Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Zvezek A

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Volume A

Uredili / Edited by

Marko Bohanec, Matjaž Gams, Vladislav Rajkovič, Tanja Urbančič, Mojca Bernik, Dunja Mladenić, Marko Grobelnik, Marjan Heričko, Urban Kordeš, Olga Markič, Jadran Lenarčič, Leon Žlajpah, Andrej Gams, Andrej Brodnik

11.–15. oktober 2010 / October 11th–15th, 2010
Ljubljana, Slovenia

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Zvezek A

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Volume A

Inteligentni sistemi
Vzgoja in izobraževanje v informacijski družbi
Izkopavanje znanja in podatkovna skladišča (SiKDD 2010)
Sodelovanje, programi in storitve v informacijski družbi
Kognitivne znanosti
Robotika
MATCOS 2010 (Mini-konferenca v uporabnem teoretičnem računalništvu)

Intelligent Systems
Education in Information Society
Data Mining and Data Warehouses (SiKDD 2010)
Collaboration, Software and Services in Information Society
Cognitive Sciences
Robotics
MATCOS 2010 (Mini-conference on Applied Theoretical Computer Science)

Uredili / Edited by

Marko Bohanec, Matjaž Gams, Vladislav Rajkovič, Tanja Urbančič, Mojca Bernik, Dunja Mladenić, Marko Grobelnik, Marjan Heričko, Urban Kordeš, Olga Markič, Jadran Lenarčič, Leon Žlajpah, Andrej Gams, Andrej Brodnik

http://is.ijs.si

11.–15. oktober 2010 / October 11th–15th, 2010
Ljubljana, Slovenia

Uredniki:

prof. dr. Marko Bohanec
prof. dr. Matjaž Gams

prof. dr. Vladislav Rajkovič
prof. dr. Tanja Urbančič
Mojca Bernik

dr. Dunja Mladenić
Marko Grobelnik

prof. dr. Marjan Heričko

dr. Urban Kordeš
prof. dr. Olga Markič

prof. dr. Jadran Lenarčič
dr. Leon Žlajpah
dr. Andrej Gams

dr. Andrej Brodnik

Založnik: Institut »Jožef Stefan«, Ljubljana
Tisk: Birografika BORI d.o.o.
Priprava zbornika: Mitja Lasič, Jana Krivec
Oblikovanje naslovnice: Miran Krivec, Vesna Lasič
Tiskano iz predloga avtorjev
Naklada: 60

Ljubljana, oktober 2010

Konferenco IS 2010 sofinancirajo
Ministrstvo za visoko šolstvo, znanost in tehnologijo
Javna agencija za raziskovalno dejavnost RS (ARRS)
Institut »Jožef Stefan«

Informacijska družba
ISSN 1581-9973

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

659.2:316.42(082)
659.2:004(082)

MEDNARODNA multikonferenca Informacijska družba (13 ; 2010 ; Ljubljana)
Zbornik 13. mednarodne multikonference Informacijska družba - IS 2010, 11.-15. oktober 2010 : zvezek A = Proceedings of the 13th International Multiconference Information Society - IS 2010, October 11th-15th, 2010, Ljubljana, Slovenia : volume A / uredili, edited by Marko Bohanec ... [et al.]. - Ljubljana : Institut Jožef Stefan, 2010. - (Informacijska družba, ISSN 1581-9973)

Vsebina na nasl. str.: Inteligentni sistemi = Intelligent systems ;

Vzgoja in izobraževanje v informacijski družbi = Education in information society ; Izkopavanje znanja in podatkovna skladišča (SiKDD 2010) = Data mining and data warehouses (SiKDD 2010) ; Sodelovanje, programi in storitve v informacijski družbi = Collaboration, software and services in information society ; Kognitivne znanosti = Cognitive sciences ; Robotika = Robotics ; MATCOS 2010 (Mini-konferenca v uporabnem teoretičnem računalništvu = MATCOS 2010 (Mini-conference on applied theoretical computer science)

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2010

V svojem trinajstem letu je multikonferenca Informacijska družba (http://is.ijs.si) znova dokazala, da je ena vodilnih srednjeevropskih konferenc, ki združuje znanstvenike z različnih raziskovalnih področij, povezanih z informacijsko družbo. V letu 2010 smo v multikonferenco povezali deset odličnih neodvisnih konferenc. V Sloveniji in po svetu mgroli konferenc. Naša multikonferenca izstopa po širini in obsegu tem, ki jih obravnava, predvsem pa po akademski odprtosti in širini, ki spodbuja nove ideje.

Multikonferenca temelji na sinergiji interdisciplinarnih pristopov, ki obravnavajo različne vidike informacijske družbe ter poglabljajo razumevanje informacijskih, komunikacijskih in družbenih storitev v najširšem pomenu besede. Na multikonferenci predstavljamo, analiziramo in preverjamo nova odkritja in pripravljamo teren za njihovo praktično uporabo, saj je njen osnovni namen promocija raziskovalnih dosežkov in spodbujanje njihovega prenosa v prakso na različnih področjih informacijske družbe tako v Sloveniji kot tujini.

Na multikonferenci bo na vzporednih konferencah predstavljenih 300 referatov, vključevala pa bo tudi okrogle mize in razprave. Referati so objavljeni v zbornikih multikonference, izbrani prispevki pa bodo izšli tudi v posebnih številkah dveh znanstvenih revij, od katerih je ena Informatica, ki se ponaša s 34-letno tradicijo odlične znanstvene revije.

Multikonferenco Informacijska družba 2010 sestavljajo naslednje samostojne konference:

- Odprta delavnica mednarodnega projekta Confidence
- Inteligentni sistemi
- Jezikovne tehnologije
- Kognitivne znanosti
- Robotika
- Rudarjenje podatkov in podatkovna skladišča (SiKDD 2010)
- Sodelovanje, programska oprema in storitve v informacijski družbi
- Soočanje z demografskimi izzivi
- Vzgoja in izobraževanje v informacijski družbi
- 3. Minikonferenca iz teoretičnega računalništva 2010.

Zanimivo je, da finančna recesija ni zmanjšala zanimanja za informacijsko družbo, saj je prispevkov primerljivo z lansko konferenco, kljub temu, da se je državno sofinanciranje močno zmanjšalo. Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi tudi ACM Slovenija. Zahvaljujemo se tudi Agenciji za raziskovalno dejavnost RS ter Ministrstvu za visoko šolstvo, znanost in tehnologijo za sodelovanje in podporo. V imenu organizatorjev konference pa se želimo posebej zahvaliti udeležencem za njihove dragocene prispevke in priložnost, da z nami delijo svoje izkušnje o informacijski družbi. Zahvaljujemo se tudi recenzentom za njihovo pomoč pri recenziranju.

V letu 2010 sta se programski in organizacijski odbor odločila, da bosta podelila posebno priznanje Slovencu ali Slovenki za izjemen življenjski prispevek k razvoju in promociji informacijske družbe v našem okolju. Z večino glasov je letošnje priznanje pripadlo dr. Tomažu Kalinu. V letu 2010 tudi prvič podeljujemo nagrado za tekoče dosežke. Za aktivno delo pri računalniških tekmovanjih in drugih računalniških dogodkih sta odbora izmed predlogov izbrala Marka Grobelnika. Čestitamo obema nagrajencema!

Franc Solina, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2010

In its 13th year, the Information Society Multiconference (http://is.ijs.si) again demonstrated that is is one of the leading conferences in Central Europe gathering scientific community with a wide range of research interests in information society. In 2010, we organized ten independent excellent conferences forming the Multiconference. There are plenty of conferences in Slovenia and all over the world. The broad range of topics and the open academic environment fostering new ideas makes our event unique among similar conferences.

The Multiconference flourishes the synergy of different interdisciplinary approaches dealing with the challenges of information society. The major driving forces of the Multiconference are search and demand for new knowledge related to information, communication, and computer services. We present, analyze, and verify new discoveries in order to prepare the ground for their enrichment and development in practice. The main objective of the Multiconference is presentation and promotion of research results, to encourage their practical application in new ICT products and information services in Slovenia and also broader region.

The Multiconference is running in parallel sessions with 300 presentations of scientific papers. The papers are published in the conference proceedings, and in special issues of two journals. One of them is Informatica with its 34 years of tradition in excellent research publications.

The Information Society 2010 Multiconference consists of the following conferences:

- Confidence Project Open Workshop
- Intelligent Systems
- Languate technologies
- Cognitive Sciences
- Robotics
- Data Mining and Data Warehouses (SiKDD 2010)
- Collaboration, Software and Services in Information Society
- Demographic Challenges in Europe
- Education in Information Society
- The Third Mini Conference on Theoretical Computing 2010.

Interestinglly, the economic recession is not affecting Information society, judging from the number of single conferences; however, the national funding significantly decreased as a result of crisis. The Multiconference is co-organized and supported by several major research institutions and societies, among them ACM Slovenia, i.e. the Slovenian chapter of the ACM. We would like to express our appreciation to the Slovenian Government for cooperation and support, in particular through the Ministry of Higher Education, Science and Technology and the Slovenian Research Agency..

In 2010, the Programme and Organizing Committees decided to award one Slovenian for his/her life-long outstanding contribution to development and promotion of information society in our country. With the majority of votes, this honor went to Dr. Tomaž Kalin. Congratulations!
In addition, a reward for current achievements was pronounced for the first. It goes to Marko Grobelnik for his support of the ACM computer competitions.

On behalf of the conference organizers we would like to thank all participants for their valuable contribution and their interest in this event, and particularly the reviewers for their thorough reviews.

Franc Solina, Programme Committee Chair
Matjaž Gams, Organizing Committee Chair

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Izrael
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Karl Pribram, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Finland
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Miklós Krész, Hungary
József Békési, Hungary

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek, co-chair
Lana Jelenkovič
Jana Krivec
Mitja Lasič

Programme Committee

Franc Solina, chair
Viljan Mahnič, co-chair
Cene Bavec, co-chair
Tomaž Kalin, co-chair
Jozsef Györkös, co-chair
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Marjan Pivka
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Tomaž Šef
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Tanja Urbančič
Boštjan Vilfan
David B. Vodušek
Baldomir Zajc
Blaž Zupan
Boris Žemva
Janez Žibert
Leon Žlajpah

KAZALO / TABLE OF CONTENTS

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Inteligentni sistemi

Intelligent Systems

Uredila / Edited by

Marko Bohanec, Matjaž Gams

http://is.ijs.si

13.–14. oktober 2010 / October 13th–14th, 2010
Ljubljana, Slovenia

PREDGOVOR

Konferenca *Inteligentni sistemi* je tudi letos, tako kot vsa pretekla leta od 1997 dalje, potekala v okviru multikonference *Informacijska družba*. To ni presenetljivo, saj se konferenca ukvarja s pomembnim vidikom informacijske družbe: inteligentnimi sistemi in inteligentnimi storitvami. Ključna tema pa so programski sistemi v informacijski družbi oziroma konkretne tehnične rešitve v inteligentnih sistemih, možnosti njihove praktične uporabe, pa tudi trendi, perspektive, nujni ukrepi, prednosti in slabosti, priložnosti in nevarnosti, ki jih v informacijsko družbo prinašajo inteligentni sistemi.

Konferenca *Inteligentni sistemi* v letu 2010 ostaja mednarodna in vseslovenska; prispevki so tako v slovenskem kot angleškem jeziku. V osrednjem konferenčnem delu je predstavljenih 13 prispevkov, v ločeni sekciji pa so predstavljeni prispevki Delavnice E9. Večina prispevkov obravnava razmeroma specifične metode, ki se uporabljajo v inteligentnih sistemih; predvsem gre za metode analize podatkov in strojnega učenja, metode modeliranja in podpore pri odločanju, pristope na osnovi agentov ter metode razpoznavanja vzorcev. Da gre res za uporabne metode, pričajo prikazani primeri praktičnih aplikacij, kot so: analiza pogodb, načrtovanje in optimizacija digitalnih komunikacijskih omrežij, krmiljenje semaforjev, klasifikacija slik in modeliranje človekovega vedenja. Vsi prispevki so bili recenzirani s strani vsaj po dveh anonimnih recenzentov, prvič s pomočjo konferenčnega sistema EasyChair. Oblikovne pripombe sva prispevala tudi predsednika konference.

Marko Bohanec in Matjaž Gams, predsednika konference

PREFACE

The conference *Intelligent Systems* remains a traditional part of the multiconference *Information Society* since its beginnings in 1997. This is not surprising, because this conference addresses an important aspect of information society: intelligent computer-based systems and the corresponding intelligent services. Specifically, it addresses technical aspects of intelligent systems, their practical applications, as well as trends, perspectives, advantages and disadvantages, opportunities and threats that are being brought by intelligent systems into the information society.

The conference *Intelligent Systems 2010* remains a national and international event and presents papers written in both English and Slovenian languages. The international part of this Proceedings includes 13 papers. In a separate section, papers from the E9 Workshop are presented. Most of them present specific methods, which are used in intelligent systems, such as: data analysis and machine learning, modeling and decision support, agent systems and recognition methods. The usefulness of these methods is illustrated through a number of practical applications: analysis of contracts, design and optimization of digital communication networks, traffic light control, image classification and modeling of human behavior. All submitted papers have been reviewed by at least two anonymous reviewers, for the first time using the EasyChair conference system. Some additional suggestions for improvements were also provided by the chairmen of the conference.

Marko Bohanec and Matjaž Gams, Conference Chairs

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Marko Bohanec, *predsednik*
Institut Jožef Stefan

Matjaž Gams, *predsednik*
Institut Jožef Stefan

Tomaž Banovec
Statistični urad Republike Slovenije

Cene Bavec
Univerza na Primorskem, Fakulteta za Management Koper; IBM

Jaro Berce
Univerza v Ljubljani, Fakulteta za družbene vede

Marko Bonač,
ARNES

Ivan Bratko
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, IJS

Dušan Caf
Telekom Slovenije

Aleš Dobnikar
Center Vlade RS za informatiko

Bogdan Filipič
Institut Jožef Stefan

Nikola Guid
Univerza v Mariboru, Fakulteta za elektrotehniko računalništvo in informatiko

Borka Jerman Blažič
Institut Jožef Stefan

Tomaž Kalin
Ministrstvo za informacijsko družbo

Thiemo Krink
University of Aarhus

Marjan Krisper
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Marjan Mernik
Univerza v Mariboru, Fakulteta za elektrotehniko računalništvo in informatiko

Vladislav Rajkovič
Univerza v Mariboru, Fakulteta za organizacijske vede

Ivan Rozman
Univerza v Mariboru, Fakulteta za elektrotehniko računalništvo in informatiko

Niko Schlamberger
Informatika

Tomaž Seljak.
Ministrstvo zašolstvo, znanost in šport

Marin Silič
Center Vlade RS za informatiko

Uroš Stanič
Institut Jožef Stefan

Peter Stanovnik
Institut za ekonomska raziskovanja

Peter Tancig
Združanje raziskovalcev Slovenije

Pavle Trdan
Lek

Iztok Valenčič
Nova Kreditna Banka Maribor d.d.

Vasja Vehovar
Univerza v Ljubljani, Fakulteta za družbene vede

Boštjan Vilfan
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

PILOT POWER OPTIMIZATION IN UMTS: A MULTI-AGENT APPROACH

Lucas Benedičič [1], *Mitja Štular* [1], *Peter Korošec* [2]
[1] Mobitel Telecommunication Services Inc., Vilharjeva 23, SI-1537 Ljubljana, Slovenia
[2] Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
e-mail: lucas.benedicic@mobitel.si

ABSTRACT

In the context of coverage planning and control, the power of the CPICH signal determines the coverage area of the cell. It also impacts the network capacity, and thus the quality of service. Pilot power is the parameter that allows us to control the strength of the CPICH signal. A higher power for pilot signals means better coverage. On the other hand, more pilot power translates in less power available to serve user traffic.

We consider the problem of minimizing the total amount of pilot power subject to a full coverage constraint. Our solution approach, based on multiple autonomous agents, gives very good solutions to the problem within an acceptable amount of time. We report the results of our experiments for three UMTS networks of different sizes based on realistic planning scenarios.

1 INTRODUCTION

Coverage planning is a key problem that all mobile operators have to deal with. Its intricacy arises from the wide range of different combinations of configuration parameters and their evaluation-time complexity. One crucial parameter, which is mainly subject of adjustment, is the transmit power of the common pilot channel (CPICH). The CPICH transmit power is common to many different planning and optimization problems in UMTS networks.

The CPICH transmits in the downlink of a UMTS cell system. The transmit power is usually between 5% and 10% of the total power available at the base station [4]. From the network point of view, the capacity of a cell is limited by the amount of available power at the base station and the interference level at the mobile terminal. The coverage area of any cell is controlled by changing its pilot power, which consequently modifies the service area of the network.

There are different approaches in the literature that are able to solve the coverage problem [6, 8]. Some of them even claim to achieve near-optimal solutions [9]. As a matter of fact, such formulations have proven useful only for small network instances and often fail when challenged with real-world networks.

The idea of using autonomous agents for optimization is not new. It has proven to be a solid optimization approach for solving different types of problems, not only within the area of mobile networks [2], but also in other fields [13].

Our optimization approach is based on a state-of-the-art mathematical model, that has been previously used to solve a comparable problem [1]. We tackle the problem of computational complexity, when dealing with big problem instances, by deploying a greater number of agents working in parallel over the service area of the network. Our approach is tested on real-world UMTS networks of different sizes. The results show that the solutions found, and more importantly their quality, are greatly improved when compared to other common planning techniques.

We begin our discussion with a description of the coverage problem and formally introduce some of its key elements. We then discuss our multi-agent approach in detail, as well as the strategy used for result comparison. Having introduced the multi-agent approach, we move on to describe the simulations and experimentation done on three real-world networks. We conclude with an overview of the achieved results and discuss future research directions.

2 PROBLEM DESCRIPTION

In the problem of optimization of pilot powers for service coverage, the objective is to find a set of pilot power settings for all cells in the network, such that the total pilot power used is minimized, and a given service coverage constraint is fulfilled. We consider the pilot power minimization problem subject to a full coverage constraint of the service area.

Because the mathematical model of the problem is not of primary interest here, we will just outline it, so that all problem elements are formally defined and represented. For additional information regarding mathematical models of comparable problems, see [6].

2.1 Problem elements

We start by considering a UMTS network of m cells and use C to denote the set of cells, i.e. $C = \{1, ..., m\}$. A pixel grid of a given resolution represents the service area for which the signal propagation predictions are known. Let n denote the total number of pixels in the service area and let S denote the pixel set, i.e. $S = \{1, ..., n\}$. We also denote g_{cs}, $0 \leq g_{cs} \leq 1$, as the attenuation factor between a cell c and a pixel s, which is calculated by performing signal propagation predictions for every pair of $c \in C$ and $s \in S$.

For every $c \in C$, we define p_c^T as the total transmission power available in cell c. This power is shared among all channels in the cell (i.e. CPICH, other common channels, and dedicated traffic channels). We define p_c as the amount of power allocated to the pilot signal of cell c, where p_c may adopt any value from a finite set of possible pilot power levels, $P_c = \{p_c^1, p_c^2, ..., p_c^T\}$. Consequently, the received pilot power of cell c in pixel s is $g_{cs}p_c$.

Considering the full coverage constraint, each pixel in the service area should have at least one cell covering it. We assume that a pixel s is under coverage of a cell c if its carrier-to-interference ratio, E_c/I_o at pixel s, is not lower than a given threshold, γ_c, i.e.

$$E_c/I_o(c,s) = \frac{g_{cs}p_c}{\sum_{i \in C} p_i^T g_{is} + \tau_0} \geq \gamma_c \qquad (1)$$

where τ_0 is the thermal noise. In (1), we are assuming that all cells in the network operate at full power, which is the worst case scenario. The same assumption has been also used in [1, 9].

The optimization problem corresponds to finding the pilot power levels p_c, for all cells $c \in C$, such that coverage of at least b pixels is guaranteed, while the total amount of pilot power used is minimized. Since we are considering full coverage, we denote $b = n$.

2.2 Optimization objective and constraints

The optimization objective is defined as follows

$$P^* = \min \sum_{c \in C} p_c; \qquad (2)$$

subject to

$$\frac{\sum_{s \in S} cov(s)}{b} = 1, \qquad (3)$$

where

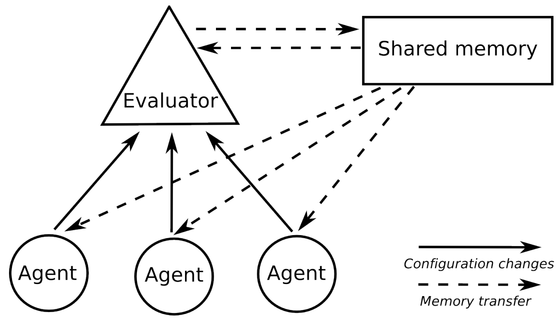
$$cov(s) = \begin{cases} 1 & \exists c \mid E_c/I_o(c,s) \geq \gamma_c \\ 0 & otherwise \end{cases} \qquad (4)$$

The definition of (4) provides us with a simple way of asserting the coverage of a given pixel, s. It follows that if the pilot signal of at least one cell c satisfies the imposed E_c/I_o threshold, γ_c, the pixel is covered and hence $cov(s) = 1$.

3 OPTIMIZATION APPROACHES

The comparison of the experimental results involve two different strategies for setting the pilot power. The first strategy is uniform pilot power, presented in [12], by which the pilot power of all cells in the network is set to the same level. The second strategy is our multi-agent approach, based on ideas inspired by two-dimensional cellular automata [7] and metaheuristics [10].

Figure 1: *Architecture of the multi-agent system.*



3.1 Uniform pilot power

This kind of pilot power setting is efficient in scenarios where the signal attenuation is essentially determined by distance (i.e. the signal propagation conditions are mostly constant). In such cases, the cell coverage areas will be approximately the same for uniformly distributed traffic and similarly-distributed site locations. On the other hand, when dealing with heterogeneous scenarios, having all cells equally set to the same pilot power, results in an unnecessary large power consumption. This not only reduces the power available to user traffic, but it also produces areas of high pilot pollution [4].

3.2 Multiple autonomous agents

In the multi-agent approach, a set of autonomous agents and an evaluator work together in order to optimize the pilot power consumption of the network. Each agent randomly moves over the service area as it proposes different changes to the pilot power of the cells. The quality of the final solution gradually improves over time, as better solutions are accepted and poor solutions are discarded by the evaluator.

Figure 1 gives an overview of the multi-agent architecture. Within this architecture, agents work in an autonomous and asynchronous manner.

3.2.1 The agents

Each agent encapsulates a set of steps that is consistently applied as it randomly moves through the service area of the network. Whenever an agent arrives at a pixel s, it asks the evaluator for the set of cells covering the current pixel, namely

$$B(s) = \{c \in C \mid E_c/I_o(c,s) \geq \gamma_c\} \qquad (5)$$

The step set applied from this point on depends on the cardinality of $B(s)$. The agent's behavior is regulated by the pseudo-code shown in Table 1. Hence, the agent applies step sets SS_0 and SS_1 based on the number of cells in $B(s)$.

If the agent's current location, at pixel s, is not covered by any cell (i.e. $|B(s)| = 0$), the step set SS_0 (shown in Table 2) is applied. It starts by defining N, the Moore neighborhood with radius 1 [7], around

Table 1: *Pseudo-code of the agent's behavior.*

Step	
	repeat
1	**if** $\|B(s)\| = 0$ **then**
2	*apply* SS_0
3	**else if** $\|B(s)\| = 1$ **then**
4	*Move randomly*
5	**else if** $\|B(s)\| > 1$ **then**
6	*apply* SS_1
	end if
	end repeat

Table 2: *Pseudo-code of step set* SS_0.

Step	
1	$N = Moore(s, radius = 1)$
2	$A = \emptyset$
3	**for each** *pixel* $r \in N$
4	$A = A \cup B(r)$
	end for
5	**if** $A \neq \emptyset$ **then**
6	$c' = Highest(A)$
7	*Adjust pilot*$(c', increase\ rate)$
	end if
8	*Move one step randomly*

the agent's current location s. The neighborhood is then analyzed, looking for cells covering any of the neighboring pixels $r \in N$ (steps 3 and 4). If a cell is found, the agent shall propose an increase in the pilot power of the cell c', that has the highest E_c/I_o in the neighborhood N (steps 6 and 7). Otherwise, if no cell is found covering any pixel within the neighborhood N, the agent moves one pixel in a random direction (step 8). It should be noted that the agent moves exclusively to adjacent pixels $r \in N$ when it reaches an uncovered pixel s. This behavior efficiently finds areas without coverage containing many uncovered pixels.

The step set SS_1 in Table 3 is applied whenever the agent's current location, at pixel s, is covered by more than one cell (i.e. $|B(s)| > 1$). The first step distinguishes the covering cell with the lowest E_c/I_o, c', from $B(s)$. The agent shall propose a decrease in the pilot power of c', and then move to a new random location (steps 2 and 3), that is generally not adjacent to the current one.

In both step sets, SS_0 and SS_1, the values *increase rate* and *decrease rate* are configuration parameters that should be given before starting the optimization process. They indicate the dB adjustment proposed to the pilot power of cell c'.

Table 3: *Pseudo-code of step set* SS_1.

Step	
1	$c' = Lowest(B(s))$
2	*Adjust pilot*$(c', decrease\ rate)$
3	*Move randomly*

3.2.2 *The evaluator*

The evaluator represents the global optimization component of the system, since it has access to the complete network, for it keeps track of the coverage area of each cell. In contrast, the agents propose their changes based exclusively on local information.

A tabu list is included in the evaluator as a mechanism to avoid solution degradation. A cell is included in the tabu list after its pilot power has been increased. While a cell is in the tabu list, any decrease of its pilot power is forbidden. The tabu list keeps track of the number of overall agents' moves since the cell was included in it. If the pilot power of a tabu cell is increased, its counter is reset to zero. By using a tabu list, the evaluator avoids coverage reduction within the context of the current solution, since increasing pilot power only occurs when uncovered pixels are discovered by the agents. The number of moves a cell should remain in the tabu list is an optimization parameter.

All changes proposed by the agents are subject to acceptance by the evaluator, that sequentially verifies the following conditions to each proposed change:

- if the change involves decreasing the pilot power setting of a cell c', then c' should not be an element of the tabu list;
- if the change involves increasing the pilot power setting of a cell c', then c' becomes/remains an element of the tabu list and its counter is set to zero;
- the new pilot power setting, after applying the agent's change, is an element of $P_{c'}$.

If the agent's proposed pilot-power adjustment breaks any of the enumerated conditions, the change is discarded without notice. This means that the agents do not know whether the proposed change has been applied or not.

4 SIMULATIONS

4.1 Test networks

All of our test networks, Net_1 of the city of Berlin (Germany), and Net_2 of the city of The Hague (Netherlands), are part of the publicly available MOMENTUM project [5]. The last network, Net'_1, is a reduced version of Net_1 of the city of Berlin. This same network has been optimized in [9]. All networks include information about site locations, path loss predictions and realistic antennas. These scenarios also contain individual requirements for E_c and E_c/I_o coverage.

Based on the data available, we have produced network configurations based on the uniform pilot power approach. These configurations represent what could be an initial network setup by common planning standards [4]. Table 4 shows some statistics about the test networks used. The parameter values used during experimentation are shown in Table 5.

Table 4: *Network statistics.*

	Sites	Cells(m)	Pixels(n)	Area	Pixel size
Net_1	65	193	22500	7500×7500	50×50
Net_2	12	36	6400	4000×4000	50×50
Net'_1	50	148	22500	7500×7500	50×50

Table 5: *Network parameters.*

Parameter	Net_1	Net_2	Net'_1
p_c^T	19.95W	19.95W	19.95W
τ_0	$1.55{\cdot}10^{-14}$W	$1.55{\cdot}10^{-14}$W	$1.55{\cdot}10^{-14}$W
γ_c	-15dB	-15dB	-15dB

4.2 Results

After considerable experimentation, we found out that a tabu list value of $\frac{n}{2}$ gives good solutions regarding power use, while keeping coverage mostly at 100%. Experimentation also gave us valuable understanding regarding the agents' behavior, and helped us set the values of *increase rate* and *decrease rate*. By setting the *decrease rate* at 0.1dB, the agents are able to find possible uncovered areas quickly. On the other hand, increasing pilot power at *increase rate* 1dB, lets the agents correct uncovered areas containing many uncovered pixels with one change only. Using this setup, coverage values never dropped under 98% during experimentation. The stopping criterion was set by limiting the total number of agents' moves. This value was set to $5n$, even though the best solutions were always found in the first quarter of the experiment. Table 6 shows the optimization parameters used for each network.

All things considered, the results achieved by our optimization approach improved the objective significantly, as it is shown in Table 7. We reduced pilot power usage in all networks and kept the service area under full coverage. Moreover, when comparing our results of network Net'_1 to those of [9], we may see that our solution achieves full coverage with even lower pilot power usage.

5 CONCLUSION

In this paper, we have addressed the problem of providing full coverage to a service area of a UMTS network by using a minimum amount of pilot power. We have introduced a multi-agent approach aimed at giving good solutions to the problem in an acceptable amount of time. The experimental results show that our approach is able to find competitive solutions, when compared to other methods from the literature [9]. The presented results also demonstrate that our algorithm is able to find high quality solutions even for large networks, that are out-of-reach for some exact techniques, like linear programming. Such solutions were found in a reduced amount of time by simply increasing the number of agents deployed during optimization. This fact reveals that our approach can be easily applied to bigger problem instances without compromising solution quality.

In any case, it would be useful to make a more extensive comparison of our experimental results with different algorithms and various heuristic and exact approaches. However, this task is not straightforward, since the results of several works (e.g. [3, 11]) depend on black-box evaluations, making experimental association very difficult, if possible at all.

Table 6: *Optimization parameters used.*

	Agents	Incr. rate	Decr. rate	Tabu value	Stopping criterion
Net_1	6	1dB	0.1dB	11250	112500
Net_2	2	1dB	0.1dB	3200	32000
Net'_1	6	1dB	0.1dB	11250	112500

Table 7: *Optimization results.*

	Uniform		Multi-agent	
	Total power (W)	Average power (W)	Total power (W)	Average power (W)
Net_1	422.226	2.1877	147.490	0.76412
Net_2	120.823	3.3562	32.621	0.9061
Net'_1	345.092	2.3317	112.125	0.7576

Acknowledgement

This project was co-financed by the European Union, through the European Social Fund.

References

[1] L. Chen and D. Yuan. Automated planning of CPICH power for enhancing HSDPA performance at cell edges with preserved control of R99 soft handover. *Proceedings of IEEE ICC'08*, 2008.

[2] G. Cheung, W. Tan, and T. Yoshimura. Real-time video transport optimization using streaming agent over 3G wireless networks. *IEEE transactions on multimedia*, 7(4):777, 2005.

[3] A. Gerdenitsch. *System capacity optimization of UMTS FDD networks.* Ph.D. dissertation, Technische Universität Wien, 2004.

[4] H. Holma and A. Toskala. *WCDMA for UMTS: Radio access for third generation mobile communications, Third Edition.* John Wiley & Sons, Inc. New York, NY, USA, 2005.

[5] IST-MOMENTUM. Models and Simulations for Network planning and Control of UMTS. `http://momentum.zib.de`, [Online; accessed 15-May-2010].

[6] M. Nawrocki, H. Aghvami, and M. Dohler. *Understanding UMTS radio network modelling, planning and automated optimisation: theory and practice.* John Wiley & Sons, 2006.

[7] P. Sarkar. A brief history of cellular automata. *ACM Computing Surveys (CSUR)*, 32(1):107, 2000.

[8] I. Siomina and D. Yuan. Pilot power optimization in WCDMA networks. 4, March 2004.

[9] Iana Siomina and Di Yuan. Minimum pilot power for service coverage in WCDMA networks. *Wireless Networks*, 14(3): 393–402, June 2007.

[10] E.G. Talbi. *Metaheuristics: from design to implementation.* Wiley, 2009.

[11] U. Türke and M. Koonert. Advanced site configuration techniques for automatic UMTS radio network design. *Proc IEEE VTC 2005 Spring*, 2005.

[12] P. Värbrand and D. Yuan. A mathematical programming approach for pilot power optimization in WCDMA networks. 2003.

[13] M. Vasile and M. Locatelli. A hybrid multiagent approach for global trajectory optimization. *Journal of Global Optimization*, 44(4):461–479, 2009.

DIFFERENT ALGORITHMS FOR TRAFFIC LIGHT TIMING AT URBAN INTERSECTION

Zoran Gacovski, Metodija Angelkoski, Ivan Kraljevski

FON University, Bul. Vojvodina, bb, Skopje, Macedonia
tel. 389 22 445-593, fax 389 22 445-550, gacovski@mt.net.mk

Abstract. *This paper presents three methods for design of traffic signal controller for an isolated intersection. The controller has the ability to make adjustments to signal timing in response to observed changes in the approach flows. Using random incoming traffic, the controller measures approach flows and estimates approach queues at regular time intervals. This information is an input in 3 different algorithms: equal-timing (pre-defined) control, statistically-based timing, and fuzzy logic procedure which decide whether to extend or terminate the current signal phase for through movements. In the first stage, the controller estimates the traffic intensity on each approach. The duration of the green is based on traffic-actuated control. We have realised simulation in Visual Studio and the results show that the fuzzy approach has significant advantage over other methods.*

Keywords: Intelligent transportation systems, Traffic light control, Fuzzy logic, Statistical model.

1. INTRODUCTION

With the development of the urban life and the growth of the automotive industry, the traffic takes an important place in the society and everyday life. This is especially valid for large cities, where the problem of vehicle flow has significant impact to the air pollution, loudness and noise problems, time waste, energy spending etc. Intelligent transportation systems, indicating the application of modern information and communication technologies in transportation systems (to increase the efficiency and safety of transportation systems and decrease air pollution and its other undesirable environmental effects), are generally composed of three important components i.e. a sensor (Loop detector), an information processor and an output device connected through a communication network. Intelligent transportation systems can be categorized into different groups, of which intelligent control systems of intersections belong to the class of advanced traffic management systems [10]. Following the appearance of computerized traffic lights in 1960s, many researchers designed traffic light control systems which were capable of coordinating the traffic lights so that at least one of the parameters e.g. the number of stops or the delay at reaching the destination would be minimized by processing some information on the current traffic conditions . In the 1980s, the introduction of SCOOT system in Great Britain and SCATS in Australia made a major breakthrough in control systems. UTCS (Urban Traffic Control Systems) has been employed in North America [4], as well as SCOOT and SCATS (Sydney Coordinated Adaptive Traffic System) in Australia, Europe, Asia and recently North America.

Traffic light is doubtlessly the most familiar, important and effective method of traffic control at intersections. Traffic lights are generally installed to ensure safety, decrease the average time of proceeding through the intersection, increase the capacity of multileg intersections, improve quality of service, quality of traffic flow and level of service for all or most traffic streams and if scheduled accurately the average delay of vehicles will be less, compared to unsignalized intersections [1]. Traffic situation, tightly tied to the cultural and social paradigms is a fuzzy concept itself. The sophistication of the real world aggravates its accurate description and definition. Despite the simple look of city intersections, they belong to this sophisticated world and thus cannot be controlled neglecting this feature. In this paper we will first study the intelligent traffic control systems and introduce the customary methods of timing control of traffic lights. Following the introduction of Fuzzy control systems, we will present the Fuzzy control model of traffic lights timing at an urban intersection and evaluate the results. In a conventional traffic light controller, the lights change at constant cycle time, which is clearly not the optimal solution. Fuzzy logic can be a better method than conventional control methods, especially in the case of highly uneven traffic flow between different directions [8], [5]. It would be more feasible to let more cars pass at the green light if there is less number of cars behind the red lights. A mathematical model for this decision is difficult to find but fuzzy logic simplifies the task. Once well working and appropriate rules are formulated for our four way intersection, it is not too difficult to modify the fuzzy rule base so that the program can be applied to any given intersection.

The paper is organized as follows – first we explain the equal-timing and statistically based control of traffic lights timing. Then, fuzzy logic control is introduced and applied, and simulation results are given. Finally we give a conclusion and directions for a future work.

2. EQUAL-TIMING AND STATISTICALLY BASED TIMING

Traffic light timing using the incoming traffic conditions can be accomplished in different ways. In the pre-timed mode, each phase period and cycle duration is determined based on some predetermined values by some statistics. In traffic prediction (Actuated Signals), the future mode is estimated and decided by sensors based on the measured situation. In the pattern matching method, the information obtained by the sensors is adapted by a set of mathematical operations with the existing information, the closest pattern to the current conditions is then selected and appropriate time values are applied to the traffic lights accordingly.

Traffic light optimization is a complex problem. Even for single junctions there might be no obvious optimal solution. With multiple junctions, the problem becomes even more complex, as the state of one light influences the flow of traffic towards many other lights. Another complication is the fact that flow of traffic constantly changes, depending on the time of day, the day of the week, and the time of year. Roadwork and accidents further influence complexity and performance. In practice most traffic lights are controlled by fixed-cycle controllers. A cycle of configurations is defined in which all traffic gets a green light at some point. The split time determines for how long the lights should stay in each state. Busy roads can get preference by adjusting the split time. The cycle time is the duration of a complete cycle. In crowded traffic, longer cycles lead to better performance. The offset of a cycle defines the starting time of a cycle relative to other traffic lights. Offset can be adjusted to let several lights cooperate, and for example create green waves.

Fixed controllers have to be adapted to the specific situation to perform well. Often a table of time-specific settings is used to enable a light to adapt to recurring events like rush hour traffic. Setting the control parameters for fixed controllers is a lot of work, and controllers have to be updated regularly due to changes in traffic situation. Unique events cannot be handled well, since they require a lot of manual changes to the system. Fixed controllers could respond to arriving traffic by starting a cycle only when traffic is present, but such vehicle actuated controllers still require lots of fine-tuning.

Most research in traffic light control focuses on adapting the duration or the order of the control cycle. In our approach we do not use cycles, but let the decision depend on the actual traffic situation around a junction, which can lead to much more accurate control. Of course, our approach requests that information about the actual traffic situation can be obtained by using different sensors or communication systems. Statistically-based timing is used after set of measurements at the particular intersection and making a database with a number of vehicles during the different parts of the day.

The variable that should be minimized is the average waiting time of the vehicles. So, we have obtained the two input parameters of the system:

1. The timing for the whole cycle;
2. The density of vehicles on all 4 sides at the intersection.

The output parameter will be the duration of „green light“, which will determine the duration of the red period on the other three sides of the intersection.

In the semi-actuated control mode - all times for different routes excluding the main route can be set pre-timed. The traffic light at the main line remains green as long as the sensors of the off-line can detect a car at the intersection. But in the full-actuated control mode, all the times of the conflicting volumes can be programmed by sensors. Full-actuated control is mostly employed where the traffic volumes of both intersecting lines are approximately equal. Full actuated control is used here to predict the future traffic flow by conforming to the designed fuzzy functions.

Obviously, a minimum amount of knowledge is required for the practical implementation of this system; the explicit and tacit knowledge can be respectively obtained from a series of minor computations on the statistical output of some traffic control software systems, and through people-to-document approach in codification strategy for knowledge management. This technique is mostly applied to cases facing similar problems and requiring reuse of a validated solution. Efforts are made to reveal and code the hidden knowledge of people and eventually store it in knowledge databases to act as a reference for similar future attempts.

But in this paper, due to lack of the tools and appropriate statistical data, required cases have been specified approximately and subjectively having no repercussion on the outcome of the system.

3. FUZZY LOGIC CONTOLLED TIMING

Signal timing control, considering the native features of driving in the city (Skopje), has been designed by taking two parameters in mind: back of queue length, as the maximum extent of the queue in give-way lines at red time according to the number of stops, and average waiting time along the route. These parameters are calculated at red time which provides static traffic conditions and not during green display with dynamic traffic; the results are then applied in the next green phase. Therefore the resulting values of the parameters are more acceptable and usable since the route traffic conditions and obstacles are ineffective on the values of the parameters. As opposed to similar preceding systems, which merely involved the parameters of one approach at green time, intelligent timing control of intersections through this method is done by taking the parameters of both approaches with the offset of a cycle, so that peak hour coefficient, main and minor streets, and the capacities of the lines of a given intersection - have been implicitly incorporated in the parameters of this system. In a way that peak hour is when the queue length is Max, at least one of the approaches reaches its maximum; and as a result, this system allows the maximum extension to the green phase [11].

Involving the queue length in place of the traffic volume causes more traffic volume during green time, depending on the width of the two approaches in the main street i.e. the street with more lanes. But the importance and capacity of an approach are not taken into account in the parameter of traffic volume and number of vehicles.

To compare and evaluate the traffic effect of different vehicles, passenger cars are usually chosen as the unit of measurement and vehicle traffic streams are converted to an equivalent passenger-car volume in measuring the capacities of intersections and queue length based on the number of vehicles. These coefficients are multiplied by 1.75 when used for left-turn adjustment factors. Determining the average waiting time parameter in each approach requires the definition of a function that represents the total waiting time of all vehicles entering the intersection during the period t . For this purpose function F(t) can be stated in short discrete intervals (e.g. 5 sec) and the product of the number of vehicles N entering the intersection during the remaining time to the end of red phase based on the previous cycle.

$$F(t_i) = N(i) * (TR_i - 1 - ti)$$

In order to determine the traffic flow in each direction, two incremental sensors are to be placed in each direction at a distance of 60 meters from each other, for a total of eight sensors. The number of cars that is accumulated behind a traffic light is calculated by taking the difference between the two sensor readings.

The input of this model consist of:

1. Current cycle time of light.
2. Accumulation of cars behind the Red Light on more crowded street side.
3. Accumulation of cars behind the Green Light on more crowded street side.

The output parameter is the probability of change of the current cycle time.

The input and the output parameters are defined by overlapping linear membership functions.

The knowledge base consists of 40 fuzzy rules, as for example:

1. If traffic accumulation behind red is very minimal and green is very minimal and cycle time is short then change is NO.
2. If traffic accumulation behind red is moderate and green is moderate and cycle time is long then change is Probably YES.

Some of the rules have been evaluated by using the fuzzy logic formulas:

$\mu_{B'}(y) = \max_{x1\ldots xn} [\min(\mu_{A1}(x_1)\ldots\mu_{An}(x_n)), \mu_{A1\ldots An\Rightarrow B}(x_1,\ldots, x_n, y)]$

$\mu_{A1\ldots An\Rightarrow B}(x_1,\ldots, x_n, y) = \max((1-\min(\mu_{A1}(x_1)\ldots\mu_{An}(x_n)), \mu_B(y))$

Test cases are being generated and the error value for n test cases is calculated using the formula:

$$\Sigma\ 1/n * (\text{simulated output} - \text{desired output})^2$$

The error value is very small and negligible, which allows us to conclude that the rules are strong and have very small error values.

Waiting time parameter is considered in the interval of [0 200] with three membership functions i.e. low [0 0 100], medium [0 100 200] and high [100 200 200] and queue length parameter is taken into account in the interval of [0 200] with 3 membership functions i.e. low in [0 0 25 75], medium in [25 75 125] and high in [75 125 200 200]. Consequently, inference rules and membership functions are designed depending on the system input in every moment, appropriate fuzzy results are obtained for green time variable in the interval of [-200 200] seconds for every route and therefore the red time for the opposing approach in their pertaining phases. On the basis of that, appropriate decision is made after center of gravity defuzzification for selecting any of the membership functions of decrease plus in the interval of [-200 -200 -100], decrease in the interval of [-200 -100 0], no change in the interval of [-100 0 100], increase in the interval of [0 100 200], increase plus in the interval of [100 200 200].

Therefore, this technique operates on the basis of changes in traffic flow conditions in this interval.

By utilizing the proposed fuzzy model, dramatic improvement to shortening the average waiting time and queue length, shown respectively in figures 2 and 3 has been observed which explains the high efficiency of the proposed model.

4. SIMULATION RESULTS

We have simulated several different conditions on a single intersection. The first case is for a dense traffic (50-100 vehicles/minute) and slight change of intensity (±10%). The second case covers the situation with very dense traffic (>100 vehicles/minute) and rapid change of intensity in both directions (±50%), while the third situation is for average dense traffic (<50 vehicles/minute) and rapid change of intensity in both directions (±50%). The simulation results are given on the following figures.

Table 1 – Waiting time for dense traffic and slight change of intensity (±10%)

Time past	Equal-time logic	Statistic logic	Fuzzy logic controller
after 60 s.	17.6 sec	17.1 sec	13.0 sec
after 120 s.	27.7 sec	26.8 sec	21.9 sec
after 180 s.	33.6 sec	31.1 sec	27.1 sec
after 240 s.	37.1 sec	35.2 sec	28.2 sec
after 300 s.	-	-	-

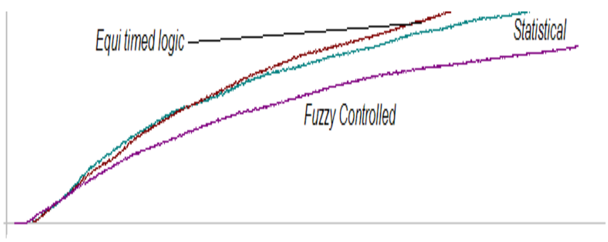


Fig. 1. *Waiting time for dense traffic (50-100 vehicles/ minute) and slight change of intensity (±10%)*

Table 2 – Waiting time for very dense traffic and rapid change of intensity (±50%)

Time past	Equal-time logic	Statistical logic	Fuzzy controller
after 60 s.	17.5 sec	15.1 sec	14.0 sec
after 120 s.	23.3 sec	22.1 sec	22.9 sec
after 180 s.	26.6 sec	24.0 sec	30.0 sec
after 240 s.	29.0 sec	27.9 sec	30.3 sec
after 300 s.	-	-	-

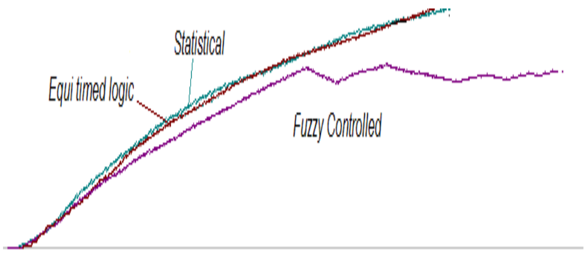


Fig. 2. *Waiting time for very dense traffic (>100 vehicles/minute) and rapid change of intensity in both directions (±50%)*

Table 3 – Waiting time for average dense traffic and rapid change of intensity (±50%)

Time past	Equal-time logic	Statistical logic	Fuzzy controller
after 60 s.	17.4 sec	17.1 sec	7.8 sec
after 120 s.	26.6 sec	28.4 sec	14.1 sec
after 180 s.	33.4 sec	35.6 sec	17.2 sec
after 240 s.	36.3 sec	39.9 sec	17.1 sec
after 300 s.	-	-	-

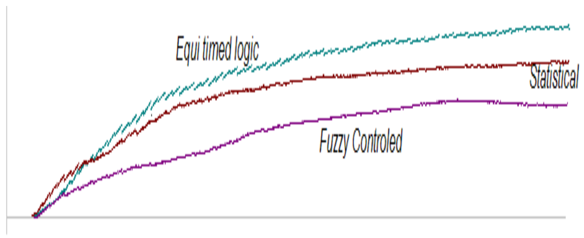


Fig. 3. *Waiting time for average dense traffic (<50 vehicles/minute) and rapid change of intensity in both directions (±50%)*

5. CONCLUSIONS

In this paper we have investigated the traffic light timing, which is the most familiar, important and effective method of traffic control at intersections. Traffic lights are generally installed to ensure safety, decrease the average time of proceeding through the intersection, increase the capacity of multileg intersections, improve quality of service, quality of traffic flow and level of service for all or most traffic streams and if scheduled accurately the average delay of vehicles will be less, compared to unsignalized intersections.

