER resources spread across various sites and collections accessible from a single site dedicated to Computer Science resources.

## 4 Conclusion

Availability and open access to resources is an important factor in open content development, but not a sufficient solution by itself. Open educational practices need to be fostered by an appropriate supportive environment, including easily discoverable, and shareable content, combined with tools and services for adaptation and re-distribution. The active engagement of users in OER repositories will promote their uptake. As an attempt to respond to these challenges, in this paper we propose a repository architecture supporting two key aspects: customized contextual search where resources are naturally discovered and accessed by potential users; integration of suitable Web 2.0 functionality extending repositories from a storage system to an open platform where users can participate and contribute.

Whilst OER repositories are having a growing impact on the amount of resources available for sharing, a key challenge is discovering the impact of these resources for both instructors and students - a goal included in the agenda of our future research.

**Acknowledgement.** The work reported here is supported by NSF, Grant DUE-1044224 “Social Bookmarking for Digital Libraries: Improving Resource Sharing and Discoverability”.



## References

1. Geser, G. (ed.): Open Educational Practices and Resources – OLCOS Roadmap 2012, [http://www.olcos.org/cms/upload/docs/olcos\\_roadmap.pdf](http://www.olcos.org/cms/upload/docs/olcos_roadmap.pdf) (retrieved May 15, 2011)
2. <http://www.jorum.ac.uk/> (retrieved May 15, 2011)
3. <http://www.merlot.org/merlot/index.htm> (retrieved May 15, 2011)
4. <http://ocw.mit.edu/index.htm> (retrieved May 15, 2011)
5. <http://sofia.fhda.edu/> (retrieved May 15, 2011)
6. [http://wikieducator.org/Main\\_Page](http://wikieducator.org/Main_Page) (retrieved May 15, 2011)
7. <http://www.citidel.org/> (retrieved May 15, 2011)
8. <http://www.ariadne-eu.org/index.php> (retrieved May 15, 2011)
9. <http://www.edna.edu.au/edna/go> (retrieved May 15, 2011)
10. McGreal, R.: A typology of learning object repositories. In: Handbook on Information Technologies for Education and Training. Springer, Heidelberg (2008)
11. <http://cnx.org/> (retrieved May 15, 2011)
12. Richards, G., Marshall, S.: Open Educational Resources for Development of University Courses. In: Open Ed 2010 Proceedings. UOC, OU, BYU, Barcelona (2010)
13. Brown, J.S., Adler, R.P.: Minds on fire: Open education, the long tail, and learning 2.0. *EDUCAUSE Review* 43(1), 16–33 (2008)
14. Gaver, W.: Technology affordances. In: Proceedings of CHI 1991. ACM, Louisiana (1991)
15. Kreijns, K., Kirschner, P., Jochems, W.: The sociability of computer-supported collaborative learning environments. *Educational Technology & Society* 5(1), 8–22 (2002)
16. Matzan, J.: Free software and the gift economy, <http://www.thejemreport.com/mambo/content/view/125/> (retrieved May 15, 2011)

# Teachers' Training Design Model and Decision Support System

Eliza Stefanova<sup>1</sup> and Svetla Boytcheva<sup>2</sup>

<sup>1</sup> Sofia University "St. Kl. Ohridski", Faculty of Mathematics and Informatics,  
5, James Bourchier Blvd., Sofia 1164, Bulgaria  
email: eliza@fmi.uni-sofia.bg

<sup>2</sup> State University of Library Studies and Information Technologies  
119, Tzarigradsko Shosee, 1784 Sofia, Bulgaria  
email: svetla.boytcheva@gmail.com

**Abstract.** This paper discusses the problems in the design of teacher's training as a form of professional development. The research is focused on design of support tools for teachers helping them to apply in effective way digital technologies in their practice. The complexity of such instructional design goal contributed to design of fuzzy logic based model. On its base was developed expert system (ES) to support decisions during the course design. In this paper we present the experts methodology for knowledge formalization. Further we describe the theoretical framework including main components in the model, their characteristics and relations. Implemented prototype of the fuzzy logic based expert system Open Virtual World (OVW) is presented, as well.

**Keywords:** Adaptive teachers training design, Fuzzy logic, Expert System, Decision support system, Intelligent systems.

## 1 Introduction

In last decade learning environment drastically was changed: a huge amount of Information and Communication Technologies (ICT) tools appears in schools. Many researchers and politicians hope that ICT itself will dramatically change the education. But it is not enough just ICT to be available in the schools. Often ICT are not effectively used, and in some cases not used at all. One of the conclusions of the Institute for Prospective Technological Studies report [1] is that in order to have effective use of technology in the school, it is necessary teachers to be trained appropriately. Looking on past experience, in many countries massive teachers training on ICT were done in recent years (Bulgaria, Romania, etc.). In other countries like UK teacher professional development is embedded in the school systems. Then, why the expected change is still not visible? One of the reasons for ineffectiveness of the ICT use in schools is related to the design for teachers training.