Using random incoming traffic, the controller measures approach flows and estimates approach queues at regular time intervals. This information is an input in 3 different algorithms: equal-timing (pre-defined) control, statistically-based timing, and fuzzy logic procedure which decide whether to extend or terminate the current signal phase for through movements. In the first stage, the controller estimates the traffic intensity on each approach. The duration of the green is based on traffic-actuated control. We have realised simulation in Visual Studio and the results show that the fuzzy approach has significant advantage over other methods.

Future improvements can be done by applying the intelligent agent technology, i.e. not only modeling a single isolated intersection, but several intersections combined. Also, we can apply wireless sensor networks, or RFID/ GSM technologies in future.

REFERENCES

[1] Gartner N.H., 1983. OPAC: A demand-responsive strategy for traffic signal control. *Transportation Research Record, No. 906*, Washington DC, pp. 75–81.

[2] Nakatsuyama M., Nagahashi H., Nishizuka N., 1984. Fuzzy logic phase controller for traffic junctions in the oneway arterial road. In: *Proceedings of the IFAC Ninth Triennial World Conference. Pergamon Press*, Oxford, pp. 2865–2870.

[3] Chiu S., 1992. Adaptive traffic signal control using fuzzy logic. In: *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 98–107.

[4] Kelsey R.L., Bisset K.R., 1993. Simulation of traffic flow and control using fuzzy and conventional methods. In: *Jamshidi, M. (Ed.), Fuzzy Logic and Control: Software and Hardware Applications*, *Prentice-Hall, Englewood Cliffs*, NJ, pp. 262–278.

[5] Chiu S., and Chand S., Rockwell International Science Center, 1993. Self-Organizing Traffic Control via Fuzzy Logic, *Proc. 32nd IEEE Conf. On Decision & Control*, San Antonio, Texas, pp. 1897- 1902.

[6] Kronborg P., Davidson F., 1993. MOVA and LHOVRA: Traffic signal control for isolated intersections. *Traffic Engineering and Control*, pp. 195–200.

[7] Driankov D., Hellendoorn H. and Reinfrank M., 1996. An Introduction to Fuzzy Control, *Springer Ed.*

[8] Lin Q., Kwan B.W., and Tung L.J., 1997. Traffic Signal Control Using Fuzzy Logic, *Proc. IEEE Int. Conf. on Fuzzy Syst. 1997,* pp 1644-1649.

[9] Murat Y.S., 2001. Fuzzy Logic Traffic Responsive Signal Cycle Model for Isolated Signalized Intersections, *Ph.D. Thesis, Istanbul Technical University,* Turkey, 179 p.

[10] Mancinelli E., Cohen G., Gaubert S., Quadrat J. P., and Rofman E., 2001. On Traffic Light Control of Regular Towns, *Institut National de Recherche en Informatique et en Automatique*, pp. 342-347.

[11] Azimirad E., Pariz N., Sistani M.B.N., 2010. A Novel Fuzzy control model of Traffic light timing. *IEEE Jour. of Systems, Vol. 4, no. 1*, pp. 107-111.

USING ACCELEROMETERS TO IMPROVE POSITION-BASED ACTIVITY RECOGNITION

Dmitry Gimon, Hristijan Gjoreski, Boštjan Kaluža, Matjaž Gams

Jožef Stefan Institute, Dept. of Intelligent Systems,
Jamova 39, 1000 Ljubljana, Slovenia.

Email: dmitry.guimon@gmail.com, hristijan_g@yahoo.com,
{bostjan.kaluza, matjaz.gams}@ijs.si

Abstract. This paper presents the results of the research conducted on wireless accelerometers in fall detection and activity recognition. This research is a part of the Confidence project, whose goal is to provide a health monitoring system for the elderly. Normally, position-based body tags are used to detect postures and activities. This paper reports the results of using accelerometers as both a supplement to and substitute of the position tags. It introduces a combined approach based on machine-learning and wave analysis. Preliminary results indicate an important increase in activity recognition accuracy when using both acceleration and position tags.

1 INTRODUCTION

The research presented in this paper is a part of the Confidence project [1], which aims to create a remote care system to detect health problems of the elderly by monitoring their posture and activities. Currently, wearable position tags are in use for behavior analysis in the project. These tags provide three-dimensional coordinates in space and time. Using a real time stream of input data, filtering and machine learning [2], the system classifies activities and detects alarming situations, such as a slip or a fall. However, position tags are problematic for several reasons. First, they are expensive. Second, the data from position tags are noisy and require a set of filters, which cause a delay of the system. The goal of our research was to investigate the possibility and advantages of using a relatively cheap and precise accelerometer tag system together with the position tag system. Two areas of behavior analysis, fall detection and activity recognition, were explored.

Accelerometers are wireless wearable devices that provide three-axes projection of the acceleration vector with respect to the tag body. When the accelerometer does not move, the measured acceleration is equal to gravity and the angle to the horizon can be calculated. Otherwise, the obtained value is the sum of the acceleration vector and gravity. Accelerometers used in our research provided precise inertial data approximately seven times per second. Let us call a set of acceleration projections and the time of the measurement $(a_{x,i}, a_{y,i}, a_{z,i}, t_i)$ a snapshot *i*. In our experiments we used up to four accelerometers attached to chest, belt and ankles. The most significant results were achieved by analyzing chest tag data only.

Since activity recognition is a typical classification problem, the goal of the system therefore is to classify every snapshot coming in from the sensors. In the Confidence project, the following activities are distinguished: walking/standing, sitting, lying, the process of sitting down, the process of lying down and falling [2]. In our research we did not use machine learning classification for fall detection because the very short duration of this activity could be confused with sitting or lying, and the importance of being able to correctly detect falls requires a more accurate classification.

2 FALL DETECTION

In the Confidence project falls are defined as lying on the floor for a certain time [4]. While the analysis allows one to detect the actual moment of a fall by analyzing the length of the acceleration vector, it does not provide information about the location of the person. A graph of the length of the acceleration vector in time shows a low value (actual fall) followed by a high value that is the impact of the surface (such as a floor, bed, etc.). We decided to measure the second component because a person can sit or lie down fast, while an impact without control over the body can show abnormal activity better.

Formalization of the inertial fall detection was the following. If $|A| = \sqrt{a_x^2 + a_y^2 + a_z^2} > A_{max}$, then a fall is detected, where the threshold value A_{max} was obtained empirically.

In our research, 103 tests of 4 people falling were performed in order to compare the results of using position and acceleration tags. Using inertial data allows the system to react faster while the accuracy of the fall detection is relatively low (76 % of falls were detected including falls on the bed, compared to 100 % detection with position tags). More promising approach is to evaluate position-based fall detection with inertial data to define the pace of going down.

3 MOVING FILTER

An important step in classifying activities is the ability to detect whether the person is moving. It helps to distinguish between static activities (e.g. lying, sitting) and dynamic ones (e.g. standing up, going down). The moving detection used in our approach uses the data sent by the chest accelerometer only, and it is very similar to the one introduced in [3]. The change of differences in the acceleration vector length from one obtained measurement to another is used to define moving detection as well as an attribute for machine learning. Mathematical formalization of this principle is represented as:

$$M=\frac{\sum_{i=0}^{n-2}|a_{i+1}^{dif}-a_{i}^{dif}|}{t_n-t_0}, where \qquad (1)$$

$$a_i^{dif}=\sqrt{a_{x,i+1}^2+a_{y,i+1}^2+a_{z,i+1}^2}-\sqrt{a_{x,i}^2+a_{y,i}^2+a_{z,i}^2}$$

Empirically, n=15 was chosen. The time interval is used to preserve the accuracy of the calculation if the event data is lost. If $M>P$, where P-is a threshold value, then the tag is moving; otherwise it is considered as static.

4 ACTIVITY CLASSIFICATION USING MACHINE LEARNING

The activity recognition described in this paper is a process that can be explained in several phases shown in Fig. 1.

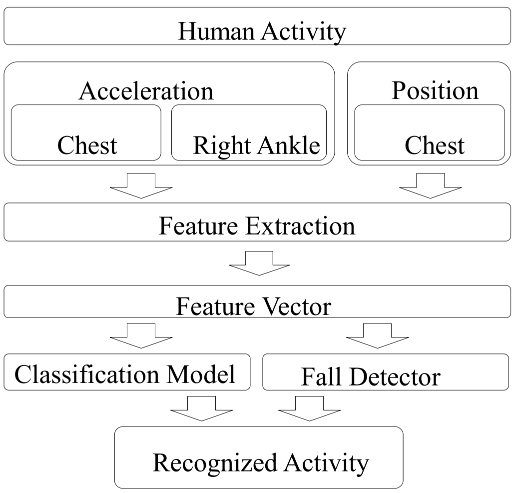


Fig.1. *Steps of activity recognition.*

First, the sensors send the raw data to the system. The software then analyzes the raw data and extracts 22 attributes (features). There are 20 attributes based on acceleration data of the chest sensor, including lengths of the acceleration vectors, directions of the acceleration vectors (angles), and changes in the difference of the acceleration vector length. The direction of the acceleration vector was calculated as an angle between the acceleration vector and one of the axes (here, the y-axis). We also calculated traditional statistical attributes using the given data such as: mean values, standard deviation, and root mean square. The binary attribute for moving was used in the classification procedure as well.

Besides inertial data, 2 attributes of the position data were used. The first attribute is the value of the z-coordinate, which is the height of the sensor. This made activities occurring at different heights easily distinguishable. The other attribute is the difference between the current and previous values of the z-coordinate. This attribute is helpful in recognizing the direction and the speed of the tag's movement along the vertical axis. Additional low pass filtering for the position attributes is used because of the position sensors' inaccuracy (about 10-15 cm).

After the process of feature extraction is complete, all calculated attributes are collected together in one feature vector. This vector is analyzed with two modules: the classification model and fall detector. If a fall is not detected, then the machine learning classification model is used to recognize activities. For machine learning, we used a built-in freeware Weka toolkit [5].

We used a longer training sequence of snapshots (approximately 7000 instances) recorded during a person's performed various activities to create the classification model and then test it with a shorter one (2000 instances). We tried to obtain an equivalent distribution of the activities in both sequences. However, some activities are very short compared to others (e.g. standing up and lying) and their occurrence is relatively rare.

Our results are presented in Table 1. The accuracy of the classifiers for each activity together with the overall accuracy are shown in this table. The best overall accuracy was archived with the Weka implementation of Support Vector Machine called Sequential Minimal Optimization (SMO) and Random Forest (94.07 % and 93.13 % respectively). The results as shown in Table 1 indicate the high importance of using position attributes to recognize certain activities. As expected, sitting and standing could not be classified correctly without the position data, because both acceleration vector length (absolute value of gravity) and the angles of gravity (orientation of the chest) are the same for these activities.

	Acceleration	**Acceleration and position**	
	Random forest		**SMO-PolyKernel**
Standing	97.28%	99.50%	99.63%
Going down	20.22%	48.24%	52.87%
Sitting	64.91%	89.26%	94.74%
Standing up	40.94%	64.60%	68.00%
Lying	91.46%	99.79%	99.58%
Sitting on the ground	89.76%	92.20%	95.59%
On all fours	98.55%	100.00%	100.00%

Table 1. *Accuracy of activity recognition.*

The position data allowed us to increase the accuracy of classification for the short activities such as sitting down or standing up. But even with all the attributes the accuracy is inadmissible. One cause of this problem is the use of filtered data to calculate the majority of the attributes. While use of filtered data is acceptable for stable activities (e.g standing), short activities are not recognizable. Another reason behind the problem is that inaccurate manual-labeled data were used to train the algorithm that is critical for short activity recognition.

5 WAVE ANALYSIS

Analyzing the graphs of acceleration length, we found that there are similarities in the shape of the areas responsible for the short activities. Segmentation of the graph to comparable areas for the same activity (ignoring of graph fluctuations) is necessary to use these similarities in the recognition process. We called such parts of the graph *waves*. The key idea of this approach is to compare not the actual values of the function but to use characteristics of its trends. To simplify formalization of the rules we decided to define a wave for discrete graphs fluctuating around 0 as a maximum set of successive points with the same sign. In Fig.2. a wave partition of a graph is shown.

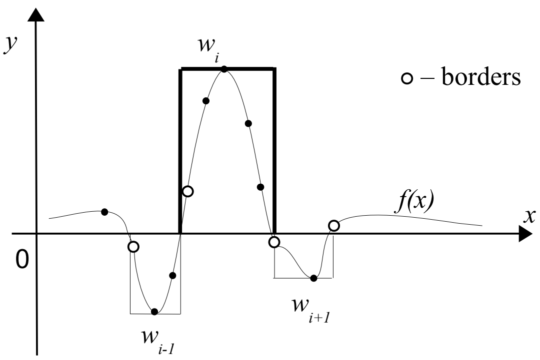


Fig.2. *Wave definition.*

Formally to define a wave of the discrete graph *f(x)* given as a set of points *F*, we call points $b_i=(x_i^b, y_i^b), b_i \in F$ where $y_i^b * y_{i-1}^b < 0$ as borders. The wave W_i is a set of points $p_k^i=(x_k^i, y_k^i), p_k^i \in F$, where $x_i^b \leqslant x_k^i < x_{i+1}^b$. Each wave has length $l_i = x_{i+1}^b - x_i^b$ and the peak value w_i. These are parameters we use in machine learning.

$$w_i = y_{max}^i, where |y_{max}^i| = max|y_k^i|, \quad \forall x_i^b \leqslant x_k^i < x_{i+1}^b \qquad (2)$$

To use the suggested approach for analyzing the acceleration length parameter fluctuating around gravity length we subtract it (*g*).

$$a_n(t) = \|a(t)\| - \|g\| \qquad (3)$$

When the tag is not moving, the function $a_n(t)$ is 0. Fig. 3 shows the graph of the function $a_n(t)$ for the part of the recording with the marked waves.

To use waves with machine learning for activity classification, let us consider that one wave is a part of the only one activity of the person. However, each activity could consist of several waves. For example, the process of sitting down usually consists of three significant waves. Therefore, we relabeled training data so that all the points of one single wave were labeled as one activity. Since any activity is a sequence of waves to classify each point, we used not only the attributes of the wave it is a part of but also the attributes of the previous and the next wave. For the real-time recognition this causes a delay lasting one wave length that is usually less than a second.

This hypothesis was tested on the recording of a person wearing one chest accelerometer only. The data was then partitioned and labeled. The Random Forest algorithm was used to detect the effect of using wave attributes together with some attributes listed in the previous section (e.g. moving attribute). Table 2 shows the comparison between

a	b	c	d	<-- classified as
270	2	2	0	a=standing
6	5	0	0	b=going down
11	0	166	0	c=sitting
0	0	8	5	d=standing up

a. *Confusion matrix of a classifier without wave attributes.*

a	b	c	d	<-- classified as
271	3	0	0	a=standing
2	9	0	0	b=going down
5	2	170	0	c=sitting
0	0	1	12	d=standing up

b. *Confusion matrix of a classifier using wave attributes.*

Table 2. *Confusion matrix of activity recognition.*

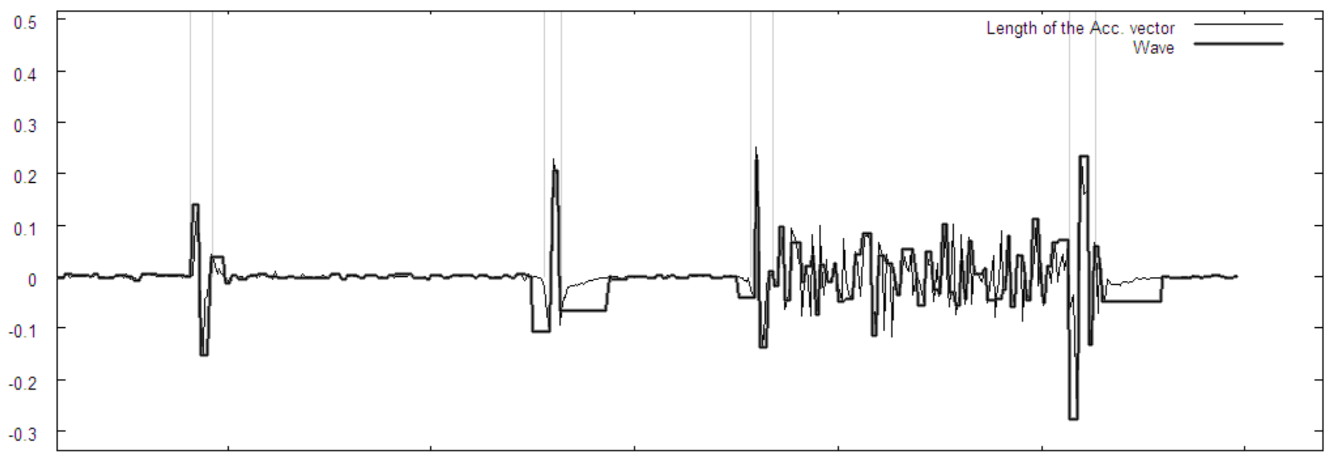


Fig.3. *Normalized length of the acceleration vector used for activity recognition.*

the confusion matrices of the classifier when the wave attributes were not used (a) and when the wave attributes were used (b).

The results show that the accuracy of short activity classification increased from 41 % to 87 % using the wave approach in this example.

6 CONCLUSION AND FUTURE WORK

The paper addressed the question of using acceleration-based system as an addition to or substitution for position-based system. The results indicate that even one accelerometer can increase the accuracy of activity recognition while using solely accelerometers does not give the acceptable accuracy in fall detection and activity recognition for the Confidence project.

Using a partition of the graph with wave analysis to improve machine learning is a promising field of study. The significant results were archived for the short activities. Further research might explore using this approach for statistical analysis of long activities, such as walking.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Framework Programme FP7/2007–2013 under grant agreement N 214986. Operation was partially financed by the European Union, European Social Fund. The authors would like to thank Božidara Cvetković, Violeta Mirchevska, Erik Dovgan and Michelle Lin for their inestimable help.

Reference

[1] Confidence. http://www.confidence-eu.org, 2010-10-01.

[2] M. Lustrek, B. Kaluza, E. Dovgan, B. Pogorelc, and M. Gams, ‘Behavior analysis based on coordinates of body tags’, Proceedings of the European Conference on Ambient Intelligence, pp. 14–23, (2009).

[3] Chin-Feng Lai, Yueh-Min Huang, Jong Hyuk Park, Han-Chieh Chao, "Adaptive Body Posture Analysis for Elderly-Falling Detection with Multisensors," IEEE Intelligent Systems, vol. 25, no. 2, pp. 20-30, Mar./Apr. 2010, doi:10.1109/MIS.2010.39.

[4] M. Lustrek, and B. Kaluza, ‘Fall detection and activity recognition with machine learning’, Informatica 33(2), pp. 197–204, (2009).

[5] Weka Toolkit. http://www.cs.waikato.ac.nz/~ml/weka/, 2010-10-01.

DIGITALNA TELEVIZIJA V MULTIMEDIJSKEM PODSISTEMU INTERNETNEGA PROTOKOLA

Aleksander Kovačić, Luka Zebec, Andrej Kos
Laboratorij za telekomunikacije, Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška cesta 25, 1000 Ljubljana, Slovenija

POVZETEK

Prispevek obravnava izvedbo integracije digitalne televizije v multimedijski podsistem internetnega protokola IMS (IP Multimedia Subsystem). Predstavljene so prednosti integracije in njena arhitektura. Realiziran je bil pilotski sistem za nudenje digitalne televizije v omrežju IMS, ki prikaže nekatere prednosti takšne združitve.

1 UVOD

Digitalna televizija je v uporabi že več let. Razvijati se je začela s prenosom prizemne in satelitske digitalne televizije, danes pa lahko dostopamo do nje tudi preko različnih fiksnih in mobilnih omrežij. Prenos digitalne televizije z uporabo internetnega protokola sta omogočila povečana pasovna širina novih dostopovnih omrežij in izboljšanje algoritmov kodiranja večpredstavnostnih vsebin.

Tudi komunikacijske storitve se selijo iz tokokrogovno komutiranih omrežij v paketno komutirano omrežje IP in ponudniki storitev že nekaj časa nudijo trojček storitev (triple play), ki vsebuje dostop do IPTV, VoIP in interneta. Vendar pa te storitve še vedno delujejo preko ločenih omrežij, kar posledično pomeni višje stroške in težje upravljanje s storitvami, kot če bi vse storitve delovale preko skupnega omrežja. Pojavi se vprašanje, kakšno bi moralo biti to skupno omrežje. Kot tista omrežna arhitektura, ki bo v prihodnosti temelj za hiter razvoj in nudenje vseh storitev naslednje generacije, vključno s storitvami televizije IP, se uveljavlja multimedijski podsistem internetnega protokola IMS (IP Multimedia Subsystem).

2 MULTIMEDIJSKI PODSISTEM INTERNETNEGA PROTOKOLA

IMS je uvedla organizacija 3GPP, za nadzor paketnih storitev v mobilnih 3G omrežjih. Deloval naj bi na posodobljenih GGSN (Gateway GPRS Support Node) in SGSN (Serving GPRS Support Node) strežnikih kot podsistem znotraj omrežja UMTS [1]. V istem času se je odvijal tudi prehod tradicionalnih fiksnih omrežij na omrežja IP, znan pod imenom NGN (Next Generation Networking).

Nenaden razmah internetnih storitev (VoIP, hipna sporočila, storitve prisotnosti, elektronska pošta, itd.) je v fiksna omrežja prinesel do tedaj nepoznane koncepte:

- uporabniki se lahko prijavijo v omrežje in dostopajo do svojih aplikacij iz različnih lokacij,
- prijava v omrežje ni pogojena z uporabo točno določene naprave,
- aplikacije delujejo na različnih vrstah naprav,
- združevanje storitev in uporabo različnih tipov medija.

Ti koncepti pa so bili že dobro znani v mobilnih omrežjih in postalo je jasno, da je potreben enak nadzor seje ne glede na uporabljeno dostopovno omrežje uporabnika (npr. ADSL, Wi-Fi, 3G, itd.).

Prava vrednost arhitekture IMS se je pokazala, ko je postalo jasno, da lahko ta podsistem nudi združen nadzor nad storitvami za različne vrste dostopa, s čimer združuje internetne in tradicionalne komunikacijske storitve ter tudi mobilna in fiksna omrežja. Pri današnji globalizaciji storitev je zelo pomembno povezovanje med ponudniki storitev. IMS v ta namen definira standarde za vzajemno delovanje z že obstoječimi omrežji (PSTN, ISDN, GSM, itd.) kot tudi z ostalimi ponudniki storitev IP. Že od samega začetka je bil IMS zamišljen kot logičen razvoj dotedanjih omrežij, ki ne vnaša radikalnih sprememb. Sestavljen je na osnovi obstoječih metod in protokolov in ne ustvarja novih. Veliko principov in protokolov je posvojil od interneta.

Operaterji vidijo IMS kot eno njihovih zadnjih možnosti za uspešno konkuriranje mnogim storitvam iz javnega interneta.

3 FIKSNO-MOBILNO ZLIVANJE IN ČETVERČEK STORITEV

Veliko operaterjev ponuja storitev trojčke, ki vsebujejo telefonijo, internet in IPTV. Delovanje vsake od teh storitev omogoča samostojen sistem, ki je ločen od ostalih, kar zelo otežuje medsebojno povezovanje storitev. Uporabnik zato v realnosti opazi povezavo med temi storitvami le na mesečnem računu. Uporaba skupnega nadzornega sistema IMS omogoča resnično združevanje vseh storitev, kar privede do nastanka novih storitev, ki so prilagojene vsakemu uporabniku. IPTV je v sistemu IMS samo ena izmed mnogih povezanih storitev.

Sistem IMS ima že vgrajene osnovne komunikacijske storitve, ki so sestavljene iz telefonskih storitev (VoIP, video klic, konferenčni klic, identiteta klicočega itd.) in internetnih storitev (hipna sporočila, storitev prisotnosti, elektronska pošta itd.). Z zlivanjem teh in televizijskih storitev je možno ustvariti nove konvergenčne storitve, ki si jih uporabnik prilagaja preko televizijskega ekrana [2]:

- obvestilo o dohodnem klicu na televizijskem ekranu s prikazom identitete klicočega, ki ga lahko uporabnik preko televizijskega vmesnika sprejme ali zavrne,
- preusmeritev klicev v telefonski predal med gledanjem priljubljene oddaje,
- samodejno utišanje televizije (ali začasna prekinitev predvajanja), ob sprejetem telefonskem klicu,
- storitev klik za klic (angl. click-to-dial) preko televizijskega vmesnika je zelo uporabna pri oglaševanju, saj med ogledom oglasa uporabniku omogoči, da s pritiskom na gumb (ali ikono) vzpostavi telefonsko zvezo, kjer pridobi dodatne informacije o izdelku ali ga naroči,
- kreiranje seznama prijateljev, ki so lahko iz različnih operaterjev, in dopisovanje s hipnimi sporočili,
- informacije o prisotnosti, lokaciji ali trenutno izbranem televizijskem kanalu prijateljev,
- pošiljanje povabil prijateljem za preklop na isti kanal,
- pošiljanje plačljivih vsebin prijateljem v obliki daril (preko hipnih sporočil, elektronske pošte itd.),
- predvajanje večpredstavnostnih vsebin iz interneta (Facebook, YouTube, Twitter, internetni radio, itd.).

Združitev storitev pod nadzor okolja IMS vsem storitvam doda tudi zmožnost mobilnosti – govorimo o četverčku storitev (quad play). To pomeni, da je dostop do vsake storitve mogoč iz vsake naprave (računalnik, mobilni telefon, televizijski sprejemnik, itd.) neodvisno od uporabljene tehnologije dostopa. S tem se fiksna in mobilna omrežja zlijejo v eno (FMC – Fixed Mobile Convergence) in odprejo se nove možnosti uporabe [2]:

- prenos aktivne seje iz televizijskega sprejemnika na mobilni telefon ali računalnik,
- nadzorovanje dogajanja na televizijskem sprejemniku preko mobilnega telefona (starši lahko nadzorujejo, kaj se predvaja na televizijskem sprejemniku, med tem ko jih ni doma, otroci pa lahko preko televizijskega vmesnika pošljejo staršem na mobilni telefon prošnjo za odobritev gledanja plačljivih vsebin,
- pošiljanje opomnikov o začetku predvajanja priljubljene oddaje na mobilni telefon.

4 INTEGRACIJA STORITEV IPTV V OKOLJE IMS

V standardizacijo sistema IPTV za delovanje preko omrežja IMS se je vključilo veliko standardizacijskih teles in industrijskih konzorcijev, kot so ATIS, ETSI, ITU-T in Open IPTV Forum [3]. Vse organizacije imajo podobne pristope standardiziranja in delujejo v navezah. Med najbolj aktivnimi sta organizaciji ETSI in Open IPTV Forum, ki v kombinaciji zelo nazorno specificirata delovanje celotnega sistema IMS IPTV [4].

4.1 Razlogi za prenos storitev IPTV v omrežje IMS

Sistemi IPTV, ki jih uporabljajo današnji operaterji, so lastniško zaščitene rešitve, zato prenos teh sistemov v okolje IMS ni trivialno opravilo. Ker pa ti sistemi že obratujejo, se operaterji upravičeno sprašujejo, ali je prenos storitev IPTV v omrežje IMS smiselno dejanje. Obstaja več razlogov, ki utemeljujejo prenos storitev IPTV v okolje IMS [5]:

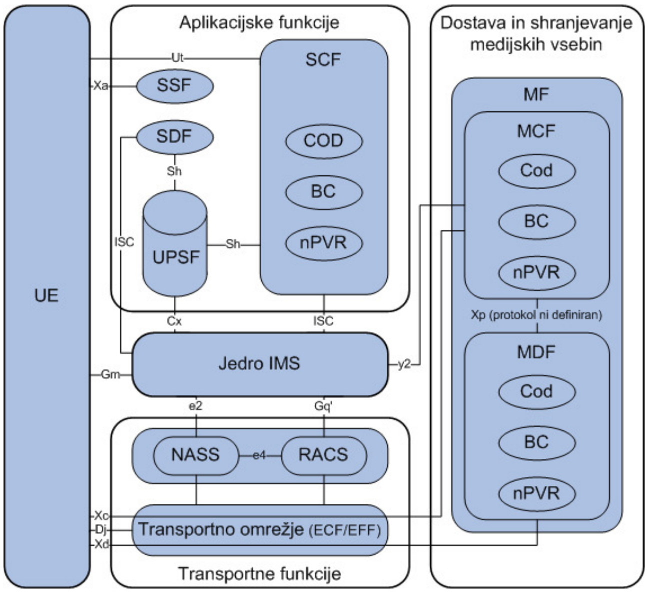
- prenos bi omogočil izboljšanje kvalitete storitev, saj se le ta v arhitekturi IMS ne izvaja več samo na transportnem sloju, ampak jo zagotavlja celoten postopek nudenja storitev,
- IMS skrbi za boljšo varnost in avtentikacijo uporabnika in omogoča enotno prijavo za uporabo vseh storitev znotraj omrežja IMS,
- uporaba enotne baze podatkov za vse storitve znotraj omrežja IMS zmanjša možnost podvajanja podatkov, operater pa si lažje ustvari celoten profil naročnika, ki je ključnega pomena pri nudenju kvalitetnih storitev prilagojenih vsakemu uporabniku,
- operaterji bi poleg klasičnih modelov zaračunavanja (npr. predplačniško ali naročniško) pridobili možnost novih modelov zaračunavanja (npr. hibridno naročniško zaračunavanje s predplačniškimi opcijami in obratno) [6],
- operaterji bi pridobili možnost hitrega razvoja novih konvergenčnih storitev z združevanjem telefonskih, internetnih in IPTV storitev, ki bi bile dostopne na vseh uporabniških napravah (televizijski sprejemnik, mobilni telefon, računalnik, itd.).

4.2 Arhitektura omrežja IPTV

Arhitektura za nudenje storitev IPTV preko omrežja IMS, povzeta po specifikaciji TISPAN [7], je prikazana na Sliki 1. Prikazani vmesniki na arhitekturi uporabljajo sledeče protokole:

- SIP/SDP (Gm, ISC, y2),
- HTTP (Xa, Ut),
- RTP/RTCP (Xd, Dj),
- RTSP (Xc, Dj),
- Diameter (Sh, Cx, e2, Gq', e4),
- IGMP/MLD (Dj).

Funkcija odkrivanja storitve SDF (Service Discovery Function) in funkcija izbire storitve SSF (Service Selection Function) nudita uporabniški napravi UE (User Equipment) informacije, ki jih potrebuje za zagon storitve. SDF vsakemu uporabniku, glede na njegov uporabniški profil, kreira in posreduje naslove SSF (v obliki naslova URI ali IP), kjer se nahajajo storitve. SSF nudi podatke o vsebinah, ki so na voljo za posamezno storitev (EPG, COD, itd.). Uporabnikom so glede na uporabniški profil lahko na voljo različne vsebine.



Slika 1: TISPAN arhitektura omrežja IPTV

Strežnik uporabniških profilov UPSF (User Profile Server Function) poseduje uporabniški profil in specifične uporabniške podatke za vse storitve, vključno z IPTV. Do teh podatkov dostopajo SDF, SCF in jedrno omrežje IMS.

Funkcije nadzora nad storitvami SCF (Service Control Functions) vsebujejo glavno odločitveno logiko za nudenje storitev IPTV. Osnovna naloga, ki jo opravlja SCF je avtorizacija uporabnikov med vzpostavljanjem in spreminjanjem seje ter sam nadzor vzpostavljene seje. Odobritev vzpostavitve seje se izvrši na podlagi uporabniškega profila. Ostale zadolžitve SCF so še npr.: nadzor nad medijskimi funkcijami MF (Media Functions), vrivanje uporabniku prilagojenih oglasov, nudenje uporabniškega vmesnika za spreminjanje nastavitev in sporočanje ustreznih informacij entitetam za zaračunavanje.

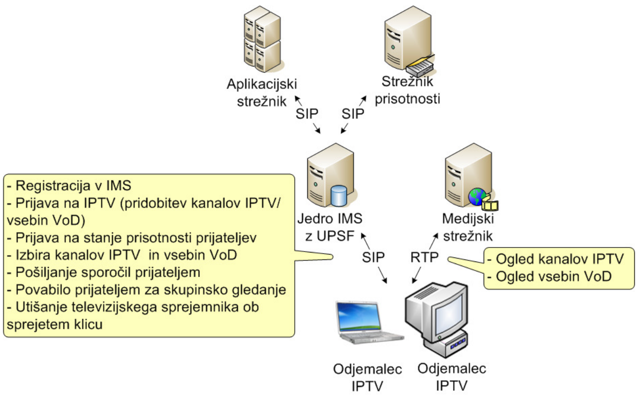
Za nadzor in dostavo medija sta zadolženi funkcija za nadzor medija MCF (Media Control Function) in funkcija za dostavo medija MDF (Media Delivery Function).

Elementa NASS (Network Attachment Sub-System) in RACS (Resource and Admission Control Subsystem) skrbita za krmiljenje omrežnih virov in dostopa. Sta ključna elementa pri zagotavljanju kvalitete storitev.

5 PILOTSKA INTEGRACIJA

V pilotskem projektu je bil realiziran sistem za nudenje televizije IPTV preko omrežja IMS. Arhitektura pilotskega projekta je prikazana na Sliki 2. Sestavljena je iz sledečih arhitekturnih elementov:

- aplikacijski strežnik GlassFish/SailFin [8] (strežniška aplikacija razvita s tehnologijo SIP Servlet),
- strežnik prisotnosti (emuliran strežnik prisotnosti v razvojnem okolju Ericsson SDS [9]),
- medijski strežnik (SIOL strujanje multicast IPTV, VLC strujanje VOD),
- jedro IMS z UPSF (emulacija jedra IMS v razvojnem okolju Ericsson SDS),
- odjemalec IPTV (uporabniška aplikacija razvita v okolju .NET).



Slika 2: Arhitektura pilotskega projekta

5.1 Uporabniška aplikacija

Realizirana uporabniška aplikacija IPTV (Slika 3) uporabniku omogoča ogled vsebin IPTV in VOD (Video On Demand) na osebnem računalniku ali televizijskem komunikatorju (na Windows platformi) ter nekaj konvergenčnih storitev:

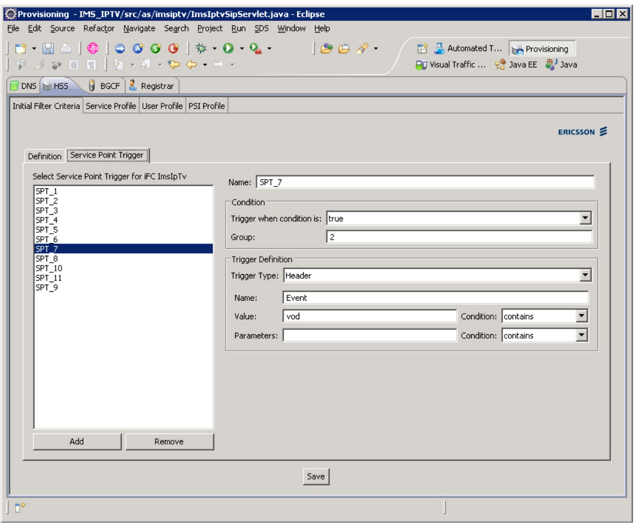
- spremljanje stanja prisotnosti svojih kontaktov (vključno s televizijskim kanalom, ki ga gledajo) ,
- interaktivno izmenjevanje hipnih sporočil s kontakti,
- pošiljanje (prejemanje) povabil svojim kontaktom za preklop na isti kanal ,
- utišanje televizijskega sprejemnika med aktivnim telefonskim klicem.

Slika 3: Pilotska uporabniška aplikacija IPTV

Uporabniška aplikacija je razvita v .NET okolju. Grafični uporabniški vmesnik temelji na tehnologiji WPF (Windows Presentation Foundation) [10], poslovna logika pa je napisana v programskem jeziku C#. Predvajanje medijskih vsebin je izvedeno s pomočjo prosto dostopnega projekta VLC Renderer [11], za podporo SIP pa je v aplikaciji uporabljen SIP .NET API [12].

5.2 Jedro IMS z UPSF in strežnik prisotnosti

Jedro IMS, UPSF in strežnik prisotnosti so komponente Ericsson SDS (Service Development Studio) okolja za razvoj, vpeljavo in testiranje storitev IMS (Slika 4). SDS je namenjen razvoju strežniških aplikacij na osnovi SIP Servlet API in klientskih aplikacij na osnovi ICP ali IJCU API. SDS temelji na Eclipse IDE (Integrated Development Environment).



Slika 4: Nastavljanje uporabniškega profila v bazi UPSF

5.4 Aplikacijski strežnik

Strežniška aplikacija IPTV, ki teče na GlassFish/Sailfin aplikacijskem strežniku, je razvita s tehnologijo SIP Servlet. Komunikacija med uporabniško in strežniško aplikacijo poteka preko jedra IMS. Strežniška aplikacija IPTV izvaja osnovne funkcije SDF, SSF in SCF, kar pomeni, da uporabniški aplikaciji sporoča informacije o medijskih vsebinah in posreduje naslove URI medijskih tokov za želene vsebine.

Strežniška aplikacija spremlja uporabniško komunikacijo VoIP. V primeru ko uporabnik sprejme klic strežniška aplikacija signalizira uporabniški aplikaciji naj utiša glasnost predvajanja medijske vsebine. Ko uporabnik zaključi s klicem, strežniška aplikacija pošlje zahtevo uporabniški aplikaciji naj povrne glasnost na prvotno vrednost.

5.5 Medijski strežnik

Medijski strežnik predstavljata ločeni entiteti za strujanje multicast televizijskih kanalov in strujanje VOD. Strujanje multicast televizijskih kanalov je izvedeno s pomočjo SIOL multicast televizije IP, za strujanje VOD pa je bil uporabljen program VLC. V obeh primerih je medijski tok v omrežju na voljo v vsakem trenutku za vse, ki poznajo naslov IP oddaje. V realnem sistemu bi to predstavljalo pomanjkljivo varnost, ki bi jo bilo potrebno odpraviti z uporabo entitete RACS.

6 ZAKLJUČEK

Televizija IP je za ponudnike storitev trenutno zelo zanimiva tema, saj že imajo delujoče sisteme IPTV, vendar bi tako njim kot tudi uporabnikom prenos teh sistemov v arhitekturo IMS omogočil do sedaj nepoznane funkcionalnosti. Uporaba sistema IMS za nudenje storitev IPTV omogoča lažji razvoj novih konvergenčnih storitev, hkrati pa se izboljša kvaliteta in varnost storitev. Operaterji tudi pridobijo boljši pregled nad uporabniškimi aktivnostmi in nove možnosti zaračunavanja. Ali bodo ponudniki storitev v teh razlogih videli zadosti potenciala za investiranje v izgradnjo sistema IMS, ki bo nudil tudi storitve IPTV, bo pokazal čas.

Literatura

[1] R. Copeland, Converging NGN Wireline and Mobile 3G Networks with IMS, CRC Press, Boca Raton, 2009.

[2] D. Opsenica, A. Kos, Converged TV over IMS, VITEL, 2009.

[3] A. Kovačić, Integracija televizije v multimedijski podsistem internetnega protokola: Diplomsko delo, Ljubljana, 2009.

[4] J. Hjelm, Why IPTV?, John Wiley & Sons, Chichester, 2008.

[5] T. Levent, 7 Reasons for Integrating IPTV with IMS, [http://blog.radvision.com/voipsurvivor/2008/09/08/7-reasons-for-integrating-iptv-with-ims/], 2.10.2009.

[6] Billing for IMS, [http://www.eurocomms.com/features/111512/Billing_for_IMS.html], 15.8.2010.

[7] Draft ETSI RTS 182 027, V3.3.0, 2009-08, IPTV functions supported by the IMS subsystem.

[8] SailfinProject, [https://sailfin.dev.java.net/], 2.10.2009.

[9] Ericsson Service Development Studio (SDS), [http://www.ericsson.com/developer/sub/open/technologies/ims_poc/tools/sds_40], 2.10.2009.

[10] The Official Microsoft WPF and Windows Forms Site, [http://windowsclient.net/], 2.10.2009.

[11] VLC Renderer, [http://www.codeplex.com/VLCRenderer], 2.10.2009.

[12] SIP .NET, [http://www.independentsoft.de/sip/index.html], 2.10.2009.

RECOGNITION OF INDIVIDUAL ANIMALS BY THEIR VOCALIZATION

Ivan Kraljevski[1], Solza Grceva[1], Igor Stojanovic[2], Zoran Gacovski[1], Biljana Spireva[1]
[1]Faculty for ICT, FON University, bul. Vojvodina bb, Skopje, Macedonia
[2]Faculty of Computer Science, University Goce Delcev, Toso Arsov 14, 2000 Stip, Macedonia
E-mail: {ivan.kraljevski, solza.grceva, zoran.gacovski, biljana.spireva}@fon.edu.mk, igor.stojanovik@ugd.edu.mk

ABSTRACT

The recent advance in development of digital signal processing (DSP) and analysis techniques in human speech recognition systems has very significant influence on the field of bioacoustical research on animals [1]. In this paper, we present our approach in recognition of animal vocalizations with techniques for voice recognition. Vocalizations induced by physiological characteristics of the vocal tract of individual animals were encoded by Time Encoded Signals (TES). Further on, Artificial Neural Networks (ANN) was used for voice samples classification. The main benefit of the proposed technique is the application of such systems on low complexity DSP hardware due to low processing requirements [2]. In this paper, a system for individual "speaker" recognition of dogs barking samples was created and evaluated.

1 INTRODUCTION

During evolution many animal species created some complex forms of communications by vocalization. By that, the animals usually express their internal physiological state (hunger, thirst, matting, anxiety, etc.) as well as particular organism's condition influenced by various environmental factors (temperature, atmospheric pressure). By appropriate analysis of animal's vocalization, particular anatomical characteristics could be estimated. For example, the vocal tract length (VTL) is specific for different species and breeds, even for individual animals. The standard speaker and speech recognition techniques implemented in voice driven applications for individual animal recognition by their vocalization, is based on the assumption that the most mammal vocalizations (but not all) match frequency ranges of human vocalization [1].

This paper outlines a system for automatic individual "speaker" recognition of dogs by their barking samples, based on anatomical characteristics related with particular vocalization. The system uses Time Encoded Signals (TES) and Artificial Neural Networks (ANN) for classification of individual animals.

2 METHODS

There are many acoustical and anatomical similarities in the process of vocal sounds production - vocalization in mammals. The primary acoustical signal is created in the vocal folds - glottal source, with mechanical oscillations of the folds, caused by the flowing air from the respiratory system - the lungs. With folds opening and closing, the air flow is modulated across the glottal opening, thus producing time varying acoustical signals.

Many mammal species produce almost periodical signals in the larynx with basic frequency and multiple harmonics, which can be relatively easy distinguished by Fourier transform or other speech analysis techniques in frequency domain. On the other hand, narrower, but non-oscillating larynx produces turbulent noise. Thus, the most appropriate way of description of the produced animal vocalization would be by the general discrete-time model for voice (speech) production.

2.1 Formants and Vocal tract length

According to acoustical theory central frequencies of the formats are related and depend on the anatomical measurements of the vocal tract, on the length and on the shape (particular variations in the cross sections). Moreover, the length is the most important anatomical measure which influences the formant frequencies [3] (Figure 1). The resonant frequencies of the uniform air tube with non-varying sections can be used as the first approximation of the formats given by:

$$F_i = \frac{(2i+1)\cdot c}{4\cdot VTL}$$

With the both ends closed:

$$F_i = \frac{i\cdot c}{2\cdot VTL}$$

Where *i*=1 is the number of the format, *c* is the speed of the sound (350 m/s), VTL - vocal tract length (in meters) and- F_i the frequency of *i-th* format. Despite the difference in the equitation regarding the status of the ends (opened or closed), it is obvious that the distance between two consecutive formants is constant and is given by:

$$F_i - F_{i-1} = \frac{c}{2\cdot VTL}$$

The distance between two formants is constant and directly depends on the vocal tract length. Larger value for VTL gives lower formant frequencies. It can be assumed that different breeds of dogs have different or dissimilar anatomical features (VTL), produces different vocalization pattern, regarding formant frequencies. This apply on individual subjects as well, each animal has different anatomical characteristics related to age, sex, size and other factors.

Therefore it will be possible to classify the vocalization sequences in certain categories. Such classification system will be able to recognize the species, breed, and individual animals (in these case dogs) by their vocalization or to estimate the possibility of presence of anatomical deviation in the vocal tract length in animals of the same breed [4].

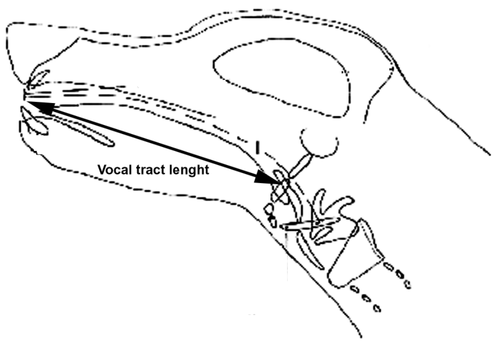


Figure 1: *Vocal tract length measurement*

In order to achieve accurate classification of the dog vocalization, well-known techniques for human speech recognition may be used [1]. This paper presents a method for classification of dog vocalization sequences based on Time Encoded Signals, for signal representation, and Artificial Neural Networks for their classification.

2.2 Proposed system

The overall system architecture for domestic dog vocalization classification for speaker recognition is shown on Figure 2. The system consists of blocks for acquisition and preprocessing of signals, A/D conversion block, TES coder where digitized signals are coded into time domain with array of symbols. Block where S-matrix with fixed length is created and represents the input vector of the neural network classificatory. This system could be implemented as a part of low-end DSP hardware system, as well.

2.3 Recording, digitalization and preprocessing

Vocalization sequences of the subjects in a state of excitement were recorded with ordinary microphone. Figure 3 presents characteristic spectrogram of the barking sequence of a German Sheppard dog breed. Voice signal was digitized with 20.05 KHz sample rate and 16 bit resolution, providing quality in processing and more discriminate classification.

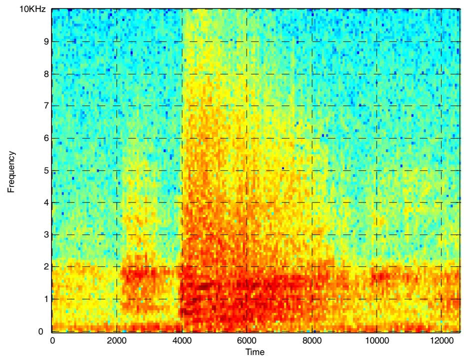


Figure 3: *Spectrogram of "German Sheppard" vocalization*

2.4 TES - Time Encoded Speech

The choice of the techniques for animal voice signals recognition is most important for designing a reliable system. There are several methods that could be applied to solve the stated problem. One possible solution is to use vector classification of the PSD (Power Spectrum Density) function. To consider the time varying nature of the vocal signal, DTW (Dynamic Time Warping) technique could be used on arrays of MFCC (Mel Frequency Cepstral Coefficients) [4]. But, despite the algorithm simplicity, this method could not provide efficient and reliable system. Other possible solution is to use HMM (Hidden Markov Model) for representing vocal sequences with more complex training process. Regarding the nature of the application and the known characteristics of voice signals, this system uses Time Encoded Signal Processing and Recognition [5] for signal parameterization and Neural Network for classification. This approach provides low complexity and computing requirements that will be well suited for robust and embedded hardware applications.

TES coding is based on precise mathematical description of waveforms, involving the polynomial theory that shows how band limited signals (such as animal vocalization) may be completely described by the locations of their real and complex zeros. The real and complex zero descriptors of TES and the time-bandwidth data produced by a Fourier transform are mathematically equivalent and both produces equal number of digital sample points. This technique is well known and it has been used for a long time in speech coding in telecommunication systems [5]. The interval between two adjacent zero-crossings is called an epoch and, for every epoch, three parameters are derived: duration of the epoch, shape (S) and the magnitude (M) of the signal. D is the number of samples and S is the number of positive or negative local maxima and minima, and M is the largest value of samples in the given period.

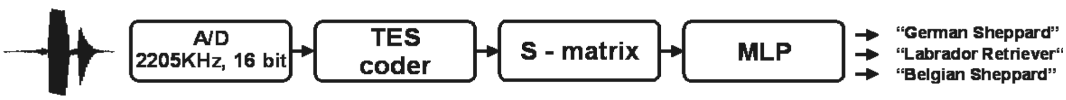


Figure 2: *Simulation system architecture*

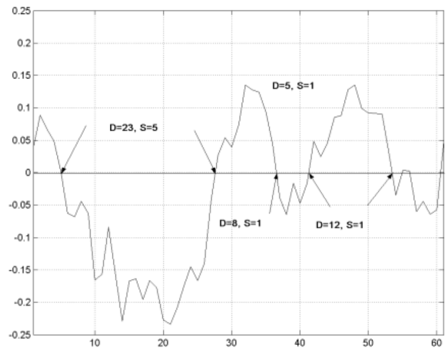


Figure 4: *Part of vocalization waveform divided in TES epochs*

These parameters are encoded with assigning a unique symbol for certain combination of the epoch duration (D) and its shape (S) (Figure 4). Thus the signal is transformed into time encoded stream of discrete numerical descriptors – TES symbols. The statistical analysis showed that number of samples between zero crossings lies between 1 and 37 for speech signals sampled with 20.05 KHz [5] and that the duration of an epoch above 13 samples is very unlikely to occur.

Using vector quantization and K-means algorithm, generalized code-book was created – TES alphabet. Standard symbol alphabet consists of 28 different symbols, and it has been proved to be quite sufficient for the representation of speech and other band limited signals. The symbols table is established according to statistical analysis and the likely occurrences of the (D, S) pairs for given typical signals. In case of pair that could not be found in the symbol table, it could be linked to the symbol that best represents the epoch for the given pair. These strings of numerical descriptors can be easily converted to TES matrices with fixed dimension. A histogram of the signal array with 28 possible symbolic descriptors can be produced, forming so called S-matrix with fixed dimension 1x28 (Figure 5), which carry information about symbols frequency.

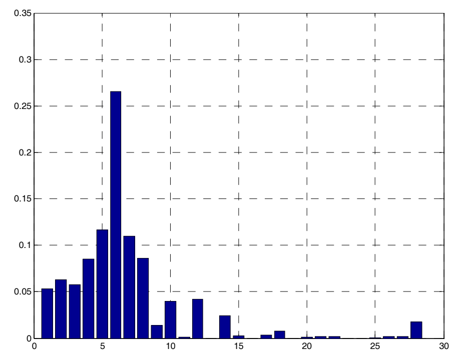


Figure 5: *S-matrix of "German Sheppard" vocalization*

TES coder transforms the original signal in array of symbolical descriptors without losing the quality. It can be easily reversed back into analog signal. Because of their fixed length, S-matrices are ideal form of input vector for artificial neural network classificatory.

The biggest advantage of the S-matrices compared to other descriptors in the frequency domain (eg. MFCC) is that regardless of the signal length they have fixed dimensions. It allows easy template creation for the system training phase, and allows use of wide range of different classification procedures.

2.5 Classification with Multi Layer Perceptron

Because of their fixed length, S-matrices are ideal form of input vector for neural network classificatory. This system uses Multi Layer Perceptron (MLP) with 3 layers, where the number of input layer nodes is equal with the length of the S-matrix – 28 and the number of outputs by the number of different classes, in this case 11 individual animals. The hidden layer contains 30 neurons with logsig transfer function and the output layer 11 neurons with also logsig function. The choice of the number of neurons in the hidden layer is made empirically comparing the neural network performance.

The choice of transfer function in the hidden layer was made according to the characteristics of the S - matrices whose elements have only positive values in the range between [0, 1]. This causes output neurons to produce values also in this range but represents probabilities of recognition for given input in output classes, in this case – that is individual animal (dog).

In case where the number of the neurons in the hidden layer is too small, the network is not trained to classify input samples appropriately (under fitting). Otherwise, increasing the number of neurons in hidden layer may introduce an effect of saturation (over fitting) and the neural network is well trained to classify or recognize only the training set of input samples. The training set is composed of input – output pairs of S-matrices and outputs values 0 and 1 on the corresponding node. For the training of the ANN, Resilient Back propagation was used for determining the direction of change of weight functions, but not the weight values. They were determined by a special variable which is increased by a factor delta_inc = 1.2 whenever that value has the same sign for two consecutive iterations and is reduced by a factor delt_dec = 0.5 whenever the derivative with respect to weight values sign changes to the previous iteration

Whenever an oscillation in the direction of change of weight function values appears, the variable decreases, and if for some period its sign does not change, the variable value steadily increases [6]. The process of the training is stopped in the moment when the target MSE is achieved or maximum number of training epochs are reached.

3 SYSTEM'S PERFORMANCE ESTIMATION

Barking samples of eleven individual police dogs were used for system performance estimation: three different subjects of the breed "Belgian Sheppard - Malinoi" with 47 sequences (Robi, Niki and Gor), five subjects of "German Sheppard" with 88 sequences (Go1, Amor, Dog, Gando, and

Rex) and three subjects of "Labrador Retriever" (LB1, Ago and Ari) with 25 sequences. That gives total number of 160 barking samples of 11 animals of 3 different dog breeds. Rough segmentation of minimum of 10 samples per each individual dog was taken and training set (Figure 6) was produced with 110 vocalization sequences. The system is implemented using the MATLAB software package.

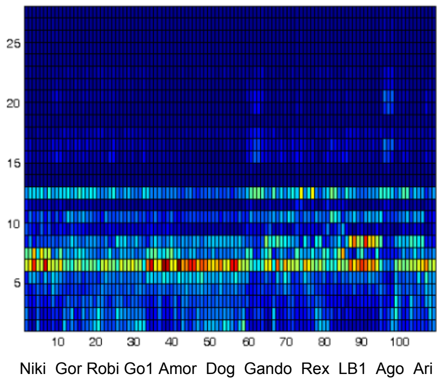


Figure 6: *Training set contained total of 110 sequences*

The system was tested for individual dog recognition by their vocalization. The test set consists of 162 samples, the neural networks topology consists of 28 input neurons, 30 hidden neurons and 11 output neurons, target mean square error was set to 1e-6, and the ANN was trained for 200 epochs.

	Niki	Gor	Robi	Go1	Amor	Dog	Gando	Rex	LB1	Ago	Ari
Niki	8	0	1	1	0	0	0	0	0	0	0
Gor	0	3	2	0	0	1	1	0	0	0	0
Robi	0	0	29	1	0	0	0	0	0	0	0
Go1	0	0	0	10	1	0	1	0	0	1	0
Amor	0	0	1	0	24	0	0	0	0	0	0
Dog	0	0	1	0	0	7	1	3	0	0	0
Gando	0	0	2	0	0	0	19	1	2	2	0
Rex	0	0	0	2	0	0	0	12	0	0	0
LB1	0	0	0	0	0	0	0	0	12	0	0
Ago	0	0	1	0	0	0	0	0	0	5	0
Ari	1	0	0	0	0	0	1	1	0	0	4

Table 1: *Resulting confusion matrix for the individual animal recognition*

The results of the performance estimation of 162 samples recognition are:

Training set: 94.68 % with: 110 samples
Test set: 80.25 % with: 162 samples

During training process, it is noticeable that after 100 epochs - the MSE decreases slowly, so higher value for MSE or shorter time for training could be set. This will produce better generalization of the neural network and increase the reliability of the system for samples that were not included in the training phase.

4 CONCLUSIONS

The proposed system uses Time Encoded Signals (TES) and Artificial Neural Networks (ANN) for classification of individual animals. It was tested and evaluated with vocalization samples of 11 individual subjects (three different police dog breeds, "Belgian Sheppard - Malinoi" "German Sheppard" and "Labrador Retriever"). It is shown that this system is able to successfully recognize subjects with different anatomical characteristics (different individuals) of the vocal tract by their vocalization samples. Such a system could be successfully used in various on-field applications implemented on low complexity smart sensor DSP hardware: monitoring and research on animal behavior of different species, their communications, automatic species detection, small animal practice in veterinary medicine, monitor individual animals behavior on limited space (e.g. Zoo and National parks) and others.

References

[1] J. G. Harris, M. D. Skowronski, "Automatic Speech Processing Methods For Bioacoustics Signal Analysis: A Case Study Of Cross-Disciplinary Acoustic Research" IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Toulouse, France, vol. 5, pp. 793-796, May 15-19, 2006

[2] Phipps, T.C., King, R.A. "A Low-Power, Low-Complexity, Low-Cost TESPAR-based Architecture for the Real-time Classification of Speech and other Band-limited Signals" ICSPAT at DSP World, Dallas, Texas, October 2000

[3] T. Riede, T. Fitch: Vocal Tract Length And Acoustics Of Vocalization In The Domestic Dog (Canis Familiaris), The J .Of Exp. Biology 202, 2859–2867 (1999)

[4] Kraljevski I., Mihajlov D., Arsenovski S., „Determination of Farm Animals Condition by Their Vocalization", Proceedings of the 24th ITI, A-7, Cavtat, Croatia, June 24-27, 2002.

[5] J. Holbeche, R.D. Hughes and R.A. King: Time Encoded Speech (TES) descriptors as a symbol feature set for voice recognition systems, IEEE Int. Conf. Speech Input/Output; Tech. and App., pp. 310-315, March 1986.

[6] Matlab Product Documentation

INTEGRATING DEXI EVALUATION MODELS INTO DECISION DECK d2 SOFTWARE

Xiaobin Li [1], *Marko Bohanec* [2,3]
[1] Jozef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[2] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[3] University of Nova Gorica，Vipavska 13, 5000 Nova Gorica, Slovenia
Tel: +386 41 925035
e-mail: xiaobin.li2006@gmail.com

ABSTRACT

DEXi is an educational Multi Criteria Decision Aid (MCDA) computer program aimed at interactive development of qualitative multi-criteria and hierarchical decision models and the evaluation of options. The Decision Desktop software, or d2 for short, is one of the software modules being developed in the Decision Deck project. It is a desktop, client/server application which has integrated several MCDA methods in order to improve the interaction and compatibility between different MCDA methods. In this paper we present our proposal to integrate qualitative DEXi decision models into d2 as a plugin.

1 INTRODUCTION

Multi-criteria (or multi-attribute) decision aiding (MCDA), also termed multi-criteria decision analysis and multi-criteria decision making (MCDM), is a discipline which aims at supporting decision makers who are faced with making numerous and conflicting evaluations [1]. Many different MCDA methods have been developed and implemented [2], but these software programs were developed in an uncoordinated way: they have very different operational requirements, they use different and incompatible representations of data and decision models, and are, consequently, very difficult to integrate.

Decision Deck is an initiative to implement functionalities of a large range of MCDA methods within a common Open Source platform [4]. In particular, it provides a component, called d2, which facilitates a unified implementation of different MCDA methods. d2 is a desktop, client/server application, meaning that it is designed to be installed locally (it is not a web application), and uses a database to store application data, thereby enabling multiple users.

DEXi [3] is a stand-alone computer program for multi-attribute decision making. It facilitates interactive development of qualitative multi-attribute and hierarchical decision models, and the evaluation of options. DEXi has been used in many real-life decision problems in the areas such as selection and evaluation of computer hardware and software, evaluation of companies and business partners, personnel management, project evaluation, land-use planning, risk assessment in medicine and health-care.

DEXi falls in a broad category of MCDA methods and is therefore a good candidate for the integration in the Decision Deck platform. In this paper we propose a way to integrate DEXi into d2 and present the current implementation.

The structure of this article is as follows. In section 2, we first present the tools we used: DEXi, Decision Deck and d2. In section 3, which is the main part of the paper, we present the methods of integrating DEXi models into d2. Section 4 illustrates the current implementation of the DEXi plugin through an example of car evaluation. Section 5 concludes the paper and suggests further work on DEXi integration into Decision Deck.

2 MCDA METHODS AND TOOLS

In this section, we introduce the multi-criteria decision making methods which were used in this work: (1) DEXi as a method that is being integrated into Decision Deck, (2) Decision Deck platform in general, and specifically (3) its d2 component.

2.1 DEXi

DEXi [3] is a MCDA computer program which differs from most conventional MCDA tools in that it uses qualitative (symbolic) attributes instead of quantitative (numeric) ones. Also, aggregation (utility) functions in DEXi are defined by 'if-then' decision rules rather numerically by weights or some other kind of formula. Consequently, a DEXi model includes the following four components:

- attributes: qualitative variables that represent decision sub-problems;
- scales: ordered or unordered sets of symbolic values that can be assigned to attributes;
- tree of attributes: a hierarchical structure representing the decomposition of the decision problem;
- utility functions: rules that define the aggregation of attributes from bottom to the top of the tree of attributes.

Bellow is an example of a simple DEXi multi-attribute model for the evaluation of cars taken from DEXi user's manual [3]. Figure 1 shows the hierarchical attribute tree, Figure 2 shows the scales of all the attributes, and Figure 3 shows the utility function of attribute CAR in this model.

Attribute	Description
CAR	Quality of a car
├**PRICE**	Price of a car
│├BUY.PRICE	Buying price
│└MAINT.PRICE	Maintenance price
└**TECH.CHAR.**	Technical characteristics
├**COMFORT**	Comfort
│├#PERS	Maximum number of passengers
│├#DOORS	Number of doors
│└LUGGAGE	Size of the luggage boot
└SAFETY	Car's safety

Figure 1: *Hierarchical attribute tree of car evaluation.*

Attribute	Scale
CAR	unacc; acc; good; *exc*
├**PRICE**	high; medium; *low*
│├BUY.PRICE	high; medium; *low*
│└MAINT.PRICE	high; medium; *low*
└**TECH.CHAR.**	bad; acc; good; *exc*
├**COMFORT**	small; medium; *high*
│├#PERS	to_2; 3-4; *more*
│├#DOORS	2; 3; 4; *more*
│└LUGGAGE	small; medium; *big*
└SAFETY	small; medium; *high*

Figure 2: *Attribute scales in the car evaluation model.*

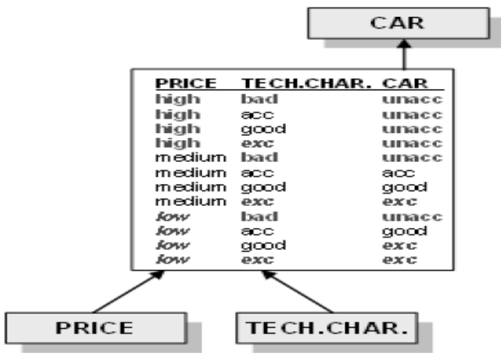


Figure 3: *Utility function of attribute car.*

In the stage of evaluation and analysis of decision options, DEXi facilitates [3]: description of options, evaluation of options, analysis of options and reporting. The analysis of options includes 'what-if' analysis, 'plus-minus-1' analysis, selective explanation and comparison of options.

DEXi models, which are created and/or edited in DEXi, are stored in DEXi ('.dxi') files. The basic format of '.dxi' files is XML (eXtensible Markup Language) [6].

2.2 Decision Deck

Decision Deck [4] is a platform for implementing various MCDA methods. Decision Deck provides several software modules (d2, d3, Diviz, XMDCA), which offer a common framework for MCDA methods. The emphasis is on developing multiple software resources that are able to interact. Consequently, several complementary efforts focusing on different aspects contribute to Decision Decks various goals [4]:

- d2: a rich open source Java software containing several MCDA methods;
- d3: an open source rich internet application for XMCDA web services management;
- diviz: an open source Java client and server for designing, executing and sharing MCDA methods, via the composition of XMCDA web services;
- XMCDA: a standardised XML recommendation to represent objects and data structures issued from the field of MCDA. Its main objective is to allow different MCDA algorithms to interact and be easily callable;
- XMCDA web services: distributed open source computational MCDA resources, using the XMCDA standard.

2.3 d2

d2 [4] was the first software to be developed in the Decision Deck project. It is rich open source Java software containing several MCDA methods as plugins, e.g. Iris, Rubis, Vip, and Weighted Sum. However, all of them are quantitative MCDA methods, dealing only with linear utility aggregation. Figure 4 shows a screenshot of d2 which includes two methods Iris and Weighted Sum. In the left column, there is a Navigation for different MCDA methods. Under the menu Global Settings, users can assign the Alternatives and Criteria, and under the Evaluations menu, users can enter specific values of each Alternative for the corresponding criterion. The right part of the figure shows the results of scoring of some decision alternatives (a1–a7 and P1–P3) using the method Weighted Sum.

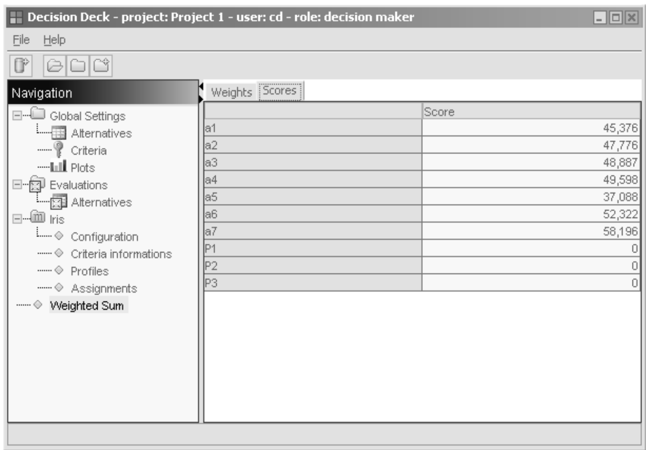


Figure 4: *Screenshot of d2.*

3 DEVELOPING DEXi PLUGIN

In this section, we present our implementation of the DEXi plugin in d2 software. First we introduce the mechanism of the d2 plugin; then we explain the methods we used when developing the plugin; finally, we demonstrate and discuss the results of our work.

3.1 OSGI

A d2 plugin is nothing more than an OSGI bundle [4]. An OSGI bundle is a package that encapsulates classes, resources, native files, etc. It can do nothing alone and is intended to be deployed inside an OSGI environment. It is a framework that enables service oriented platforms. It provides a number of infrastructure services (it takes care of bundle isolation, integrates bundle versioning, deals with security concerns, etc.) and allows to dynamically plug and unplug services (e.g., a MCDA computation service).
A typical d2 plugin usually contributes to the user interface (UI) and refines the common model provided by d2. It is usually composed of the following layers :

- model layer;
- data access layer;
- service layer;
- UI layer.

3.2 Methods

d2 is known to work against these versions of software: Eclipse 3.3.1.1, Java 5+ and MySQL 5. For developing the DEXi plugin in d2, we employed the following methods:

- Technology framework: Swing + Hibernate. Java Swing has been used to design the interface, Hibernate and MySQL have been used to handle with data;
- MVC: Model, View and Controller three layers have been developed to carry out the basic DEXi functionalities in d2 by using Java.

3.3 DEXi Plugin

So far the d2 software focused on dealing with linear MCDA methods rather than hierarchical models. DEXi, on the other hand, is based on hierarchical models, and the current version of d2 was not powerful enough to handle DEXi tree-structured models. Therefore, during the process of DEXi plugin designing, we designed the model, data access, service and UI layers only for the DEXi plugin, rather than modeling through the Alternatives and Criteria functions under Global Settings in d2. This to some extent reduces the interactivity and compatibility between the DEXi plugin and the other MCDA methods, but appears to be the only solution in the current d2 framework.
Figure 5 shows the DEXi plugin in d2. In the left column, there is a navigation menu of different MCDA methods which have been integrated into d2, here we added the DEXi icon together with three sub-menus: Model, Options and Evaluation, which implement the basic functionalities of DEXi:

- *Model*: When the Model sub-menu is selected, the main plugin window (Figure 5, center) displays the current hierarchical DEXi model. The user can edit the model. In the rightmost column, the user can edit attribute scales (Scale) and decision rules (Utility Function) that corresponds to the currently selected attribute from the model.
- *Options*: When the Options sub-menu is selected, the corresponding right window shows a table where the user can define and edit different decision options (see section 4.2 for an example).
- *Evaluation*: Here, the right part of the pulgin window displays evaluation result of the options (section 4.3).

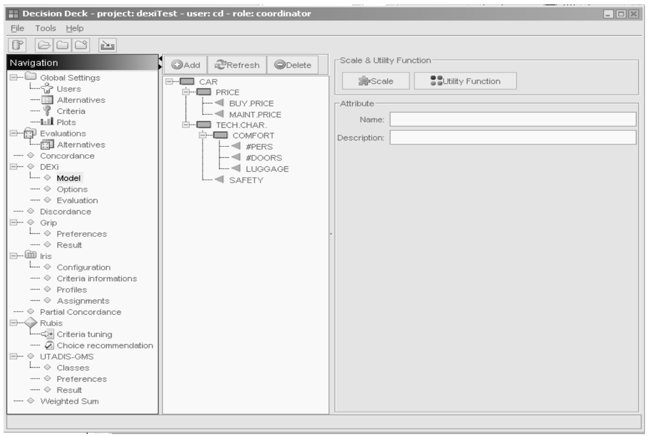


Figure 5: *DEXi plugin in d2.*

3.4 Data Storage

d2 uses MySQL to store the data of MCDA models. This is different from DEXi, which uses an XML format. Therefore, the DEXi plugin could only construct one model at the same time. Figure 6 shows the database VM where data of all d2 methods stored by MySQL. For DEXi plugin, we created 4 tables: dexi_option, dexi_rule, dexi_scale and dexi_treenode to handle with the car evaluation model. Importing the existing models and exporting models are not implemented at the moment. This is somehow a shortcoming of our work, we are planning to solve when the d2 version is upgraded.

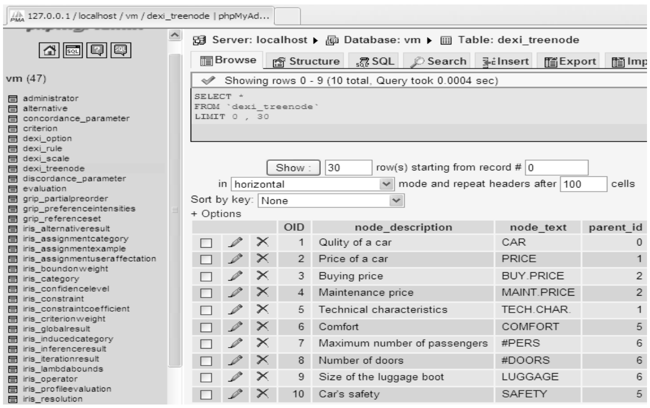


Figure 6: *Data storage in d2.*

4 AN EXAMPLE OF CAR EVALUATION

In order to test the DEXi, we use an example of car evaluation model, introduced in section 2.

4.1 Car Evaluation Model

Figure 7 shows the model of car evaluation, which we used to test the DEXi plugin and illustrate its capabilities. When we click the PRICE attribute, the plugin shows the details of this attribute on the right. By clicking the Scale and Utility Function buttons, the user can edit the scale and utility function of this attribute, respectively.

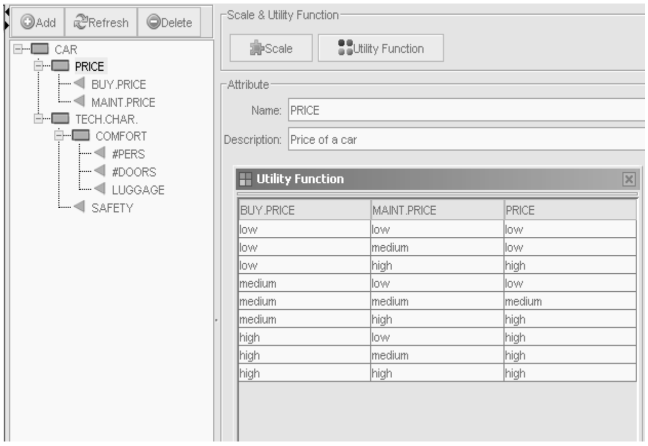


Figure 7: *Model of car evaluation.*

4.2 Options of Car Evaluation Model

Figure 8 shows seven different cars (decision options), represented with different values of attributes. These values were entered by the user under the Options sub-menu.

Option	BUY.PRICE	MAINT.PRICE	#PERS	#DOORS	LUGGAGE	SAFETY
Car1	medium	low	more	4	big	high
Car2	medium	medium	more	4	big	medium
Car3	low	medium	to_2	4	medium	high
Car4	high	medium	more	more	big	medium
Car5	medium	low	more	more	medium	high
Car6	low	medium	more	more	big	high
Car7	medium	medium	3_4	3	medium	medium

Figure 8: *Options of car evaluation model.*

4.3 Evaluation of Car Evaluation Model

In figure 9, the field CAR shows the evaluation results of seven cars. For example, "Car1" is evaluated as "exc" (excellent), because of its "PRICE" is "low", and "COMFORT" is high; "Car3" is evaluated as "unacc" (unacceptable), because its "PRICE" is "low", but its "COMFORT" is "small". These values are determined according to the utility functions defined by the user.

Option	.CAR	..PRICE	...BUY.PRICE	...MAINT.PRICE	..TECH.CHAR.	...COMFORT	....#PERS	....#DOORS	....LUGGAGE	...SAFETY
Car1	exc	low	medium	low	exc	high	more	4	big	high
Car2	good	medium	medium	medium	good	high	more	4	big	medium
Car3	unacc	low	low	medium	bad	small	to_2	4	medium	high
Car4	unacc	high	high	medium	good	high	more	more	big	medium
Car5	exc	low	medium	low	exc	high	more	more	medium	high
Car6	exc	low	low	medium	exc	high	more	more	big	high
Car7	acc	medium	medium	medium	acc	medium	3_4	3	medium	medium

Figure 9: *Evaluation of cars.*

5 CONCLUSION

We integrated DEXi qualitative MCDA program into the Decision Desktop platform as a d2 plugin, supporting the basic functionalities. This is the first hierarchical MCDA method that has been integrated into d2. This has shown that the Decision Deck platform, in addition to 'flat' quantitative models, can support a qualitative decision making method which deals with the hierarchically structured data. We believe that both d2 and DEXi have benefited from this integration: d2 by being extended by a first hierarchical qualitative MCDA methodology, and DEXi by becoming a member of a widely recognized open source decision-support software platform.

The current implementation of DEXi plugin has some shortcomings. As the current version of d2 focuses on linear MCDA methods, the platform does not provide framework to implement hierarchical methods. We had to develop our own framework for the DEXi plugin; this reduces the interaction and compatibility with the other MCDA methods to some extent.

d2 uses MySQL to store the data rather than XML files which DEXi uses, therefore our DEXi plugin could only construct one evaluation model at the same time, meanwhile it could not import and export models. This will be overcome with the upgrading of d2 platform and our subsequent work.

References

[1] Belton, V. and Stewart, T.J. 2002. Multiple Criteria Decision Analysis: An Integrated Approach. Boston: Kluwer Academic Publishers.

[2] Bisdorff, R. and Meyer, P. January 2008. The Decision-Deck Project - Developing a Multiple Criteria Decision Analysis Software Platform, ERCIM News 72.

[3] Bohanec, M. 2008. DEXi: Program for Multi-Attribute Decision Making, User's Manual, Version 3.00. http://kt.ijs.si/MarkoBohanec/pub/DEXiManual30r.pdf (accessed 24-8-2010).

[4] Decision Deck 2010. Decision Deck Project. http://ernst-schroeder.uni.lu/decision-deck (accessed 24-8-2010).

[5] Maxwell, D.T. 2008. Decision Analysis: Find a Tool that Fits. OR/MS Today, October 2008. http://lionhrtpub.com/orms/orms-10-08/frsurvey.html (accessed 24-8-2010).

[6] W3C 2008. Extensible Markup Language (XML) 1.0 (fifth edition). W3C Consortium, http://www.w3.org/TR/2008/REC-xml-20081126 (accessed 27-8-2010).

NON-LINEAR METHODS FOR RANKING QUALITATIVE NON-MONOTONE DECISION PREFERENCES

Biljana Mileva-Boshkoska[1,2], *Marko Bohanec*[1,3]
[1]Department of Knowledge Technologies, Jožef Stefan Institute,
Jamova 39, 1000 Ljubljana, Slovenia
[2]Faculty for Electrical Engineering and Information Technologies,
Karposh 3, bb, 1000 Skopje, Macedonia
[3]University of Nova Gorica, Nova Gorica, Slovenia
Tel: +386 1 4773145;
e-mail: biljana.mileva@ijs.si, marko.bohanec@ijs.si

ABSTRACT

One approach to option ranking in qualitative decision making is first to automatically construct a quantitative evaluation model using the qualitative and quantitative (QQ) method and then use the quantitative model for ranking. However, the quantitative model constructed by QQ, which uses linear techniques, fails to provide consistent and complete option ranking of non-monotone decision preferences. In this paper we investigate alternative methods for consistent and complete option ranking of non-monotone preferences by using non-linear techniques. Results show that non-linear methods are superior to linear techniques, when dealing with non-monotone two-attribute decision problems.

1 INTRODUCTION

Qualitative decision problems appear in our everyday life all the time. We manage to successfully decide in cases when we have to make a few qualitative decisions. However, when we have to make many decisions, we saturate in a way that we cannot make consistent decisions for all possible situations that occur. It was shown that when dealing with problem of classification, humans face the natural upper limit capacity to distinguish among five different classes [1]. Trained users may distinguish among seven different classes, and really highly skilled users may achieve to distinguish nine different levels. This limit to distinguish up to nine classes, constrains us in the way we perform decision making when faced with the problem of evaluating many options. Additionally, when we evaluate many qualitative options, it often happens that several options belong to the same qualitative output class which means that they are almost equally preferred [2]. In order to distinguish among the options that belong to the same class, ranking of options within classes has to be performed. Therefore, in order to consistently rank qualitative options we seek for algorithms that would support our decisions in the process of qualitative decision making.

2 DESCRIPTION OF THE PROBLEM

One of the methodologies that deals with qualitative multi attribute decision problems is DEX [3]. Attributes in DEX are represented with discrete or qualitative values, while the inference is presented with if-then decision rules given in tabular format. In addition to the qualitative description of options in DEX, we need a numeric utility for ranking of options that belong to the same qualitative class. For this purpose, we use combined qualitative and quantitative (QQ) method [4] to obtain the numerical utility. To find the numerical utility in QQ, first a mapping of qualitative variable into quantitative variable is performed. In this process, each of the values of a qualitative variable is substituted with ordinal numbers. For example, let a qualitative variable has preferentially ordered values such as {good, better, the best}, where the decision maker has the preference of "the best" $\succ$ "better" $\succ$ "good", and where the sign $\succ$ denotes "is strictly more preferable than". Then the qualitative values are substituted with ordinal numbers, for instance {1,2,3}, where number 1 represents "good", number 2 represents "better" and number 3 represents "the best". The variables in the quantitative domain are compared with the relation "is greater than" or simply >. This mapping ensures that the greater the numerical value, the larger the preference of the decision maker.

The next step is to quantitatively evaluate the options in a way that the evaluated ranking describes the preferences of the decision maker as precisely as possible. To quantitatively evaluate the options within classes, we use the additive value function, which is a well-known method and it is easily understandable. It has the form of

$$y_i = \sum_i \omega_i \alpha_i \qquad (1)$$

where the coefficients ω_i are weighting factors (weights) and α_i are values of attributes. Such linear functions are used in many areas like in economy, commerce and operational research. For instance, the entire theory of linear programming is based on the assumption that decision makers' preferences may be represented by a linear value function [5]. The main problem with this kind of representation is how to choose the weights properly so that we can correctly rank the options that describe the decision maker's preferences. Different methods carry out this task differently, as described below in section 4.

3 EXPERIMENTAL SETUP

In our study, we have evaluated several methods for determining the weights from numerical tables that represent the decision rules used in DEX methodology. These rules are represented in the form of decision tables (DTs) that have the format given in Table 1. The first column in the decision table is the number of the option that has to be evaluated; the second and third columns are values of the two attributes of the option; the fourth column is the class to which the option belongs. All the remaining columns are evaluations of the options obtained with different methods. In our experiment, all the numeric attributes in the decision tables may acquire three different values: {1,2,3}.

Each decision table comprises all the possible combinations of the attribute values, i.e., without missing options. Each decision table represents a possible decision maker's preference, from which we try to determine the weights of the evaluation model (1).

4 OVERVIEW OF THE USED METHODS

To determine the weights, in our study we examined the performance of the following methods: QQ (qualitative-quantitative approach), different definitions of Gini index (Gini index defined by Breiman which we refer further as Gini; Gini Covariance; Gini based on the population which we refer as Gini Population), Information Gain and χ^2.

4.1 QQ Method

The Qualitative-Quantitative or QQ method [4] maps qualitative attributes into quantitative and then uses multiple linear regression to determine the weights of the additive value function. It first calculates the value of weights ω_i by using the relation

$$g = \sum_i \omega_i \alpha_i + \omega_0 \qquad (2)$$

and then it constraints the outputs of the options into intervals $c \pm 0.5$, where c is the class to which the output belongs. The final output of the QQ method is a set of functions. For each class c, the corresponding ranking function is

$$f_c = n_c g + k_c \qquad (3)$$

Here, n_c and k_c are parameters that are different for each class and that ensure that the final output of the function is in the interval $c \pm 0.5$ This means that qualitative and quantitative evaluations are always consistent: if an option belongs to a qualitative class c, then the numerical evaluation is in the interval $c \pm 0.5$. That way, when we look at a certain numerical evaluation, we immediately know the class of the option (except for borderline numerical evaluations, such as 2.5).

There are two important properties of quantitative rankings: completeness and consistency. The ranking is *complete* if there are no two options that receive the same evaluations, so that the options can be uniquely ranked from best to worst. To determine the *consistency* of ranking, we observe the differences between two options. For all pairs of options whose values of attributes differ by the same amount (i.e., the same vector), the signs of the difference of their numerical evaluations have to be the same.

The QQ method has been designed to cope with decision tables that are monotone (the class always increases or remains constant with the increasing values of attributes) and close to linear (they can be sufficiently well approximated by a linear function). Therefore, the disadvantage of QQ is that in general it cannot consistently rank non-monotone decision tables. For this reason, we have to look for other methods to perform consistent ranking.

4.2 Gini Coefficient, Information Gain and χ^2

The Gini coefficient (or Gini index) was first proposed by Italian statistician Corrado Gini in 1912 as a measure of income inequality [6]. It is mathematically defined as a ratio between the Lorenz curve that plots the income of population versus population and perfect equality of income. In later works it is defined as second order of Shannon's Entropy [7]. Since its first proposal, the Gini index has been used in many different areas to measure different kinds of distributions. In machine learning it is used for making splits in decision trees [8] and for representation of the performers of different classifiers [9].

In this paper we examine different estimates for Gini index: the definition of Gini index as introduced by Breiman et al. [8], the Gini covariance approach [6] and Gini population approach [10].

Information gain has its origin in information theory [11] and it is frequently used in decision tree learning for determining the attribute that gives most information regarding some splitting criteria. It is defined as the difference between the original information and the information obtained after using an attribute to split the decision tree.

χ^2 distribution has its origin in statistics and was devised as a test of goodness of fit [12] of an observed distribution to a theoretical one.

In this paper we exploited the Gini index, Information Gain and χ^2 for the calculation of weights ω_i in (1) for non-monotone decision tables. Unlike QQ, which uses multiple linear regression for determining weights ω_i in (2), these methods use non-linear calculations to obtain some measure of influence of each of the attributes on the output class. This measure is used to determine the weights ω_i in (1). As soon as the weights are obtained, the evaluation and ranking of options is exactly the same as in QQ. We continue to determine the value of the function y_i as given in (1), and finally we constrain the output rankings to the interval $c \pm 0.5$ by using (3).

5 RESULTS AND DISCUSSION

We evaluated the selected methods on a complete set of all the decision tables which map two three-valued attributes (x_1 and x_2) to a three-valued class – in total, there are 3^9=19.683 different tables.

As evaluation criteria, we used the performance of consistent rankings of each method within classes. However, there are decision tables for which exist multiple consistent and complete ranking solutions. Such a decision table with two ranking solutions is presented in Table 5. Namely, in Table 5 in class two, ranking of options with numbers 4, 5 and 6 is different when Gini Population method is used compared to the one when other methods are used. In that case, as evaluation criteria we use the sum of relative absolute error (RAE) over all table rows. We choose as the best method, the one with the smallest RAE.

Although QQ method was originally developed for ranking monotone decision tables, we evaluated its performances on all the decision tables. Results show that QQ provides a complete ranking of 13 % of the whole set of decision tables. The rest of the decision tables are not consistently and/or completely ranked with QQ.

Ranking with Gini methods provides better results than ranking with QQ for non-monotone cases, however there are differences in rankings depending on the used estimator for the Gini methods. In general, ranking with Information Gain and χ^2 is better than ranking with QQ, Gini and Gini Covariance, but worse that ranking with Gini Population method. The percentage of completely ranked decision tables with each method is given in Figure 1.

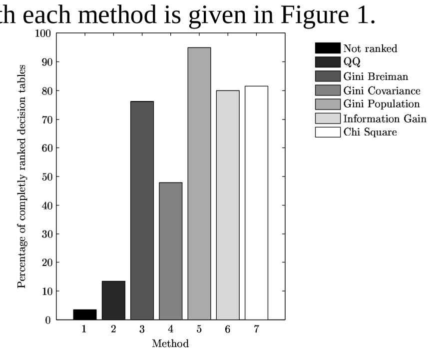


Figure 1: *Distribution of DTs that are completely ranked with different methods*

As shown in Figure 2, results are divided into five groups:

- Group 1: decision tables that cannot be completely ranked by any of the methods (an example is given in Table 1 and calculated weights are given in Table 2);
- Group 2: decision tables that are completely and consistently ranked only by Gini Population method (example of a decision table of this kind is given in Table 3 and calculated weights are given in Table 4);
- Group 3: decision tables that are completely and consistently ranked with QQ method, but also with other methods;
- Group 4: decision tables that are improperly ranked by QQ, but all other methods perform flawlessly (an example is given in Table 5 and weights in Table 6);
- Group 5: decision tables that are completely and consistently ranked only by Gini, Information Gain and χ^2.

For the decision tables in group 1 we have to look for algorithms other then those included in this research. These decision tables have in common that:

Table 1: *Example of DT for which all methods fail to provide complete ranking*

No.	x_1	x_2	class	All methods
1	1	1	1	0.83
2	2	1	1	1.16
3	1	2	1	1.16
4	3	1	2	2.00
5	2	2	2	2.00
6	1	3	2	2.00
7	3	2	3	2.83
8	2	3	3	2.83
9	3	3	3	3.16

Table 2**:** *Weights obtained for the DT given in* Table 1

weights	QQ	Other methods
ω_0	33.33	n/a
ω_1	33.33	50
ω_2	33.33	50

Table 3**:** *Example of DT for which complete ranking is possible only with the method Gini Population*

No.	x_1	x_2	class	Gini Population	All other methods
1	1	1	1	0.73	1.00
2	1	2	1	1.00	1.00
3	1	3	1	1.26	1.00
4	2	1	2	1.69	1.75
5	2	2	2	1.91	1.75
6	3	3	2	2.30	2.25
7	3	1	3	2.79	3.25
8	3	2	3	3.12	3.25
9	2	3	3	3.20	2.75

Table 4: *Weights obtained for the DT given in* Table 3

weights	QQ	Gini Population	Other methods
ω_0	0.00	n/a	n/a
ω_1	71.42	56.74	0
ω_2	28.57	43.25	100

Table 5: *Example of DT for which only Gini Population provides different complete ranking compared with other methods and RAE is smallest for Gini Population*

No.	x_1	x_2	class	Gini Population	Information Gain	Other methods
1	1	1	1	0.73	0.80	0.81
2	1	2	1	1.00	1.00	1.00
3	1	3	1	1.26	1.19	1.18
4	2	1	2	1.81	1.79	1.79
5	3	1	2	2.09	2.20	2.20
6	2	2	2	2.18	1.98	1.95
7	3	2	3	2.81	3.01	3.04
8	2	3	3	2.90	2.79	2.79
9	3	3	3	3.18	3.20	3.20

Table 6: *Weights obtained for the DT given in* Table 5

weights	QQ	Gini Population	Information Gain	Other methods
ω_0	22.22	n/a	n/a	n/a
ω_1	55.55	56.74	31.36	28.57
ω_2	22.22	43.25	68.63	71.42

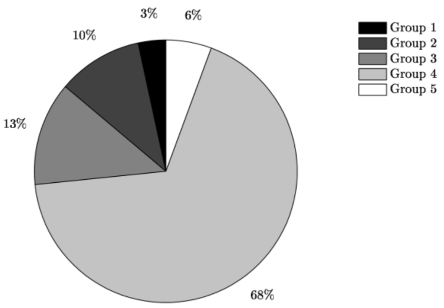


Figure 2: *Distribution of DTs ranking results in five groups*

- the methods provide equal weights for the two attributes or
- the methods choose that only one attribute is important and weight that attribute 100 %, while they weigh the second attribute 0 %.

As a general instruction for ranking of decision tables with two three-valued attributes, we propose:

- to first rank using Gini Population method, and
- if Gini Population fails, rank with any of the three methods: Gini, Information Gain or χ^2.

If none of the above provides a complete ranking, then the decision table belongs to the small group of decision tables that are not fully ranked with any of the discussed methods. In this case, we have to accept the incomplete ranking or seek for a different approach.

6 CONCLUSION

In this work we modeled 19.683 decision tables that consist of two three-valued attributes and three-valued class, using 6 different non-linear techniques for determining the weights of the additive weighting function model. We have shown that the QQ method may be used for ranking in 13 % of the whole set of decision tables. We manage to rank the options in most of the decision tables when weights in the model are determined by using different estimates of Gini coefficients, from which the most prominent one is Gini Population. Additionally, for one group of decision tables, the full ranking is possible only when using Gini, Information Gain and χ^2. Furthermore, when multiple complete rankings exist for a DT, we propose to use the one with smalles RAE. In addition 3 % of the set of non-monotone decision tables are not fully ranked with any of the methods. For them we have to further investigate other methods. In future we want to investigate the applicability of the described methods for different kinds of decision tables, for example for decision tables with more than two attributes, with different domains of the attributers and different number of options.

Acknowledgments

The research of the first author was supported by Ad futura Programme of the Slovenian Human Resources and Scholarship Fund.

References

[1] T.L.Saaty, Why the magic number seven plus or minus two, *Math. Comput. Modell.*, vol 38. pp 233-244, 2003.
[2] M. Bohanec, *Odločanje in modeli. DFMA*, 2006.
[3] M. Bohanec, DEXi: Program for multi-attribute decision making, User's manual, Version 3.00. IJS Report DP-9989, Jožef Štefan Institute, Ljubljana, 2008, http://kt.ijs.si/MarkoBohanec/pub/DEXiManual30r.pdf
[4] M. Bohanec, B. Urh, V. Rajkovič, Evaluation of options by combined qualitative and quantitative methods, *Acta Psychol*, vol. 80, pp. 67-89, 1992.
[5] S. French, *Decision theory: an introduction to the mathematics of rationality*. NY, Halsted Press, 1986.
[6] K. Xu, How has the literature on Gini index evolved in the past 80 years, Dalhousie University, Department of Economics, 2004.
[7] L. Wehenkel, On uncertainty measures used for decision tree induction, *P IPMU*, pp. 413-418, Granada, Spain 1996.
[8] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Belmont, California, USA Wadsworth Publ Company, 1984.
[9] D. J. Hand, R. J. Till, A simple generalization of the area under the ROC curve for multiple class classification problems, *Mach Learn*, vol. 45 no. 2, pp. 171-186, 2001.
[10] F. Noorbakhsh, International convergence or higher inequality in human development? Evidence for 1975-2002, in (G. Mavrotas and T. Shorrocks, eds.), *Adv Dev: Core Thms in Glob Econ*, pp. 149-167, 2007
[11] L. E. Raileanu, K. Stofiel, Theoretical comparison between the Gini index and Information Gain criteria, *Ann Math Artif Intel*, vol. 41, pp. 77-93, 2000.
[12] R. A. Fisher, On a Distribution Yielding the Error Functions of Several Well Known Statistics, *P Int Congr Math, Toronto*, vol. 2, pp 805-813, 1924.