Teacher training is one of the four forms of the professional development [6]. In-service teachers' courses format is appropriate and very effective when educators need to obtain information about new programs, new instructional approaches, or

changes in school policy and regulation, as well as some innovations are introduced and small number of people is well informed about them [2]. Teachers training in field of integration of technology across curricula can be referred to this format.

During the design of teacher training, the characteristics of professional developments of adult should be taken in account. It is not enough to build the knowledge for a technology per se. The effective teaching of technology requires an understanding of how technology relates to the pedagogy and content. The designers of teachers training should aim to build Technological Pedagogical Subject Knowledge [3]. We need a model to support decisions during the teacher training course design. The characteristics of teachers training in such model are very complex. Furthermore the model should be adaptable to different technologies, users and their objectives.

In section 2 we present describe the theoretical framework including main components in the model, their characteristics and relations. The decision support system prototype based on Fuzzy Logic is presented in section 3. The conclusion briefly sketches some further steps in the research.

## 2 Theoretical Framework

The domain model is too complex and there is no consensus of the experts in the area. That is why abstract model OVW based on teachers and teacher trainers' opinion and experience was developed. For formal modeling in such cases, it is appropriate a Fuzzy Logic [8, 9] based Expert System [4] to be developed. The design process starts with collection and conversion of the experts' knowledge to the conceptual abstract model. Further the implemented system can be used by course designers to derive the conclusions from the model through the expert system. All these steps are described below.

**Components identification** is based on collecting experts' understanding on importance of the factors related to teachers training in digital technologies for education. Only those of them that have great impact in effective use of ICT in school practice are took in consideration. In this phase 23 experts from Bulgaria were involved. They are mainly experts in the field of training teachers for effective integration of ICT in education. Methodology used to collect experts opinion follow the structured participative approach called Group Concept Mapping, used for similar research [7]. Through the analysis of the collected results four top factors, rated by participants, are identified to be main components of the model. Namely: Methodology, Objectives, User, and Technology. The listed by participants main reasons related to each of the factors is detected as important characteristic of the component. On their base main variables of the each component are drawn [5].

**Variables values identification** was done through the collection of expert opinions. For instance, the Table 1 represents the learner's activity values.

**Relations between components characteristics definition** based on expert knowledge was next phase of the model development. The survey was used also to collect the experts' knowledge on relations between variables. On that base the rules were proposed. An excerpt of list of extracted rules from defined relations is presented on Figure 1.

**Table 1** Sample of methodology linguistic variable *learner's activity* values set

Linguistic variable: <i>Learners activity</i> – la			
Linguistic value	Notation	Numerical Range (normalized)	Fuzzy Sets of la
Very Low	VL	[0, 0.3]	
Low	L	[0.1, 0.4]	
Intermediate	I	[0.3, 0.8]	
High	H	[0.7, 0.9]	
Very High	VH	[0.8, 1]	

```

IF meth.la=VH THEN
    (u.m=V OR u.m=EX) AND
    (obj.s=A OR obj.s=P) AND
    (u.q=I OR u.q=A)
IF tech.cp=H THEN tech.u=L
IF meth.po=H THEN
    (u.m=V OR u.m=EX) AND
    (obj.c=A OR obj.x=P) AND
    (u.cm=A OR u.cm=P)
IF tech.c=H THEN tech.u=L
IF meth.ti=L THEN u.m=L AND tech.u=L
    
```

**Fig. 1** Sample rules in OVW

**Model testing** was performed through four pilot trainings, designed and conducted by teachers and teacher educators. The input data for the design of these trainings are available and used to generate inference based on fuzzy logic centroid technique. To compare the results inferred by the model with reality the surveys with the participants in each of these trainings were conducted.

### 3 System Architecture and Design

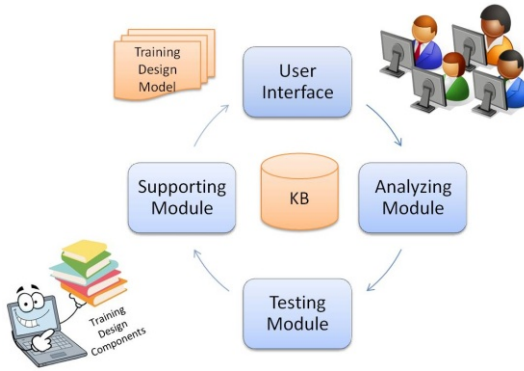
On the base of the model, a Fuzzy Logic based Expert System was developed. The system contains four main modules (Figure 2) with the following functionality:

**User interface module** supports user’s registration, maintains all users’ activities, and provides tools for design of training model development, update and storage.

**Analyzing module** evaluates the training design model and provides feedback for Methodology, Technology, User, and Objectives, based on Knowledge base with fuzzy logic rules. It provides features for testing the user’s expectations about the calculated values and the result provided by system – for OVW system evaluation purposes and for tuning the fuzzy logic rules.