INTELIGENTNI SISTEM ZA PODPORO KONSTRUIRANJU PLOŠČE STISKALNICE

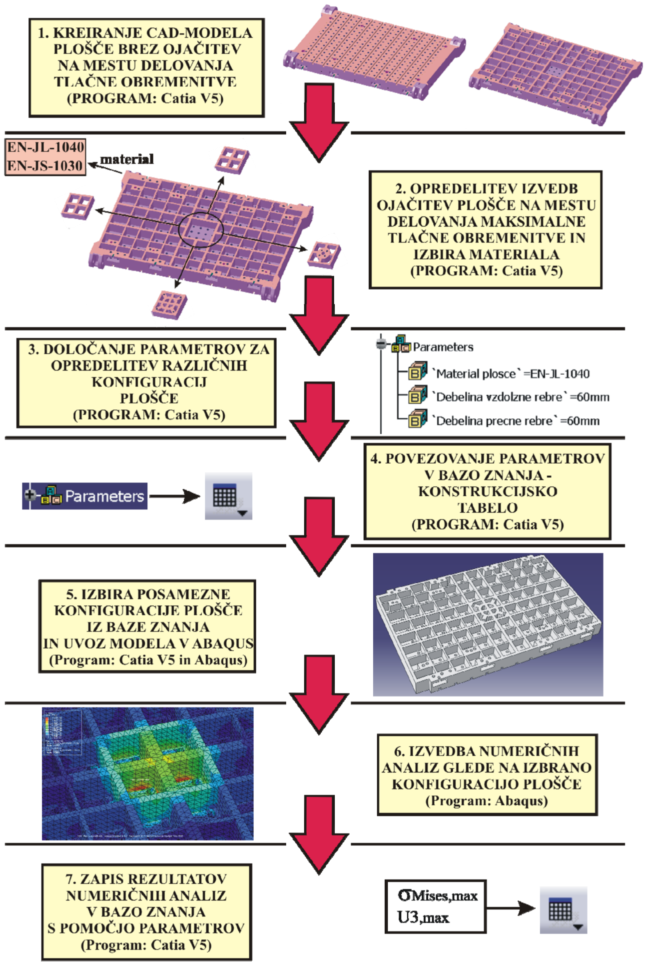
David Potočnik[1], *Stanko Stepišnik*[1], *Bojan Dolšak*[2]
[1]EMO – Orodjarna d. o. o., [2]Fakulteta za strojništvo v Mariboru
Bežigrajska 10, 3000 Celje, Slovenija
Tel: +386 3 4282100; fax: +386 3 4282101
e-mail: david.potocnik@emo-orodjarna.si

POVZETEK

V prispevku predstavljamo inteligentni sistem, ki je konstrukterju (oblikovalcu) sposoben podati kakovostno podporo pri sprejemanju odločitev z vidika oblikovanja ojačitve plošče stiskalnice na mestu delovanja maksimalne tlačne obremenitve ter pri izbiri ustreznega materiala plošče. Na podlagi izbrane kombinacije ojačitve in materiala sistem uporabnika seznani z velikostjo in položajem največje primerjalne napetosti ter največjega pomika v smeri delovanja obremenitve. Sistem deluje na podlagi vgrajenega znanja strokovnjakov s področja konstruiranja, izbora materialov ter izvajanja numeričnih analiz po metodi končnih elementov (krajše MKE), ki smo ga s pomočjo parametrov zapisali v bazo znanja znotraj CAD-sistema Catia V5.

1 UVOD

Danes je računalniško podprto oblikovanje ali računalniško podprto konstruiranje (angl.: Computer Aided Design = CAD) tako razširjeno, da večina podjetij večji del konstrukcijskega procesa izvede z uporabo CAD-sistemov. Obstaja pa tudi mnenje, da so prednosti, ki jih CAD-orodja prinašajo, pod pričakovanji. Menimo, da leži razlog za to v trenutni neustreznosti obstoječih CAD-sistemov v smislu nudenja strokovne pomoči inženirju med procesom konstruiranja novega izdelka. Sodobni CAD-programski paketi so osredotočeni le na grafični prikaz izdelka, ne pomagajo pa konstrukterju pri kreativnem delu, kot sta sprejemanje odločitev ali iskanje konstrukcijskih rešitev [4].
Dokazano je, da povečanje inteligence obstoječe računalniške programske opreme, kot je CAD-sistem, vodi k pomembnemu napredku na področju učinkovitosti in zanesljivosti pri opravljanju različnih inženirskih nalog, tudi konstruiranju [8]. Prav tako je eden izmed ključnih problemov vsakega podjetja, ki se ukvarja z razvojem izdelkov, zajemanje in ponovna uporaba obstoječega znanja. Vsako tovrstno podjetje namreč dnevno ustvarja podatke in informacije o izdelkih, pretvarjanje teh podatkov v dostopno bazo znanja pa predstavlja bolj zahtevno opravilo [7].
Na temo inteligentnih sistemov za podporo konstruiranju v strojništvu je objavljena vrsta prispevkov [1, 2, 3, 5], v pričajočem pa predstavljamo inteligentni sistem, ki je namenjen konstrukterski uporabi v orodjarstvu. Namen izgradnje inteligentnega sistema za podporo konstruiranju plošče stiskalnice je bil ravno odpravljanje predhodno omenjene problematike, in sicer ozkega grla CAD-sistema Catia V5 v smislu podpore konstrukterju, kar smo dosegli z uporabo znanja in izkušenj strokovnjakov za povečanje »inteligence« CAD-sistema.



Slika 1: *Zaporedni koraki izgradnje inteligentnega sistema za podporo konstruiranju plošče stiskalnice*

Skozi prispevek podrobno opisujemo izgradnjo omenjenega inteligentnega sistema, ki je potekala v več zaporednih korakih (Slika 1), njegovo delovanje ter podamo njegovo uporabno vrednost v smislu podpore neizkušenim konstrukterjem.

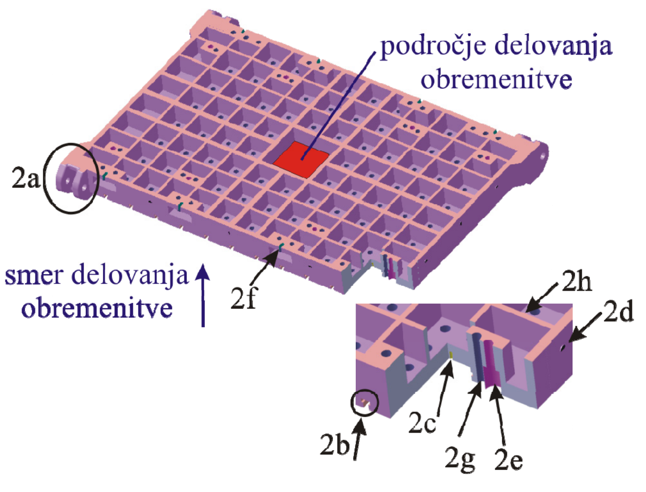
2 DELOVNA FUNKCIJA PLOŠČE STISKALNICE IN CILJ IZGRADNJE SISTEMA

Funkcija, ki jo opravlja plošča stiskalnice, je prenašanje delovnih obremenitev (predvsem tlačnih) s stiskalnice na orodja ter omogočanje pritrditve preoblikovalnih orodij. Željeno je, da je plošča čim lažja, hkrati pa morajo biti napetosti, ki se pojavijo v plošči pri maksimalni obremenitvi, manjše ali enake dopustnim. Slednje so odvisne od vrste izbranega materiala in konstrukcijske izvedbe ojačitve na mestu delovanja maksimalne obremenitve.
Neizkušeni konstrukterji pogosto nimajo zamisli in znanja, kako oblikovati ojačitve na mestu delovanja maksimalne obremenitve ali se celo ne zavedajo, da je od oblike in dimenzij teh odvisna velikost notranjih napetosti, ki se v plošči pojavijo ob delovanju obremenitve. Enake težave se pojavijo pri izbiri ustreznega materiala ter pri razumevanju posledic izbire določenega materiala.
Preko uporabe sistema smo jim želeli omogočiti zapolnitev teh vrzeli tako, da jim bo ta pomagal pri iskanju omenjenih konstrukcijskih rešitev in sprejemanju odločitev.

3 IZGRADNJA SISTEMA

3.1 Kreiranje parametričnega CAD-modela plošče znotraj programa Catia V5

V prvi fazi smo znotraj programa Catia V5 zgradili CAD-model plošče stiskalnice s popolnoma opredeljeno geometrijo, vendar brez vsebnosti ojačitev na mestu delovanja maksimalne tlačne obremenitve (Slika 2).

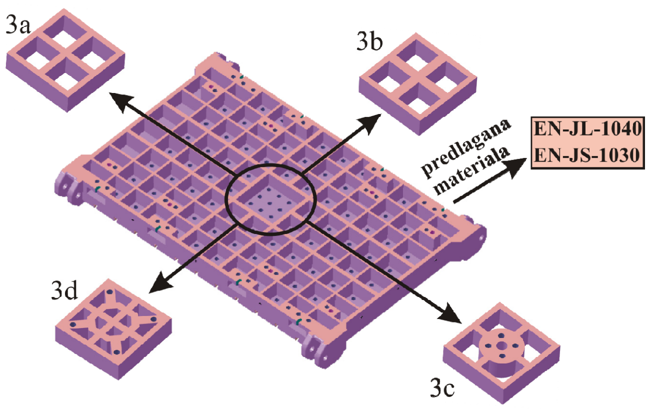


Slika 2: *Sestavni gradniki plošče*

Na ta način je plošča vsebovala naslednje gradnike oziroma značilnosti, ki so detaljno prikazani tudi na Sliki 2: obešala za transport (2a), T-utore (2b) in navojne izvrtine (2c) za pritrditev preoblikovalnih orodij, izhodišča namenjena strojnim obdelavam (2d), poglobljene izvrtine (2e) in stranske utore (2f) namenjene pritrditvi plošče na stiskalnico, utore za sornike vlečne blazine (2g) ter ojačitve v obliki reber (2h) na mestih, kjer ne deluje maksimalna obremenitev. Parametrično zgrajen model plošče omogoča enostavno spreminjanje modela v realnem času.

3.2 Opredelitev izvedb ojačitev na mestu delovanja maksimalne tlačne obremenitve in izbira materiala plošče

Nadaljevali smo z opredelitvijo konstrukcijskih različic ojačitev na mestu delovanja maksimalne tlačne obremenitve, ki smo jih v parametrični model plošče vključili na podlagi priporočil strokovnjakov. Pri tem je bila upoštevana zmožnost izdelave (upoštevanje standardov in priporočil za oblikovanje odlitkov) ter predhodne izkušnje pri oblikovanju tovrstnih elementov.
S strani strokovnjakov za konstruiranje so bile predlagane štiri različne izvedbe ojačitev prikazane na Sliki 3, in sicer: križna ojačitev z debelino reber 60 mm (3a), križna ojačitev z debelino reber 80 mm (3b), ojačitev v obliki valja s premerom 160 mm (3c) ter ojačitev, ki ima profil oblike šestkotnika in stranicami debeline, debeline 60 mm (3d). Vse naštete različice ojačitev smo v programskem paketu Catia V5 glede na podana priporočila tudi zmodelirali.
Za material plošče so strokovnjaki predlagali dve možnosti, in sicer EN-JL-1040 (siva litina z lamelarnim grafitom) ter EN-JS-1030 (nodularna litina s krogličastim grafitom). Izbiro so utemeljili z dejstvom, da je odlika sive litine tlačna trdnost in jo izkoriščamo le tam, kjer je material obremenjen na tlak, ne sme pa biti obremenjen na upogib. Nodularna litina pa po svojih statičnih in dinamičnih lastnostih ustreza jeklu, obenem pa zadrži lastnosti, ki sta pri sivi litini najbolj cenjeni: dušilnost in obrabno trdnost. Predlagana materiala torej z vidika zagotavljanja funkcionalnosti plošče predstavljata ustrezno izbiro.



Slika 3: *Prikaz s strani strokovnjakov predlaganih izvedb ojačitev plošče na mestu delovanja maksimalne tlačne obremenitve in izbire materialov*

3.3 Določanje parametrov za opredelitev različnih konfiguracij plošče

Sledilo je določevanje parametrov CAD-modelu plošče, na osnovi katerih smo določili različne konfiguracije plošče za izvedbo numeričnih analiz. Na ta način smo ustvarili tri uporabniške parametre, in sicer večvrednostni parameter tipa string za opredelitev materiala plošče ter dolžinska parametra, ki bosta opredeljevala debelino središčne vzdolžne in prečne rebre plošče, s čimer bomo lahko v bazi podatkov, tj. konstrukcijski tabeli, na enostaven način opredelili obe predlagani križni ojačitvi. Poleg omenjenih, smo preučili tudi notranje Boolove parametre (ti zavzemajo vrednosti true ali false), ki predstavljajo posamezne izvedbe ojačitev plošče. Te bomo v nadaljevanju glede na zamišljeno konfiguracijo plošče za izvedbo numeričnih analiz v bazi znanja aktivirali ali deaktivirali.
Parametri predstavljajo lastnosti dokumenta Catia. Definirani so kot gradniki ter vsebujejo vrednost. Parametre uporabljamo za popis lastnosti modela in jih lahko definiramo z relacijami ali jih v relacijah uporabimo kot argumente. Podrobnejši opis parametrov znotraj programa Catia V5 je podan v [7].

3.4 Povezovanje parametrov v bazo znanja

Parametre smo smiselno vključili v konstrukcijsko tabelo, ki predstavlja bazo znanja, pri tem pa s pomočjo deaktivacije nekaterih aktivnosti izključili geometrijo, ki bistveno ne bi vplivala na rezultate numeričnih analiz (obešala, izhodišča za strojne obdelave), ampak bi njena prisotnost le občutno podaljšala računske čase. Tako opredeljeni parametrični modeli predstavljajo kakovostno izhodišče za opravljanje inženirskih analiz, saj lahko s pomočjo parametrov, ki smo jih vključili v konstrukcijsko tabelo, enostavno izbiramo med lastnostmi modela, ki jih želimo uporabiti za izvedbo analiz.
Glede na s strani strokovnjakov predlaganih izvedb ojačitev plošče na mestu delovanja tlačne obremenitve in glede na predlagana materiala smo s pomočjo parametrov v konstrukcijski tabeli opredelili 8 različnih konfiguracij plošče (Slika 4). Pri tem velja poudariti, da vsak stolpec tabele definira vrednosti parametra, vsaka vrstica pa njegovo konfiguracijo.
Na ta način smo do sedaj v bazo znanja vključili znanje s področja oblikovanja ojačitev, znanje s področja izbire ustreznega materiala ter znanje priprave CAD-modela plošče stiskalnice za kakovostno izvedbo numeričnih analiz.

3.5 Izbira konfiguracije plošče in uvoz modela v program za izvajanje numeričnih analiz Abaqus

Preko izbire v konstrukcijski tabeli smo posamezne konfiguracije plošče izvozili v program za izvajanje numeričnih analiz Abaqus. Z njegovo uporabo smo želeli pridobiti rezultate o največjih primerjalnih napetostih in največjih pomikih, s pomočjo katerih bomo lahko kasneje nadgradili obstoječo bazo znanja.

3.6 Izvedba numeričnih analiz glede na izbrano konfiguracijo modela plošče

Inženirske analize nam omogočajo, da bolje razumemo obnašanje konstrukcije in nam dajejo smernice za optimizacijo konstrukcije [6]. Vendar pa neizkušeni konstrukterji velikokrat premalo poznajo osnovna načela inženirskih analiz in zaradi tega smo želeli rezultate izvedenih analiz vključiti v bazo znanja Catia V5, s čimer bi lahko tudi konstrukterji brez znanja izvajanja analiz dobili kakovostno informacijo o tem, kako vpliva izbira ojačitve in materiala plošče na pojav notranjih napetosti in pomikov.
Numerične analize po MKE so izvedli strokovnjaki iz tega področja, pri čemer smo jim posredovali vse potrebne podatke, ki so podani v nadaljevanju.
Za opredelitev izotropnega materiala sta bila potrebna zgolj modul elastičnosti in Poissonovo število, zaradi upoštevanja težnosti plošče pa so bili v materialni model vključeni tudi podatki o gostoti obeh materialov (Razpredelnica 1).

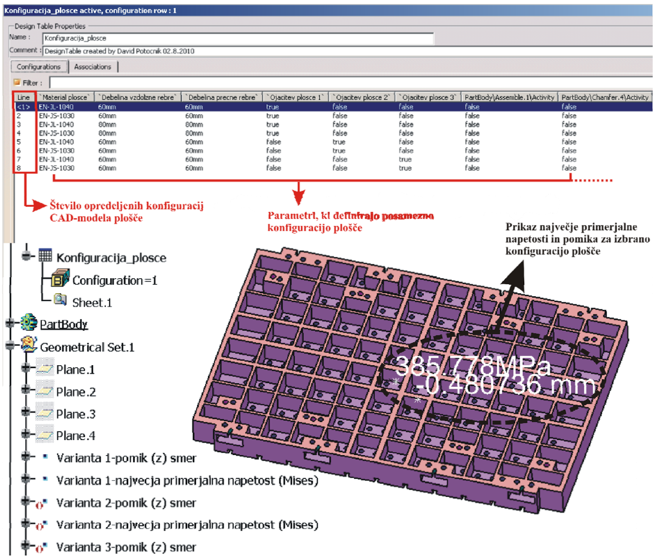
Razpredelnica 1: *Prikaz materialnih podatkov plošče za izvedbo numeričnih analiz*

Material	**Modul elastičnosti (E)**	**Poissonovo število (ν)**	**Gostota (ρ)**
EN-JS-1030 (nodularna litina)	169000 N/mm²	0,275	$7.1 \cdot 10^{-6}$ kg/mm^2
EN-JL-1040 (siva litina)	110500 N/mm²	0,26	$7.2 \cdot 10^{-6}$ kg/mm^2

Pri modeliranju smo izhajali iz predpostavke, da bodo orodja vedno nameščena v središču plošče. Najmanjša površina orodja, na katero bo delovala stiskalnica, bo znašala 250000 mm^2 (500 mm x 500 mm), pri čemer bo pritisk na ploščo pri maksimalni sili stiskalnice 30 MN znašal 120 N/mm². Poleg omenjene obremenitve je bila upoštevana tudi teža plošče, kjer je bila za velikost zemeljskega pospeška vzeta vrednost $9{,}81 \cdot 10^3$ mm/s². Modeli so bili zamreženi s tetragonalnimi volumskimi paraboličnimi elementi globalne velikosti 40. Definiranju robnih pogojev vpetja sta sledila pogon in spremljanje analize. Izvedenih je bilo 8 statičnih materialno-linearnih analiz, in sicer za vsako posamezno konfiguracijo plošče iz konstrukcijske tabele programa Catia V5.

3.7 Zapis rezultatov numerične analize v bazo znanja Catia V5

Na podlagi izvedbe numeričnih analiz smo pridobili rezultate o največjih dopustnih napetostih po Misesu ter o maksimalnemu pomiku vozlišča na plošči. Rezultati so pokazali, da so največje primerjalne napetosti pri vseh konfiguracijah plošče znotraj dopustnih. S pomočjo točk in tekstovnih zapisov smo te podatke opredelili znotraj programa Catia V5. Zatem smo parametre, ki predstavljajo omenjene gradnike, povezali s pripadajočo konfiguracijo plošče ter s tem v bazo znanja vključili še rezultate numeričnih analiz. Podrobne podatke o izvedbi analiz in interpretaciji njihovih rezultatov smo za posamezno konfiguracijo pregledno zapisali tudi v obliki tekstovne datoteke in grafičnih izpisov tako, da lahko uporabnik kadarkoli dostopa do omenjenih podatkov.



Slika 4: *Izbira konfiguracije plošče iz baze znanja (zgoraj) in z znanjem opremljen CAD-model plošče (spodaj)*

4 UPORABA IN DELOVANJE SISTEMA TER MOŽNOSTI NJEGOVEGA NADALJNJEGA RAZVOJA

Inteligentni sistem za podporo konstruiranju plošče stiskalnice se uporablja kot podpora neizkušenim konstrukterjem znotraj podjetja Emo – Orodjarna. Sistem deluje na podlagi vgrajenega znanja strokovnjakov s področja konstruiranja, izbora materialov ter izvajanja numeričnih analiz po metodi končnih elementov, ki je pomočjo parametrov zapisano v bazo znanja. Določena kombinacija parametrov predstavlja posamezno konfiguracijo plošče, do katere lahko uporabnik enostavno dostopa preko drevesnega menija programa Catia V5. Preko aktivacije določene konfiguracije plošče v konstrukcijski tabeli sistem uporabniku ponudi konstrukcijsko rešitev ojačitve plošče na mestu delovanja maksimalne tlačne obremenitve, ponudi mu ustrezen material za izgradnjo plošče ter ga seznani, kako oblika ojačitve in material vplivata na velikost in položaj največje primerjalne napetosti ter največjega pomika plošče v smeri delovanja obremenitve.

Možnosti nadaljnje nadgradnje sistema se kažejo v:

- vključitvi dodatnih variant ojačitev plošče ter nekaterih drugih materialov primernih za izgradnjo tlačno-obremenjene plošče,
- izgradnji vmesnika, ki bi medsebojno primerjal rezultate numeričnih analiz posameznih konfiguracij plošče,
- izvedbi dinamičnih analiz ter analiz kontaktnih problemov vezanih na ploščo stiskalnice ter v zapisu pridobljenih rezultatov v bazo znanja.

5 ZAKLJUČEK

Razvili smo inteligentni sistem namenjen podpori konstruiranju plošče stiskalnice. Ta neizkušenemu konstrukterju: omogoča ustrezno izbiro materiala plošče, mu poda konstrukcijske rešitve v smislu oblikovanja ojačitve plošče na mestu delovanja maksimalne tlačne obremenitve ter ga seznani, kako oblika ojačitve in vrsta materiala vplivata na velikost in položaj največje primerjalne napetosti ter največjega pomika plošče v smeri delovanja obremenitve. Uporaba sistema uporabniku omogoča da pride do spoznanj, ki so ključnega pomena za konstruiranje plošče stiskalnice.

6 Viri

[1] Dagli, C. H., Kusiak, A.: *Intelligent Systems in Design and Manufacturing*. ASME Press, New York, 1994.

[2] Dixon, J. R.: *Artificial Intelligence and Design*: A Mechanical Engineering View. Proceedings of 5th National Conference On AI. Los Altos, 1986. Str. 872–77.

[3] Dixon, J. R., Simmons, M. K.: *Expert System for Design: Standard V-belt Driven Design as an Example of the Design-Evaluate-Redesign Architecture*. Proceedings of International Computers in Engineering Conference. Las Vegas, Nevada, 1984. Str. 332–37.

[4] McMahon, C., Browne, J.: *CADCAM: from principles to practice*. Wokingham: Addison-Wesley, 1993.

[5] Mostow, J.: *Toward Better Models of the Design Process*. AI Magazine, št. 6 (5), 1985. Str. 44–7.

[6] Novak, M.: *Inteligentni računalniški sistemi za pomoč pri optimiranju konstrukcij.* Doktorska disertacija. Maribor, 2004.

[7] Potočnik, D., Pesan, B., Ulbin M.: Parametrično-asociativni pristop h konstruiranju preoblikovalnega orodja v programskem paketu Catia V5. V *Vir znanja in izkušenj za stroko /Industrijski forum IRT. Škofljica,* PROFIDTP, 2010.

[8] Sancin, U., Kaljun, J., Dolšak, B.: Intelligent support to specific design aspects. *WSEAS transactions on information science and applications*. Feb. 2008, št. 5. Str. 192–201.

Zahvala:

»Operacijo delno financira Evropska unija, in sicer iz Evropskega socialnega sklada.«

ROC CURVES COMPARISON OF INFERRED GENE REGULATORY NETWORKS

Blagoj Ristevski[1], *Suzana Loskovska*[2]
[1]Department of Information Systems Management
Faculty of Administration and Information Systems Management
St. Kliment Ohridski University – Bitola, Macedonia
[2]Department of Computer Science and Informatics
Faculty of Electrical Engineering and Information Technologies
Ss. Cyril and Methodius University – Skopje, Macedonia
e-mail: [1]blagoj.ristevski@uklo.edu.mk, [2]suze@feit.ukim.edu.mk

ABSTRACT

Gene regulatory networks are utilized to decipher and reveal the fundamental cells' gene regulatory mechanisms. They are composed of nodes representing genes, transcription factors and other components and edges representing interactions among nodes. In this paper we compare the obtained ROC curves of the gene regulatory networks, inferred from gene expression microarray data. As network inference models we have used graphical Gaussian models, dynamic Bayesian networks and Boolean networks. We have generated three artificial different size gene expression datasets using a simple autoregressive process. We illustrated and compared the ROC curves as a validation measure, too. At the end of this paper, some directions toward a novel method for GRNs inference including prior knowledge are suggested.

Keywords: bioinformatics, gene regulatory networks, ROC curves, graphical Gaussian models, dynamic Bayesian networks, Boolean networks.

1 INTRODUCTION

The gene regulatory networks (GRNs) provide an understandable view of gene regulatory mechanisms and can uncover the reasons for many diseases. Their components are nodes which represent the genes, metabolites, proteins or modules, and edges which correspond to the interactions between nodes.

The central dogma in molecular biology is presented by two processes: transcription and translation. Genes are transcribed into mRNA during transcription and after that proteins are produced by translation. When the protein is synthesized the corresponding coding gene is expressed. The gene expression levels correspond to the approximate number of produced RNA copies from the gene, i.e. gene expression corresponds to the amount of produced proteins. The microarray technology provides gene expression data as a measurement of gene expression under particular experimental conditions or different time points [8].

The reverse engineering of gene regulatory mechanisms from experimental data is called GRNs inference. There are many models for GRNs inference which are usually based on Boolean networks, Bayesian networks, dynamic Bayesian networks, graphical Gaussian models, linear and nonlinear differential and difference equations systems.

In this paper, we apply Boolean networks, dynamic Bayesian networks and graphical Gaussian models for GRNs inference and we validate the models using ROC curves.

The rest of this paper is structured as follows. In Section 2 we depict the Boolean networks. Bayesian Networks and dynamic Bayesian networks are described in the following section. In Section 4 we present the graphical Gaussian models, their assumptions and application in the GRNs inference and explain the partial correlation coefficients and their meaning for network inference. The receiver operating characteristic (ROC) curves as a validation measure of inferred networks are described in Section 5. In Section 6 we describe the simulated gene expression data which are used as input data for GRNs inference. The ROC curves which validate the inferred networks are shown and the performances of three models depending on the datasets size are compared. Finally, the concluding remarks and future directions are given in the last section.

2 BOOLEAN NETWORKS

Boolean Networks are simple models for GRNs inference. They are composed of set of nodes and edges. The nodes correspond to the genes whereas the edges between the nodes represent the gene interactions. In Boolean networks, gene expression levels are represented by two-level states: 1 and 0.

The graph representing a Boolean network gives information about the wires between genes, but it is insufficient for understanding dependencies between genes. The aim of the reverse engineering by Boolean networks is to find a Boolean function of every gene in the network, so that discretized values of gene expression can be presented by the function. But, the small changes in the gene expression levels cannot be covered by two levels discretization, which leads to big information loss. Another shortcoming of Boolean networks is the super-exponential number of all possible networks 2^{2^n} depending on the number of genes n.

Liang et *al*. (1998) have introduced REVerse Engineering Algorithm (REVEAL) which constructs a Boolean network of given gene expression data by setting the gene in-deegree k [9]. The number of all possible networks is calculated by the Eq. 1:

$$\left(2^{2^k} \frac{n!}{(n-k)!}\right)^n \tag{1}$$

This algorithm extracts minimal network structures from the state transition tables of the Boolean network using the mutual information approach.

3 BAYESIAN NETWORKS

Bayesian networks (BNs) are a special case of graph model defined as a triple (G, F, θ), where G denotes the structure of the graph, F is a set of conditional probability distributions, and θ is a set of parameters for the graph structure [10]. The graph structure G is consisted of set of n nodes x_1, x_2, ..., x_n and set of directed edges between the nodes. The nodes correspond to the random variables whereas the directed edges show the conditional dependences between the variables (genes). Edges and nodes together have to create a directed acyclic graph (DAG).

The joint probability distribution is given by Eq. 2:

$$p(x) = \prod_{i=1}^{n} p\left(x_i \middle| x_{\{1,\cdots,i-1\}}, \theta, G\right) \tag{2}$$

Lets pa_i denotes the parent nodes of the node x_i which means that the state of each variable x_i depends on the states of its parent pa_i so previous equation can be presented as:

$$p(x) = \prod_{i=1}^{n} p\left(x_i \middle| pa_i, \theta, G\right) \tag{3}$$

BNs can deal with noisy and incomplete data, and stochastic nature of gene expression data. The small number of data points (samples) and large number of genes are common problems for BNs learning. Another disadvantage is that BNs cannot handle feedback loops, although they exist in the real GRNs. Because BNs represent probabilistic relations between genes at the same time they cannot deal with time lag between different genes.

To overcome these shortcomings of BNs, dynamic Bayesian networks (DBNs) are used to infer gene regulations. DBNs can deal with stochastic variables, time series gene expression data, feedback loops, missing values, hidden variables and can include prior knowledge [11]. The hidden nodes (variables) can capture effects that cannot be directly measured in the microarray experiments.

If x_t^i represents the i-th node at time point t, the joint probability distribution is given by Eq. 4:

$$p\left(x_t \middle| x_{t-1}\right) = \prod_{i=1}^{n} p\left(x_t^i \middle| pa(x_t^i), \theta, G\right) \tag{4}$$

The GRNs inference is followed by structure and parameter learning of the BNs from training data D [7]. For given data D, the goal is finding of posterior distribution of the network structure M. Then from this distribution, the structure M^* which best fits the data should be found according to Eq. 5:

$$M^* = \arg\max_M \{P(M \mid D)\} \tag{5}$$

For an optimal network structure M^* and given data D, it needs to find posterior distribution of parameters q by Eq. 6:

$$q^* = \arg\max_q \{P(q \mid M^*, D)\} \tag{6}$$

Because the number of DAGs super-exponential depends on the number of nodes n, the BNs learning is NP-hard task. Thus BNs and DBNs are inappropriate for inference of large networks [12].

4 GRAPHICAL GAUSSIAN MODELS

Graphical Gaussian models (GGMs) are commonly used very computationally efficient models for GRNs inference based on gene expression data [3]. GGMs as graphical probabilistic models can identify conditional independence relations between the nodes. The main assumption which they make is that the input gene expression data follow a multivariate Gaussian distribution [6].

In GGMs the nodes represent genes and the edges represent conditional dependences between nodes. The absence of an edge between two genes means that the corresponding genes are conditionally independent given other genes in the model.

Let Y be the input gene expression data matrix with G columns, corresponding to the number of genes, and with N rows that correspond to the number of samples (time series data points or other experimental conditions) [3]. It is supposed that matrix Y follows a multivariate normal distribution $N_G(\mu, \Sigma)$, where $\mu = (\mu_1, ..., \mu_G)'$ is the mean vector, and $\Sigma = (\sigma_{ij})_{1 \le i, j \le G}$ is the positive definite covariance matrix. $\sigma_{ij} = \sigma_i \sigma_j$ are covariance parameters between genes i and j, and σ_i^2 refer to the variance terms for gene i.

The estimation of the covariance matrix of the data distribution is a basis for the GGMs inference.

First, in the GGMs inference, reliable estimation of the partial correlation matrix $\tilde{\mathrm{P}} = (\tilde{\rho}_{ij})$ is required [4]. This matrix is associated to the inverse matrix of the covariance matrix Σ. The straightforward estimator is given by the following Eq. 7:

$$\tilde{r}_{ij} = -\frac{\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}} \tag{7}$$

where

$$\hat{\Omega} = (\hat{\omega}_{ij}) = \hat{\Sigma}^{-1} \tag{8}$$

The partial correlation coefficients $\tilde{r}_{ij}$ describe the correlation between nodes/genes Y_i and Y_j [5]. Partial correlation between two genes measures the degree of correlation remaining after removing the effects of the other genes which differs from Pearson correlation coefficients [1] [6].

The above mentioned procedure is appropriate when *N* is larger than the number of genes *G*, otherwise the covariance matrix is not positive-definite and cannot be inverted. In microarray data, the sample size *N* is usually much smaller than the number of genes *G*. For that reason, using of shrunk estimation of the covariance matrix is suggested. The goal is to construct well conditioned positive-definite matrix so that the inverse matrix can be found. If λ is a shrinkage coefficient so that $0 \le \lambda \le 1$, then shrunk covariance matrix Σ^* is computed by Eq. 9:

$$\Sigma^* = \lambda T + (1-\lambda)S \tag{9}$$

where $\hat{S}$ is the estimated empirical covariance matrix.

The shrinkage parameter λ is chosen to minimize the mean-square error and it is determined analytically given by Eq. 10:

$$\lambda^* = \frac{\sum_{i \ne j} \mathrm{var}(r_{ij})}{\sum_{i \ne j} r_{ij}^{\ 2}} \tag{10}$$

After computing the partial correlation coefficients $\tilde{r}_{ij}$, the distribution of $|\tilde{r}_{ij}|$ is checked and the edges with significantly small values of $|\tilde{r}_{ij}|$ are removed from the network [2]. The second stage of the GRNs inference is model selection, i.e. assigning statistical significance to the edges from the GGMs network.

5 VALIDATION OF GRNs INFERENCE

To validate the inferred networks, they should be assessed by comparison with the referent network. Commonly used criterion for validation is the Receiver Operating Characteristics (ROC) curve. In a graph between two nodes it might be an edge or it might be no edge which means that each edge (instance) of the network belongs to either positive (p) or negative (n) class, respectively.

For a two-class classifier and test samples, four cases are possible:

- TP (true positive), if the instance is positive and it is classified as positive;
- FN (false negative), if the instance is positive and it is classified as negative;
- TN (true negative), if the instance is negative and it is classified as negative and
- FP (false positive), if the instance is negative and it is classified as positive.

Based on the defined TP, FN, TN and FP, the following rates are defined:

tpr true positive rate (recall):

$$tpr = \frac{TP}{P} = \frac{TP}{TP+FN} \tag{11}$$

fpr false positive rate:

$$fpr = \frac{FP}{N} = \frac{FP}{FP+TN} \tag{12}$$

The ROC curve (Fig. 1) is a chart of the ratio between *tpr* sensitivity and *fpr* (1-specificity), where sensitivity corresponds to a proportion of the actual positive edges which are correctly identified, whereas specificity is proportion of negative edges which are correctly identified [13] [14]. The ROC curve should be above the line *y*=*x* and when closer to the upper left corner it is, the better inference performances are.

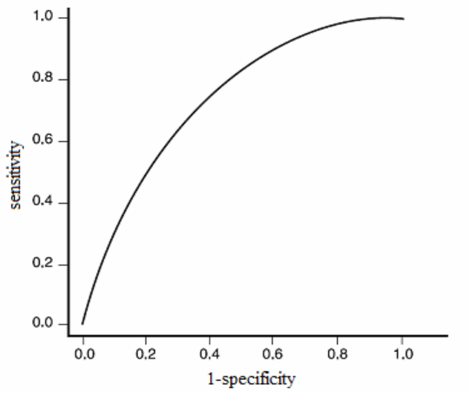


Figure 1. *ROC curve*.

6 DATA SIMULATION AND RESULTS

To infer GRNs and then to validate above described models Boolean networks, DBNs and GGMs, we generated artificial gene expression data by a simple first order autoregressive process given by Eq. 13:

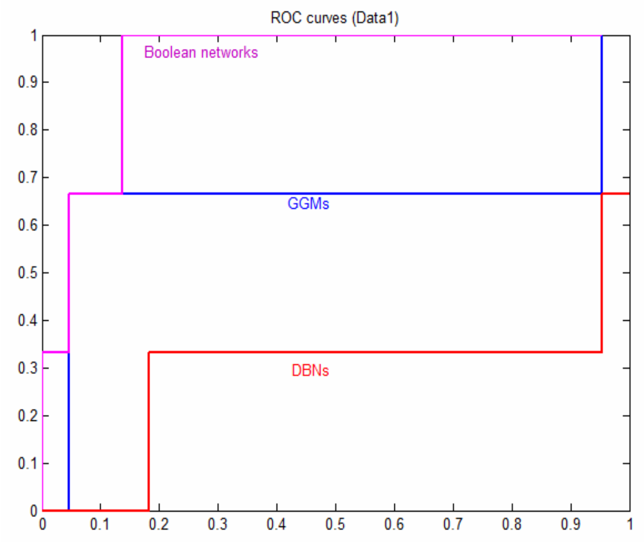
$$X(t) = Ax(t-1) + B + \varepsilon(t) \tag{13}$$

where $\varepsilon(t)$ is a vector distributed by zero-centered multivariate Gaussian distribution with diagonal variance matrix.

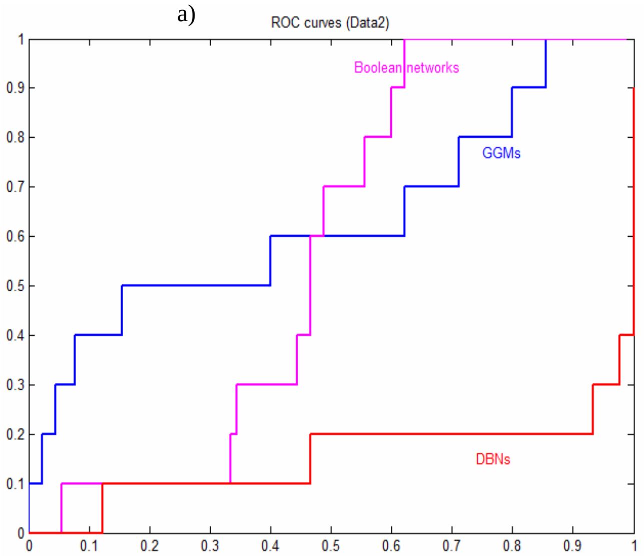
As a result of the simulation, three different size datasets are obtained. The first dataset Data1 consists of simulated gene expression data referring to 5 genes and 50 time points. The dataset Data2 corresponds to 10 genes and 50 time points, and the number of genes in the third dataset Data3 is 15 measured in 100 time points. For inference of GGMs, DBNs and Boolean networks we have used the following R packages: GeneNet, edbNet and BoolNet, respectively.

The obtained ROC curves are illustrated on Fig. 2. The ROC curves referring to Boolean networks are colored in magenta, GGMs - in blue and curves which regard to DBNs are dyed in red. Fig. 2 a) regards to the inferred GRNs networks from Data1, Fig 2 b) to the networks from Data2 and ROC curves which regard to the inferred networks from Data3 are shown on Fig. 2 c).

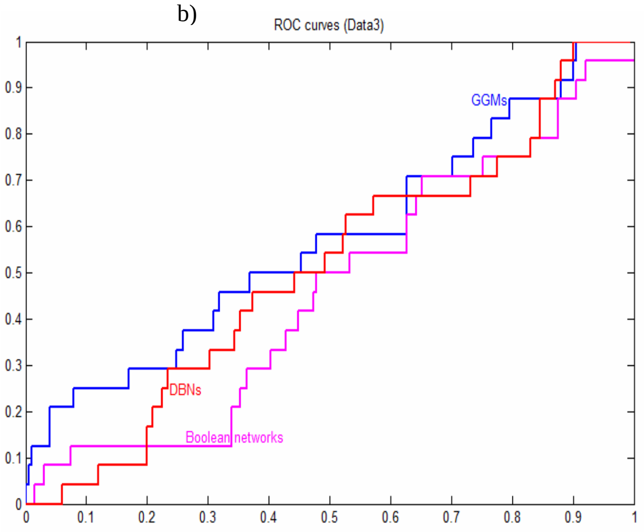
Comparing these ROC curves, we can conclude that for smaller datasets Boolean networks model has better performances than GGMs and DBNs. For medium and large datasets, GRNs inference performed by GGMs is better than the other models.



a)



b)



c)

Figure 2. *ROC curves for a) Data1, b) Data2 and c) Data 3.*

7 CONCLUSIONS AND FURTHER DIRECTIONS

The illustrated ROC curves obtained by GRNs inference using different models and datasets have shown that for datasets containing time series for larger number of genes, GGMs surpass the Boolean networks and DBNs, excluding the case where time series refer to small number of genes. For small number of genes (in our case-5 genes), the Boolean network model shows better inference performance compared to GGMs and DBNs, whereas DBNs model shows worst inference properties. According to the obtained results we suggest using the GGMs as a first stage of the novel improved approach for GRNs reconstruction. The second inference stage will be Bayesian structure learning followed by Markov Chain Monte Carlo (MCMC) method to correct the bias introduced by the prior knowledge gained from the first inference stage.

Reference

[1] N. Kraemer, J. Schaefer, A.-L. Boulesteix, *Regularized Estimation of Large-scale Gene Association Networks Using Graphical Gaussian Models*, Technical Report, Department of Statistics, University of Munich, 2009.

[2] A. V. Werhli, M. Grzegorczyk, D. Husmeier, *Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks*, Bioinformatics, Vol.22, no. 20, 2006.

[3] F. Jaffrezic, G. Tosser-Klopp, *Gene Network Reconstruction from Microarray Data*, BMC Proceedings, 2009.

[4] J. Schaefer, R. Opgen-Rhein, K. Korbinian, *Reverse Engineering Genetic Networks using GeneNet Package*, R-News 6/5:50-53, 2006.

[5] S. Ma, Q. Gong, H. J. Bohnert, *An Arabidopsis Gene Network Based on the Graphical Gaussian Model*, Genome Research, 2009.

[6] J. Schaefer, K. Strimmer, *Learning Large-Scale Graphical Gaussian Models from Genomic Data*, American Institute of Physics, Vol. 776, 2005.

[7] B. Ristevski, S. Loskovska, *Bayesian Networks Application for Representation and Structure Learning of Gene Regulatory Networks*, 12th International Conference on Computers and Information Technology ICCIT '09. , Dhaka, Bangladesh; 2009.

[8] B. Ristevski, S. Loshkovska, S. Dzeroski, I. Slavkov, *A Comparison of Validation Indices for Evaluation of Clustering Results of DNA Microarray Data*, Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008, China.

[9] S. Liang, S. Fuhrman and R. Somogyi, *REVEAL, a general reverse engineering algorithm for inference of genetic network architectures*, Pacific Symposium on Biocomputing 3, 1998,

[10] N. Friedman and M. Goldszmidt, *Learning Bayesian Networks with Local Structure*, Proceedings of the NATO Advanced Study Institute on Learning in graphical models, 1998.

[11] R. E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, 2003.

[12] M. Grzegorczyk and D. Husmeier, *Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move*, Machine Learning, 71: 265-305, 2008.

[13] L. M. de Campos, *A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests*, Journal of Machine Learning Researc 7, 2006.

[14] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters 27, 2006, pp. 861-874.

ANALIZA PODATKOV LIZING POGODB S PROGRAMSKIM PAKETOM WEKA

Peter Starbek, študent
Univerza v Ljubljani, Ekonomska fakulteta
e-pošta: peter.starbek@gmail.com

POVZETEK

Za uspešno ter učinkovito poslovanje v podjetju se je potrebno pravilno odločati. Za pravilne odločitve pa ima velike zasluge podrobna analiza vseh vrst informacij ter z njimi povezani podatki, ki nastopajo tokom poslovanja. V prispevku želim prikazati primer analize podatkov lizing podjetja X s pomočjo programskega paketa Weka. Podjetje X je eno od pomembnejših ponudnikov lizing storitev v Sloveniji. Zaradi kakršnekoli zlorabe rezultatov raziskave, je bila želja podjetja, da v referatu ne razkrijem njihovega pravega naziva. Za analizo so bili uporabljeni podatki lizing pogodb od leta 2007 do konca leta 2009. V tem obdobju je bilo sklenjenih ter aktiviranih okoli 7000 pogodb, ki so zajete v analizo. Določeni vzorci, ki so se pokazali, so splošni in ugotovitve lahko smiselno impliciramo tudi na ostala lizing podjetja. Raziskava je potrdila nekaj predhodnih predvidevanj ter obenem pokazala neke vzorce obnašanja, ki jih je smiselno upoštevati.

1 UVOD

Podjetja za uspešno ter učinkovito poslovanje potrebujejo vedno več znanja. Vse pomembnejši vir znanja so obstoječe baze podatkov, ki jih podjetje generira preko uporabe različnih računalniških programov v okviru svojega poslovanja. Žal pa so surovi podatki le malokrat neposredno uporabni ter koristni. Iz njih je potrebno informacijo izluščiti. V ta namen se v svetu informacijskih tehnologij pojavlja vse več programskimi orodji ter rešitev, ki omogočajo lažje obdelave in tudi vse bolj intelegentne analize podatkov ter s tem pridobitev koristnih informacij. Eno izmed teh orodij je tudi programski paket Weka, katerega sem uporabil za analizo nabora podatkov lizing pogodb.

Uporabljeni podatki so pridobljeni iz poslovnega sistema lizing podjetja X. Vsebujejo osnovne informacije lizing pogodb, ki jih je lizing podjetje X sklenilo z svojimi strankami ter na podlagi katerih poslovni sitem izračuna plačilni načrt odplačila glavnic ter obresti posamezne pogodbe. V analizo je bilo vključeno okoli 7000 pogodb, ki so bile aktivirane v obdobju od začetka leta 2007 do konca leta 2009.

2 NAMEN IN CILJ

Namen referata je pridobitev čim večje število koristnih informacij iz podatkov lizing pogodb s pomočjo strojnega učenja. Le to ne pomeni učenje na pamet, ampak pridobivanje znanja na podlagi podatkov, ki opisujejo pretekle dogodke, odločitve ali aktivnosti znotraj podjetja.[2] Programski paket Weka predstavlja enoten programski vmesnik različnim algoritmom za strojno učenje.[4] Odkrivanje znanja iz podatkov je netrivialen proces odkrivanja implicitnega, doslej neznanega in potencialno uporabnega znanja iz podatkov.[3] Faze procesa odkrivanja znanaja so: priprava podatkov, podatkovno rudarjenje ter interpretacija, vrednotenje in predstavitev dobljenega znanja.[1]

Cilj referata je potrditi ali zavrniti naslednje postavljene hipoteze:

1. Ženske, kot komercialistke v lizing podjetju, pri ponudbi dosegajo višje marže kot moški. S svojo šarmantnostjo naj bi ženske naredile več posla za lizing podjetje ter pri tem dosegale višje marže.
2. Bolj tvegani predmeti najema dosegajo višje marže kot manj tvegani predmeti najema. Tvegana oprema je oprema, kateri cena na trgu močno variira ali je to opremo po odvzemu zaradi nezmožnosti plačevanja, težko prodati. Veliko težje se proda oprema, ki je zelo specifična - laboratoriji, posebni stroji ipd., zato je ta oprema tudi bolj tvegana.
3. Višje marže se dosega pri pogodbah višjih vrednosti pri izbranem predmetu najema. Ker je znesek višji, je s tem tudi bolj tvegan.
4. Največ lizing pogodb vsebuje predmet najema osebno vozilo. Osebno vozilo je predmet, ki je v današnjem času zahteva že vsakega posameznika. V Sloveniji ima osebno vozilo še posebni pomen, ker se še vedno nanj gleda kot na statusni simbol.
5. Trajanje lizing pogodb je običajno krajši od 5 let. Največ lizing pogodb se sklene za nižje zneske, katere imajo temu ustrezno krajši čas odplačevanja.

3 PRIPRAVA PODATKOV

Podatki za analizo so bili pridobljeni iz poslovnega sistema lizing podjetja X. Nahajajo se v večih medsebojno povezanih tabelah podatkovne baze, zato je bilo potrebno najprej opredeliti vse tabele, ki so povezane z pojmom entitete lizing pogodb. Oblika nekaterih podatkov ni bila najbolj primerna za namen analize. Poleg tega je bilo atributov mnogo preveč, kot jih dejansko potrebujem. Zato je bilo poleg prenosa podatkov iz vseh opredeljenih tabel potrebno opraviti tudi selekcijo ter transformacijo podatkov. Zaradi velike količine podatkov je bilo potrebno določiti reprezentativno delovno podmnožico, ki ne sme povzročiti popačenja rezultatov. Tako sem se omejil na podatke vseh lizing pogodb, ki so bile aktivirane od leta 2007 do konca leta 2009. S tem sem dobil dovolj veliko množico pogodb, ki pa izraža tudi dovolj ažurno stanje na področju sklepanja lizing pogodb. Izmed množice 50. atributov, ki jih je zajemala tabela pogodb, sem izbral 18 atributov, ki so bili zanimivi za obravnavo ter uporabni pri dokazovanju na začetku opisnih hipotez. Izbrani atributi so:

Length – trajanje pogodbe
tip: nominalni, zaloga vrednosti: {1Y, 1Y-3Y, 3Y-5Y, 5Y-10Y, 10Y-20Y, >20Y }
Atribut je bil za potrebe analize iz numeričnih vrednosti preslikan v vrednosti izbranih intervalov. Intervali so bili določeni na način, da je bilo število posameznih zapisov znotraj intervala čim bolj uravnoteženo. Tako so bili izbrani intervali vrednosti do 1 leta, od 1. do 3. let, od 3. do 5. let, od 5. do 10. let, od 10. do 20. let ter več kot 20 let.

ValueNum – vrednost pogodbe
Tip: numerični
Predstavlja vrednost pogodbe v EUR. Ker so bili zneski podani v različnih valutah, je bilo zaradi medsebojne primerjave potrebno vse vrednosti preko tabele tečajev preslikati v vrednost EUR.

Value – vrednost pogodbe
Tip: nominalni, zaloga vrednosti: {<5k, 5k-10k, 10k-15k, 15k-20k, 20k-30k, 30k-100k, 100k-500k, >500k}
Atribut je bil za potrebe analize preslikan iz numeričnih vrednosti v nominalne. Definicija intervalov je temeljila na večkratnem razseku numeričnih vrednosti glede na vrednosti mediane ter ustrezne zaokrožitve. Tako so bili Izbrani intervali manj kot 5000 EUR, od 5000 do 10000 EUR, od 10000 do 15000 EUR, od 15000 do 20000 EUR, od 20000 do 30000 EUR, od 30000 do 100000 EUR, od 100000 do 500000 EUR ter več kot 500000 EUR.

CurrencyID – valuta pogodbe
Tip: nominalni, zaloga vrednosti {EUR, CHF}
Valuta v kateri je bila sklenjena pogodba. Atribut zavzema dve vrednosti. Pogodba je lahko v valuti EUR ali švicarskih frankov.

RevTypeID – tip referenčne obrestne mere
Tip: nominalni, zaloga vrednosti {EUR1, EUR3, EUR6, CHF1, CHF3, CHF6}
Atribut predstavlja bolj natančno opredelitev atributa CurrencyID in nam podaja tip obrestne mere lizing pogodbe. Lahko zavzame eno od vrednosti šifranta referenčnih obrestnih mer.

InterestMargin – marža
Tip: numerični
Atribut podaja višino marže, ki jo pogodbi določi prodajna oseba. Je eden glavnih atributov, ki ga želim opazovati.

AverageMargin – povprečna marža
Tip: numerični
Atribut podaja povprečno maržo preko vseh pogodb. Izračunana je po formuli $\frac{\sum ValueNum*InterestMargin}{\sum ValueNum}$. Na ta način imajo marže pogodb z višjim zneskom večji vpliv na povprečno maržo kot pogodbe z nižjo vrednostjo. To je ključnega pomena saj nam pogodba z višjo vrednostjo ob isti marži prinese veliko večji prihodek. Uporablja se pri izarčunu ostalih atributov.

AverageByType – povprečna marža po predmetu najema
Tip: numerični
Podobno kot atribut AverageMargin tudi ta podaja povprečno maržo ki jo prodajalci dosežejo pri določenem predmetu najema. Izračunana je po formuli $\frac{\sum ValueNum*InterestMargin}{\sum ValueNum}$ po pogodbah z določenim predmetom najema. Uporablja se pri izračunu ostalih atributov.

HighMarginFromAverage – višja marža od povprečne
Tip: nominalni, zaloga vrednosti {0,1}
Atribut podaja ali je marža pogodbe višja od povprečne marže vseh pogodb.

HighMargin – višja marža
Tip: nominalni, zaloga vrednosti {0, 1}
Atribut nam poda ali je marža, pri določeni pogodbi z določenim predmetom najema, višje od povprečne marže po predmetu najema. Atribut se izračuna glede na vrednost atributa AverageByType.

FullName – prodajna oseba
Tip: nominalni
Atribut vsebuje ime in priimek osebe, ki je v lizing podjetju sklenila pogodbo s stranko. Zaradi velikega nabora vrednosti se ga običajno v analizi ne upošteva.

Gender – spol
Tip nominalni, zaloga vrednosti {M, W}
Podatek v sistemu ni bil prisoten. Na podlagi imena osebe sem ročno določil spol osebe.

ObjectTypeName – vrsta predmeta najema
Tip: nominalni
Atribut predstavlja in opisuje tip najema predmeta. Je zelo natanačna delitev predmeta najema.

GroupID – skupina vrste najema
Tip: nominalni, zaloga vrednosti {OSV, POP, KOV, NPR, LVL, OST, RAO}. Bolj groba delitev tipa predmeta najema. Zaradi manjšega nabora vrednosti je bolj uporaben od svojega predhodnika.

Critical – tvegan predmet najema
Tip: nominalni, zaloga vrednosti {1, 0}
Atribut opisuje ali je predmet najema tvegan. Tvegana oprema je oprema kateri cena na trgu močno variira ali je to opremo po odvzemu zaradi nezmožnosti plačevanja, težko prodati. Veliko težje se proda oprema, ki je zelo specifična - laboratoriji, posebni stroji ipd.

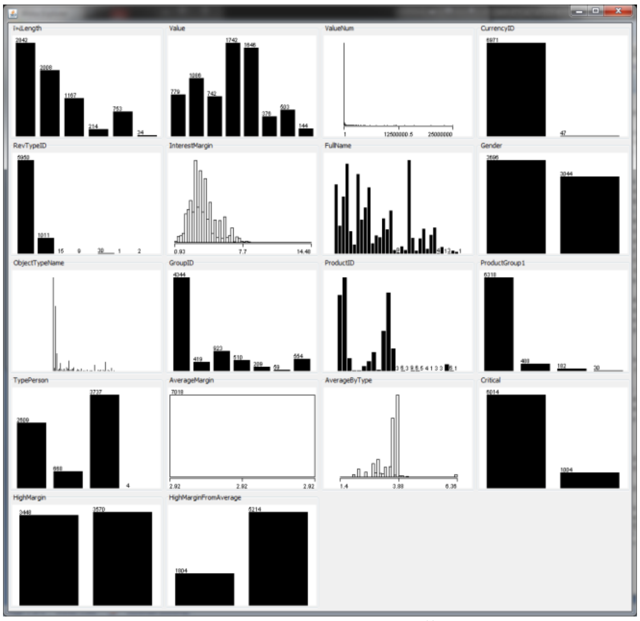
ProductID – vrsta lizinga
Tip: nominalni
Atribut podaja vrsto lizinga z vidika produkta. Nabor vrednosti zajema 25 različnih vrst produktov lizinga.

ProductGroup1 – skupina vrste lizinga
Tip: nominalni, zaloga vrednosti {finančni, operativni, posojilo}
Bolj groba delitev vrst lizinga.

TypePerson – vrsta stranke

Tip: nominalni, zaloga vrednosti { P1, P2, FI, JS}
Atribut opisuje vrsto osebe, ki je najela lizing. Oseba je lahko večja pravna oseba, samostojni podjetnik, fizična oseba ali javna služba.

Ker WEKA uporablja ARFF format, sem moral podatke ustrezno pretvoriti. Podatke sem iz SQL strežnika izvozil v CSV datoteko. WEKA za ločila med posameznimi polji uporablja vejico, izvoz iz SQL strežnika pa nam da podpičja. Zato sem na dobljeni CSV datoteki iz SQL strežnika vsa podpičja zamenjal z vejicam. Tako pripravljeno datoteko sem odprl v WEKI.



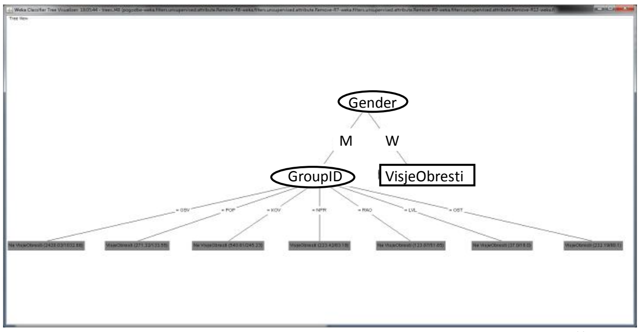
Slika 1: Porazdelitev vrednosti atributov množici 7018 uporabljenih učnih primerov.

4 OBDELAVA PODATKOV TER INTERPRETACIJA REZULTATOV

Po opravljeni pripravi podatkov sem se lotil »rudarjenja«. Pri tem sem uporabil Weka Explorer, ki je eden izmed delovnih okolij Weke. Za modeliranje podatkov se lahko odločamo med večimi algoritmi strojnega učenja, toda za preizkus postavljenih hipotez sem izbral klasifikacijo J48 (bolj znan kot C4.5) ob različnih naborih atributov za posamezne hipoteze.

4.1 Ženske dosegajo višje marže kot moški

Za analizo sem izločil atribut CurrencyID ter RevTypeID, ker je večina pogodb bila narejenih v EUR valuti z EUR6 referenčno obrestno mero. Izločil sem tudi atribute AverageByType, ValueNum, InterestMargin, GroupId ter Critical, ki so že bili prisotni v analizi preko drugih atributov. Da drevo nebi bilo preveliko ter da se nebi preveč prilegalo učnim primerom sem iz analize odstranil tudi atribute FullName ter ObjectTypeName, ProductID, ki imajo velike nabore vrednosti.
Tako sem sprva v analizi upošteval naslednje atribute: Length, Value, Gender, GroupId, ProductGroup1, TypePerson, HighMargin. Pri tem je eden najpomebnejših atributov na katerega sklepamo atribut HighMargin, ki je podrobneje opisan v razdelku priprave podatkov. Dobljeno drevo je imelo 82% klasifikacijsko točnost in 529 vozlišč. Tudi poskus spreminjanja parametrov algoritma (minNumObj) nam drevo ni bistveno oklestilo. Ob odstranitvi dodatnih atributov sem našel drevo, ki je dovolj majhno vendar pa se je njegova klasifikacijska točnost bistveno zmanjšala – 66%.
Iz rezultata prvega drevesa ne morem potrditi, da ženske v splošnem dosegajo višje marže kot moški. To velja le pri določenih predmetih najema, ob določenih trajanjih pogodbe, kar pa je preveč selektivno, da bi hipotezo lahko posplošil.

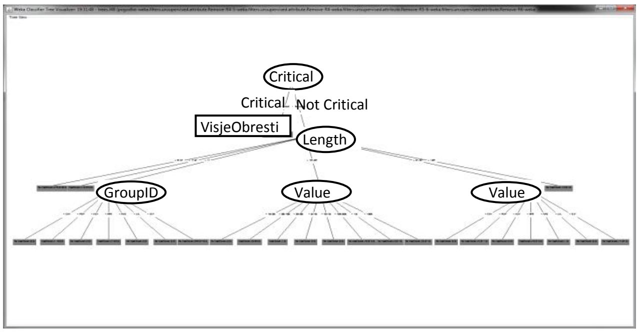


Slika 2: Poenostavljeno drevo za dokazovanje hipoteze, da ženske dosegajo višje marže kot moški

Poenostavljeno drevo nam na drugi strani prikaže, da ženske, ne glede na vse ostale atribute, dosegajo višje marže. Čeprav je na drugi strani točnost poenostavljenega drevesa bistveno manjša ter tudi ob pogledu na ostala vozlišča, ki govorijo, da moški za določene predmete najema dosegajo višje marže, lahko na podlagi tega ob klasifikacijski točnosti večje od 65% v splošnem potrdimo, da ženske v večini primerov dosegajo višje marže ne glede na ostale atribute. Tako lahko sklepamo, da je bila postavljena hipoteza morda pravilna..

4.2 Pri bolj tveganih predmetih najema nastopajo višje marže

Glavna atributa za dokazovanje navedene hipoteze sta Critical ter HighMarginFromAverage. Natančnejši opis atributov je podan v razdelku priprava podatkov. Tudi pri dokazovanju te hipoteze sem iz analize izločil atribute, ki bistveno ne vplivajo na samo klasifikacijsko točnost. Z poizkušanjem sem prišel do naslednjega nabora atributov: Length, Value, Gender, GroupID, Critical ter HighMarginFromAverage. Tako sem dobil drevo z 31 vozlišči ter klasifikacijsko točnostjo 78%. Drevo nam potrdi hipotezo, da v primeru, ko je predmet najema tvegan so največkrat tudi marže višje od povprečne marže. Po drugi strani v primeru, ko predmet najema ni tvegan, pa je višina marže odvisna tudi od trajanja pogodbe, višine zneska ter od samega predmeta najema. Tako lahko sklepam, da je tudi navedena hipoteza po vsej verjetnosti pravilna.



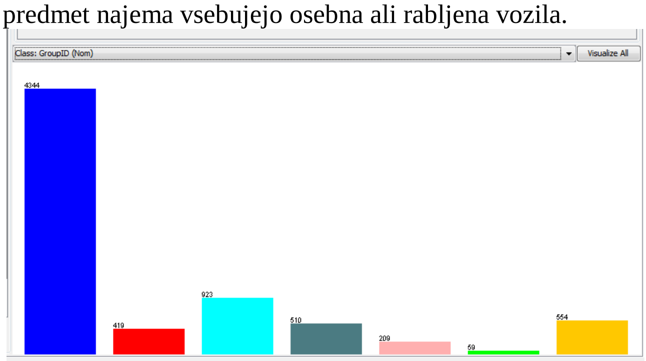
Slika 3: Drevo za dokazovanje hipoteze, da pri bolj tveganih predmetih najema nastopajo višje marže

4.3 Višje marže se dosega pri pogodbah višjih vrednosti za določen predmet najema

Za dokaz ali zavrnitev hipoteze sem v opazovano množico atributov vključil HighMargin, Value ali ValueNum ter ObjectTypeName, ki se neposredno nanašajo na dokazovanje hipoteze. Poleg teh sem z preizkušanjem nabora različnih atributov ter opazovanjem klasifikacijske točnosti vključil še atributa Length ter GroupID. Dobljeno drevo, ki vsebuje 189 vozlišč ter ob klasifikacijski točnosti 59%, zavrne hipotezo. Ne glede na znesek pogodbe se višja marža pojavlja zelo neodvisno od vseh vsebovanih atributov. Temu vprid govori tudi nizka klasifikacijska točnost. Vzrok za to vidim v tem, da na eni strani višja marža res pomeni večje tveganje, po drugi strani pa je marža podana v odstotkih od zneska pogodbe ter tako že sama po sebi pri višjih zneskih dejansko dosega višjo denarno vrednost, ki pokrije to tveganje.

4.4 Najbolj pogost predmet najema je osebno vozilo

Hipotezo enostavno potrdimo z pogledom porazdelitve po atributu GroupID, ki nam predstavlja bolj grobo delitev predmetov najema. Iz slike se vidi, da z veliko prednostjo kot najpogostejši predmet najema nastopa osebno vozilo. Podobno nam potrdi tudi porazdelitev po atributu ObjectTypeName, ki je bolj natančna opredelitev predmeta najema. Tudi tu močno prevladujejo pogodbe, ki za predmet najema vsebujejo osebna ali rabljena vozila.



Slika 4: Porazdelitev atributa GroupId. (moder stolpec – osebna vozila)

4.4 Trajanje lizing pogodb je običajno krajši od 5 let

Tudi to hipotezo enostavno potrdim ali zavrnem z porazdelitvijo atributa. Porazdelitev atributa Length nam pokaže, da močno prevladujejo pogodbe, kjer je trajanje krajše od 5 let.

5 ZAKLJUČEK

Zaposleni pri svojem vsakodnevnem delu proizvajajo velike količine podatkov. Ti ponavadi služijo le kot evidenca, da se ve kaj, kje in kdaj se je zgodil nek poslovni dogodek. Takšni podatki sami po sebi nimajo velike vrednosti. Na podlagi teh podatkov ter z ustreznimi orodji za odkrivanje znanja pa lahko ti podatki nudijo veliko število informacij ter novega znanja, ki ga s pridom izkoristimo pri učinkovitejšem ter uspešnejšem poslovanju. Na podlagi podatkov lizing pogodb sem poskušal ponazoriti kako si lahko s sodobnimi programskimi orodji pomagamo najti odgovore na številna vprašanja, ki se pojavljajo tekom optimizacije poslovnega procesa. Za pridobitev novega znanja pa je poleg zmogljivih orodij bistvenega pomena razumeti metode, ki nam jih ponujajo ter same podatke, ki jih raziskujemo. Srečujemo se z številnimi atributi in če bi samo vse uporabili naenkrat bi dobili povsem nepregledne in zelo kompleksne rezultate. Zato je zelo pomembno poznavanje povezav med atributi, preoblikovanje določenih atributov v obliko, ki je primerna za analizo ter navsezadnje pravilen izbor atributov. Zelo pomembna je kasneje tudi interpretacija rezultatov, saj ravno pri tem lahko hitro pridemo do napačnih zaključkov, še posebej če ne poznamo dobro orodij ter podatkov samih. Zato se je za izvedbo takih vrst analiz potrebno najprej ustrezno izobraziti.

Literatura

[1] Matjaž Gams: Predavanja Intelegentni programski sistemi 2008/2009
[2] Han, J., Kamber, M. (2001): Data Mining: Concept and Techniques, Morgan Kaufman
[3] Strojno učenje. [URL: http://sl.wikipedia.org/wiki/Strojno_u%C4%8Denje], 23.1.2010
[4] WEKA; University of Waikato, [URL: http://www.cs.waikato.ac.nz/~ml/weka/], 03.01.2010
[5] Eibe, F.; WEKA Explorer User Guide for Version 3-4-5; Department of Computer Science, University of Waikato, New Zeland; [URL: http://weka.sourceforge.net/manuals/ExplorerGuide.pdf], 02.01.2010
[6] C4.5 algoritem [URL: http://en.wikipedia.org/wiki/C4.5_algorithm], 22.12.2009
[7] Burian, S., Černelč, M., Gams, M.; Analiza zastav vrednostnih papirjev s programom Weka; KDD d.d., Autocommerce d.d., Inštitiut Jožef Štefan, [URL: http://ai.ijs.si/Mezi/pedagosko/seminar.EF6.doc], 12.12.2009

PRIMERJAVA NEVRONSKIH MODELOV VEDENJA NA PRIMERU LOTERIJE

Damjan Strnad, Nikola Guid
Fakulteta za elektrotehniko, računalništvo in informatiko
Univerza v Mariboru
Smetanova 17, 2000 Maribor, Slovenija
Tel: +386 2 220-7472
e-mail: damjan.strnad@uni-mb.si

POVZETEK

V članku predstavljamo rezultate primerjave učinkovitosti treh arhitektur nevronskih mrež pri modeliranju kontekstno odvisnega človeškega vedenja. Obravnavana oblika vedenja je udeležba pri loteriji, ki jo je možno kvantificirati z višino vplačil v odvisnosti od višine nagradnega sklada, dneva žreba in periodike glavnih dobitkov. Nadzorovano učenje standardne večplastne nevronske mreže, nevronske mreže z zakasnitvijo vhodov in Elmanove rekurzivne nevronske mreže poteka s pomočjo podatkov o preteklih udeležbah, ki predstavljajo vedenjske vzorce. Testiranje povprečne in optimalne učinkovitosti mrež pri napovedovanju oz. simulaciji vedenja pokaže, da upoštevanje pretekle dinamike konteksta odločanja pripomore k večji natančnosti modela kot samo upoštevanje aktualnega konteksta. Najuspešnejše zajemanje te dinamike doseže Elmanova nevronska mreža.

1 UVOD

Človeško vedenje je težko predvidljivo zaradi velikega števila čustvenih in fizioloških faktorjev, ki tvorijo kontekst sprejemanja odločitev pri posamezniku. Tvorbe modelov človeškega odločanja in vedenja v specifičnih situacijah so se raziskovalci v preteklosti že lotevali [1,2,3]. Namen modelov je zajemanje osnovnih komponent vedenja, kar bi omogočilo njegovo kasnejšo simulacijo ali napovedovanje. Veliko težavo pri učenju modelov vedenja predstavlja pomanjkanje podatkov. Vedenje je namreč težko kvantificirati, še posebej če je opazovanje osredotočeno na posameznika, katerega odločitve so v vsakem trenutku odvisne od nemerljivih komponent njegovega notranjega stanja. V naši raziskavi se ukvarjamo z modeliranjem vedenja skupine ljudi, za katero je značilno:

- odločitve posameznikov v skupini so binarne narave in pomenijo izvajanje ali neizvajanje določene vidne akcije,
- kontekst odločitve je podoben pri vseh posameznikih ali vsaj vsebuje iste primarne komponente,
- razlike med individualnimi vedenjskimi karakteristikami posameznikov se zlijejo oz. izničijo v povprečni reakciji skupine, ki jo je mogoče izmeriti in
- združeni učinek vedenja skupine je podan kot delež njenih posameznikov, ki so akcijo izvedli, in se nesposredno izraža v rezultatu meritve.

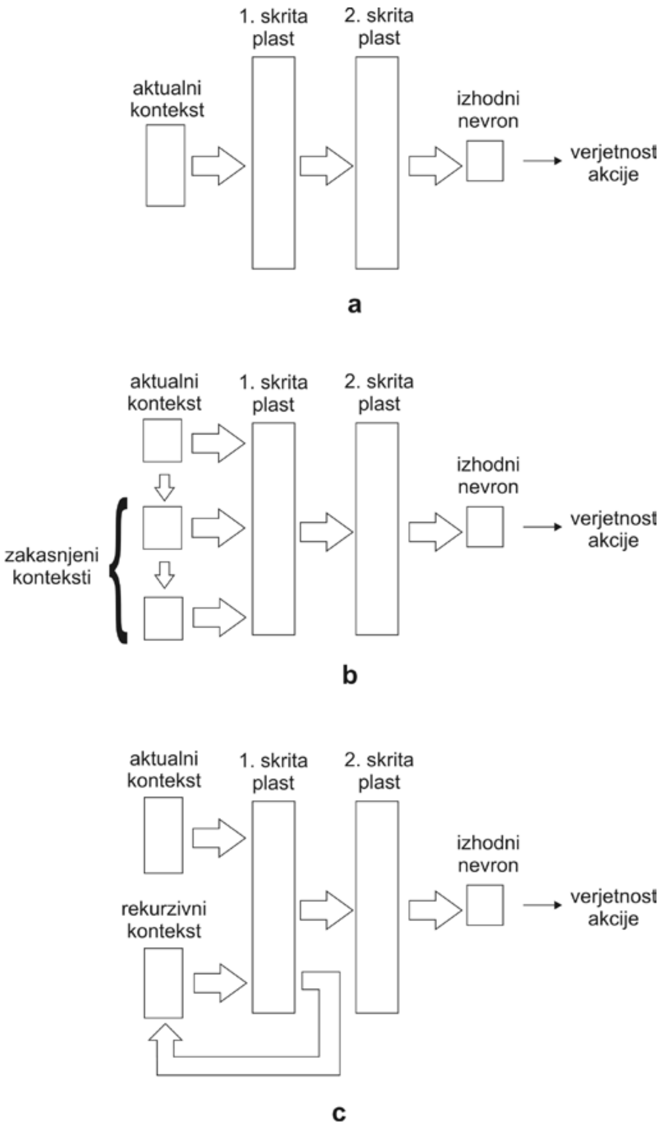
Model vedenja skupine je možno preslikati na posameznega agenta, če izhod modela obravnavamo kot verjetnost, da bo agent izvedel določeno akcijo oz. se vedel na določen način. Če takšno interpretacijo modela uporabimo na homogeni množici agentov, bo rezultat pravilno pričakovano vedenje te množice. Vhodi v model so numerični opisi tistih lastnosti okolja, ki tvorijo kontekst sprejemanja odločitve in s tem osnovo za določeno vedenje.

2 NEVRONSKI MODELI VEDENJA

V naši raziskavi smo za opis modela udeležbe pri loteriji uporabili tri arhitekture nevronskih mrež, ki se razlikujejo v načinu vzdrževanja konteksta odločanja. To so:

- triplastna polno povezana nevronska mreža (*multi-layer perceptron* - MLP), pri kateri kontekst odločanja sestoji izključno iz trenutno aktivnih vhodov (slika 1a),
- triplastna nevronska mreža s časovno zakasnitvijo nekaterih vhodov (*time-delay neural network* - TDNN) [4], pri kateri kontekst odločanja vključuje določeno število (t.i. časovni okvir) prejšnjih vhodov (slika 1b) in
- triplastna Elmanova rekurzivna nevronska mreža (*recurrent neural network* - RNN) [5], ki s povratno vezavo izhodov prve skrite plasti na vhod vzdržuje kontekst v obliki procesirane celotne zgodovine vhodnih podatkov (slika 1c).

Bistvena strukturna prednost zadnjih dveh modelov pred MLP je v tem, da omogočata upoštevanje dinamike konteksta pri določanju izhodov. Medtem ko je TDNN pri tem praktično omejen z velikostjo časovnega okvirja, RNN te omejitve nima. Vendar pa RNN pri tem nima dostopa do originalnih vhodnih podatkov, pač pa le do njihovega akumuliranega učinka na kontekst.



Slika 1: Modeli vedenja z a) standardno večplastno nevronsko mrežo, b) nevronsko mrežo z zakasnjenimi vhodi, c) rekurzivno nevronsko mrežo

3 PRIMER LOTERIJE

Cilj naše raziskave je primerjati učinkovitost omenjenih modelov pri napovedovanju skupne višine vplačil v posamezen krog loterijske igre Loto 7/39, ki jo organizira Loterija Slovenije [6]. Udeležba pri loteriji je primer prej opisanega vedenja skupine ljudi, pri kateri se posameznik odloča za ali proti sodelovanju. Čeprav je višina posameznega vplačila diskretna spremenljivka z realno zalogo vrednosti, jo lahko modeliramo kot seštevek več posameznih binarnih odločitev. V našem primeru je kontekst odločanja o udeležbi določen z naslednjimi podatki, ki postanejo vhodi v nevronsko mrežo:

- višina sklada za dobitke, ki jo aproksimiramo z višino prenosa iz prejšnjega kroga žrebanja,
- dan žrebanja (sreda ali nedelja),
- izplačilo glavne nagrade v prejšnjem krogu.

Izhod nevronke mreže lahko konceptualno interpretiramo kot verjetnost, da se bo agent, ki uporablja takšen model odločanja, odločil za sodelovanje pri loteriji in vplačal loterijski listek (tj. izvedel akcijo). Če isti model apliciramo na primerno veliki skupini agentov, bo skupna višina vplačila ustrezala predvideni udeležbi oz. modeliranemu vedenju skupine.

Za namene testiranja smo pridobili podrobna poročila za zadnjih 70 žrebanj lota [6], ki vključujejo vse potrebne podatke. Že bežen pogled nanje pove, da loterijska vplačila rastejo skladno z višanjem nagradnega sklada, kar je tipičen primer vedenja z občutnim emotivnim kontekstom. Vplačil je opazno več ob nedeljah kot ob sredah, zato je dan žreba kot binarna spremenljivka vključen v kontekst odločanja. Kot opcijske komponente konteksta smo vključili še binarni podatek o morebitnem izplačilu najvišje nagrade v prejšnjem krogu. Izplačilo najvišje nagrade izniči prenos sklada v naslednji krog in je zato samo drugačna interpretacija vhodnega podatka o višini prenosa, vendar je možno, da ima perioda izplačil vpliv na višino vplačil. Število krogov med dvema glavnima dobitkoma se je v opazovanem obdobju gibalo od 1 do 15, dobitnih krogov je bilo 10. Višina prenosa sklada je bila med 0 in 1.08 milijona eurov, kar smo uporabili kot normalizacijski interval za pripadajočo vhodno vrednost. Razpon višin vplačil v trenutni krog žrebanja je bil v istem obdobju med 334 tisoč in 958 tisoč euri, zato smo ciljne izhodne vrednosti prav tako izrazili v normalizirani obliki kot delež milijona eurov. Če predpostavimo, da je cena loterijskega listka 1 euro, bi bil milijon zato ustrezna velikost skupine agentov, na kateri bi model udeležbe pri loteriji uporabili v simulaciji vedenja.

4 UČENJE IN EVALUACIJA MODELOV

Zbrano zaporedje podatkov o gibanju vplačil predstavlja vzorec vedenja, ki smo ga z nadzorovanim učenjem nevronskih modelov poskušali zajeti in posplošiti na prihodnje situacije. Učenje je potekalo s standardno metodo vzvratnega prenosa napake, pri čemer smo za minimizacijo napake uporabili srednjo kvadratično napako. Iz celotnega vzorčnega zaporedja smo izločili pet podzaporedij dolžine 50 od začetnega vzorca 1, 6, 11, 16 in 21 naprej. V vsakem podzaporedju smo prvih 40 vzorcev (80 %) uporabili za učenje, naslednjih 5 (10 %) za prečno vrednotenje in zadnjih 5 (10 %) za testiranje. Parametre učenja smo prilagodili z začetno fazo eksperimentiranja in za končno primerjavo uporabili naslednje vrednosti:

- hitrost učenja η=0.3,
- momentni faktor λ=0.1,
- maksimalno število epoh je 20000,
- interval prečnega vrednotenja je 1000,

- maksimalna dopustna rast napake pri prečnem vrednotenju je 20 % minimalne dosežene napake,
- interval primerjave napake pri učenju je 1000,
- minimalna potrebna sprememba napake pri učenju od zadnje primerjave je 10^{-6} in
- prag napake za zaključek učenja je 10^{-4} (t.j. povprečna napaka pri napovedi je manjša od 10 tisoč eurov).

Pri testiranju smo spreminjali naslednje parametre nevronskih mrež:

- število skritih nevronov H, ki je enako v obeh skritih plasteh (H=1..4) in
- velikost časovnega okvirja T pri mrežah s časovno zakasnitvijo (T=2..5).

Da bi zmanjšali vpliv posamičnih neuspešnih izvedb učenja iz neugodno inicializiranih vrednosti uteži, smo postopek učenja in testiranja ponovili v 30 ciklih. Naj bo C=30 število ciklov učenja, L=5 število učnih podzaporedij in M=5 število testnih vzorcev v vsakem podzaporedju. Napaka v c-tem ciklu pri testiranju na l-tem zaporedju je definirana kot:

$$E_c^l = \sum_{m=1}^{M} |d_m - y_m|,$$

kjer sta d_m in y_m ciljna in dejanska izhodna vrednost nevronske mreže pri m-tem testnem vzorcu. Minimalna povprečna napaka na l-tem podzaporedju je

$$E_{\text{min}}^l = \frac{1}{M} \text{min}_c\{E_c^l\},$$

povprečna napaka pa

$$E_{\text{pov}}^l = \frac{1}{M \cdot C} \sum_{c=1}^{C} E_c^l.$$

Skupna minimalna in povprečna napaka nevronske mreže sta seštevka napak po podzaporedjih, tj.:

$$E_{\text{min}} = \sum_{l=1}^{L} E_{\text{min}}^l$$

in

$$E_{\text{pov}} = \sum_{l=1}^{L} E_{\text{pov}}^l.$$

Skupna minimalna napaka odraža optimalno doseženo učinkovitost nevronske strukture pri modeliranju vzorčnega vedenja, skupna povprečna napaka in njen standardni odklon pa njeno pričakovano srednjo učinkovitost in stabilnost učenja na naključnem vzorcu. Zato smo oba podatka uporabili kot merili za primerjavo modelov.

5 REZULTATI

Testiranje je pokazalo, da vključitev vhodnega podatka o izplačilu glavnega dobitka v predhodnem krogu po pričakovanjih ne vpliva na učinkovitost napovedovanja pri MLP, ker gre za podvojitev vhodnega podatka o prenosu. Pri TDNN, ki sicer razloči samo dobitne serije znotraj omejenega časovnega okvirja, in pri RNN, ki detektira poljubno dolga obdobja brez dobitka, sta bili skupna povprečna in minimalna napaka ob uporabi dodatnega vhodnega podatka nižji do 6 % (TDNN) oz. do 8 % (RNN). Slednje kaže na to, da periodika dobitkov nosi koristno sekundarno informacijo, čeprav je njen vpliv relativno majhen. V poročilih zato podajamo rezultate s tremi vhodnimi podatki.

Tabela 1 prikazuje dosežene skupne povprečne napake in standardne odklone testiranih mrež. Iz nje je razvidno, da večina mrežnih struktur daje najboljše rezultate pri dveh skritih nevronih. Napake v tabeli so izražene v deležu milijona eurov, torej ima najboljša nevronska mreža (t.j. RNN) skupno povprečno napako približno 116 tisoč eurov. Sicer so različne mrežne strukture po tem kriteriju precej izenačene.

Tabela 1: Skupna povprečna napaka testiranih mrež in njen standardni odklon (v oklepaju) v milijon eurih

Nev. mreža	H=1	H=2	H=3	H=4
MLP	0.174 (0.013)	0.125 (0.007)	0.126 (0.007)	0.127 (0.006)
TDNN (T=2)	0.171 (0.020)	0.132 (0.009)	0.136 (0.010)	0.137 (0.010)
TDNN (T=3)	0.179 (0.026)	0.126 (0.014)	0.143 (0.009)	0.147 (0.012)
TDNN (T=4)	0.157 (0.014)	0.127 (0.012)	0.132 (0.011)	0.142 (0.011)
TDNN (T=5)	0.187 (0.028)	0.143 (0.011)	0.141 (0.012)	0.137 (0.013)
RNN	0.138 (0.015)	0.116 (0.016)	0.133 (0.014)	0.143 (0.014)

Višji standardni odklon povprečne napake pri TDNN in RNN pove, da zaradi večjega števila prostostnih stopenj (tj. uteži) stabilnost učenja niha bolj kot pri MLP. Zato ni presenetljivo, da je optimalna učinkovitost teh dveh tipov mrež opazno boljša kot pri MLP. To je razvidno iz tabele 2, ki prikazuje doseženo skupno minimalno napako testiranih mrež. Tudi v tem primeru se najbolje odreže RNN, katere skupna minimalna napaka na petih podzaporedjih znaša 67 tisoč eurov ali manj kot 14 tisoč evrov na podzaporedje. Ker je povprečna ciljna vrednost na testnih podzaporedjih enaka 0.452 (tj. povprečno vplačilo 452 tisoč eurov), je napaka najboljšega modela znotraj meje 3 %, kar ocenjujemo kot zelo dober rezultat.

Skupna minimalna napaka pri TDNN ima opazen trend upadanja z večanjem časovnega okvira in hkratnim povečevanjem števila skritih nevronov. Slabost pri tem je sočasen porast nestabilnosti učenja, saj zaradi kompleksnejše ploskve napake učenje pogosteje obtiči v slabem lokalnem optimumu.

Tabela 2: Skupna minimalna napaka testiranih mrež v milijon eurih

Nev. mreža	*H*=1	*H*=2	*H*=3	*H*=4
MLP	0.134	0.103	0.104	0.108
TDNN (*T*=2)	0.127	0.101	0.101	0.105
TDNN (*T*=3)	0.109	0.083	0.109	0.107
TDNN (*T*=4)	0.125	0.092	0.097	0.106
TDNN (*T*=5)	0.117	0.108	0.096	0.090
RNN	0.112	0.067	0.090	0.104

6 ZAKLJUČEK

V članku smo izvedli primerjavo različnih tipov nevronskih mrež, ki modelirajo kontekst odločanja o udeležbi pri loteriji. Slednje je zgled množičnega vedenja, ki se neposredno odraža v višini vplačil, njegov kontekst pa je možno zadostno definirati že z majhnim številom parametrov. Pri modeliranju vedenjskih vzorcev skupine postanejo individualni vzgibi nepomembni, saj obravnavamo množico kot homogeno skupino agentov, ki delujejo po povprečnem skupinskem vzorcu. Simulacijo takega vedenja dosežemo s preslikavo modela vedenja na to skupino, pri čemer model za posameznega agenta določa verjetnost binarne akcije (nakupa loterijskega listka), s katero se vedenje manifestira. Akumulirani učinek vedenja množice je potem pričakovana višina vplačila v loterijski žreb.

Poleg udeležbe pri loteriji med zanimive vedenjske vzorce skupin spadajo še obiski tekmovanj in kinematografov ter odločitve o nakupu določenega produkta. Pri modeliranju in simulaciji omenjenih vedenjskih vzorcev nastopita dve praktični oviri:

- problem definicije primarnih komponent konteksta, ki vplivajo na odločitev posameznika in so skupne množici (npr. kakovost ekip, zadnji rezultati, lokalno rivalstvo in pomembnost tekme pri odločitvi za njen ogled),
- problem pomanjkanja numeričnih podatkov, ki ta kontekst kvantificirajo.

Rezultati raziskave kažejo, da je pri modeliranju vedenja kontekst definiran ne samo s trenutnim opisom stanja, pač pa tudi z njegovim dosedanjim razvojem. Čeprav je povprečna učinkovitost MLP, TDNN in RNN pri simulaciji udeležbe pri loteriji primerljiva in razlike statistično nepomembne, pa optimalna učinkovitost potrjuje prednost arhitektur, ki vzdržujejo dinamiko konteksta. Slednja je pokazatelj emotivnega ali motivacijskega trenda v okolju, s pomočjo katerega lahko izboljšamo napovedi oz. simulacijo vedenja množice. Naš cilj v prihodnosti je validacija ugotovitev na drugih primerih vedenja in raziskave v smeri aplikacije dinamike konteksta z nevronskimi mrežami višjega reda [7].

References

[1] T. Jin, Y. Son, H. Hashimoto. Mobile Robot Control Using Fuzzy-Neural-Network for Learning Human Behavior. *Proc. of 13th International Conference on Neural Information Processing - ICONIP*, pp. 874-883, 2006.

[2] Z. Zhang, F. Vanderhaegen, P. Millot. Prediction of Human Behaviour Using Artificial Neural Networks. *Proc. of 4th International Conference on Machine Learning and Cybernetics - ICMLC*, pp. 770-779, Guangzhou, China. Hong Kong, China, 2005.

[3] A. Shelestov, S. Skakun, O. Kussul. Complex Neural Network Model of User Behavior in Distributed Systems. *Proc. of 13th International Conference Knowledge-Dialogue-Solutions*. pp. 42-49, Varna, Bulgaria, 2007.

[4] K. J. Lang, A. H. Waibel, G. E. Hinton. A Time-Delay Neural Network Architecture fior Isolated Word Recognition. *Neural Networks*. vol. 3, no. 1, pp. 23-43, 1990.

[5] J. L. Elman. Finding structure in time. *Cognitive Science*. vol. 14, no. 2, pp. 179-211, 1990.

[6] Loterija Slovenije, poročila o igri Loto, http://www.loterija.si/LOTERIJA,,igre_z_zrebanji,loto, porocila.htm.

[7] M. Zhang. Artificial Higher Order Neural Networks for Economics and Business. IGI Global. Hershey, USA, 2008.

HIERARCHICAL CLASSIFICATION OF MAGNETIC RESONANCE IMAGES

Katarina Trojacanec, Gjorgji Madjarov, Suzana Loskovska, Dejan Gjorgjevikj
Faculty of Electrical Engineering and Information Technologies
Ruger Boskovik bb., 1000 Skopje, Republic of Macedonia
Tel: +389 2 3099191; fax: +389 2 3064262
e-mail: katarina.trojacanec@feit.ukim.edu.mk

ABSTRACT

The objective of the paper is to explore classification on magnetic resonance images (MRI). In our work on MRI classification, two types of classification (flat and hierarchical) are addressed and explored. The examination is conducted on the dataset of magnetic resonance images that have hierarchical organization. All images are described by using Edge histogram descriptor for the feature extraction process. We compared the experimental results obtained from the hierarchical classification to the results provided by flat classification using different classifiers, such as SVM methods, k nearest neighbors, C4.5 algorithm and artificial neural networks. As a result, we concluded that the hierarchical classification technique outperforms all other explored classifiers for the examined dataset of magnetic resonance images.

1 INTRODUCTION

Following the rapid development of sophisticated medical devices and digital systems, the number of the generated digital medical images is continuously growing. This rapidly increasing number of medical images at the same time increases the necessity for automatic and efficient methods for image annotation and retrieval [1].

Magnetic Resonance Imaging (MRI) has become a useful modality in clinical and surgical environment [2]. It is very important, powerful and, very often, irreplaceable medical diagnostic technique. The large volume of MRI to be analyzed makes manual image classification impractical, labor intensive, time consuming and often inaccurate. Hence, the demand for efficient and automated analysis and classification of magnetic resonance images is continuously challenging problem for researchers and scientists.

Magnetic resonance images have been widely researching [3][4][5]. Different classification techniques have been applied and examined by research such as: Support Vector Machines (SVM) [6], k nearest neighbors [7], Artificial Neural Networks (ANN) [8]. A method for Automated Segmentation and Classification of Brain MRI using SVM classifier is proposed in [9]. Advanced classification techniques based on Least Squares Support Vector Machines (LS-SVM) are proposed and applied to brain image slices classification using features derived from slices in [3]. In [10], the authors show the results of their algorithm on the classification of gray and white matter along with surrounding cerebral spinal fluid in brain MRI scans. In [11] support vector machines classifier is applied on breast multi-spectral magnetic resonance images.

The rest of the paper is organized as follows: section 2 briefly describes the basic concepts beside the flat and hierarchical classification. Section 3 gives details on the organization of the dataset used for examination in our work, while section 4 reports the obtained results. Conclusions and summaries are given in section 5.

2 FLAT VERSUS HIERARCHICAL CLASSIFICATION

MRI classification is very sensitive and challenging problem. Two types of classification of MRIS are addressed and explored in our work: the flat classification, examined in our previous work [12][13], and tje hierarchical classification, the main goal of this paper.

In its bases, classification addresses problems of assigning newly, previously unseen samples to one or more pre-existing classes. If the predefined classes are separately treated and there is no structure defining the relationships among them (or that structure is not treated if it exists), those problems are addressed by flat classification. Otherwise, hierarchical classification refers to assigning samples to a suitable class from a hierarchical class space [14]. By utilizing the previously defined hierarchical architecture, the classification problem can be decomposed into a smaller set of problems corresponding to the hierarchical structure and splits within it [14] [15]. In such architecture, a distinguishing between classes at the first (top) level is performed at the beginning. Once this distinguishing is accomplished, the lower level distinctions are performed, but only taking into account the subclasses of the appropriate top level class. This approach in hierarchical classification is referred as top-down level-based approach [16]. Another, big-bang approach for hierarchical classification exists [17]. For our purposes we use the first approach, namely, the top-down level-based approach. In this approach the classification is accomplished with the cooperation of classifiers built at each level of the tree. One of the obvious problems with top-down approach is that a misclassification at a parent class may force a sample to be misrouted before it can be classified into child classes [14].

3 DATASET DESCRIPTION

The examined dataset in this paper consists of magnetic resonance images obtained from [18] and [19].
Because the explored images did not have any organization, we organized them in a hierarchical architecture. In fact, we separate the images, firstly, according to the body part they represent. According to this, the whole dataset can be divided into three classes: brain, abdomen, and gynecology. Each of these classes can be additionally divided on the bases of pathology present in the image. The hierarchy that represents this classification is depicted on Figure 1[13].

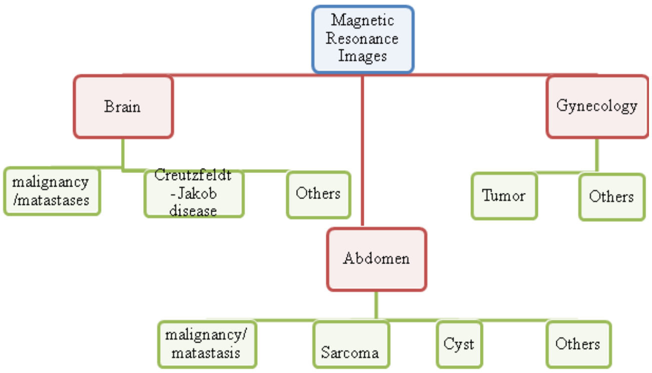


Figure 1: *Hierarchical organization of the dataset.*

According to Figure 1, at the first level of the hierarchy three categories could be distinguished: Brain, Abdomen and Gynecology. The Abdomen class is additionally divided into four subclasses. The first subclass of the Abdomen class contains images with presence of malignancy, metastases or tumor in the abdominal part of the human body, while the second subclass represents the images with presence of sarcoma. The third subclass includes MRIs that denote presence of cysts in the abdominal part of the examined patients. The fourth subclass consists of all other abdominal MRIs.
There are three subclasses that could be distinguished in the Brain class. The first one contains images where malignancy, metastases or tumor is present. The second subclass represents MRIs taken from patients in whom Creutzfeldt-Jakob disease has been diagnosed. The last subclass in the Brain class, the subclass Others, includes images with none of the mentioned pathologies and/or images where no pathological region has been detected.
We separated two subclasses in the Gynecology class, according to the presence or absence of tumor, respectively. Following the leaf nodes in the hierarchy depicted on Figure 1, the examined magnetic resonance images are classified into nine classes. There are 1870 magnetic resonance images in the dataset in total. The training set consists of 1247 MRIs, while the test set consists of 623 MRIs. Table 1 depicts the distribution of the number of images through the classes [13].