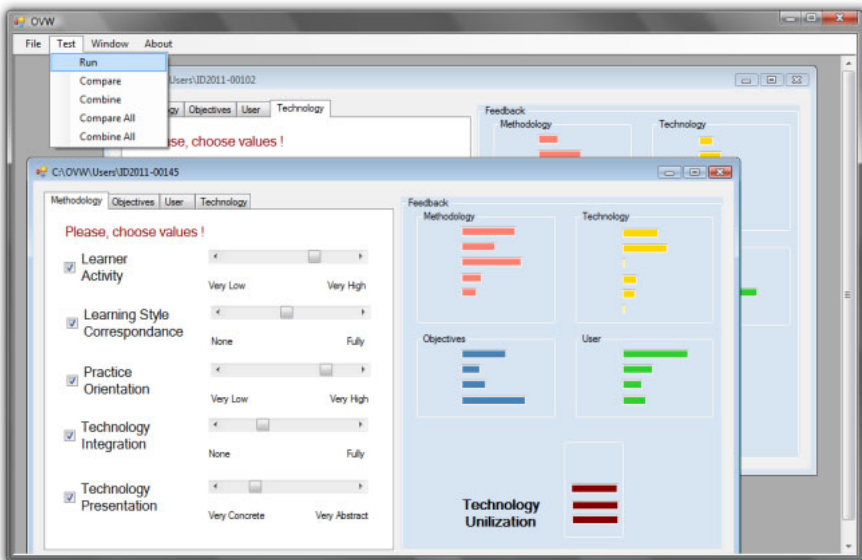
**Testing module** can be used to compare two designs of training or models, to aggregate with common (or more general) feature of two trainings’ models in order to provide the individual and group suitable training.

**Supporting module** realizes searches in the existing repository of appropriate materials according to training model and supports the training design process.



**Fig. 2** System Architecture

The important decisions during the system design are related to **the user interface**. Two main issues are: (1) how to collect the input values of the variables from the training designers; (2) how to present the system's inference to them. In both cases system should communicate with them in comprehensive and clear language and style. The decisions taken in these two directions are presented below.



**Fig. 3** User interface of the system

- User interface solution for entering variables' values by training designers**  
 Many systems based on fuzzy logic collect values of linguistic variables directly from the environment through sensors when the event happens (e.g. temperature becoming high). In the prototype of the OVW system this solution is not possible, because the values should be collected in advance of the event: training is still in

the phase of design and all values of variables related to it could just be planned by training designers. As the variables should be collected from the user, it is important to make it as easy and intuitive as possible. Therefore each components of the model is presented at separate tab, each variable of the component is presented at separated row and the values of the variables are entered by sliders, not as numbers. The solution takes into account human-computer interaction issue related to the user, who prefers to slide between two extreme values instead of entering numbers. This makes the input of the values really easy and intuitive.

- **User interface solution for presenting the information and the inference to training designers in order to support decisions**

Graphical representation of the information enhances readability of the inferred results. The primary objective of user interface in OVW system is to present graphically the inferred value for technology utilization (Figure 3). The generated value is based on the user input for linguistic variables, using rules and applying fuzzy logic centroid technique.

## 4 Conclusion

In current paper we describe work in progress. Performed experiments and tests with the designed model and the developed prototype of system so far demonstrate that they can be applicable in practice. The prototype of the system will be further tested with instructional designers of teachers' trainings and improved based on the results from these experiments.

## References

1. Cachia, R., Ferrari, A., Ala-Mutka, K., Punie, Y.: Creative Learning and Innovative Teaching: Final Report on the Study on Creativity and Innovation in Education in EU Member States, Institute for Prospective Technological Studies, (2010), <http://ftp.jrc.es/EURdoc/JRC62370.pdf> (last visited May 26, 2011)
2. Guskey, T.R.: Evaluating Professional Development. Corwin Press (2010)
3. Mishra, P., Koehler, J.M.: Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record* 108(6), 1017–1054 (2006)
4. Negnevsky, M.: Artificial Intelligence: A Guide to Intelligent Systems. Pearson Education, London (2002)
5. Stefanova, E., Boytcheva, S.: Theoretical Framework of Open Virtual Worlds for Professional Development. In: *Proceeding of S3T Conference, Bulgaria*, pp. 79–83 (2010)
6. Sparks, D., Loucks-Hotsley, S.: Five models of staff development of teachers. *Journal of Staff Development* 10(4), 40–57 (1989)
7. Wopereis, I., Kirschner, P.A., Paas, F., Stoyanov, S., Hendriks, M.: Failure and success factors of educational ICT projects: a group concept mapping approach. *British Journal of Educational Technology* 36(4), 681–684 (2005)
8. Zadeh, L.A.: Fuzzy Set. *Information and Control* 8, 338–353 (1965)
9. Zadeh, L.A.: Outline of new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3(1), 28–44 (1973)

# A Framework for Scaling Up Distributed Minds

Oisín Mac Fhearaí, Mark Humphrys, and Ray Walshe

School of Computing, Dublin City University, Dublin, Ireland

email: {omacfh, humphrys, Ray.Walshe}@computing.dcu.ie

**Abstract.** An important topic in artificial intelligence is the modular decomposition of problems and solutions into smaller ones. Modular, “hybrid” solutions tend to be built from components mostly created by the same programmer or the same team. What if we already have many solutions online - programs created by a number of different authors and hosted as services - and wish to compose them into a larger, distributed hybrid program which performs better than the individual components? Can this be done automatically? The World-Wide Mind (W2M) project attempts to scale up artificial intelligence by distributing action-selecting agents (which we call “minds”) and problem environments (which we call “worlds”) as services on the internet, and by allowing minds to call other minds and thus facilitate building hybrid minds from many programs which may have been written by many authors. This paper gives a general overview of the W2M architecture and ongoing work examining the possibility of automatically constructing hybrid minds.