Table 1: *Distribution of the number of images through the classes*

Level 1	Level 2	Class No.	Training set	Test set	Total
Abdomen	malignancy /metastases	0	67	34	101
	Sarcoma	1	28	14	42
	Cyst	2	36	18	54
	Others	3	455	228	683
Brain	malignancy /metastases	4	53	27	80
	Creutzfeldt - Jakobdisease	5	13	7	20
	Others	6	343	171	514
Gynecol ogy	Tumor	7	56	27	83
	Others	8	196	97	293
Total			**1247**	**623**	**1870**

4 EXPERIMENTAL RESULTS

The experimental results obtained from the flat classification and hierarchical classification are presented in this subsection. The ultimate goal is to signify that the hierarchical classification shows better results over all explored flat classifiers on the bases of classification error for the examined dataset of MRIs.
The examination focuses on two main processes: the feature extraction process and the classification process.

4.1 Feature extraction

Due to the fact that color features does not have very expressive power for medical images [1], texture and shape descriptors are usually researched as descriptors for medical images. According to this and according to our previous work in which Edge Histogram Descriptor (EHD) showed the best results compared to six other descriptors [12][13], for the feature extraction process in this work we used exactly EHD.
Once the feature vector for each image of the dataset is generated, the normalization process is conducted. For this purpose, the min-max normalization is used.

4.2 Flat Classification

For the flat classification purposes, we distinguish nine classes in the dataset of magnetic resonance images, described in the previous subsection. In fact, each of the leaf nodes from the hierarchy depicted in Figure 1 is a separate class used for the flat classification. Thus, we do not take into account the real connection between the classes for the purpose of the flat classification. Table 2 presents the classification error obtained from the flat classification process when several classifiers were evaluated [13][14], such as: SVM classifier based on one-against-one and one-against-all strategy, SVM classifier in binary tree architecture, SVM utilizing binary decision tree, SVM utilizing balanced binary decision tree [20][21], as well as, artificial neural networks, k nearest neighbor and C4.5 algorithm [22]. SVM classifiers extended to address multiclass classification problem, as well as the multilayer perceptron with one hidden layer and 25 units within it are implemented using the Torch library [23]. For the k nearest neighbor classifier and C4.5 algorithm, we used Weka implementation [24].

Table 2: *Classification error obtained from the flat classification for each classifier separately*

Classifier	Classification error (%)
SVM One vs. All	17.66
SVM One vs. One	18.14
SVM – BTA	18.62
SVM – BDT	18.78
SVM - BBDT	18.46
ANN	25.20
k-nn	18.29
C4.5	43.02

4.3 Hierarchical Classification

The aim of the paper is to examine the hierarchical classification applied to the dataset of magnetic resonance images and to compare the results obtained from the hierarchical classification of MRIs to the results obtained from the flat classification. The ultimate goal is to choose the most appropriate one for this kind of images and for the way we organized the whole dataset. The idea to apply the hierarchical classification is reasonable because of the hierarchical organization of the data that we provided.

The hierarchical classification architecture used for analysis in this work is similar and proper to the hierarchical organization of the image given by Figure 1. Because each node in the hierarchy could possibly have more than one branch, classifiers that are able to address multiclass problems are needed. According to this, one-against-all multiclass SVM strategy is used for solving the classification problems in each node.

The top node (the first level of the tree) consists of a multiclass SVM classifier based on one-against-all strategy that is trained to make a distinction between the images from the three classes. These classes, in fact, represent MRIs of the three body parts: abdominal, brain, and gynecological part. The classifier at the top node is trained with the whole training set (1247 MRIs).

For each of the three classes, a separate multiclass SVM classifier based on one-against-all strategy is trained in each node at the second level. Thus, a separate multiclass SVM classifier based on one-against-all strategy is trained to make a distinction between the subclasses of the classes from the previous level. These subclasses represent the diseases in each body part. In fact, the first node at level 2 represents the class Abdomen. It consists of SVM one-against-all classifier trained to distinguish images that belong to its four subclasses: Malignancy/metastases, Sarcoma, Cyst, and Others. This classifier is trained with 586 images (the number of images that represent the abdominal body part). The second node at the same level represents the class Brain. It contains SVM one-against-all classifier that separates the subclasses derived from the Brain class, namely, Malignancy/metastases, Creutzfeldt – Jakob disease, and Others. The number of the training samples in this case is 409. Finally, the last node from the second level of the hierarchy has multiclass SVM classifier based on one-against-all strategy trained with that images of the training set that belong to the Gynecology class (252 MRIs). It is trained to make distinction between the two subclasses Tumor and Others.

During the testing phase, each test sample is passed through the classification hierarchy. At the beginning, the test image is considered to which body part it belongs. For example, if the classifier decides that it is an image that represents the abdominal part, then it is passed to the leftmost branch. The appropriate classifier at the second level then classifies the image in one of the subclasses of the abdominal class. In such hierarchical classification architecture, it is obvious that if the test example is erroneously classified at the first level, it will be erroneously classified at the second level as well.

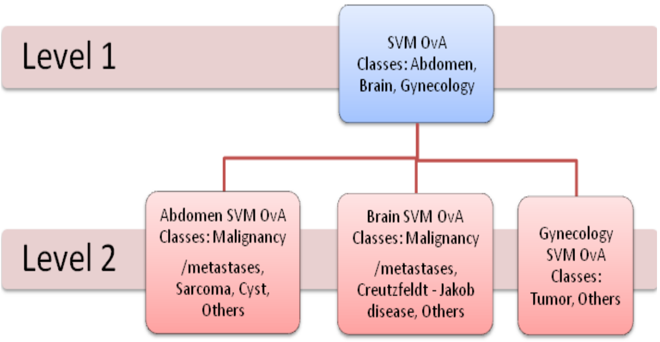


Figure 2: *Hierarchical classification architecture.*

The classification error obtained from the classification of magnetic resonance images provided by the described hierarchical classification architecture is 16.53%. In fact, from all test images, 23 images are misclassified at the first level. At the second level, 22 misclassifications were obtained at the first node, and 34 and 24 misclassifications at the second and third node. Comparing with the results obtained from the flat classification depicted in Table 2, we can conclude that for the examined dataset of magnetic resonance images which we organized hierarchically, the hierarchical classification outperforms all classifiers used for flat classification in our work.

5 CONCLUSION

We applied hierarchical classification technique to provide classification of magnetic resonance images. The hierarchical classification architecture is built of SVM classifiers based on one-against-all strategy at each node of the hierarchy. The multiclass classifier at the first level is trained to make a distinction between the images that represent different body part (three classes are available at the first level). The multiclass classifier in each node at the second level is trained to separate the images into subclasses that denote the possible diseases in each body part. The analysis in this work was conducted to the dataset of magnetic resonance images that we organized in a hierarchical way.

The experiments performed for the purpose of the paper showed that the hierarchical classification gives better results in comparison to the flat classification provided by SVM for multiclass classification, k nearest neighbors, C4.5 algorithm and artificial neural networks. In fact, the lowest classification error obtained from the flat classification is 17.66% (provided by SVM classifier based on one-against-all strategy) and with the hierarchical classification we gain lower classification error, 16.53%. According to this, we can conclude that the hierarchical classification is more appropriate for the examined dataset of magnetic resonance images.

References

[1] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler. A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *International Journal of Medical Informatics* 73, pp. 1-23. 2003.

[2] E. M. Haacke, R. W. Brown, M. R. Thompson, R. Venkatesan. Magnetic Resonance Imaging, Physical Principles and Sequence Design. *Wieley-Liss*, New York. 1999.

[3] H. Selvaraj, S. Thamarai Selvi, D. Selvathi, L. Gewali. Brain MRI Slices Classification Using Least Squares Support Vector Machine, *International Journal of Intelligent Computing in Medical Sciences and Image Processing*, Vol. 1, No. 1, Issue 1, pp. 21-33. 2007.

[4] Collins, D., Montagnat, J., Zijdenbos, A., Evans, A., Arnold, D., Automated estimation of brain volume in multiple sclerosis with BICCR, In: Insana, M. F., Leahy, R. M. (Eds.), Proc. of IPMI 2001. Vol. 2082 of LNCS. Springer-Verlag, pp. 141-147, Jun. 2001.

[5] MacDonald, D., Kabani, N., Avis, D., Evans, A. C., Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12 (3),340-56, 2000.

[6] Zijdenbos, A. P., Forghani, R., Evans, A. C., Automatic 'pipeline' analysis of 3D MRI data for clinical trials: Application to multiple sclerosis. IEEE Trans Med Imaging 21 (10), 1280-91, Oct. 2002.

[7] Wareld, S. K., Kaus, M., Jolesz, F. A., Kikinis, R., Adaptive, template moderated, spatially varying statistical classication. Med Image Anal 4 (1), 43-55, Mar 2000.

[8] Zijdenbos, A. P., Forghani, R., Evans, A. C., Automatic 'pipeline' analysis of 3D MRI data for clinical trials: Application to multiple sclerosis. IEEE Trans Med Imaging 21 (10), 1280-91, Oct. 2002.

[9] L. Antonie. Automated Segmentation and Classification of Brain Magnetic Resonance Imaging. *C615 Project.*

[10] J. Melonakos, Y. Gao, A. Tannenbaum. Tissue Tracking: Applications for Brain MRI Classification. *Proceedings of SPIE Medical Imaging*. 6512:651218-1. 2007.

[11] Chuin-Mu Wang, Xiao-Xing Mai, Geng-Cheng Lin, Chio-Tan Kuo, Classification for Breast MRI Using Support Vector Machine, Proceedings of IEEE 8th International Conference on Computer and Information Technology Workshops, 2008.

[12] K. Trojacanec, I. Kitanovski, S. Loskovska. Comparison of Classification Techniques Applied to Magnetic Resonance Images. *Proceedings of the 7th International Conference for Informatics and Information Technology*. 2010.

[13] K. Trojacanec, G. Madzarov, D. Gjorgjevikj, and S. Loskovska. Classification of magnetic resonance images. *Proceedings of the 32nd International Conference on Information Technology Interfaces (ITI)*. 2010.

[14] A. Pulijala, S. Gauch. Hierarchical Text Classification. *International Conference on Cybernetics and Information Technologies*. Orlando, FL. 2004.

[15] S.Dumais, H.Chen. Hierarchical Classification of Web Content. *SIGIR*. 2000.

[16] A. Sun, E. Lim, and W. Ng. Performance Measurement Framework for Hierarchical Text Classification. *Journal of the American Society for Information Science and Technology*. 54(11):pp.1014-1028. 2003.

[17] Claire and R. D. King. Predicting Gene Function in Saccharomyces Cerevisiae. *Bioinformatics*, 19(2):42-49, 2003.

[18] http://www.imageclef.org/ImageCLEF2008 [03.09.2010].

[19] http://www.info-radiologie.ch/index-english.php [03.09.2010].

[20] Gjorgji Madzarov, Dejan Gjorgjevikj and Ivan Chorbev, A Multi-class SVM Classifier Utilizing Binary Decision Tree. *Informatica* 33, pages 233-241. 2009.

[21] G. Madzarov, D. Gjorgjevikj, I. Chorbev. Multi-class Classification using Support Vector Machines in Binary Tree Architecture. *International Scientific Conference*, Gabrovo. 2008.

[22] B. E. Boser, I. M. Guyon, V. N. Vapnik. A training algorithm for optimal margin classifiers. *Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, Pittsburgh. 1992.

[23] R. Collobert, S. Bengio, J. Mariethoz, Torch: a modular machine learning software library, Technical Report IDIAP-RR 02-46, IDIAP, 2002.

[24] http://www.cs.waikato.ac.nz/ml/weka/ [03.09.2010].

WINDOW RECOGNITION ON IMAGES OF BUILDING FAÇADES

Petar Vračar, Igor Kononenko
Faculty of Computer and Information Science
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
e-mail: {petar.vracar, igor.kononenko}@fri.uni-lj.si

ABSTRACT

We describe a method for learning and recognizing windows as basic structural elements of façades. The method begins with the segmentation of the input image into a hierarchical structure of window candidates. We developed four heuristic functions for estimating the quality of generated window candidates. The generation of the final hypothesis uses a greedy approach, which favors homogeneous and aligned candidates. The results are satisfactory but for the practical use the method should be improved further.

Keywords: façade segmentation, window detection, urban environment

1 INTRODUCTION

In recent years there has been a growing interest in the field of modeling, recognition and interpretation of urban environment. Urban planning, 3-D city modeling, simulation of emergency population evacuation, driving simulators, visual navigation systems and computer games are examples of applications that, at least to some extent, require the visual interpretation of urban scenes. Basic functions required are detection, reconstruction and architectural interpretation of buildings as the main structural elements of urban environment. Very important role in the realization of these functions have procedures for detection and interpretation of the façade as the carrier of the building's visual identity.

In this paper we have focused on the window recognition as a basic symbol in the description of the façade. The recognition process is difficult because of great diversity in the windows appearance. It could be affected by partially or fully lowered blinds, closed curtains or other decorative items such as flowers on the window shelves. On the same façade windows may be present in various shapes and sizes. Glass is reflective material and therefore the windows on the same façade may vary due to the reflection of the sky and other surrounding buildings. Façades can be multicolored with various ornaments that are, like windows, periodically repeated which further complicates the detection process.

One could use the fact that buildings are often divided into floors which are further divided into rooms. The consequence of such a building organization is the presence of the windows align arrangement in horizontal (for individual floors) and vertical (for individual rooms on different floors) directions. This is a key property for the detection of window candidates.

Although in general they vary in shape and size, the windows on the same façade are often similar to each other, which can be used in the process of recognition. In other words, multiple repetition of a certain part of the façade image increases the probability that this part corresponds to a window candidate.

Finally, it is possible to use the machine learning methods with appropriately selected set of features to develop a window detector based on learned window appearance.

The rest of the paper is structured as follows. Section 2 describes the segmentation of input images which generates a hierarchical structure of window candidates. Four heuristic functions for estimating the quality of generated candidates are presented. Section 3 provides the experimental results of our method. The paper concludes with Section 4 in which we give some ideas for further work.

2 FAÇADE SEGMENTATION

Façades, like many man-made objects, usually have a highly ordered structure. On the other hand, the photographs of façades do not retain parallelism and orthogonality between structural elements due to the perspective transformation. In order to simplify further processing, it is necessary to remove the perspective projection from the input image by reconstructing the orthogonal view to the plane that models the spatial orientation of the façade's surface. After the rectification we assume that we are dealing with rectified input images that retain parallelism and orthogonality of the façade elements' edges (more details about image rectification can be found in [1]).

3.1 Creation of candidates

For generating the hypotheses about windows on the façade we assume also that windows are rectangular in shape and have their edges aligned with the image coordinate axes. Therefore, we can expect more vertical edges in the image sections corresponding to the horizontal bands of windows of the same floor, and fewer vertical edges in the image sections corresponding to the wall between floors. Similarly, more horizontal edges can be expected in image areas that correspond to columns of the façade windows and fewer between them. Lee and Nevatia [2] have proposed a method for windows detection based on these findings. They presented the concept of a projection profile that represents the histogram of the edge distribution in the input image. Vertical projection profile is calculated by projecting the vertical edges to the vertical axis. The vertical edges of the aligned windows are projected to the same part of the vertical axis. The accumulation of these projections gives a histogram of the vertical edge distribution. The horizontal projection profile can be calculated analogously, by accumulating the projections of the horizontal edges to the horizontal axis. The window candidates may be segmented by intersecting the strips that correspond to the peaks in the horizontal and vertical projection profile histograms.

We have chosen a slightly different approach which is similar to [3]. In order to use as much of local information as possible, we decided to segment the façade image using a greedy method which recursively divides the façade in the global minima of the local projection profiles. In each iteration, the input image is split into three parts: the central part (corresponding to the valley around the global minimum of the both projection profiles) and the two marginal parts – one on each side (see Figure 1).

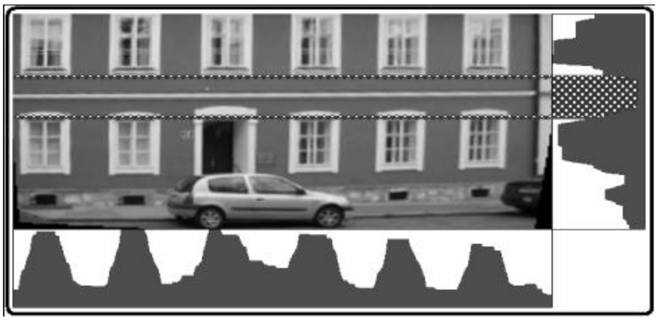


Figure 1: *Façade segmentation using projection profiles. The hatched area denotes the valley around the global minimum in both histograms that corresponds to a part of the façade with no windows.*

We assume that the central part does not contain any façade's structural elements and therefore it is excluded from the further analysis. Each marginal part is recursively divided by the same procedure while taking into account only the part of the input image that belongs to the observed area. In this way we build a hierarchical structure of the observed façade using local information. Subdivision into smaller parts stops when the size of the observed area becomes smaller than a predetermined minimum, or when the projection profiles are evenly distributed so that subsequent divisions are meaningless.

Using the described procedure we can build a structural decomposition of the façade in the form of a tree (see Figure 2). Each node of that tree covers some part of the input image. The root node corresponds to the whole image. All internal nodes have three children, according to the division procedure. From our construction it follows that the internal nodes of the tree are increasingly framed due to the fact that in each recursive step we eliminate parts of the image without edges. The construction of the tree also provides rectangular nodes with sides that are aligned with the image axes, which is in accordance with the simplified definition of a window. Therefore, each node of the generated tree structure is a candidate for the window on the observed façade.

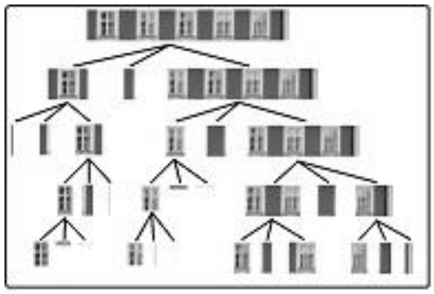


Figure 2: *Structural decomposition of the façade in the form of a tree.*

3.2 Evaluation of candidates

For the final segmentation of the façade it is necessary to evaluate all candidates and to declare the best rated as windows. We've evaluated candidates using heuristics based on position in the tree structure, edge distribution, similarity and the learned window appearance. The final evaluation of the quality of hypotheses is obtained as a linear combination of the individual heuristic evaluations.

The simplest heuristic is the one based on the position in the tree structure. It is derived directly from the way the hierarchy was built. Each internal node of the tree has three children and the middle child covers the least amount of edge pixels. From this we can conclude that the middle son covers a wall of the façade, so we can assign a negative rating to it and thus reduce the expectation that it corresponds to a window. The heuristics is neutral to the other two child nodes.

The second heuristic evaluation is based on the edge distribution and is similar to the method of Lee and Nevatia. We want to positively evaluate candidates that are located at the intersections between histogram peaks of the projection profiles. With this objective we set up a 2D accumulator array of the same size as the input image. Each histogram peak votes for its vertical or horizontal strip in the accumulator array, depending on which projection profile it belongs to. The evaluation of a candidate is defined as the average value of accumulator array cells from the area that corresponds to that candidate.

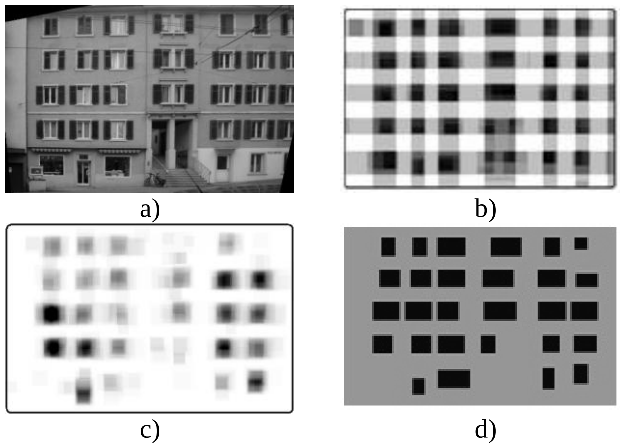
The idea of the next heuristic follows from the finding that the individual windows on the same façade are similar to each other. The quality of each candidate is assessed as the degree of visual similarity with other candidates which can be measured, for example, using the normalized correlation coefficient. The nice property of this measure is that it's not sensitive to the changes in image brightness and contrast, which is a common phenomenon in the façade images.

Viola and Jones [4] have developed a framework to train a cascade classifier for face detection task, which is due to its generality suitable for the detection of arbitrary types of objects. We used this framework to learn the visual appearance of façade windows. We have created a learning set using the reference image libraries ZuBuD[1] and TSG-60[2]. The positive example set consists of 1529 manually marked windows. The negative example set contains 4668 objects that often appear in the input images but are not windows (e.g. façade ornaments, traffic signs, trees, parked vehicles, passers-by etc.). Using the described learning set we have built 18-stage cascade classifier. Again, we used an accumulator array that has been incremented in regions where the classifier detected the window patterns. As before, the evaluation of a candidate is defined as the average value of its accumulator array cells.

The last step in forming a hypothesis about the location and size of windows on the façade image is the selection of the best evaluated candidates. We have used a simple greedy method that in each iteration first selects the best evaluated candidate and then increases the scores of candidates that have approximately the same size and are approximately aligned with the selected candidate. The entire process is repeated until all remaining candidates have scores less than a predefined threshold. When the process stops, we accept only those selected candidates that are not covered by any other selected candidate. Figure 3 shows the forming of the final hypothesis for an input image.

4 EMPIRICAL EVALUATION

We implemented the described methods in C++ using open source library OpenCV. The window recognition performance was tested on 60 photographs from the eTRIMS image database [5] and 96 photographs of building façades taken in Slovenia. All photographs were manually annotated for the ground truth window locations. In the experiments we measured the sensitivity and precision of generated hypotheses on position and size of windows in the input images. We define the sensitivity of generated hypothesis as the proportion of actual windows pixels which are correctly identified as such. Precision is defined as the proportion of pixels covered by the generated hypothesis that actually belong to façade windows.



a) b) c) d)

Figure 3: *Creation of the final hypothesis on windows. a) input image. b) and c) cell values in the accumulator array after applying the heuristics evaluation based on the edge distribution and learned classifier, respectively; darker dots correspond to higher values. d) final hypothesis; dark rectangles correspond to individual windows recognized in the input image.*

Figure 4 shows the sensitivity and precision of the final hypotheses measured on the eTRIMS and Ljubljana image datasets. The average sensitivity of hypotheses on eTRIMS dataset is 52.34%, while the average precision is 72.43%. On Ljubljana dataset we have found that the average sensitivity of the final hypotheses is 60.92% and the average precision equals to 77.84%. The results show that our method is conservative. The generated hypotheses cover fewer pixels that actually belong to windows (lower sensitivity) but they are therefore more precise.

For a 640x480 input image the average processing time was 2.92s (measured on standard 2.4GHz CPU). Image rectification and candidates generation typically requires 0.19s and 0.84s, respectively. The rest of computation time goes on heuristic evaluation of the generated candidates. Heuristics based on similarity between candidates and on learned window appearance are computationally the most expensive.

5 CONCLUSION

We have described a method for learning and recognizing windows as basic structural elements of façades. The proposed method has proven successful in analyzing the highly structured façades with large number of aligned windows. This is understandable, as the process of generating candidates and the heuristic evaluation function are based on the projection profile histograms, which are more pronounced in the case of aligned edges. Figure 5 shows some examples of final hypotheses.

---

[1] http://www.vision.ee.ethz.ch/datasets/

[2] http://dib.joanneum.at/cape/TSG-60/

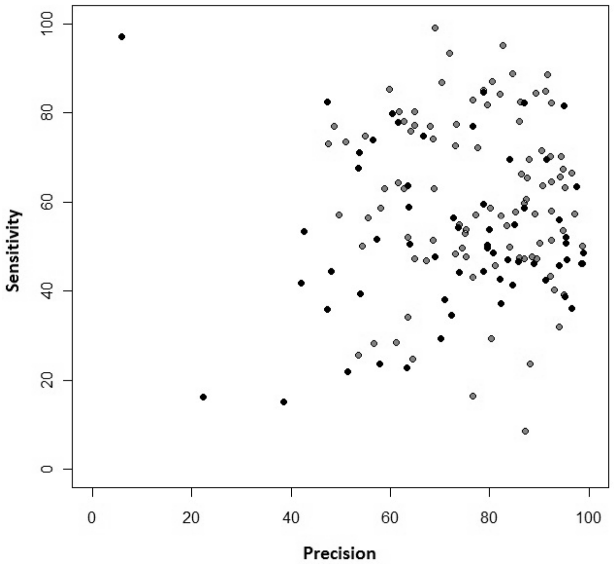


Figure 4: *Sensitivity and precision of the final hypotheses measured on the eTRIMS (black dots) and Ljubljana (grey dots) image datasets.*

The disadvantage of the proposed method is the formation of the final hypotheses solely using the generated candidates. If the image is poorly segmented (e.g. due to a significant level of noise) then most candidates will be inadequate and hence poorly evaluated, which leads to bad hypotheses. To deal with such situations we need more explicit knowledge on the structure of façades.

A promising approach is to use the stochastic context-free grammars, whose production rules provide symmetry, alignment and repetitive patterns of the final hypothesis, as in [6]. They play a role of a guiding mechanism in the search space of all possible structural descriptions of the input façade image. The excessive time complexity is the main problem in using them. The resulting hypothesis generated by our procedure may represent a good starting point for a stochastic grammar, which would drastically limit the search space and thereby significantly shorten the processing time.

Acknowledgements

We wish to thank Marko Robnik-Šikonja and Dean Pertinač for their contribution.

References

[1] P. Vračar. Learning and recognition of façade's structural elements from digital images. MSc Thesis. University of Ljubljana. 2010.

[2] Sung Chun Lee, Ramakant Nevatia. Extraction and integration of window in a 3D building model from ground view image. CVPR. pp. 113-120. 2004.

[3] J. E. Haugeard, S. Philipp Foliguet, F. Precioso, J. Lebrun. Extraction of windows in façade using kernel on graph of contours. SCIA. pp. 646-656. 2009.

[4] P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features. Proc. CVPR. pp. 511-518. 2001.

[5] F. Korč, W. Förstner. eTRIMS Image Database for interpreting images of man-made scenes. Technical Report TR-IGG-P-2009-01. April 2009. URL http://www.ipb.uni-bonn.de/projects/etrims db/.

[6] Nora Ripperda, Claus Brenner. Reconstruction of facade structures using a formal grammar and rjmcmc. DAGM-Symposium, volume 4174 of Lecture Notes in Computer Science, pp. 750-759. Springer, 2006.

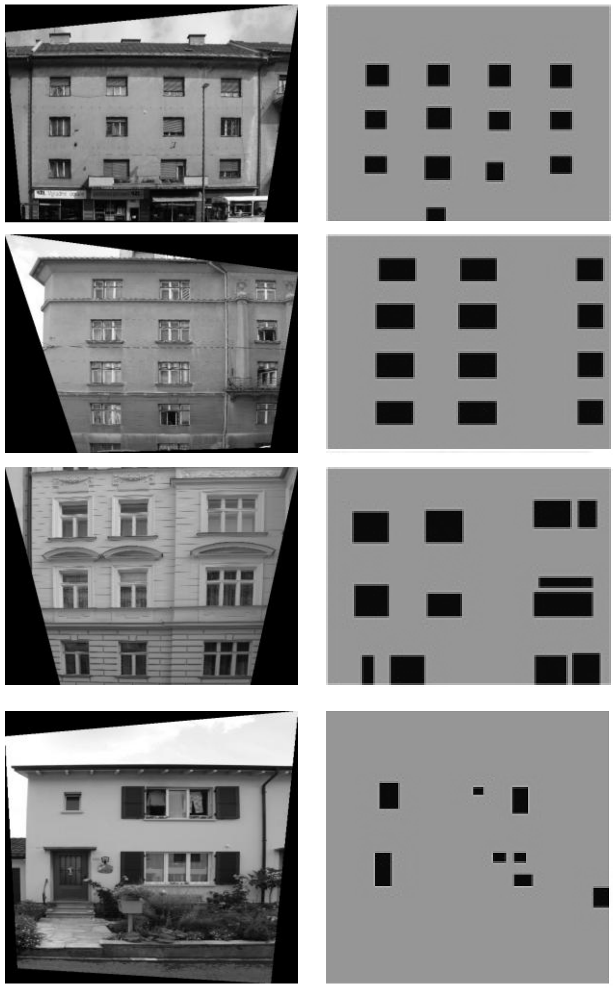


Figure 5: *Examples of final hypotheses. The actual window alignment also improves the procedure of forming the final hypotheses, which favors candidates structured in grid. The results of the proposed method are poor on the facades with very few windows, since in such cases, the evaluation heuristics (especially those based on the distribution of edges and similarity) are usually tenuous. An inaccurate final hypothesis will be generated for the facade with windows that stick close together.*

MERITVE VPLIVA KAKOVOSTI GOVORA V VOWLAN NA SAMODEJNO RAZPOZNAVANJE GOVORCEV

Robert Blatnik, Tomaž Šef
Odsek za inteligentne sisteme, Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana, Slovenija
robert.blatnik@ijs.si, tomaz.sef@ijs.si

POVZETEK

Članek predstavlja metode, orodja in standarde za meritve kakovosti prenosa govora v brezžičnem lokalnem omrežju (VoWLAN). Izdelano je bilo testno okolje za analizo uspešnosti sistemov za samodejno razpoznavanje govorcev v odvisnosti od kakovosti prenosa glede na različne obremenitve omrežja s prometom VoIP. Prvi rezultati so pokazali, da se napaka razpoznavanja EER povečuje v odvisnosti od števila sočasnih VoIP kanalov, s čimer smo potrdili ustreznost okolja.

1 UVOD

Merjenje kakovosti prenosa govora je predvsem namenjeno razvoju novih tehnologij in vzdrževanju telekomunikacijske infrastrukture. Vendar pa nam podatki o kakovosti prenosa govora preko različnih komunikacijskih poti med drugim omogočajo tudi boljši nadzor nad različnimi viri variabilnosti, ki lahko vplivajo na uspešnost oz. zanesljivost sistemov za razpoznavanje govorcev [1].

Kakovost govora v brezžičnih omrežjih je močno odvisna od pogojev znotraj samega radijskega omrežja in od zunanjih pogojev, kot so šum iz ozadja in nelinearni akustični odmev. Za izboljšanje kakovosti so v mobilnih telefonih vgrajeni različni dušilci šuma in izločevalniki odmeva, ki se razlikujejo glede na procesno moč aparatov in kapaciteto baterije. Na zmanjševanje šuma in kakovost zvoka močno vpliva uporaba raznih kodekov, zelo pomembna pa je tudi pokritost signala oziroma število baznih postaj.

Pri samodejnem razpoznavanju govorcev nas še posebej zanimajo vplivi najrazličnejših telefonij in njihovih kombinacij na končne/pričakovane rezultate sistemov razpoznavanja.

2 METODE, ORODJA IN STANDARDI

Do 90. let prejšnjega stoletja je bilo zaznano kakovost govora v telefoniji mogoče meriti le s poslušanjem. Govorimo o tako imenovanih **subjektivnih meritvah kakovosti zvoka**.

Metoda MOS (Mean Option Score) je prva splošno priznana metoda za merjenje kakovosti zvoka v telekomunikacijah [2]. Standardizirana je v priporočilu ITU-T P.800. Rezultat meritve MOS predstavlja statistično povprečje kakovosti posnetega testnega govora, ki je ocenjena s strani določenega števila ocenjevalcev (ca. 20-50 poslušalcev). Ocenjevalna lestvica MOS se giblje med 1 za popolnoma nesprejemljivo kakovost in 5 za odlično kakovost govornega zvoka (glede na določene kriterije). Metoda ne predvideva primerjave popačenega signala na koncu telekomunikacijske povezave z referenčnim izvornim signalom. Pomanjkljivost metode je nezmožnost avtomatizacije izvajanja meritev in posledično velike potrebe po človeških virih ter subjektivnost rezultatov.

Metode za samodejno ocenjevanje kakovosti zvoka na osnovi primerjave zajetega referenčnega in merjenega degradiranega signala do določene mere odpravljajo zgoraj navedene pomanjkljivosti. Prvi je ta koncept predstavil Karjalainen [3] in predstavlja osnovo večine kasnejših objektivnih metod za ocenjevanje kakovosti govora. Te pokrivajo naslednje pomembnejše indikatorje kakovosti govora: zakasnitve in trepetanje, odmev, porezovanje in izguba paketov.

V tabeli 1 je predstavljeno razmerje lestvice subjektivne uporabnikove ocene in objektivno pridobljenega rezultata PESQ MOS.

Uporabnikova ocena	**Rezultat PESQ MOS**
Zelo zadovoljen	4.1 – 4.5
Zadovoljen	3.7 – 4.1
Nekateri uporabniki so zadovoljni	3.4 – 3.7
Veliko uporabnikov je nezadovoljnih	2.9 – 3.4
Skoraj vsi uporabniki so nezadovoljni	2.4 – 2.9
Ni priporočljivo	1.0 – 2.4

Tabela 1: Razmerje med uporabnikovo subjektivno oceno in rezultatom PESQ MOS

Metoda PESQ [4] (standardizirana v priporočilu ITU-P P.862) je nastala kot odgovor na pomanjkljivosti metode PSQM (priporočilo P.861) v pogojih variabilnih zakasnitev, ki so značilni za paketne prenose (npr. VoIP telefonija) [4,5,6]. Najnovejše izboljšave v metodologijah za merjenje kakovosti zvoka pa je prinesla **metoda E-model** (priporočilo G.107), ki lahko spremlja kakovost zvoka na telefonskih povezavah v realnem času. Govorimo o tako imenovanem nevrinjenem merjenju kakovosti zvoka, kjer je meritev izvedena šele po tem, ko je prenos dejansko že izveden [5].

Na razpolago je večje število orodij za merjenje kakovosti zvoka po navedenih metodologijah ITU-T. Večinoma so predraga za akademsko okolje (še posebej tista, ki so osnovana na namenski strojni opremi). Iz tega razloga je bila sprejeta odločitev za uporabo cenovno sprejemljivega **programskega orodja Opera** [6]. Z njim je možno izračunavati kakovost zvoka iz zajetega digitalnega posnetka po metodah PESQ, PSQM in PEAQ. Vključimo lahko tudi module za generiranje klicev v telefonskem omrežju, specializirane zvočne vmesnike za zajem in analizo zvoka ter vmesnike za direktni zajem VoIP prometa v omrežju IP. Sistem primerja in na osnovi izbranega testnega algoritma analizira referenčno in popačeno zvočno datoteko zajetega signala.

3 MERJENJE KAKOVOSTI PRENOSA VoWLAN IN USPEŠNOSTI SISTEMA ZA SAMODEJNO RAZPOZNAVANJE GOVORCEV

3.1 Testno okolje za merjenje kakovosti prenosa VoWLAN

Za meritve kakovosti prenosa govora v brezžičnem lokalnem omrežju (ang. Voice over WLAN) VoWLAN **smo delno prilagodili testno okolje**, ki smo ga zgradili **v okviru mednarodnega projekta Windect** (6. okvirni program) [7].

Slika 1 prikazuje osnovno shemo testnega okolja za analizo uspešnosti sistemov za samodejno razpoznavanje govorcev (SRG) v odvisnosti od kakovosti prenosa VoWLAN. Posneti izvorni govorni signal na vhodni strani telefonske povezave preko D/A pretvorbe pripeljemo na telefon T1 in ga preko vzpostavljene povezave VoIP v WLAN omrežju pošljemo na telefon T2. Popačeni testni signal iz telefona T2 po A/D pretvorbi zakodiramo v stereo WAV datoteko skupaj z izvornim nepopačenim signalom za D/A pretvorbo.

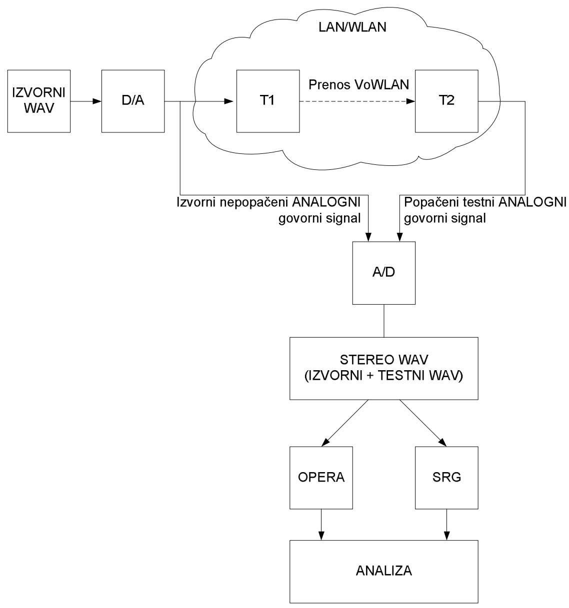
V uporabljeni verziji testnega okolja Opera ni vključena A/D in D/A pretvorba govornega signala, zato smo to izvedli z lastnim programom in gonilnikom vgrajene zvočne kartice, s katerim smo hkrati predvajali referenčni posnetek in ga degradiranega preko testnega WLAN omrežja ponovno zajeli na isti zvočni kartici, kar nam je omogočilo strogo sinhronizacijo začetka predvajanja in začetka snemanja zajetega signala pod redom milisekunde. Dodatno smo lahko izbrali ustrezno frekvenco vzorčenja, število posnetih kanalov, zaznavali popačenja zajetega signala zaradi previsoke amplitude (ang. clipping), poljubno dolžino posnete tišine pred začetkom predvajanja vzorca in podaljšanje snemanja po izteku predvajanja vzorca za primere večjih zakasnitev na telefonski liniji.

Prednost izbranega pristopa je v neodvisnosti od tipa telefonske povezave, saj popačeni testni govorni signal zajemamo v analogni obliki na izhodu za slušalko poljubnega telefonskega aparata.

Zajeti govorni signal v stereo WAV datoteki nato uporabimo za analizo z orodjem Opera in ločeno za meritve uspešnosti sistema SRG, kot je opisano v poglavju 3.2.

Za testni sistem SRG smo uporabljali komercialni sistem za verifikacijo in identifikacijo govorcev [8].

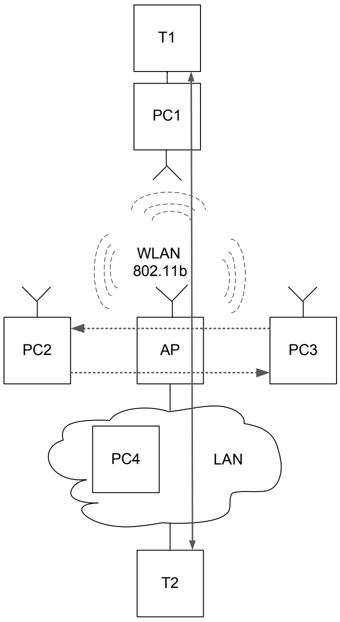
Pomemben del testnega okolja je sistem za generiranje podatkovnega prometa, ki predstavlja simulacijo obremenitev WLAN omrežja z več sočasnimi VoIP povezavami. **Na sliki 2 je** prikazana **izvedba VoWLAN okolja za vzpostavitev testne povezave VoIP in generiranje podatkovnega prometa.**



Slika 1: Osnovna shema testnega okolja

Za brezžično lokalno omrežje smo uporabili dostopovno točko WLAN po standardu 802.11b. Telefonsko povezavo VoIP smo vzpostavili preko dveh telefonov SIP, pri čemer je bil telefon T1 vključen v fiksno lokalno omrežje LAN, drugi, prav tako fiksni telefon T2, pa je bil povezan z omrežjem WLAN preko prenosnika PC1 z vzpostavljeno mostiščno povezavo (ang. Bridged LAN). Podatkovni tok testne VoIP povezave je označen z neprekinjeno črto med T1 in T2. Podatkovne obremenitve omrežja WLAN pa smo izvajali z dvema klientoma WLAN (PC2 in PC3), ki sta si medsebojno izmenjevala rafalni podatkovni promet po protokolu RTP (ang. Real-time Transport Protocol). Promet RTP smo generirali preko ene do dvajset sočasno vzpostavljenih povezav v omrežju WLAN. Tok podatkovnega prometa RTP je označen s črtkanima črtama med PC2 in PC3. Na ta način smo simulirali obremenitve z več sočasnimi povezavami VoIP preko istega RF kanala v omrežju WLAN. Računalnik PC4 smo uporabljali za izvajanje programskih skript s katerimi smo avtomatizirano vklapljali generiranje prometa med PC2 in PC3, predvajali in zajemali govorne posnetke v vzpostavljeni testni povezavi VoIP ter izvajali analize kakovosti govora z orodjem Opera.

Orodje Opera omogoča izvajanje analize PESQ na govornih datotekah, ki praviloma niso daljše od dvajset sekund. Zato smo za verifikacijo ustreznosti testnega okolja prve meritve izvajali z isto namensko testno datoteko, ki je dolga 12 sekund. V nadaljevanju postopka meritev pa smo preko omrežja WLAN popačili tudi testno množico govornih datotek, ki so bile dolge približno 5 minut. Te testne popačene posnetke smo v nadaljevanju uporabili za analizo uspešnosti sistema SRG, ki ga podrobneje opisujemo v naslednjem poglavju.



Slika 2: Izvedba okolja VoWLAN za generiranje prometa RTP

3.2 Postopek meritve uspešnosti sistema SRG

Tipičen sistem SRG podaja nek rezultat istovetnosti govornega vzorca neznane osebe (recimo mu testni vzorec) z govornim vzorcem znane osebe (recimo mu učni vzorec). Bolj kot so karakteristike testnega vzorca statistično blizu karakteristikam učnega vzorca, višji je rezultat v prid potrditvi istovetnosti. V nasprotnem primeru je rezultat ustrezno nižji; običajno pravimo, da sistem zavrača istovetnost.

Uspešnost sistema SRG je tem večja, čim manjkrat se sistem zmoti v potrditvi ali zavrnitvi istovetnosti. V primeru, kadar sistem nepravilno zavrača resnično istovetne testne vzorce (recimo jim klienti) govorimo o napaki napačno zavrnjenih FR (ang. False Rejection), v primeru, ko sistem nepravilno potrjuje istovetnost za resnično neistovetne testne vzorce (recimo jim vsiljivci) govorimo o napaki napačno sprejetih FA (ang. False Acceptance). Uspešen sistem bo torej v čim večji meri pravilno zavračal *vsiljivce* in pravilno sprejemal *kliente*.

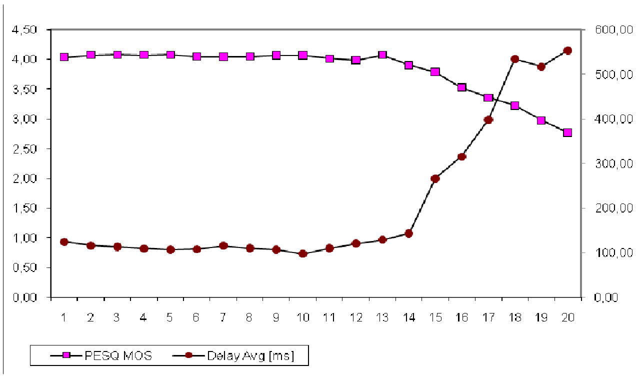
V postopkih meritev uspešnosti sistemov SRG običajno izvajamo določeno število razpoznavanj nad vzorci, za katere vnaprej poznamo pravilen rezultat. Na osnovi rezultatov istovetnosti za testne vzorce klientov in vsiljivcev, ki jih primerjamo z učnimi vzorci klientov, lahko izračunamo pričakovano uspešnost sistema SRG. Pogosto uspešnost sistema SRG izrazimo z enotno stopnjo napake EER (angl. Equal Error Rate). EER predstavlja odstotek napake, ki je enak odstotku napačno sprejetih vsiljivcev in odstotku napačno zavrnjenih klientov.

Prve meritve uspešnosti sistema SRG smo izvedli v 3 serijah, v vsaki seriji smo izvedli po 43 razpoznav za testne vzorce klientov in 1892 razpoznav za testne vzorce vsiljivcev. Prvo serijo razpoznav smo izvedli na datotekah, ki smo jih popačili v neobremenjenem omrežju WLAN. Preostali dve seriji pa smo izvedli na istih govornih datotekah, ki smo jih predhodno popačili v omrežju WLAN pri 5 in nato še pri 10 sočasno vzpostavljenih povezavah VoIP.

4. REZULTATI MERITEV PESQ MOS in EER

4.1 Rezultati meritev PESQ

Na sliki 3 so predstavljeni povprečni rezultati PESQ MOS in povprečne zakasnitve za 32 ponovitev meritev s postopnim povečevanjem obremenitev kanala WLAN z 1 do 20 sočasnih VoIP povezav. Povprečni rezultat PESQ MOS se stabilno giblje okoli rezultata 4 za obremenitve od 1 do 13 povezav VoIP, nato se s povečevanjem števila povezav prične spuščati do rezultata 2,5 pri 20 povezavah. Pri 13 povezavah pa se občutno pričnejo povečevati tudi zakasnitve.

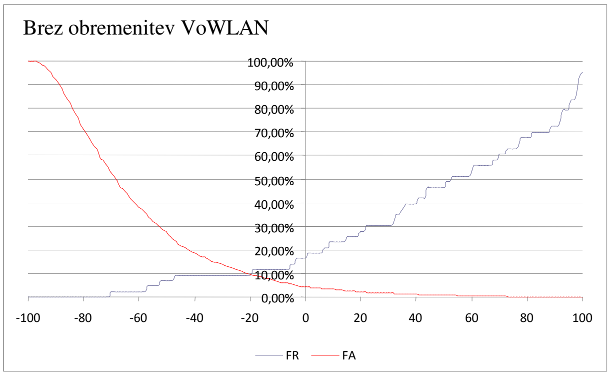


Slika 3: Rezultati meritev PESQ MOS

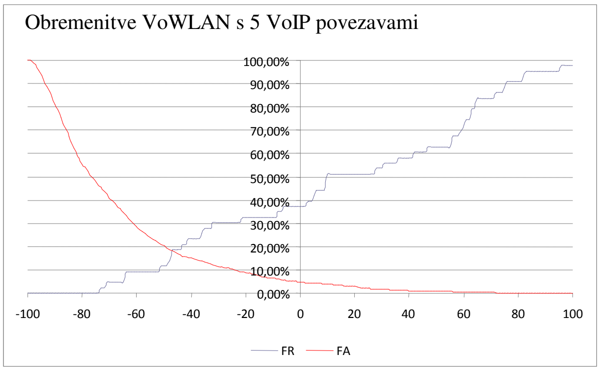
Uspešna izvedba meritev in primerljivost rezultatov s sorodnimi viri potrjuje ustreznost testnega okolja za nadaljnje izvajanje meritev v različnih konfiguracijah merjenega omrežja.

4.2 Rezultati meritev uspešnosti EER

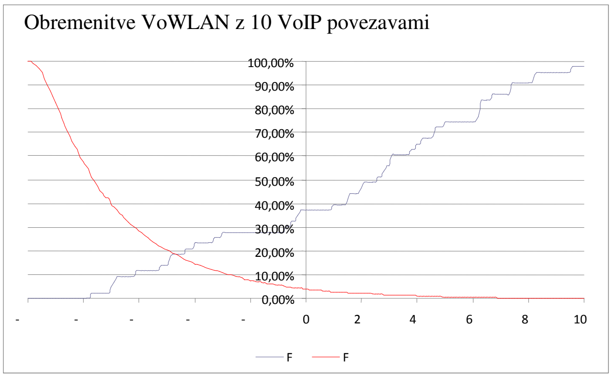
Na slikah 4, 5 in 6 so prikazani rezultati meritev uspešnosti sistema SRG. Modre črte prikazujejo gibanje napake FR, rdeče črte pa prikazuje gibanje napake FA (ordinatna os) v odvisnosti od normaliziranih rezultatov sistema SRG na abscisni osi. Napako EER odčitamo na mestu, kjer se krivulji FA in FR sekata. Napaka razpoznavanja EER sistema SRG je ob neobremenjenem omrežju SRG okoli 10%. S postopnim povečevanjem obremenitev se napaka pričakovano povečuje, in sicer od 10% za 5 sočasnih VoIP povezav do skoraj 20% za 10 sočasnih VoIP povezav (slika 4, slika 5 in slika 6).



Slika 4: Nepopačeni testni podatki.



Slika 5: 5 sočasnih VoIP povezav na WLAN kanalu.



Slika 6: 10 sočasnih VoIP povezav na WLAN kanalu.

5 VPLIV KAKOVOSTI PRENOSA GOVORNEGA SIGNALA NA SAMODEJNO RAZPOZNAVANJE GOVORCEV

Izvedli smo primerjavo uspešnosti sistema SRG v odvisnosti od obremenitev omrežja WLAN s povečevanjem števila sočasnih VoIP kanalov na istem WLAN kanalu ter posledično manjše kakovosti prenosa govora. Opazimo lahko, da se napaka EER s povečevanjem obremenitev WLAN omrežja postopoma povečuje do skoraj 20%. Po drugi strani pa se kakovost PESQ MOS občutno poslabša šele nad 13 sočasno vzpostavljenih kanalov VoIP.

Ob primerjavi rezultatov uspešnosti sistema SRG z izmerjeno kakovostjo prenosa govora v WLAN lahko ugotovimo, da se sistem SRG prične občutno prej odzivati na obremenitve WLAN omrežja, kot lahko to ugotovimo z meritvami kakovosti prenosa govora po metodi PESQ.

Potrebno pa je poudariti, da so meritve PESQ MOS izvedene na eni govorni datoteki, meritve napake EER pa na celotni testni množici govornih podatkov, kar ima lahko za posledico nekoliko slabšo odzivnost meritev PESQ MOS glede na izmerjeno napako EER.

6 SKLEP

Kakovost komunikacijske poti in zaznana kakovost prenesenega signala sta eden od pomembnih elementov pri razvoju sistemov SRG in snemanju govornih korpusov. V ta namen smo zgradili opisano testno okolje. Prvi rezultati meritev so potrdili ustreznost metode.

Testirali smo uspešnost sistema SRG v odvisnosti od postopnega obremenjevanja omrežja WLAN z več vzpostavljenimi VoIP povezavami na istem WLAN kanalu.

V nadaljevanju nameravamo izvesti meritve PESQ MOS nad istimi podatki, kot jih uporabljamo za meritve uspešnosti sistema SRG, kar nam bo omogočilo neposredno opazovanje korelacij med rezultati EER in PESQ MOS. Okolje nameravamo razširiti še za meritve zanesljivosti sistema v različnih telefonijah (in njihovih kombinacijah) ter na različnih telefonskih aparatih in mikrofonih. Tako bo omogočena kakovostnejša analiza in kvantitativno ovrednotenje razlik v različnih snemalnih pogojih.

Literatura

[1] P. Rose, *Forensic speaker Identification*, Taylor & Francis, 2002.

[2] Mean Opinion Score (MOS) Terminology, ITU-T Recommendation P.800.1. 2003.

[3] Karjalainen, M. A new auditory model for the evaluation of sound quality of audio systems, IEEE ICASSP, 608–611, 1985.

[4] PESQ: An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech CODECs, ITU-T Recommendations, str..862, 2001.

[5] Rix, A. W. Beerends, J. G. Hollier, M.P. Hekstra, A. P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. IEEE ICASSP 2001. Salt Lake City, UT, USA. 2001.

[6] http://www.opticom.de/

[7] http://www.windect.ethz.ch/

[8] http://www.persay.com/pdf/SPID_V6_DataSheet.pdf

PRILAGAJANJE MODELA ZA RAZPOZNAVANJE AKTIVNOSTI ČLOVEKA Z AKTIVNIM DELNO NADZOROVANIM UČENJEM

Božidara Cvetković, Boštjan Kaluža, Matjaž Gams
Department of Intelligent Systems
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773419
e-mail: boza.cvetkovic@ijs.si

POVZETEK

Fizične karakteristike človeka definirajo način izvajanja aktivnosti, saj se razmerja dolžin med posameznimi deli telesa ob izvajanju aktivnosti razlikujejo od človeka do človeka. Pri prepoznavanju aktivnosti s pomočjo strojnega učenja se natančnost modela lahko znatno zmanjša, ko uporabimo model za prepoznavanje aktivnosti človeka s specifičnimi fizičnimi karakteristikami. V tem članku predstavimo postopek prilagajanja modela končnemu uporabniku.

Metodo sestavljajo trije koraki: (i) zajemanje posnetka uporabnika pri treh osnovnih aktivnostih, (ii) skaliranje vrednosti atributov, ki so povezani z višino človeka, in (iii) aktivno učenje osnovnega modela z uporabo novih, delno označenih podatkov. Rezultati kažejo, da predlagana metoda lahko poveča natančnost modela tudi za 11 odstotnih točk.

1. UVOD

Razpoznavanje aktivnosti uporabnika je ena ključnih komponent v sistemu Confidence [1], katerega namen je podaljšati neodvisno življenje starejših, ki se še lahko samostojno gibajo in bi želeli to samostojnost obdržati brez vsakodnevne pomoči družinskih članov ali medicinskega osebja. Sistem Confdence za razpoznavanje aktivnosti uporablja model opisan v [2], a starejši imajo precej individualen način premikanja. Da bi bil model bolj uspešen pri prepoznavanju aktivnosti, je izredno pomembno da se prilagodi končnemu uporabniku. S prilagoditvijo lahko a) modelu bolj zaupamo in b) pridobimo kvalitetnejše podatke o aktivnostih, ki se nato uporabijo za detekcijo zdravstvenih težav.

V prispevku je predstavljena metoda za prilagoditev modela za prepoznavanje človeških aktivnosti končnemu uporabniku. Prilagoditev se vrši z uporabo aktivnega učenja. Cilj je razpoznati osem krajših aktivnosti, ki predstavljajo bistvene elemente daljših opravil. Te aktivnosti so stanje ali hoja, sedenje, ležanje, usedanje, vstajanje, padanje, sedenje na tleh in aktivnost na vseh štirih. Metoda, ki smo jo uporabili je sestavljena iz treh korakov: (1) zajemanje podatkov uporabika pri treh osnovnih aktivnostih (sedenje, hoja,ležanje); (2) skaliranje modela na višino uporabnika; [3] uporaba meta modela za delno nadzorovano aktivno učenje.

V primeru, da se zadnji korak izteče uspešno in smo instanco opredelili za informativno, jo vzamemo za nov učni primer in s tem omogočimo aktivno učenje. Na ta način dobimo nove učne primere za vseh osem aktivnosti in postopoma izboljšamo uspešnost modela.

2. SORODNA DELA

Največkrat uporabljene metode aktivnega učenja v literaturi so tokovno selektivno vzorčenje (*stream-based selective sampling*), poizvedba po stopnji članstva instance (*membership queries*), vzorčenje iz bazena podatkov (*pool –based sampling*) [4].

Izbira metode je odvisna od domene problema in dosegljivosti podatkov. Izkazalo se je, da se v literaturi največkrat uporablja metoda vzorčenja iz bazena podatkov za aktivno učenje, ne glede na domeno. V tem primeru se podatki nahajajo v bazi podatkov iz katere algoritem izbere instanco in se odloči ali jo bo označil ali ne [6],[7]. V okviru prepoznavanja aktivnosti je navedena metoda uporabljena v [5], kjer so z delno nadzorovanim učenjem, dva komplementarna modela, označevali vsebino video posnetkov.

3. PRILAGAJANJE MODELA

Sistem je postavljen v okolje, kjer lahko raziskave potekajo na podatkih v realnem času. Podatki, ki jih sistem Confidence sprejema v procesiranje, so pozicije senzorjev Ubisense [8], ki jih ima človek ves čas pripete na obleko in obutev. Uporabili smo 4 senzorje: vrat, pas, leva noga in desna noga. Podatki o pozicijah vseh značk in izračunane vrednosti atributov [2] v realnem času vstopajo v sistem in vsaka instanca se klasificira z modelom za prepoznavanje aktivnosti.

Predstavljena metoda za prilagajanjen končnemu uporabniku je osnovana na ideji tokovno selektivnega vzorčenja, saj moramo v vsakem trenutku poznati aktivnost uporabnika, torej moramo prav vsako instanco klasificirati.

3.1 Modeli za prepoznavanje aktivnosti

V postopku prilagajanja modela končnemu uporabniku smo uporabili tri ločene modele: splošni model, ki ne vsebuje podatkov o gibanju končnega uporabnika, model osebe, ki se zgradi tekom inicializacije sistema in meta model, ki uporablja tako klasifikacije prvih dveh kot tudi vse razpoložljive učne podatke. Shema modelov je prikazana na sliki 1.

Splošni model

Osnovni model [2] se je izkazal za uspešnega na večini testiranih uporabnikov, zato smo ga privzeli kot *splošni model*. Model je naučen iz podatkov aktivnosti treh po fizičnih lastnostih različnih ljudi. Za učenje smo uporabili algoritem Random Forest in ga testirali z metodo »leave-out« validacije na način izpusti enega človeka. Tako smo testirali model dveh oseb na podatkih osebe, ki ni zajeta v modelu. Test smo izvedli trikrat, tako da smo vsakič izpustili eno osebo iz modela in na njej testirali. Rezultat validicacije je 86% klasifikacijska točnost modela. Klasifikacijska točnost modela na četrti osebi oz. končnemu uporabniku, katerega učni podatki niso bili vključeni v osnovni model, je bila 73%.

Model osebe

V času inicializacije sistema Confidence se zajame podatke osnovnih aktivnosti končnega uporabnika. Zajemanje poteka tako, da se posname po 30 sekund aktivnosti ležanja, hoje in sedenja. Ker to prestavlja sorazmerno majhno množico podatkov, se podatke pomnoži štiri-krat. Za učenje uporabimo algoritem Random Forest. Testiranje uspešnosti modela je potekalo na posnetku dolgem 30 minut z označenimi osmimi aktivnostmi. Izmerjena klasifikacijska točnost je bila 69%. Uspešnost modela je relativno visoka, saj se osnovne zajete aktivnosti izvajajo večkrat in daljši čas, kot tiste, ki jih *model osebe* ne pozna. Te so vsedanje, vstajanje, padanje, sedenje na tleh in aktivnost na vseh štirih.

Meta Model

Namen meta modela je sprejeti odločitev, kateri izmed zgornjih dvem modelov je bolj točen za klasifikacijo določene instance. Meta model smo zgradili na dveh posnetkih (2x30min) končnega uporabnika. Uporabili smo rezultate obeh modelov in statistične podatke, ki sta jih vrnila oba modela.

Atributi so bili izbrani na podlagi vrste tesiranj uspešnosti meta modela. Preizkusili smo tri skpine atributov, kot je prikazano v tabeli 1, in različne algoritme strojnega učenja. Pri tem zaupanje določenega modela predstavlja odstotek dreves, ki so določeno instanco enako klasificirala. Rezultati testiranja, ki je potekalo z 10-kratnim prečnim preverjanjem, so prikazani v tabeli 2. Kot najboljša kombinacija se je izkazala izbira atributov iz prve in tretje množice ter učni algoritem Random Forest.

Množica	Atributi
1	- klasifikacija splošnega modela C_1 - kasifikacija modela osebe C_2 - zaupanje splošnega modela $M_s(C_1)$ - zaupanje modela osebe $M_o(C_2)$ - ali je vrnjen razred osnovnega modela osnovna aktivnost $C_1 \in R_o$ - ali sta razreda ista $C_1 = C_2$
2	- z koordinate vseh senzorjev - razlika po z koordinati med vratom in povprečjem z koordinate nog - razlika po z koordinati med vrato in pasom
3	- zaupanje splošnega modela v razred modela osebe $M_s(C_2)$ - zaupanje modela osebe v razred splošnega modela $M_o(C_1)$

Tabela 1. Tri različne množice atributov, ki smo jih uporabili za izbiro najuspešnejšega meta modela

	Snap + 1	1	1+ 2	2	1 + 3
SMO	86.6 %	92.9 %	87.8 %	88.9 %	88.3 %
C4.5	96.8%	95.4 %	96.6 %	96.1 %	95.9 %
Random forest	90.9 %	95.9 %	96.9 %	96.6%	97.4 %
Naive Bayes	61.0 %	75.7%	68.8 %	70.1 %	82.3 %
AdaBoost	88.6 %	84.8 %	84.6 %	84.6 %	79.0 %
Bagging	96.9%	94.7 %	96.2 %	95.8 %	95.8 %

Tabela 2. Rezultati testiranj algoritov na petih različnih množicah atributov. Najboljši rezlutat smo dobili z algoritmom Random Forest in uporabo atributov iz množice ena in tri (Tabela 1).

3.2 Pol avtomatično označevanje in aktivno učenje

Adaptacija je setavljena iz treh korakov, v katerih smo zgradili popolnoma nov model, skalirali osnovni model glede na višino uporabnika in uporabili meta model za izbiro modela za označitev na podlagi katere smo prilagodili vse aktivnosti splošnega modela uporabniku.

Korak

Cilj prvega koraka je zajeti podatke o osnovnih aktivnostih uporabnika in zgraditi *model osebe*. Postopek zajema aktivnosti je opisan v sekciji 0. Ta model je ključnega pomena pri uporabi *meta modela*, saj prav tako klasificira in poda oceno zaupanja v svojo odločitev.

Predvidevali smo, da bo *model osebe* z večjim zaupanjem klasificiral instance osnovnih aktivnosti (hoja, sedenje, ležanje) in prepustil *splošnemu modelu* klasifikacijo ostalih aktivnosti.

Korak

Višina uporabnika je pomemben parameter pri uspešnosti modela. V drugem koraku smo podatke iz *splošnega modela*, ki so povezani z višino uporabnika, skalirali na višino končnega uporabnika. Privzeta višina uporabljena v osnovnem modelu je izračunana kot povprečna višina vseh oseb, katerih podatki so uporabljeni v modelu.

Korak

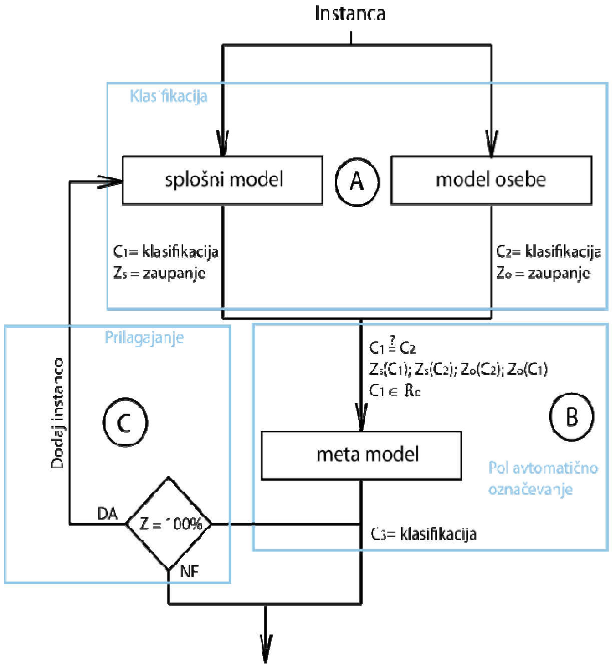
Tretji korak poteka vzporedno z delovanjem sistema. Vsaka instanca, ki jo sistem sprejme, se klasificira s splošnim in z modelom osebe (Slika **1**, proces A). Vsak od modelov vrne tudi zaupanje v svojo napoved. Rezultata modelov in ostale statistične vrednosti ter logične funkcije postanejo vhodni podatki meta modela (Slika **1**, proces B). Izhod modela je odločitev, kateri model bomo uporabili za označevanje instance in zaupanje v svojo odločitev. Zaupanje v odločitev modela katerega je izbral meta model in zaupanje meta modela morata biti 100%, da se odločimo da je trenutna instanca dovolj informativna za prilagajanje. To instanco dodamo kot nov učni primer in ponovno zgradimo splošni model (**Error! Reference source not found.**, proces C). Z vsakim popravljanjem splošnega modela se ta prilagodi trenutnemu uporabniku.

REZULTATI EKSPERIMENTA

Z aktivnim učenjem smo izvedli prilagajanje splošnega modela uporabniku. Klasifikacijska točnost *splošnega modela* je glede na izbranega končnega uporabnika pred prilagajanjem 73%. *Splošni model* pogosto narobe klasificira ravno osnovne aktivnosti, kot sta sedenje (43% napačno klasificiranih) in stanje (23% napačno klasificiranih). Med inicializacijo sistema Confidence, kjer sistem seznanimo z novim uporabnikom, smo za vsako od osnovnih aktivnosti (sedenje, stanje, ležanje) zajeli po 30 sekund podatkov. Natančnost novega *modela osebe* je 69%, saj *model osebe* ne pozna nobene druge aktivnosti razen osnovnih.

Končni uporabnik, ki je sodeloval v eksperimentih, je bil nižji od oseb katerih podatki so sestavni del *splošnega modela*, zato smo že s samim skaliranjem vrednosti atributov *splošnega modela* izboljšali klasifikacijsko točnost na 79%. Za tem korakom je končni uporabnik pričel uporabljati sistem.

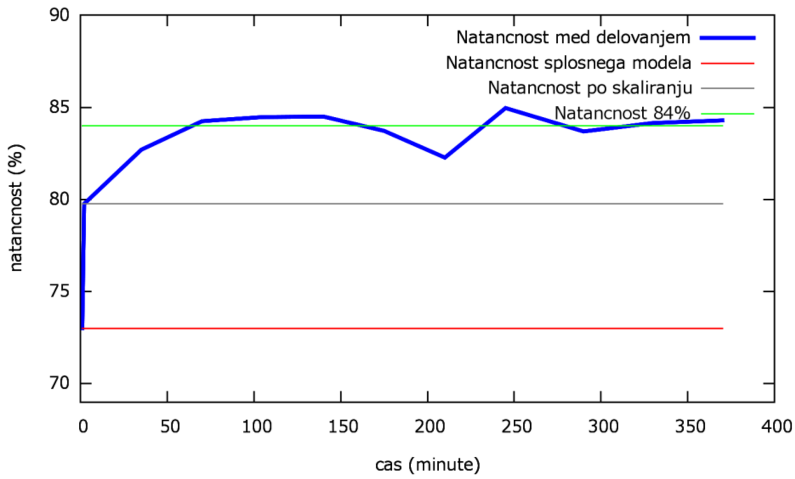
Uporabili smo tri 60-minutne posnetke z neoznačenimi podatki.



Slika 1. Diagram poteka aktivnega učenja A) Klasifikacija obeh modelov. B) Meta model vrne odločitev, kateri model vzamemo za označevanje. C) V primeru da je vrednost zaupanja modela, ki ga vrne meta model 100% se izvrši dodajanje instance v učno množico in s tem prilagajanje splošnega modela.

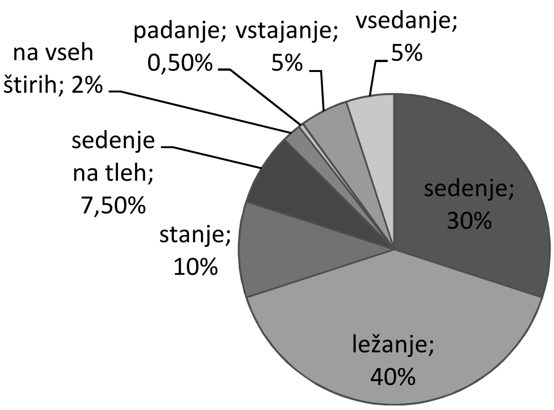
V posnetkih smo zajeli vse aktivnosti uravnoteženo, kot bi jih starejši uporabnik izvajal tekom dneva (slika 2). Vsak posnetek smo uporabili dvakrat.

Vsako instanco, ki je vstopila v sistem smo klasificirali z obema modeloma, splošnim in osebnim. Poleg vrnjenih razredov aktivnosti, smo uporabili tudi vektor zaupanja modela v razrede, ki jih napoveduje. Uporabili smo tako zaupanje modela v razred, ki ga je sam napovedal, kot tudi zaupanje modela v rezred, ki ga je napovedal drugi model (Tabela 1, atributi 3). *Meta model* na podlagi statističnih podatkov, vrenjenih razredov in logičnih enačb vrne razred, ki nam pove katerega od modelov naj vzamemo za klasifikacijo trenutne instance. V primeru, ko je zaupanje meta modela in izbranega modela enako 100% smo instanco označili za prenos v splošni model.



Graf 1. Graf natančnosti modela med prilagajanjem. Krivulja se dvigne s 73% na 79% s skaliranjem splošnega modela. Nato se z aktivnim učenjem zaniha na natančnost 84%.

Vsako dodano instanco smo v splošni model dodali štiri-krat. Vsakih pet minut podatkov se je splošni model ponovno zgradil. V nadaljni klasifikaciji se je uporabil prilagojen splošni model.



Slika 2. Razporeditev aktivnosti v enem dnevu

Klasifikacijska točnost se je postopoma dvignila iz 79% najprej na 85% in nato ustalila pri 84% (Graf 1). Točnost modela smo testirali na ločenem testnem posnetku. Aktivno učenje smo ustavili 30 minut pred koncem zadnjega posnetka, saj je bila vsaka instanca klasificiran le s *splošnim modelom*. To pomeni, da je *splošni model* prilagojen končnemu uporabniku. Projekcija na realen dan bi pomenila, da je *splošni model* uspešnejši od *modela osebe* po 90 minutah.

ZAKLJUČEK

Z uporabo skaliranja in aktivnega učenja, kjer smo si za označevanje novih podatkov izbrani delno nadozorovano učenje, smo model, ki je na izbranem uporabniku deloval z natančnostjo 73%, zvišali na 84%.
V nadaljnjem delu bomo *meta model* nadgradili z novimi atributi, ki bodo opisovali razmerje med zajetimi podatki, zgodovino klasifikacij *osnovnega modela* in *modela osebe* ter kompleksnejšo statistiko. Predvidevamo, da bi z implementacijo staranja podatkov lahko bolj natančno predvideli uporabnikovo aktivnost, saj se pri starejših ljudeh lahko karakteristike gibanja iz dneva v dan spreminjajo. Nadaljnje delo vključuje tudi razširjeno testiranje metode z različnimi končnimi uporabniki.

Reference

[1] Confidence: Ubiquitous Care System to Support Independent Living, http://www.confidence-eu.org

[2] M. Luštrek, B. Kaluža, Fall Detection and Activity Recognition with Machine Learning, Informatica, pp. 205-212, 2009

[3] M. Luštrek, B. Kaluža, E. Dovgan, B. Pogorelc, M. Gams, Behavior Analysis Based on Coordinates of Body Tags, Ambient Intelligence, pp. 14-23, 2009

[4] B. Settles, Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009

[5] M. Stikic, K. Van Laerhoven, B. Schiele, Exploring semi-supervised and active learning for activity recognition, 12th IEEE International Symposium on Wearable Computer, 2008

[6] N. Abe and H. Mamitsuka., Query learning strategies using boostingand bagging , *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998

[7] T. Zhang V. S. Iyengar, C. Apte, Active learning using adaptive resampling, *Sixth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 92–98, 2000

[8] Ubisense. http://www.ubisense.net, 2008-09-15.

SEARCHING FOR DRIVING STRATEGIES ACCORDING TO CONFLICTING OBJECTIVES

Erik Dovgan (1)*, Matija Javorski* (2)*, Bogdan Filipič* (1)
(1) Department of Intelligent Systems, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
(2) Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva cesta 6, 1000 Ljubljana, Slovenia
Tel: +386 1 4773393; fax: +386 1 4773131
e-mail: erik.dovgan@ijs.si

ABSTRACT

This paper presents a multiobjective optimization approach to search for nondominated driving strategies regarding two conflicting objectives: consumed time and fuel. It describes the developed vehicle simulator, driving simulation and an algorithm for finding the strategies based on the popular Non-dominated Sorting Genetic Algorithm II (NSGA-II). It also reports on preliminary tests and results, and describes the ideas for future work on this subject.

1 INTRODUCTION

The price of goods significantly depends on the transportation costs. If, for example, the goods are produced in the countries with low labor costs, the largest proportion of their price on the European market comes from the transportation costs. Besides the transportation costs, the transportation time is also important. The two objectives are in conflict with each other, i.e., minimizing one of the objectives deteriorates the other.

When computing the transportation costs, a set of variables have to be taken into account. If local transportation is addressed, e.g., transportation in Slovenia, the set of variables is reduced since the goods can be transported only with road vehicles. The transportation costs include the labor costs that are linearly proportional to the transportation time, and the price of the fuel needed for the transportation.

To minimize both or even additional objectives, we have defined and implemented a system that searches for a set of nondominated driving strategies. A strategy is nondominated if there exists no other strategy better in at least one objective and not worse in all other objectives. When a set of nondominated strategies is found, higher-level knowledge has to be used in order to select the preferred strategy with respect to the consumed time and fuel. For example, if fast transportation is needed, a strategy with low transportation time can be used. On the other hand, if the transportation time is not critical, a slower transportation strategy is beneficial as it decreases the fuel consumption.

The developed system can be used by transportation companies. Besides, its usage can be extended to the individuals that are interested in their time and fuel consumption management when driving their cars. If we concentrate on the transportation companies, their vehicles usually travel repeatedly on the same routes due to the fixed positions of the company storages. Therefore, the driving strategy search system can be focused on single routes thus finding the best strategies for these routes. However, if the number of routes is very large, the single route approach is not feasible. To cope with this problem, the strategy search must take into account all the routes or even a larger set of routes, e.g., all routes of a country that can be used for transportation, thus providing one set of strategies for all routes. The latter use case is also potentially useful for individual users.

This paper presents a system for discovering driving strategies, the preliminary tests and results, and the ideas for future work.

2 A SYSTEM FOR FINDING DRIVING STRATEGIES

The described problem requires a multiobjective optimization algorithm [1] to find nondominated strategies for vehicle driving.

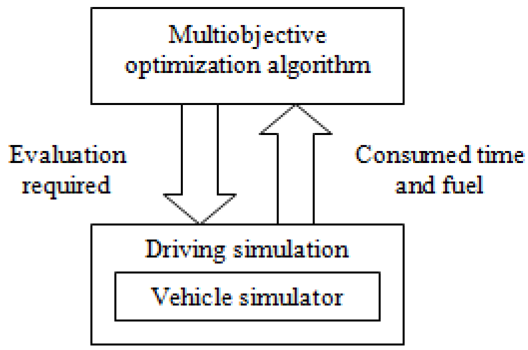


Figure 1: *System architecture*

$\varepsilon_{throttle}$	throttle percentage $[0,\ldots,1]$
$\varepsilon_{braking}$	braking percentage $[0,\ldots,1]$
I	gear $[I_{vehicle,min},\ldots,I_{vehicle,max}]$

Figure 2: *Vehicle control parameter*

When testing a found strategy, vehicle driving on one or more routes has to be simulated, where a route is divided into segments defined by their length, turning radius, vertical angle and speed limit. Therefore, a numerical vehicle simulator included in a vehicle driving simulation is required. This structure can be seen in Figure 1. The following subsections describe the vehicle simulator, its role in the driving simulation, and the multiobjective algorithm for finding driving strategies.

$$F_{moving} - F_{braking} = F_{wheel} + F_a + F_{aero} + F_g$$
$$F_{moving} = \frac{T_{wheel}}{r_{wheel}}$$
$$T_{wheel} = i_{gearRatios}(I) i_{differentialRatio} T \eta$$
$$F_{moving} = \frac{T i_{gearRatios}(I) i_{differentialRatio}}{r_{wheel}} \eta$$
$$T = \varepsilon_{throttle} T_{max} = \varepsilon_{throttle} f_{T_{max}}(n)$$
$$n = n_{wheel} i_{gearRatios}(I) i_{differentialRatio} = \frac{v}{2\pi r_{wheel}} i_{gearRatios}(I) i_{differentialRatio}$$
$$F_{braking} = \mu m g \cos\alpha \varepsilon_{braking}$$
$$F_{wheel} = c_r m g \cos\alpha$$
$$F_{aero} = 0.5 \rho v^2 A_x c_x$$
$$F_g = m g \sin\alpha$$
$$F_a = ma = F_{moving} - F_{braking} - F_{wheel} - F_{aero} - F_g$$
$$a = \frac{F_{moving} - F_{braking} - F_{wheel} - F_{aero} - F_g}{m}$$
$$v = v + a\Delta t$$
$$P = 2\pi T n$$
$$\Delta c = f_c(T, n) P \Delta t$$
$$\Delta s = v\Delta t + \frac{a\Delta t^2}{2}$$

Figure 3: *Vehicle simulator formulas*

2.1 Vehicle simulator

Vehicle simulator [2] simulates the vehicle driving for one time step on the current route segment. The vehicle control parameters are presented in Figure 2. The vehicle simulator is implemented as a set of formulas presented in Figure 3. The simulator parameters is given in Figure 4. The results of one step simulation are the new velocity, the consumed fuel and the traveled path.

2.2 Driving simulation

Vehicle driving simulation on a route runs until the whole route is traveled. At each time step we define the throttle and braking percentage, and gear. Next, we simulate one time step with the simulator and update the vehicle state and position on the route.

The vehicle driving, i.e., defining throttle and braking percentage, and gear, is done with each driving strategy. A strategy consists of if-then rules. After each time step, the rules in a strategy are checked and the first rule that fires is applied to control the vehicle in the next time step. If no rule fires, the driving control does not change. The form of rules and rule parameters are shown in Figure 5. The rules and strategies are found with a multiobjective optimization algorithm as described in the following section.

r	turning radius of the current segment
α	inclination of the current segment
F_{moving}	moving force
$F_{braking}$	tire braking force
F_{wheel}	wheel friction force
F_a	inertial force
F_{aero}	aerodynamic drag force
F_g	slope friction force
Δt	interval
T	engine torque
n	engine speed
T_{wheel}	torque at vehicle wheels
r_{wheel}	wheel radius
$i_{gearRatios}$	vehicle specific gear ratios
$i_{differentialRatio}$	vehicle specific differential ratio
η	vehicle specific transmissions mechanical efficiency
T_{max}	vehicle specific maximal engine torque
n	engine speed $\quad n \leq n_{vehicle,max}$
$f_{T_{max}}(n)$	vehicle specific maximal torque function
n_{wheel}	wheel rotations
v	vehicle velocity
μ	tire braking force percentage
m	vehicle mass
c_r	rolling resistance coefficient
ρ	air density
A_x	vehicle frontal area
c_x	vehicle specific aerodynamic coefficient
a	inertial acceleration
P	engine power
Δc	fuel consumption
$f_c(T,n)$	fuel consumption function
Δs	route traveled

Figure 4: *Vehicle simulator parameters*

2.3 Multiobjective optimization algorithm for finding driving strategies

To find a set of nondominated solutions, a multiobjective genetic algorithm based on NSGA-II [3] has been implemented. It operates on a population of candidate solutions, i.e., strategies in the form of sets of rules encoded as shown in Figure 6. The algorithm pursues two objectives: the consumed time and fuel. It is presented in Figure 7 and works as follows. Firstly, the strategies are randomly initialized. Next, the population is evolved for several generations where in each generation the strategies for the current population are selected and the others are removed. This is done with the *fastNondominatedSort* and *crowdingDistanceAssigment* functions. The function *fastNondominatedSort* sorts the strategies into nondominated fronts. Strategies of the front i dominate the solutions of the fronts $i+1, i+2, \ldots$. Besides, the strategies inside a front are sorted based on the crowding distance: strategy j has a higher crowding distance than the strategies $j+1, j+2, \ldots$.

if
$\alpha_{current,min} \le \alpha_{current} \le \alpha_{current,max}$
$\alpha_{next,min} \le \alpha_{next} \le \alpha_{next,max}$
$I_{min} \le I \le I_{max}$
$s_{toNext,min} \le s_{toNext} \le s_{toNext,max}$
$v_{current,min} \le v_{current} \le v_{current,max}$
$v_{currentLimit,min} \le v_{currentLimit} \le v_{currentLimit,max}$
$v_{nextLimit,min} \le v_{nextLimit} \le v_{nextLimit,max}$
then
$\varepsilon_{throttle,rule}$
ΔI
$\varepsilon_{braking,rule}$
$\zeta_{braking}$
end if

Rule condition parameters

Parameters	Lower bound	Upper bound
$\alpha_{current,min}$, $\alpha_{current,max}$	-0.5 radians	0.5 radians
$\alpha_{next,min}$, $\alpha_{next,max}$	-0.5 radians	0.5 radians
I_{min}, I_{max}	$I_{vehicle,min}$	$I_{vehicle,max}$
$s_{toNext,min}$, $s_{toNext,max}$	0 m	1000 m
$v_{current,min}$, $v_{current,max}$	0 km/h	140 km/h
$v_{currentLimit,min}$, $v_{currentLimit,max}$	0 km/h	140 km/h
$v_{nextLimit,min}$, $v_{nextLimit,max}$	0 km/h	140 km/h

Rule consequence parameters

Parameters	Allowed values
$\varepsilon_{throttle,rule}$	$[0,1]$
ΔI	$\begin{cases} \{0,1\} & ; I = I_{vehicle,min} \\ \{-1,0\} & ; I = I_{vehicle,max} \\ \{-1,0,1\} & ; \text{otherwise} \end{cases}$
$\varepsilon_{braking,rule}$	$[0,1]$
$\zeta_{braking}$	$\{true, false\}$

Figure 5: *The form of rules and the rule parameters*

$$[\alpha_{current,min}, \alpha_{current,max}, \alpha_{next,min}, \alpha_{next,max}, I_{min}, I_{max}, s_{toNext,min}, s_{toNext,max}, v_{current,min}, v_{current,max}, v_{currentLimit,min}, v_{currentLimit,max}, v_{nextLimit,min}, v_{nextLimit,max}, \varepsilon_{throttle,rule}, \Delta I, \varepsilon_{braking,rule}, \zeta_{braking}]$$

Figure 6: *Encoding of driving strategies*

After the strategies are sorted, the strategies with the highest front indexes and, within a front, with the highest crowing distance indexes are removed to reduce the population to a predefined size. After the population is resized, the genetic operators are applied to the strategies as indicated in Figure 7 (function *makeNewPopulation*). Initially, two parents are selected. Then we crossover the parents by exchanging the data between subsets of rules of the strategies. Next, three types of mutation are applied to each strategy. Mutation of the first type changes the sequence of rules in a strategy. By the second type mutation, each parameter of each rule in a strategy is changed with a predefined probability. The third type mutation adds a random rule and deletes a randomly selected rule. When the mutations are carried out, the new strategies are evaluated, marked as offspring and added to the population. The evaluation is done with driving simulation on one ore more routes. If more routes are used, the objective values are the summed consumed time and the summed consumed fuel over all routes. The algorithm stops after the predefined number of generations and returns the first front of strategies, namely the best found strategies that are not dominated by any other found strategy. The following section presents the algorithm testing and an example of the found nondominated strategies.

Algorithm NSGA-II
randomly initialize parent population P_0
{number of chromosomes is the population size S_{pop}}
initialize empty offspring population Q_0
$t = 0$
while stopping condition is not true **do**
unite parent and offspring population $R_t = P_t \bigcup Q_t$
$F = fastNondominatedSort(R_t)$ {returns $F = (F_1, F_2, \ldots)$ all nondominated fronts of R_t}
$i = 1$
while $\|P_{t+1}\| < S_{pop}$ **do**
{fill the new parent population}
$crowdingDistanceAssignment(F_i)$
$P_{t+1} = P_{t+1} \bigcup F_i$ {include i-th nondominated front in the new parent population}
$i = i + 1$
end while
$sort(P_{t+1}, \ge_n)$ {sort in descending order using $\ge_n$}
$P_{t+1} = P_{t+1}[0 : S_{pop}]$ {choose first S_{pop} solutions}
$Q_{t+1} = makeNewPopulation(P_{t+1})$ {use crossover and mutation to create a new offspring population Q_{t+1}}
$t = t + 1$
end while

Algorithm makeNewPopulation(P_{t+1})
for $i = 1$ to $\frac{S_{pop}}{2}$ **do**
parents selection
crossover of parents resulting in new solutions
swap mutation of new solutions
one-parameter mutation of new solutions
add-and-delete-rule mutation of new solutions
solutions evaluation
solutions insertion into offspring population
end for

Figure 7: *Multiobjective optimization algorithm*

$s_{segment}$ [m]	$r_{segment}$ [m]	$\alpha_{segment}$ [radians]	$v_{limit,segment}$ [km/h]
1000	∞	0	50
1000	600	0	70
1000	∞	0	90

Figure 8: *Test route*

3 ALGORITHM TESTING AND RESULTS

The presented algorithm has been tested on one route defined as shown in Figure 8. To test the system, the vehicle parameters have been defined as shown in Figure 9 and functions $f_{T_{\max}}(n)$ and $f_c(T,n)$ as shown in Figure 10 [4]. Furthermore, the algorithm parameters have been defined as shown in Figure 11. The results, i.e., the found front of nondominated strategies, are presented in Figure 12.

Parameter	Value
$I_{vehicle,min}$	1
$I_{vehicle,max}$	5
η	0.96
μ	0.9
ρ	1.225 kg/m^3
c_r	0.01
c_s	0.7
r_{wheel}	0.31725 m
m	1335 kg
A_x	2.10 m^2
c_x	0.31
$i_{gearRatios}$	[3.45, 1.94, 1.37, 1.03, 0.85]
$i_{differentialRatio}$	4.25
$n_{min,upperGears}$	800 /min
$n_{vehicle,max}$	6400 /min

Figure 9: *Test* v*ehicle parameters*

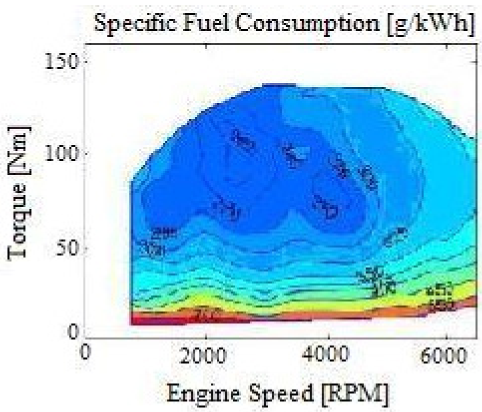


Figure 10: *Test vehicle functions* $f_{T_{\max}}(n)$ *and* $f_c(T,n)$

4 CONCLUSION

This paper presents a system that includes a vehicle simulator, simulation of vehicle driving and a multiobjective genetic algorithm based on NSGA-II, implemented to search for driving strategies that minimize two conflicting objectives: the consumed time and fuel. This system has been preliminarily tested and the results already confirm the effectiveness of the system.

The future work will include the improvements of the strategy structure, the tests on additional, more complex routes, the adjustment of the NSGA-II genetic operators, and possible extensions towards the predictive/adaptive control as an alternative approach for finding nondominated vehicle driving strategies.

Parameter	Value
S_{pop}	100
S_{tour}	2
N_{gen}	1000
N_{cr}	3
P_{cr}	0.7
P_{swMut}	0.1
P_{opMut}	0.1
P_{adrMut}	0.1

Figure 11: *Algorithm parameters*

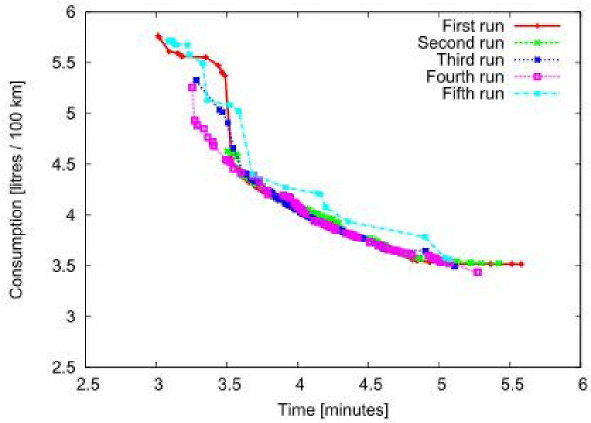


Figure 12: *Nondominated solutions found in five experimental runs*

References

[1] A. E. Eiben, J. E. Smith. *Introduction to Evolutionary Computing*. Springer. Berlin. 2003.

[2] E. Dovgan, M. Javorski, B. Filipič. *A multiobjective genetic algorithm for discovering car driving strategies.* Technical report IJS-DP 10402. Ljubljana. 2010.

[3] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*. Vol. 6. No. 2. pp. 182–197. 2002.

[4] T. Randolph. Waste heat regeneration systems for internal combustion engines. *http:// www. heat2 power.net/ downloads/ GPC2007/ 20070618 _heat2 power_ GPC_WHR_presentation.pdf*. 2007.

CONFIDENCE STAND-ALONE SW SYSTEMS

Matjaž Gams, Erik Dovgan, Božidara Cvetković, Violeta Mirchevska, Boštjan Kaluža, Mitja Luštrek
Department of Intelligent Systems, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: matjaz.gams@ijs.si

ABSTRACT

CONFIDENCE - Ubiquitous Care System to Support Independent Living - is one of the most interesting FP7 projects for The department of intelligent systems at JSI. In this paper we present WP3, the major SW component of the system that is also implemented as two stand-alone versions of the system. In this paper, the two systems are presented from the user viewpoint based on five manuals.

1 INTRODUCTION

The population in the developed countries and particularly in Europe is aging due to the increase in life expectancy and decrease in birth rate. As a consequence of this process, the number of elderly will exceed the society's capacity for taking care of them. Therefore, technical solutions are being sought by the EU other countries to ensure that the elderly can live longer independently with minimal support of the working-age population. Many of these efforts belong to the area of ambient assisted living (AAL), whose objective is to make daily life easier and safer by placing unobtrusive smart devices and services into the environment. This is also the goal of the European FP7 project CONFIDENCE – Ubiquitous Care System to Support Independent Living FP7-ICT-214986 [1].

The CONFIDENCE system will unobtrusively monitor the user in order to detect health problems, such as falls, unusual behavior and some diseases. Literature suggests that the fear of falling or being left unattended in case of trouble can lead the elderly to the refusal of mobility, isolation, decline in the ability to perform daily activities and eventually admission to institution care. The target group are the elderly aged over 65 who live on their own, and do not have serious mobility problems. With the CONFIDENCE system, such people will gain confidence and security and will have a better quality of life and a longer active participation in the society. The beneficiaries will be not only the elderly, but also their families and caregivers, since the burden on them will be reduced. In practical terms, the goal of the project is to extend the independent life of the elderly by several years, which will also save the cost of institutional care.

The system is based on tags that are attached to the user's body and sensors in the top corners of the room [2,3]. The system learns and adapts to each particular user [5,6,7].

Here we present WP3 of the EU FP7 project Confidence [1]. The main objectives of this work package are the development of SW subsystems that will

a) reconstruct the user's body in the environment
b) interpret the body posture within the environment
c) ring an alarm when hazardous situations are detected
d) issue a warning when unusual behavior is detected.

The subsystems originally have to be integrated into the complete system that must be able to gather user position and acceleration from a real-time position and acceleration systems and has to send alarm messages about the detected hazardous situations to an independent portable device. This device is used to communicate with the user and is able to decide whether a hospital or a specialized care-giving institution has to be noticed.

The described system has already been developed. It uses the Ubisense, a real-time localization system [2], and an accelerometer system developed at the Fraunhofer Society [3] in order to get the positions and accelerations of the user from dedicated body tags and sends the alarm messages to a simulated portable device. The official version of the system is already being intensively tested and presented in public.

In addition to the official Confidence version, two additional stand-alone versions were produced:

a) technical version
b) simple version.

Both share practically the same algorithms, and mainly differ in the user communication – the technical version provides information about the system at a level, appropriate for technically educated while the simple version provides only simple communication. Both versions also need only the PC that is already part of the Confidence system, i.e. the screen and the keyboard. Both versions also do not call or communicate in any other way than through the PC.

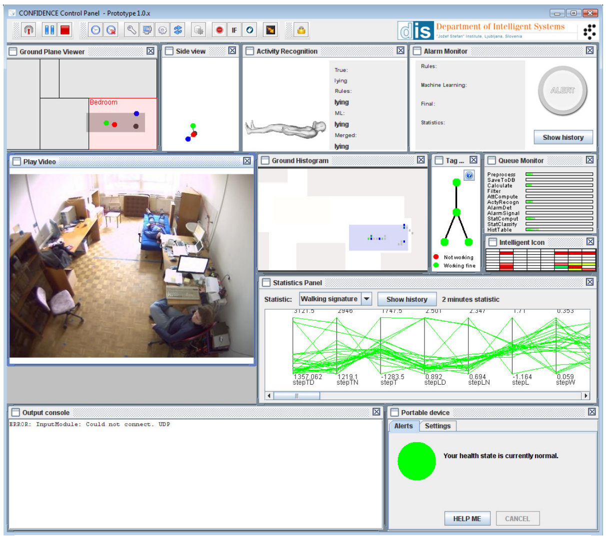


Figure 1: *The technical screen*

2 TECHNICAL VERSION

All three versions of the system: the official one, the technical and the simple one use the same algorithms, although some parameters are usually set differently.