## 1 The World-Wide Mind: Architecture and Implementation

To facilitate the creation of minds and worlds, and especially of hybrid minds which consult other minds when making decisions, we introduced a uniform interface which services must follow, representing messages that may be sent to minds (*getaction*) and worlds (*getstate* and *takeaction*). This, coupled with the ability to upload a mind to a server and have it immediately appear as a service online, makes experimentation and composition of minds simpler.

Although we consider the W2M platform to be useful for teaching and exploration of problems, our hope is that hybrid minds will be created which query other subminds for suggested actions (which may themselves be hybrids), and thus large-scale hierarchical problem-solving programs can be built from the work of many authors who may not understand how the other subcomponents function.

Since each level in the hybrid mind’s hierarchy may have many branches, it is easy to see that computational demands can rise quickly. To cope with this problem, it must be possible to distribute minds across machines and networks.

In the first implementation of the W2M server, worlds and minds were embodied as web services and assigned a URL. These services were hosted using the Tomcat application server, so that messages between services consisted of a web request and response, with the message content represented as XML. While this enabled connectivity across the internet, running a mind - especially a hybrid mind

composed of many remote minds - in a world was slow due to overheads in the underlying servlet technology and the use of HTTP to wrap messages.

To avoid this bottleneck, the server now sends the XML messages over a simple TCP protocol with very little overhead. This allowed us to take advantage of the common case where mind/world services are located on the same machine, by avoiding the network stack completely. Distribution will still be needed to scale up to bigger hybrids, but where network access can be avoided, it should be.

To further reduce latency, a scheme was implemented whereby the user asks the mind to carry out a run with the world and receives an asynchronous stream of messages from the mind with the states seen and actions taken, as well as a score object representing how well the mind performed in this run.

## 2 Evaluation and Future Work

The system was used by undergraduate computer science students taking an A.I. module. A W2M server was used to host several hundred minds for the Tyrrell animal world. Minds were ranked by their performance in this world, which prioritises mating and survival in a simulated environment.

A requirement was added that every student must submit at least one hybrid mind which delegates to one or more subminds. A call graph feature was implemented to track calls between minds, and the scoreboard at the end of the assignment showed that nine of the top ten minds called at least one other mind.

Ongoing research looks at the selection of subminds when constructing a hybrid mind. The set of minds described above was used to perform an analysis of the world state and score data. A large number of runs were performed with the submitted minds, and a record kept of all states seen and actions taken.

Metrics were created corresponding to important subgoals, for example minimising total thirst or hunger. These metrics were used to rank minds by their performance on each goal. A hybrid mind was created which consults the best mind for each subgoal in a series of simple case-action tests, for example: “if a mate is nearby, return the action suggested by the best mating mind”.

Experiments were carried out for hybrids composed of 2, 3 and 4 subminds. These tests produced a hybrid which scored 10% better than any of the submitted minds (mating 82 times versus 74 in the best existing mind).

Future work will explore ways to automatically determine which metrics most strongly influence the score, using statistical correlation rather than human intuition.

**Acknowledgments.** This work was supported by IRCSET’s Embark Initiative.



# Ontology-Based Electronic Test Results Evaluation

Veselina Jecheva and Daniela Orozova

Burgas Free University, Faculty for Computer and Technical Sciences,  
62 San Stefano Str., 8000, Burgas, Bulgaria  
email: {vessi, orozova}@bfu.bg

**Abstract.** The present paper examines an approach to automated test units evaluation, based on preliminarily created ontologies. Each test element is described by metadata, according to LOM standard. The learners' results are evaluated by intelligent agents, using ontologies.

**Keywords:** electronic tests, test elements, ontology, software agent.

## 1 Introduction

Knowledge representation and reuse is among of the key areas in the contemporary e-learning techniques. For the purposes of the present research domain ontologies were examined as a knowledge description in the field of the information security. On the other side, agent technologies have been broadly applied in the field of e-learning. Many e-learning systems contain tools for creation of questions, which require answer, completely worked out by the learner. Unfortunately, manual tests evaluation is a difficult and time-consuming task for the lecturer and systems that adequately provide automated "open" questions evaluation are still to come. This problem can be solved by creation of tools, which can automatically evaluate the learners' answers to the tests, comparing them to a preliminarily created knowledge database in a specific domain.

The purpose of the present paper is to create an e-learning system, which automatically evaluates a short answer or an essay in a specific topic. The proposed approach consists of the following stages: the first one is to represent the knowledge data by OWL domain ontology, and the second one is the agent-based system, which accesses the ontology and evaluates the user input.

## 2 Description of the Methodology

The presented methodology uses a domain ontology, which describes the organization of theoretical concepts and notions in a specific field, namely information security. An OWL-based main security ontology, described in [2], was examined for the purposes of the class *Information security* for the Computer science bachelor program in Burgas Free University. This ontology models the major concepts: assets, threats, vulnerabilities, countermeasures and their relations. It contains 88 threat classes, 79 asset classes, 133 countermeasure

classes and 34 relations between them. As the purpose of the work is to increase the automated test assessment ability, using ontology-based approach, the system should contain a database with test units and metadata, describing them. The examined test units are described by metadata, which contain keywords, applied for learners' answers evaluation, using specific search query results from the ontology. The test unit metadata could be defined in the following way:

*Test\_element([ test\_element\_id: <test element number>,  
key\_words: <keywords list>,...]).*

The proposed system is based on software agents. The major stage of the methodology is the test evaluation itself. The system input is the learner's answer to the "open" test question, which is created by the student themselves. It may contain one or more sentences, some terms enumerated, or an essay. The system input contains metadata, describing the expected answer, except the user input. Then the system creates SPARQL queries with the metadata as their arguments to the ontology, which contains the domain knowledge. The results contain the relations and concepts, related to the question keywords. This result should be used for a comparison with the learner's answer.

The system scans the learner's answer and performs a search of each sequence of  $n$  words in the ontology. The experiments were conducted with the values of  $n=1$  and  $n=2$ . With purpose of comparison the learner's answer with the results, obtained from the ontology, the  $q$ -gram metrics [1] was applied. The proposed approach calculates the degree of closeness of the concepts and relations, retrieved from the ontology, to the preliminarily defined keywords and the data, retrieved from the learner's answer. Finally, an average score is calculated, measuring the overall similarity of the answer to the knowledge database.

### 3 Results

The described methodology was applied with a purpose to evaluate the students' progress in the elective course "Information security". The obtained results reveal that the proposed methodology evaluated the students' answers with 80.76% accuracy when  $n=1$  and 84.61% accuracy for  $n=2$ . The number of students in the class was 26, and for 21 and 22 of them, respectively for  $n=1$  and  $n=2$ , the evaluation was correct and showed no significant deviation from the manual evaluation, performed by the lecturer.

### References

1. Gravano, L., Ipeirotis, P.G., et al.: Using q-grams in a DBMS for Approximate String Processing. *IEEE Data Engineering Bulletin* 24(4), 28–34 (2001)
2. Herzog, A., Shahmehri, N., Duma, C.: An Ontology of Information Security. *International Journal of Information Security and Privacy* 1(4), 1–23 (2007)

# Self-Adaptive Software Systems through Exploratory Changes

Stavros Stavru and Sylvia Ilieva

Sofia University “St. Kliment Ohridski”, Sofia, Bulgaria  
email: {stavross,sylvia}@fmi.uni-sofia.bg

**Abstract.** Most of the research in self-adaptive software systems is concerned with self-adaptation as response to change in its environment, which threatens system’s efficiency and operation. But except for avoiding and resolving system disturbances, there could be another reason for self-adaptation – the reason for exploration. This paper applies the concept of exploratory change to self-adaptive software systems and proposes a new paradigm for self-adaptation named exploratory self-adaptation.

**Keywords:** Self-Adaptive Software Systems, Software Evolution.

## 1 Introduction

The concept of exploratory change is not new and has been successfully applied in many different fields in computer science - from computation theory (metaheuristics and stochastic optimizations), artificial intelligence (machine learning and data mining algorithms, evolutionary computation and etc.) to software engineering (software prototyping, fault injection and etc.). On the other hand this concept has been barely studied by the research community working on self-adaptive software systems and it does not fit with the existing taxonomies of change [1, 2] and paradigms for self-adaptation [1, 3, 4].

Therefore the main contributions of this paper are the introduction of exploratory change in self-adaptive software systems and the specification of exploratory self-adaptation.

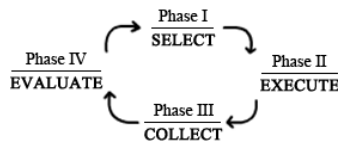
## 2 Self-Adaptive Software Systems through Exploratory Changes

Five new modeling dimensions are proposed to the existing taxonomies in order to accommodate the concept of exploratory change: (1) *Change motif* (exploitative or exploratory) describes the reason for change; (2) *Change control* (forced or voluntary) describes the possible control over the event that triggers the change; (3) *Change adoption* (threat or opportunity) describes how change is adopted; (4) *Change occurrence* (discontinuous or continuous) describes the rate of change occurrence; and (5) *Change speed* (time limited or time unlimited) describes the

time limitations of change. Based on these modeling dimensions, two different paradigms for achieving self-adaptation are identified: exploitative and exploratory.

**Exploitative self-adaptation** manages discontinuous exploitative changes, triggered by uncontrolled events, which are threatening the effectiveness and operation of the software system and are obligatory and time limited in order to avoid or suppress these negative effects in a timely manner.

**Exploratory self-adaptation** manages continuous changes, initiated by the system itself through controlled events for the purpose of exploration. By these changes the system looks for opportunities to increase its effectiveness and they are not constrained by any predefined timeframes. The adaptation loop behind exploratory self-adaptation is depicted in Fig. 1.



**Fig. 1** Adaptation loop in exploratory self-adaptation.

The adaptation loop starts with the SELECTION process. During this phase the set of variations to be introduced into the software system are defined (e.g. using random selection, rule-based selection and etc.). Within the EXECUTION process the introduction of these variations take place and the system actually changes. Then, during the COLLECTING process, information on the effectiveness of the changed software system is being collected. Based on this information, variations are further evaluated within the EVALUATION process and if the final evaluation is positive, the system remains as it is, but if it is negative, variations are rolled back.

Exploratory self-adaptation is applicable to *I know it when I see it* adaptation, when there is some degree of uncertainty on how the software system should change in order to properly self-adapt. It is not applicable when there are *strong requirements for software dependability, self-adaptation itself is critical, feedback is not accessible or rollback mechanisms are not supported*.

## References

1. Salehie, M., Tahvildari, L.: Self-adaptive software: landscape and research challenges. *ACM Transactions on Autonomic and Autonomous Systems* 4(2), 1–42 (2009)
2. Andersson, J., de Lemos, R., Malek, S., Weyns, D.: Modeling dimensions of self-adaptive software systems. In: Cheng, B.H.C., de Lemos, R., Giese, H., Inverardi, P., Magee, J. (eds.) *Software Engineering for Self-Adaptive Systems*. LNCS, vol. 5525, pp. 27–47. Springer, Heidelberg (2009)
3. Kephart, J., Chess, D.: The vision of autonomic computing. *IEEE Computer* 36(1), 41–50 (2003)
4. Dobson, S., Sterritt, R., Nixon, P., Hinchey, M.: Fulfilling the vision of autonomic computing. *Computer* 43(1), 35–41 (2010)

# Semantic Annotation in SINUS Project\*

Kamenka Staykova, Gennady Agre, Kiril Simov, and Petya Osenova

Institute of Information and Communication Technologies, BAS, Sofia, Bulgaria  
email: {kamy,agre}@iinf.bas.bg, {kivs,petya}@bultrebank.org

**Abstract.** The paper describes an approach for semantic annotation of multimedia objects stored in a Digital Library implemented as a Web Service. The Library has its own fixed annotation schema and provides a set of functions accessible as Web Service operations. The main objective of semantic annotations (supported by ontologies) is to extend both the Library functionality and the scope of the knowledge in it.

The main objective of the nationally funded research project SINUS (sinus.iinf.bas.bg) is to provide a semantic technology-based environment facilitating development of Technology-Enhanced Learning applications, which are able to reuse existing heterogeneous software systems. SINUS environment is tested on a use-case, which applies the basic principles of Technology Enhanced Learning for the process of Learning-by-Authoring [1]. The domain of Bulgarian Iconography is chosen for the SINUS Project scenario, because it presents an interesting example to apply Technology Enhanced Learning in humanities. The multimedia resources for SINUS demo-examples come from the Multimedia Digital Library “Virtual Encyclopedia of East-Christian Art” [2], which content is accessible via Web service.

*The Objects of Semantic Annotation* in SINUS Project are multimedia objects presenting information in digital form about icons, wall-paintings, miniatures and other iconographical objects; these are pictures and different texts concerning the iconographical objects, information about authors, places, dating periods, religious characters and so on. The Library has fixed annotation schema, which organizes all the resources and the available data. SINUS semantic space is on one hand based on that schema and on the other hand it pretends to present formalized knowledge of the domain of Bulgarian Iconography, to allow flexible and deep reasoning about that knowledge.

*SINUSBasic Ontology* is the main conceptual model of SINUS semantic space. The Library fixed annotation schema is taken as a ground for creation of SINUS-Basic Ontology. SINUSBasic Ontology is realized in OWL, it comprises 55 classes, 38 object properties and 28 data-type properties. Main classes are: Iconographical Object with its sub-classes Icon, Wall-Painting, Miniature, Mosaic and so on, Author, Iconographical Scene, Character, Iconographical Technique and so on. SINUS semantic space is envisioned to contain also, so called, “specialized

---

\* The research is supported by the Bulgarian Science Fund under the project No. D-002-189 “Semantic Technologies for Web Services and Technology Enhanced Learning” (SINUS).

ontologies”, which encode experts’ knowledge on particular aspect of the domain Bulgarian Iconography. At the current stage of the project the SINUSSpecTechnology ontology is realized in OWL, kept in Ontology Repository of the platform ready to be loaded to the semantic annotation space when needed. It contains 16 classes, 14 object properties and 43 ontological individuals.

*Semantic Annotation Models* of the objects presented in SINUS annotation space are mainly two (Basic Semantic Model and Extended Semantic Model), and give access to the concept of Iconographical Object. Semantic models for concepts of Author, Iconographical School, Collection will also be built. The Basic Semantic Model of Iconographical Object is supported by concepts of SINUSBasic Ontology, the Extended Semantic Model of Iconographical Object contains 14 additional features supported by the SINUSSpecTechnology ontology. Semantic Repository of SINUS platform is realized by SESAME RDF Schema querying and storage repository.

*The process of basic semantic annotation* is, in fact, Data Lifting process and concerns automated transfer of structured data from the Library to the annotation space corresponding to Basic semantic annotation model. Many SINUSBasic ontology individuals are created and made available for search and reasoning. *Additional semantic annotations made by the user* are also supported. This user-directed semantic annotation process makes it possible some new annotation features to be added to an existing annotation or a new annotation to be created. During this process the extended semantic annotation model is used.

Semantic annotation models of SINUS contain several links to descriptive texts supported by the Library. The main idea of some established experiments is to help the user in his/her attempt to annotate objects with ontological notions “visible” in these texts. Texts are preliminary semantically annotated/ tagged, which makes some ontological notions mentioned in them “sensitive” and technically prepared to be used further by the user in semantic annotations. *The semantic text annotations* are created off-line and stored, so they can be seen as indexes to the objects of annotation and used for on-line searching and retrieving the objects. Semantic text annotation mechanism is based on a model of Ontology-to-Text relation developed within [3] and [4]. *Search functionalities* of the semantic repository are available through the SINUS User Interface, which input is transformed to SPARQL queries.

The future work on SINUS Project includes the usage of the pre-prepared tags in the texts, extensive tests on the different semantic annotation processes and search process. The results will be analyzed in detail and compared with related works.

## References

1. Dochev, D., Agre, G.: Towards Semantic Web Enhanced Learning. In: Proceedings of Int. Conference on Knowledge Management and Information, Sharing, Madeira, pp. 212–217 (2009)
2. Pavlova-Draganova, L., Georgiev, V., Draganov, L.: Virtual Encyclopaedia of Bulgarian Iconography. Information Technologies and Knowledge 1(3), 267–271 (2007)
3. Simov, K., Osenova, P.: Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects. In: Workshop on NLP and Knowledge Representation for eLearning Environments, RANLP-2007, pp. 49–55 (2007)
4. Simov, K., Osenova, P.: Language Resources and Tools for Ontology-Based Semantic Annotation. In: Proceedings of OntoLex 2008 Workshop at LREC, pp. 9–13 (2008)

# A Multi-agent Approach to Development of E-Learning Facilities

Valentina Terzieva, Petia Kademova-Katzarova, and Rumén Andreev

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences Acad. G. Bonchev Str., Bl. 2, 1113 Sofia, Bulgaria

## 1 Environment for Development of E-Learning Facilities

Individuals with specific learning difficulties such as dyslexia have below-average learning performance. The state regulations require such learners to be integrated in public schools, but they need specific teaching methods, additional lessons with specially trained teachers, etc. We present a conceptual model of environment for facilities development that serves teachers in providing personalized e-teaching (Fig. 1). The model of *Subject Domain* (SD) defines a framework for educational process in two aspects – psychologists’ and educators’. The teacher’s representation of SD is goal-oriented. It determines basic requirements to e-learning facilities. The model *Cognitive Ability* shows the specific cognitive abilities necessary for achievement of main educational objectives in a concrete subject. They are fixed through mapping SD description on the basic human cognitive abilities. In that way a learner’s cognitive profile is determined by adequate to this SD learning characteristics. Model *Pedagogical Room* designs suitable methodology for learning activities in interaction with the models *Teaching Methods* and *Pedagogical Instruments*. The model *Teaching-Learning Goals*, which depends on the corresponding SD educational goals, influences on the aforesaid two models. The models are context-independent (global), so the learning facilities’ production requires contextualization. The model *Learner* serves for personalization.

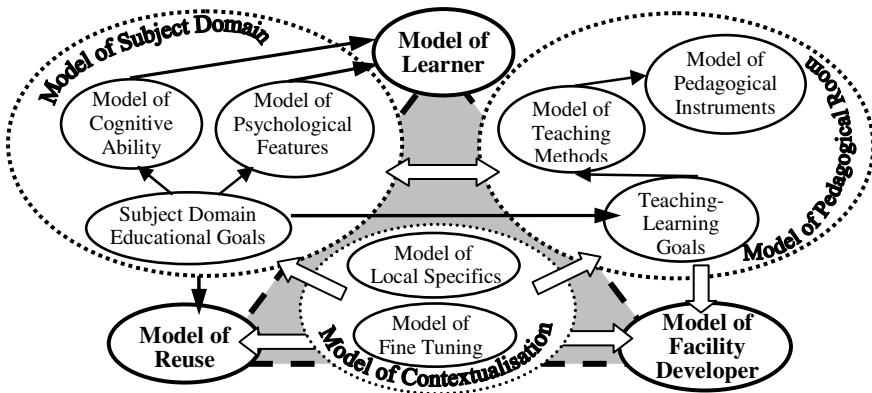


Fig. 1 Conceptual model of environment for development of e-learning facilities.

## 2 Multi-agent Approach to Design of the Environment

Since the environment for development of e-learning facilities has collaborative framework, it could be well designed through a multi-agent system. The design bases on presentation of conceptual model of the system and elaboration of detail models of each agent. The multi-agent architecture is in accordance with environment, purpose, agents' roles and communications. Each agent has exactly specified role and functions. For example, agent *Teaching-Learning Goals* determines a sequence of detailed sub-goals leading to efficient fulfilling of the educational goal, defined by *SD Educational Goals* for every subject. This sequence is individually tailored for every pupil and depends on the personal information delivered from *Learner's Model*. Agent *Cognitive Abilities & Psychological Features* determines the profile of each pupil. These agents give the parameters for agent *Learner's Model*. The latter selects the appropriate parameters for *Pedagogical Instruments & Teaching Methods* and defines the *Teaching-Learning Goals*. The last mentioned agents transfer the necessary information for agent *Facility Developer*. *Reuse* ensures the use of already existing learning facilities. Learning units are implemented by agent *Teaching/Training/Verification*. *Outcomes* decides how to proceed in the learning process, proposes suitable learning activities at home. *Feedback* reports the results of the collateral training. *Communications* connects all areas, supports cooperation among professionals and between them, parents and pupil; links the presented system and external ones (educational, repositories).

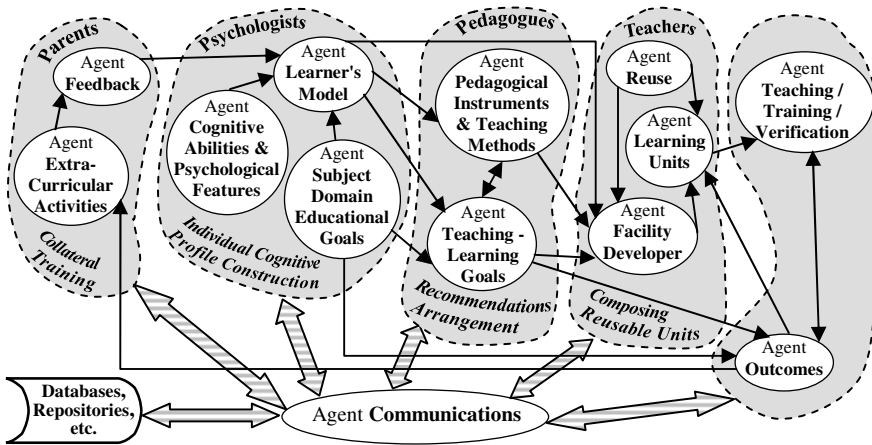


Fig. 2 Multi-agent model of a system for development of e-learning facilities



# Author Index

- Agre, Gennady 217  
Andreev, Rumén 219  
Angelov, Ivan 167  
Antonova, Albena 147, 175  
Ataide, Walker 155
- Banda, Srinivas 81  
Bhattarai, Bandana 195  
Bieliková, Mária 1  
Bittencourt, Ig Ibert 155  
Boychev, Pavel 161  
Boycheva, Svetla 205  
Brito, Patrick H.S. 155
- Caldwell, Rebecca 183  
Clonch, Curan 195  
Costa, Evandro 155
- Damova, Mariana 17  
Dannells, Dana 17  
da Silva, Alan Pedro 155  
Deters, Ralph 127  
Dichev, Christo 53, 195  
Dicheva, Darina 53, 195  
Dimov, Aleksandar 103, 139
- Eskenazi, Avram 103  
Esterline, Albert 81
- Farkas, László 47  
Fasli, Maria 33  
Fhearaí, Oisín Mac 211
- Gacovski, Zoran 133  
Ganchev, Ivan 167
- Geler, Zoltan 67  
Gourova, Elissaveta 75  
Gul, Ensar 111
- Holanda, Olavo 155  
Humphrys, Mark 211  
Hussain, Afshana S. 41
- Ilieva, Sylvia 103, 215  
Ivanović, Mirjana 67
- Jecheva, Veselina 213  
Johnsen, Lars 53  
Jones, Elva 183  
Juhász, István 47
- Kademova-Katzarova, Petia 219  
Karagiozov, Diman 119  
Khan, Koffka 59  
Kompan, Michal 1  
Koychev, Ivan 9, 25  
Kraychev, Boris 9  
Kurbalija, Vladimir 67
- Nikov, Alexander 59
- Orozova, Daniela 213  
Osenova, Petya 217
- Paulin, Alois 95  
Pavlova, Irena 103  
Pentiuc, Stefan G. 189  
Peshterliev, Stanislav 25  
Petrova-Antonova, Desislava 103

- Radovanović, Miloš 67
- Sahai, Ashok 59
- Schipor, Maria D. 189
- Schipor, Ovidiu A. 189
- Sharma, Sunny 127
- Simov, Kiril 217
- Stavru, Stavros 215
- Staykova, Kamenka 217
- Stefanov, Krassen 175
- Stefanova, Eliza 205
- Tenório, Thyago 155
- Terzieva, Valentina 219
- Todorova, Korneliya 147
- Toteva, Kostadinka 75
- Vágner, Anikó 47
- Vancov, Goko 133
- Vassileva, Julita 41
- Velev, Dimiter 89
- Walshe, Ray 211
- Wright, William 81
- Zabunov, George 89
- Zeb, Muhammad Ali 33
- Zlateva, Plamena 89