The presented system recognizes hazardous situations from user's movement and reports them to the user/caregiver. In order to do that, several modules have been developed and integrated in the final system. In the following subsections we firstly present the architecture of the whole system. Secondly, we present each module/method that has been developed. Besides, we present the interface that was developed as an extension of the portable device, namely control panel, for the advanced users, system developers and for presentation purposes.

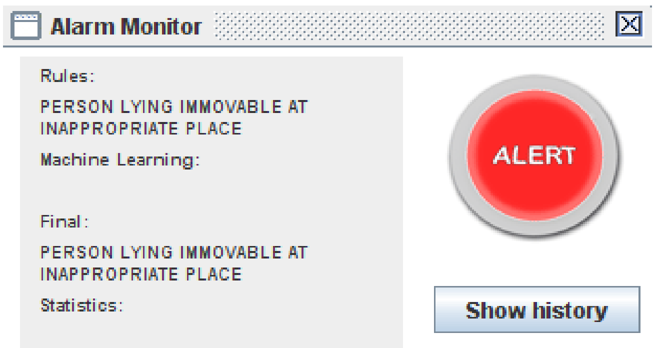


Figure 2: *Alarm monitor showing an alarm*

The system has been developed as a set of independent modules/threads. They are organized as a pipeline where a module gathers the data from the previous module(s), processes them and sends them to the next module(s) in the pipeline. The main modules are the reconstruction modules (consisting of posture modules) and interpretation modules (consisting of interpretation and prevention modules). In addition, the communication modules were also implemented that communicate with a localization system and portable device, and that show the system status in details on the computer screen.

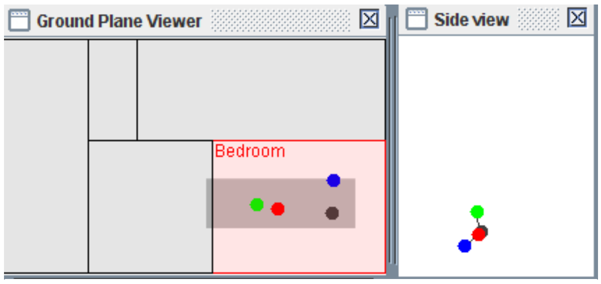


Figure 3: *Ground plane viewer and side view*

In this paper we present the system through communication with the user, mostly through the screen, presented in Figure 1. This screen can be observed in the three versions of the system through remote access, but on the physical PC screen only in the technical version of the system.

The screen consists of the top menu bar and of several windows. In the middle left is the video view of the room where the HW is installed. In the top right is the alarm window as shown in Figure 2. It starts blinking in red and is accompanied by sound alarm if the system observes an alarm. In addition, the bottom right window also reports an alarm since this is a simulation of the portable device. An alarm can be refuted or cancelled by clicking the ignore button on the portable device simulator.

Left alone, the alarm goes on until refuted or the alarm situation ends (e.g. user walking again). The list of previous alarms is stored in the left down windows and can be reviewed and thus reanalyzed.

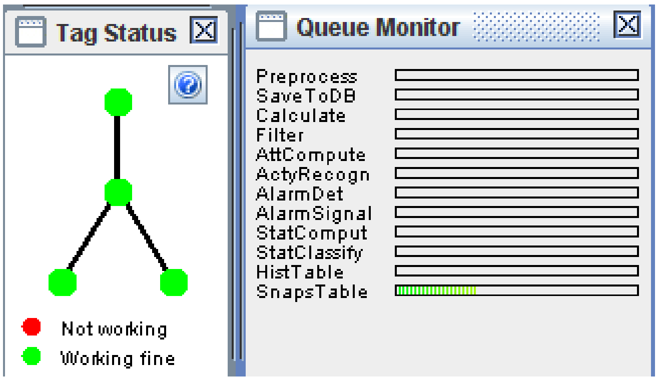


Figure 4: *Tag status and queue monitor*

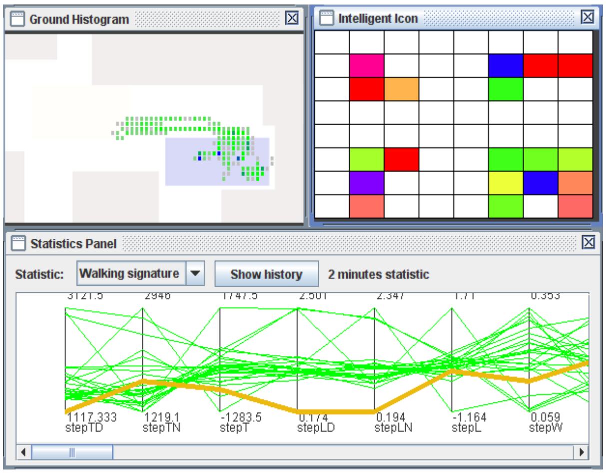


Figure5: *A warning showed in control panel with the yellow line*

The two windows at the top left are presented in Figure 3. One represents the top view of the room, divided into subspaces as set during the initialization, and the position of the bed, in grey. The four tags attached to the body are presented by four colors: Green for neck, red for belt and the two remaining for ankles. The right window in Figure 3 represents the side view of the user. It should be noted that there is currently a time delay of a couple of seconds between the video and the position in Figure 3. The alarm usually triggers 10-20s after lying at an undesired position. At the bed, only macro alarms are triggered.

The next important window is the right one middle as presented in Figure 4. It shows the status of the four legs. If a tag is not attach or not functioning, e.g. it has fallen off, the battery is empty etc., the system reports this and asks for user action (see user manual). Among possibilities is also having only three tags, but the one at the neck is obligatory. In addition to the tag status, the queue status of the system processes is reported.

Figure 6: *Normal*

Figure 7: *Alarm*

Figure 5, lower window, shows an example of a warning because of unusual behavior. This screen usually consists of green lines denoting normal movement. An orange line shows unusual movement. Each horizontal line corresponds to a particular attribute, e.g. speed of walk or step length. When averaged over several minutes these attributes resemble normal movement or not – in that case the orange line is displayed. It enables visualizing what was the cause for warning.

The top two windows show macro-level movement. The left top screen shows density of user positions in the room. The top right screen shown the graphical representation of the posture and position of the user. Both top screens by themselves enable visualization of user behavior that enables observing deviances in behavior thus enabling the care personnel to propose appropriate modifications and care actions.

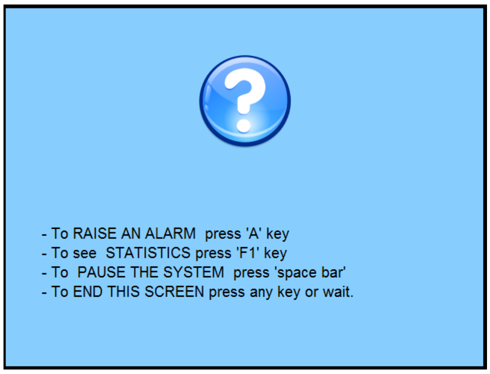


Figure 8: *Help screen*

3 SIMPLE VERSION

There are six simplified screens for elderly or non-technical users, presented in Figures 6 to 10. They report: normal status (6), an alarm (7), and a help screen (8). The alarm can be raised by pressing the A key and cancelled by pressing any key. The help screen is shown if any key but A, F1 or space bar is pressed.

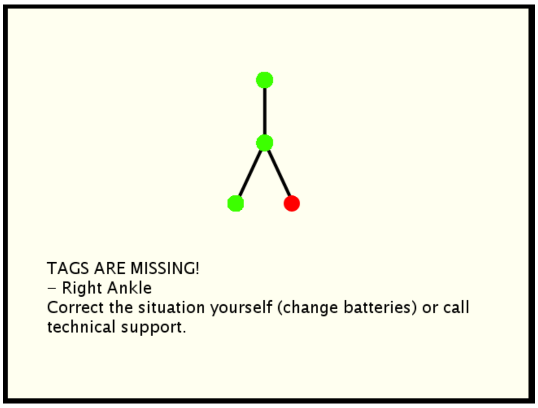


Figure 9: *Tag missing*

Figures 9, 10 and 11 show screens in cases when one tag is missing, when the system was paused (the user wants to freeze the system because of any reasons, e.g. going to a toilet), and the report in case of unusual behavior. These screens are so simple and understandable that they do not demand technical knowledge and as such appropriate for any elderly.

4 MANUALS

Several manuals have been produced describing the presented system and its functionalities. The first one it the System manual that describes the whole system and all of its modules in detail. In addition, it describes the installation of the system and the tag placement.

The second manual is the Recording instructions. It presents how a user can record the data for testing purposes step-by-step. In addition, the preferred scenarios for testing are also described.

The third and forth manual are the User manuals that describe different versions of the system, e.g. the portable device and the possible interaction with it. They show all the possible screens and messages and the keys that can be used to manipulate with the portable device.

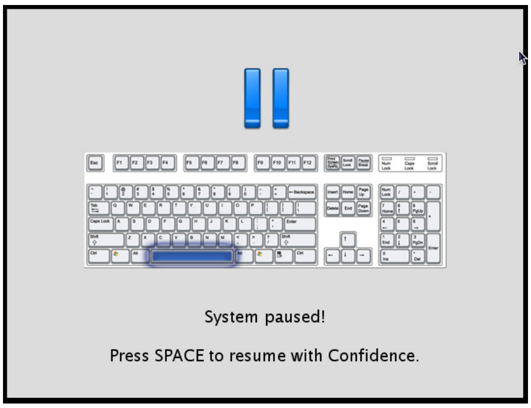


Figure 10: *System paused*

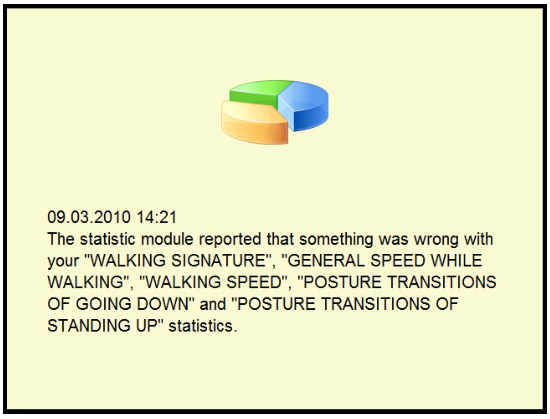


Figure 11: *Unusual behavior*

The last manual is the Init wizard. It describes the required initialization of the system when it is used for the first time. It also describes which actions have to be recorded, possible errors during the initialization, and which data must be inserted by the user.

5 CONCLUSION AND FUTURE WORK

This paper presents the two stand-alone versions of the Confidence system as a result of the EU FP7 project Confidence. It describes the screens as the most important part of the user communication. The tests show that the system is reliable and is ready to be tested in the real world.

References

[1] Confidence. http://www.confidence-eu.org.
[2] P. Steggles, S. Gschwind. Ubisense – a smart space platform. Technical report. Ubisense. 2005.
[3] Fraunhofer. http://www.fraunhofer.de.
[4] M. Lustrek, B. Kaluza. Fall detection and activity recognition with machine learning. Informatica 33(2). 2009. pp. 197-204.
[5] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. of the IEEE 77 (2). pp. 257-286.
[6] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers. 1993..
[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander. LOF: identifying density-based local outliers. Proc. of the 2000 ACM SIGMOD international conference on Management of data. 2000. pp. 93-104.

UNIVERZALNI VMESNIK INTELIGENTNEGA DOMA (UVID)

Jana Krivec, Matej Ožek, Robert Blatnik, Matjaž Gams
Department of Intelligent Systems
Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
matjaz.gams@ijs.si, jana.krivec@ijs.si

Povzetek

V vse kompleksnejšem svetu elektronskih in drugih naprav se povečuje potreba po enostavnem in funkcionalnem upravljanju le teh. V prispevku predstavljamo univerzalni uporabniški vmesnik za enostavnejše in učinkovitejše upravljanje poljubne naprave v domačem večpredstavnem središču in z njim povezanih hišnih naprav v inteligentnem domu, kar smo dosegli z uporabo komunikacije v naravnem jeziku, s prilagajanjem uporabniku ter z lokalno in daljinsko dostopnostjo.

1 UVOD

Hiša danes ni več le prostor, ki nas varuje pred zunanjimi vplivi, ampak nam zagotavlja tudi udobje. Slednje nam omogoča nabor tudi po več deset modernih naprav, začenši s televizijo, telefonom, hladilnikom ipd.. Naprave postajajo vse bolj kompleksne in jih je marsikdaj naporno upravljati, še posebej starejšim ljudem. V pomoč čim lažjemu upravljanju s hišnimi napravami in njihovemu čim boljšemu izkoristku se je razvila posebna aplikativno-znanstvena veja informacijske tehnologije imenovana Ambientalna inteligenca (AmI) in znotraj nje AAL (Ambient Assisted Living). Kot pravijo Phillipsovi razvijalci in raziskovalci na Massachusetts Institute of Technology (MIT), je AmI poroka med tehnologijami vseprisotnega računalništva in socialnih uporabniških vmesnikov. Inteligentno ambientalno okolje združuje konstantno (a nemotečo) navzočnost, inteligentno zavedanje in naravno interakcijo [7]. Vključuje metode, (elektronske) sisteme, naprave in podporo, ki omogočajo lažje, varnejše, kvalitetnejše in cenejše vsakodnevno življenje. Tehnologije, ki jih uporabljajo so prilagojene posameznemu uporabniku (njegovim potrebam in zmožnostim) in integrirane v njegovo osebno okolje. Trenutno v svetu poteka vrsta projektov ob podpori velikih korporacij in državnih skladov, ki v tem vidijo prihodnost in se problematike lotevajo na različne načine [2, 3, 5, 6]. Philips, na primer, ima poseben laboratorij, HomeLab, ki služi za razvoj in testiranje hišne AmI tehnologije. Njihov cilj je razviti "zaznavno hišno okolje", ki bo sposobno identificirati okolje, v njem locirati ljudi, določiti njihove namene in jim zagotoviti optimalno storitev [11].

Napredek na tem področju smo poskušali ustvariti tudi s projektom UVID (univerzalni vmesnik za inteligentni dom). Kot odgovor na kompleksnost upravljanja domačega zabavnega središča smo v okviru projekta razvili prototip integrirane inteligentne e-storitve za upravljanje zabavne elektronike in inteligentnega doma, predvsem smo se osredotočili na televizor in z njim povezane naprave.

Televizor je že mnogo let središče domače zabave. Z leti njegova vloga kvečjemu narašča, saj je na voljo vse več vsebin, ki jih je na njem moč prikazovati (klasična televizija, video na zahtevo) in z najrazličnejših nosilcev (pretočni video, fotografije, računalniške igre). Poleg tega lahko televizor služi tudi kot uporabniški vmesnik za predvajanje glasbe, uporabo interneta in drugih storitev in aplikacij v inteligentnem domu. Vse te vsebine kajpak zahtevajo tudi množico različnih naprav. Poleg televizorja so to televizijski vmesnik (angl. set-top box), satelitski sprejemnik, videorekorder, računalnik, ki je lahko na priklopljen neposredno ali prek večpredstavnega podaljška (angl. multimedia extender), predvajalniki DVDjev in blu-rayev. V našem primeru smo TV opremili s programsko in strojno opremo, ki omogoča prijazno uporabo, prilagajanje uporabniku in neodvisnost od lokacije.

V nadaljevanju predstavljamo integriran sistem ter posamezne module in njihovo delovanje:

1) Naravna interakcija: Sistem za komunikacijo v naravnem jeziku in modul za sintezo govora
2) Modul za prilagajanje uporabniku
3) Funkcionalna in enostavna povezava med uporabnikom in strojnim nivojem: Napredni televizijski vmesnik (Net-top box) in dlančnik
4) Integracija sistema

2 SISTEM ZA NARAVNO INTERAKCIJO

Naravna komunikacija uporabnika z napravami je ena osrednjih teženj in hkrati problem ambientalne inteligence. Z njo se ukvarjajo raziskovalci po vsem svetu. Eden takšnih sistemov so razvili v Fraunhoferju: SUI-A dialog demo [9].

2.1 Komunikacija v naravnem jeziku

V okviru projekta smo razvili storitve za interakcijo z uporabnikom, ki na osnovi podanega pisnega besedila v naravnem jeziku izvršijo ukaz za upravljanje televizijskega vmesnika ali storitev inteligentnega doma. Sistem je sestavljen iz treh modulov:

- Pogovor, pridobivanje informacij s spleta in pomoč uporabniku (npr. Kakšno je danes vreme?; Povej številko Siolovega servisa.; Kaj je na sporedu danes?; Kdaj bo danes kakšen šport?; Kako naročim video na zahtevo? ipd.)

- Upravljanje TV/računalnika/radia – Spreminjanje naravnega besedila v strojne ukaze (Daj bolj na glas, Preklopi na pop…).

- Prikaz: Odgovor sistem prikaže na dlančniku ali TV oziroma izvede ukaz.

Za konstrukcijo programa za komuniciranje v naravnem jeziku se večinoma uporabljata dva povsem različna pristopa. Prvi bazira na programu za analizo odgovora, ki vsebuje mnogo različnih metod in baz podatkov, da čim bolj obdela vprašanje, in izdela najprimernejši odgovor. Drug pristop gradi v nasprotno smer. Izbira le tiste funkcije, ki so nujno potrebne. Tako dobimo majhen program, uspešnost katerega pa je odvisna od dobro zgrajene baze odgovorov. Oba pristopa imata kvalitete in napake, zato smo v naš sistem implementirali oba. Gradili smo na dveh obstoječih sistemih:
- Sistemu Klepec, ki bazira na projektu Kolos. Gre za vzorčno voden sistem, katerega osrednji del sestavlja pomenski stavčni analizator. Modul omogoča pretvorbo vhodnega vprašanja v naravnem jeziku v ustrezen odgovor. Pri nekaterih programih je lahko izhod ustrezen logičen izraz, pri drugih kar ustrezen odgovor v naravnem jeziku.
- Sistemu Virtualni svetovalec, ki je rezultat konstrukcije po drugem pristopu. Gre za majhen program, ki na podlagi vprašanja primerno izbere v naprej pripravljen odgovor v bazi. Odgovori so lahko generirani tudi dinamično, ko je v naprej pripravljena le struktura odgovora, natančna vsebina pa se ustvari sproti iz interneta ali iz različnih baz podatkov.

Storitev UVID uporablja kombinacijo obeh odgovorov na osnovi ocene zanesljivosti, ki jo poleg odgovora poda vsak izmed podsistemov. Vprašanje/ukaz se vnese preko tipkovnice ali dlančnika. Kot »string« se pošlje v obdelavo na strežnik. Na strežniku se vprašanje obdela in najde odziv. Za odgovor se preveri lokalna baza, baza na strežniku ponudnika storitev in internet. Odgovor je lahko »string« ali ukaz za net-top box (NTB). Izhod je odvisen od vhoda, tako imamo kombinacije vhod-izhod: glas-glas, glas-TV (NTB), dlančnik-dlančnik, dlančnik-TV (NTB), tipkovnica-TV (NTB). Vsa komunikacija med komponentami poteka preko html protokola in z uporabo XML formata sporočil. Primer:

Zahtevek:	Odgovor:
http://micka.amebis.si/Micka/Ask.asp?question=Daj+na+POP	Kanal preklopljen na POP TV.@@tv:command=SetChannel(POPTV)##5

Pri tem je za @@ ukaz na NTB, za ## pa ocena zanesljivosti odgovora.

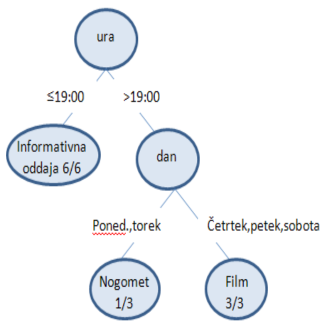
2.2 Modul za sintezo govora

Pripravili smo novo verzijo Govorca [4], ki ne naglašuje več s pomočjo za to ustreznega algoritma, ampak s pomočjo v bazi zapisanih naglasov za posamezne besedne oblike. Poleg tega smo v okviru projekta razvili prilagojeno spletno storitev, ki s pomočjo novega Govorca omogoča posredovanje govornega (zvočnega) zapisa ostalim odjemalcem v sistemu UVID, ob tem da besedilo pretvori v zvočni posnetek MP3. Na ta način lahko uporabnik svoj ukaz sistemu sporoči tudi z govorom.

3 PRILAGAJANJE UPORABNIKU

Prilagajanje uporabniku predstavlja inteligentni del sistema. Omogoča, da sistem analizira kontekst, se prilagaja ljudem, ki živijo v njem, se uči njihovega obnašanja in navad, na osnovi katerih potem tudi ustrezno reagira. V sistemu UVID imamo tri vrste funkcij za prilagajanje uporabniku, za katere smatramo, da sta najbolj uporabni:
Prilagajanje na željo uporabnika. To je primer enostavnega prilagajanja, kjer posameznik določen kompleksen ukaz, ki ga denimo večkrat uporablja zamenja z enostavnejšim ukazom. Pomembno je, da lahko uporabnik sporoči njegovo željo sistemu v naravnem jeziku in ga le-ta nato po korakih vodi do želene nastavitve.
Enostavno avtomatsko prilagajanje. Primer za to vrsto je: ko sistem zazna IP uporabnika, prilagodi nastavitve televizije (glasnost, slika…) dotičnemu uporabniku.
Kompleksno avtomatsko prilagajanje: Dosedanji sistemi ne nudijo avtomatskega prilagajanja uporabniku v pravem pomenu besede, to je da se prilagodijo vsakemu uporabniku posebej, ampak predlagajo splošno najboljše stvari (npr. filme z najvišjo oceno po IMDB). Avtomatsko prilagajanje je v našem primeru osnovano na profilu posameznega uporabnika, zgrajenega na njegovih specifičnih navadah. Iz uporabnikovega obnašanja se program uči tipične vzorce in oblikuje njegov profil, na osnovi katerega se izvajajo priporočila. Profil uporabnika se zgradi s pomočjo tehnik strojnega učenja, ki odkrivajo zakonitosti v podatkih.
Sistem najprej določen čas zbira podatke o ukazih, ki jih na televiziji izvede posamezen uporabnik. Podatki so naslednji: mesec, dan v tednu, ura, program, oddaja, ki je v tem času na programu (EPG), informacija o vrsti trenutne oddaje (film, informativna oddaja,…), informacija o izbranem filmu,…. Iz podatkov predstavljenih v obliki tabele (glej Sliko 1) zgradimo odločitveno drevo, ki predstavlja model, katerega lahko uporabimo za klasifikacijo novih primerov. Začnemo pri korenu in potujemo po ustreznih vejah do lista. Odgovor je razred ciljnega lista, v našem primeru vrsta oddaje.

MESEC	DAN V TEDNU	ČAS	PROGRAM	VRSTA_ODDAJE
februar	pon	1140	slo1	informativna_oddaja
februar	pon	1200	nat_geo	dokumentarec
februar	tor	1140	slo1	informativna_oddaja
februar	tor	1195	slo1	razvedrilna_oddaja
februar	sre	1140	slo1	informativna_oddaja
februar	sre	1195	slo2	nogomet
februar	čet	1140	slo1	informativna_oddaja
februar	čet	1195	cinemax	film
februar	pet	1140	slo1	informativna_oddaja
februar	pet	1195	pop	film
februar	sob	1140	slo1	informativna_oddaja
februar	sob	1200	TV1000	film
februar	sob	1335	Kanal_A	nanizanka
…				



Slika 1: Primer simulacije 1. dela delnih vhodnih podatkov za grajenje uporabnikovega profila ter primer simuliranega zgrajenega drevesa, na osnovi podatkov.

V primeru na Sliki 1 lahko izpeljemo pravilo:

»Če je ura manjša ali enaka 19:00, potem uporabnik navadno gleda informativno oddajo« ali »Če je ura večja kot 19:00 in dan četrtek, petek, ali sobota, potem uporabnik navadno gleda film, sicer pa nogomet«. Številke v listih nam prikazujejo število pravilno klasificiranih primerov in število vseh primerov, ki jih pravilo pokriva. Druga možnost, ki bi jo veljalo preučiti je uporaba »case-based reasoninga« za potrebe prilagajanja. Ta metoda primerja trenutno akcijo neke osebe z enim, a najbolj podobnim dogodkom iz njene preteklosti [8].

Ko imamo zgrajen osnovni profil uporabnika moramo izdelati metodo, s katero bomo izbrali primerna priporočila. V ta namen je potrebno profil uporabnika nadgraditi tako da zberemo podatke, o oddajah, ki jih uporabnik rad gleda. Za primer vzemimo, da uporabnik prižge televizijo ob uri in dnevu, ko navadno gleda film. Na različnih programih je v tem času veliko filmov. Katerega bomo uporabniku priporočili se odločimo iz njegove zgodovine do sedaj gledanih filmov. Da ugotovimo, kakšni filmi so uporabniku všeč beležimo naslednje podatke: naslov, režiser, igralci, žanr filma, letnica izdelave, IMDB ocena, ključne besede (atributi). Tudi v tem primeru uporabimo tehnike strojnega učenja, in sicer asociacijska pravila. Asociacijska pravila omogočajo identifikacijo razmerij, povezav, asociacij med vrednostmi oz. atributi v opazovani množici primerov. Tako ugotovimo, kateri izmed zgoraj navedenih atributov (oz. njihova kombinacija) se pri uporabniku pogosto pojavljajo. Ko imamo nabor za uporabnika značilnih atributov (npr. rad gleda filme, v katerih igra Brad Pitt ter drame), pregledamo, kateri izmed filmov, ki je trenutno na sporedu se najbolj ujema z naborom željenih lastnosti. Sistem se nauči, kakšne oddaje gleda uporabnik ob katerih urah, dnevih, ipd.. Na osnovi tega mu lahko ponuja naslednje praktične rešitve: seznam njemu ljubih oddaj, ki jih lahko posreduje ob vklopu TV ali med gledanjem kakšne druge oddaje. Sistem lahko uporabniku predloge posreduje tudi na uporabnikovo eksplicitno zahtevo, katero posreduje preko virtualnega vmesnika (Npr."Kje je kakšen dober film?"). Program deluje kot spletna storitev.

Slika 1:Simulacija avtomatskega priporočila med gledanjem oddaje.

4 FUNKCIONALNA IN ENOSTAVNA POVEZAVA MED UPORABNIKOM IN STROJNIM NIVOJEM

V AmI morajo naprave med seboj komunicirati, če želijo uresničiti ves svoj potencial. Velik napredek v tej smeri se je zgodil po razvoju brezžične tehnologije (WLAN, RFID, Bluetooth, ipd.). Še vedno pa to ni dovolj. Mrežna integracija mora biti enostavna, brez potrebe po preveliki pretvorbi podatkov. Družbe, ki se ukvarjajo z razvojem AmI se v veliki meri posvečajo zagotavljanju komunikacijskih standardov, ki zagotavljajo združljivost različnih objektov povezanih v mrežo [8]. Na US National Institute of Standards and Technology v ta namen razvijajo odprtokodni računalniški standard, imenovan Smart Flow [10]

4.1 Napredni televizijski vmesnik (net-top box)

Pomemben dosežen projektni cilj je implementacija Net-Top Box (NTB) naprave, ki predstavlja osrednjo komunikacijsko in predstavitveno vozlišče povezanega doma. Komunicira lahko proti prenosnemu osebnemu krmilniku in aplikacijam za vodenje v naravnem jeziku. Iniciativa za razvoj naprednega televizijskega vmesnika prihaja s trga »odprtih multimedijskih terminalov« z nazivom »Premium STB« ali »Net-top-Box« (NTB). Trend razvoja televizijskih vmesnikov gre v smeri podpore odprtih aplikativnih platform; z namestitvijo ustreznega programskega ogrodja je neodvisnim razvijalcem omogočen razvoj aplikacij s področja multimedije (dostop do vsebin na lokalnih, hišnih strežnikih, vsebin na Web), inteligentnega doma (hišna avtomatizacija) in vsebin komercialnih ponudnikov storitev (telefonija, sporočilni sistemi ipd.). Seveda osnovna funkcionalnost televizijskega vmesnika, t.j. prikaz multimedijskih vsebin iz različnih virov na TV zaslonu, ostaja osnovna funkcionalnost. Za potrebe projekta UVID smo uporabili napredni televizijski vmesnik (NTB) na PC arhitekturi (x86, Linux OS). NTB omogoča dostop do IPTV in VoD (video na zahtevo) storitev; do lokalnih multimedijskih vsebin, do internetnih (Web) vsebin (npr. elektronski vodič po programih (EPG), YouTube, Twitter,…) Odprt je za razvoj in namestitev aplikacij neodvisnih razvijalcev. Implementacija aplikativne logike sloni na intenzivni uporabi XBMC grafičnih knjižnic, skriptnih jezikov (Python) in XML ter C/C++ kot osnovni tehnologiji implementacije odprtokodnega XBMC.

4.2 Dlančnik

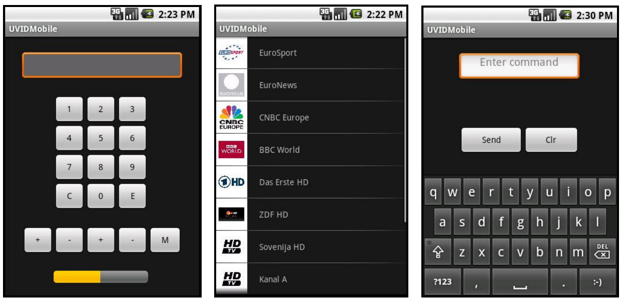
Uporaba informacijsko- komunikacijskih tehnologij v hišnih sistemih, gospodinjskih napravah in domači zabavni elektroniki prinaša poleg neslutenih možnosti razvoja novih funkcionalnosti in storitev informacijske družbe tudi nove izzive v razvoju uporabniških vmesnikov. Klasični daljinski upravljalnik ne zadošča več za enostavno in funkcionalno upravljanje naprav in sistemov. Ena izmed vodilnih nalog AmI je razvoj dobrega vmesnika med uporabnikom in strojno opremo v ozadju, ki poveča efektivnost uporabnika. V projektu UVID mora vmesnik komunicirati med uporabnikom in centralnim domačim aplikacijskim strežnikom, ki teče NTB, domačem PC računalniku ali celo pri ponudniku storitve. Komunikacija teče po IP omrežju, z uporabo http protokola in z uporabo XML formata sporočil. Nove tehnologije omogočajo drastično poenostavitev in intuitivno uporabo naprav. Razvili smo delujoč prototip vmesnika, ki podpira približno polovico standardnih ukazov klasičnih TV daljincev in je uporabniku prijazen. Podprti so naslednji ukazi (slika2):

● neposreden izbor programa po številki (numerična tipkovnica 0-9)

povečanje/zmanjšanje glasnosti

● grafični pregled programov in izbor želenega programa s pritiskom na ikono. Spisek ekranov se sproti osvežuje s podatki iz NTB in tako vedno prikazuje dejansko stanje nastavitev programske sheme.

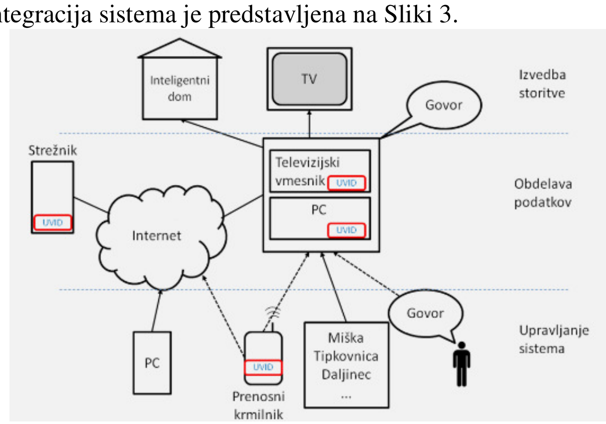
●vnos vprašanja in prikaz besedilnega ali glasovnega odgovora



Slika 2. Tri funkcionalnosti krmilnika: izbor programa po številki, izbor programa s spiska in vnos besedilnega ukaza in prikaz odgovora.

5 INTEGRACIJA SISTEMA

Integracija sistema je predstavljena na Sliki 3.



Slika 3. Shema delovanja sistema UVID.

Uporabnik preko TV daljinca, tipkovnice televizijskega vmesnika, dlančnika z ustrezno programsko opremo vnaša ukaze za dostop do želenih multimedijskih vsebin. Procesiranje uporabnikovih ukazov izvaja aplikacija, ki se lahko nahaja pri ponudniku storitve ali na lokalnem, hišnem PCju, ki prevzame vlogo multimedijskega strežnika. Vlogo multimedijskega strežnika lahko prevzame tudi napredni televizijski vmesnik. Vloga televizijskega vmesnika je procesirati odločitev aplikativne logike s ciljem prikaza želene multimedijske vsebine na TV zaslonu in omogočiti programsko ogrodje za razvoj in namestitev multimedijskih aplikacij.

6 ZAKLJUČEK

S sistemom UVID nadomeščamo običajne daljince, ki omejujejo upravljanje zahtevnejših funkcij televizorja, radia, raznih predvajalnikov, osebnega računalnika in drugih hišnih naprav pri čemer nekaterih ukazov sploh ne podpirajo. Razvili smo enostaven in funkcionalen osebni prenosni krmilnik, napredni televizijski vmesnik, modul za sintezo govora in izdelavo profila uporabnika ter sistem za komunikacijo v naravnem jeziku. Z njim lahko uporabnik poišče priljubljen TV program z ukazom v naravnem jeziku. Z vprašanji v naravnem jeziku lahko preko računalnika dostopamo do različnih informacij. Za pomoč pri uporabi storitev ali tehničnih težavah nam ni potrebno klicati servisa , saj nam sistem ponudi ustrezen odgovor. Poleg tega UVID na osnovi zgrajenega uporabniškega modela priporoča primerne vsebine. Z uporabo sistema UVID je torej upravljanje s televizorjem in njegovimi dodatnimi napravami enostavnejše in prijetnejše. Sistem UVID predstavlja platformo, na kateri je mogoče razvijati nove funkcionalnosti, kot je npr. dostop do socialnih omrežij in integracija hišnih naprav v pravi inteligentni dom.

ZAHVALA

Sistem UVID je nastal ob podpori Ministrstva za visoko šolstvo, znanost in tehnologijo v okviru Javnega razpisa za spodbujanje raziskovalno razvojnih projektov razvoja e-vsebin in e-storitev v letih 2009 in 2010. Operacijo delno sofinancira Evropska unija, in sicer iz Evropskega sklada za regionalni razvoj.Zahvaljujemo se tudi Tomažu Kompara in Janu Marinčku za pomoč pri projektu.

7 LITERATURA IN APLIKACIJE

[1] Ailisto, H., Kotila, A. and StrÃ¶mmer, E. (2003) "Ubicom applications and technologies". Presentation slides from *ITEA 2003*, Oulu, Finland

[2] CASAS Samrt Home: http://ailab.eecs.wsu.edu/casas/

[3] Confidence: http://dis.ijs.si/confidence/consortium.html

[4]GOVOREC: http://ai.ijs.si/govorec/

[5] e2 Home project: http://www.e2-home.com/Folder_16/trials.asp

[6] Evropski AAL projekti: http://www.aal-deutschland.de/europa/projekte

[7] Gupta, M. (2003) "Ambient Intelligence - unobtrusive technology for the information society". *Pressbox.co.uk*

[8] Kapetanakis, S.; Petridis,M.,Ma,J.; Bacon, L. *(2010).* Providing Explanations for the Intelligent Monitoring of Business Workflows Using Case-Based Reasoning. *In Proceeding of the Workshop on the Explanation aware Computing.* ECAI, Lisbon.

[9] Mukasa, K.S.; Graf,C.(2008). Workshop on Intelligent User Interfaces for Ambient Assisted Living Canary Islands, 13.01.2008

[10] Thibodeau**,** P. (2003) "Pervasive computing gets organized". *Computerworld*, Vol. 37, No. 2, January 13: 32

[11] Van Loenen, E.J. (2003). Towards a Natural Interaction Concept for SUI in AAL,Ambient intelligence: Philips' vision". Presentation at *ITEA 2003*, Oulu, Finland

SPROTNA AVTOMATIČNA DETEKCIJA IN KLASIFIKACIJA IZREDNIH DOGODKOV V ČASOVNIH VRSTAH

Damjan Kužnar, Domen Marinčič
Odsek za inteligentne sisteme
Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana, Slovenija
e-mail: damjan.kuznar@ijs.si

POVZETEK

Pričujoči prispevek predstavlja metodi za sprotno (ang. *real-time*) detekcijo in klasifikacijo izrednih dogodkov (v kritične in nekritične) v časovnih vrstah. Metoda za detekcijo izrednega dogodka sloni na uporabi diskretne valjčne transformacije (ang. *Disctrete Wavelet Transform*, v nadaljevanju DWT) originalne časovne vrste. Izsek časovne vrste za zaznan izredni dogodek je nato transformiran v n-gram predstavitev, ki je primerna za atributno strojno učenje v koraku za klasifikacijo, ki sloni na uporabi metode za gručenje k-Means.

Metoda je bila preizkušena na podatkih, pridobljenih z merjenjem temperature v raznih laboratorijih in komorah. Rezultati kažejo, da je metoda zmožna zanesljivo zaznati in združiti dogodke glede na tip izrednega dogodka.

1 UVOD

Časovne vrste so zelo pogosta vrsta podatkov, ki nastane z zajemanjem vrednosti iz nekega generatorja (npr. merjenje temperature, vlage, itd.) v časovnih intervalih, ki so ponavadi enako dolgi.

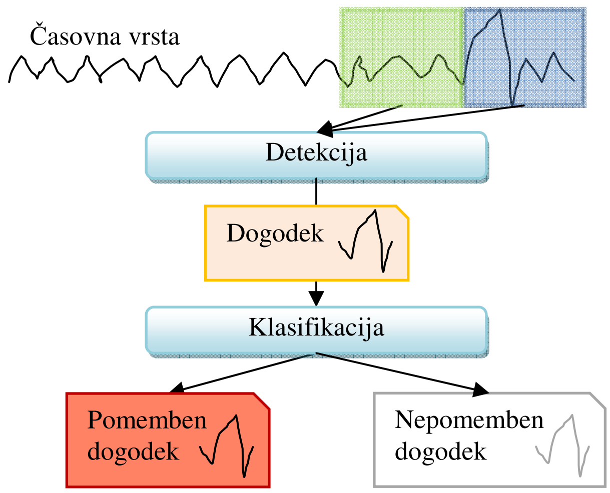
Pri sprotni analizi časovnih vrst nas zanima, kdaj je prišlo do takšnih sprememb, ki naznanijo nastop nekega dogodka (npr. okvara hladilnika in posledično padanje temperature), saj lahko tako preprečimo marsikateri dogodek, ki bi imel neželene posledice, že med samim dogajanjem.

Med izvajanjem sprotne analize pogosto zaznamo tudi dogodke, ki nimajo neželenih posledic oziroma so bili namenoma izvedeni s strani uporabnika (npr. odpiranje vrat hladilnika za kratek čas). V takšnih primerih mora imeti uporabnik možnost, da dogodek označi kot nepomemben, metoda pa imeti zmožnost prilagoditve.

2 SORODNO DELO

V literaturi je mnogo prispevkov, ki se ločeno ukvarjajo ali z detekcijo [1,2] ali klasifikacijo [3,4] časovnih vrst. Avtorji tega prispevka niso uspeli najti sorodnega dela, ki bi neposredno ukvarjalo z detekcijo in klasifikacijo hkrati, kot to prikazuje slika 1.

Prav tako je večina metod, ki se pojavljajo v literaturi, namenjena detekciji izrednih dogodkov šele po koncu zajemanja meritev – ko je celotna časovna vrsta na voljo [5]. To se razlikuje od pristopa, ki je opisan v tem prispevku, kjer se detekcija in klasifikacija neprestano vrši med samim zajemanjem meritev. Torej je potrebno dogodke zaznavati in klasificirati med samim dogajanjem dogodka, saj je v mnogo okoliščinah bistven hiter odziv na dogodek (npr. okvara naprave).



Slika 1: *Sprotno zaznavanje in klasifikacija dogodkov.*

3 DETEKCIJA IZREDNIH DOGODKOV

Detekcija izrednih dogodkov temelji na diskretni valjčni transformaciji (DWT), katere skalogram omogoča ogled časovne in frekvenčne komponente časovne vrste hkrati [6] in je primerna za iskanje nenadnih sprememb v časovnih.

Detekcija se neprestano vrši s primerjavo dveh najnovejših izsekov iz časovne vrste, kot prikazuje slika 1. Za oba izseka, *glavo* in *rep*, izračunamo DWT, katere rezultat je dvodimenzionalna matrika. Nato iz matrik izračunamo razliko po formuli (1).

$$razlika = \sum_{i=1}^{S} \sum_{j=1}^{W} (glava_{i,j} - rep_{i,j})^2 \quad (1),$$

kjer je S stopnja DWT dekompozicije in W je širina časovnega okna.

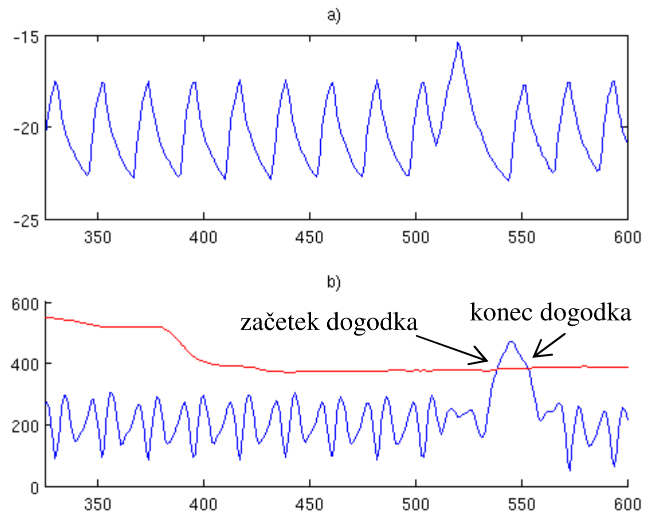
Ker izračunana metrika (1) ni omejena na zaprti interval, je potreben dodaten mehanizem, ki ugotavlja *mejno vrednost*, kdaj je *razlika* med glavo in repom prevelika. To izvedemo z uporabo povprečja in standardne deviacije *razlike* za obdobje, ki je večje od dvakratnika širine časovnega okna. Definicija mejne vrednosti je navedena v formuli (2).

$$\textit{mejna vrednost} = \textit{povprečje} + k * \textit{standardna deviacija},$$

kjer je *k* večkratnik standardne deviacije.

Izredni dogodek v časovni vrsti nastopi, ko je trenutna vrednost *razlike* večja od trenutne *mejne vrednosti* in traja vse dokler vrednost *razlike* ne pade pod vrednost *mejne vrednosti*.

Slika 2 prikazuje primer izračunane *razlike* in *mejne vrednosti* za vsako meritev v časovni vrsti.



Slika 2: *Primer izračuna razlike in mejne vrednosti: a) časovna vrsta, b) moder graf prikazuje razliko, rdeč graf prikazuje mejno vrednost.*

.

4 KLASIFIKACIJA ZAZNANIH DOGODKOV

Na podlagi predhodne detekcije izrednih dogodkov je možno z dovolj veliko natančnostjo iz časovne vrste izrezati dele signala, ki pripadajo izključno enemu dogodku. Ta izsek iz časovne vrste pa nato služi kot osnova za klasifikacijo v pomembne ali nepomembne dogodke. Nepomembni so tisti dogodki, ki jih uporabnik preko mehanizma povratne informacije zavrne oziroma ne želi, da se mu jih prikazuje (npr. ne želi obvestil o odpiranju vrat hladilnika).

Klasifikacija novega dogodka se vrši tako, da dogodke (novega in vse pretekle) najprej z metodo k-Means gručimo, nato pa na podlagi dogodkov, ki so v isti gruči kot nov dogodek določimo ali gre za pomemben ali nepomemben dogodek.

Ker se za klasifikacijo v osnovi uporablja metoda k-Means, ki sprejema podatke v atributni obliki, časovnih vrst ne moremo neposredno uporabiti. Namesto tega skale diskretne valjčne transformacije (ki so tudi časovne vrste) zakodiramo v atributni opis s pomočjo n-gram transformacije [7]. Ta v časovni vrsti poišče krajša zaporedja vrednosti ter jih uporabi kot atribute. Vrednosti teh atributov je nato število pojavitev tega kratkega zaporedja v časovni vrsti. V tem prispevku smo uporabili n-grame do dolžine 3.

Sprotna klasifikacija poteka tako, da ob vsakem zaznanem dogodku izvedemo gručenje trenutno zaznanega dogodka in vseh v preteklosti zaznanih dogodkov. Optimalno število gruč se določi z vrednostjo silhuete (ang. *silhouette*) [8]. Glede na to kateri gruči pripada trenutni dogodek po končanemu gručenju, določimo ali pripada razredu pomembnih ali nepomembnih dogodkov. Pomembnost oziroma razred dogodka se določi po formuli (3).

$$\textit{razred} = \begin{cases} \textit{pomemben} & z < 0.5 \\ \textit{nepomemben} & z \geq 0.5 \end{cases} \quad (3)$$

$$z = \frac{\#\,\textit{zavrnjenih dogodkov v gruči}}{\#\,\textit{vseh dogodkov v gruči}}$$

5 EVALUACIJA

Evaluacija je bila izvedena na časovnih vrstah, ki jih je zajelo podjetje Lotrič d.o.o. (laboratorij za meroslovje) pri merjenju temperature v različnih napravah ter v različnih prostorih. Skupaj je bilo na voljo 16 časovnih vrst, kjer vsaka vsebuje vsaj en izredni dogodek. Časovne vrste je ročno označila (začetek in konec dogodkov) oseba, ki ni strokovnjak na področju meroslovja.

Nastavitve parametrov detekcije so bili naslednji:

- širina časovnega okna (ČO): 16 meritev
- večkratnih ČO za izračun povprečne vrednosti: 15
- večkratnih standardne deviacije: 3

6 REZULTATI

Razpredelnica 1 navaja rezultate za detekcijo.

	Označeni	Neoznačeni
Zaznani	41	6
Nezaznani	15	/

Razpredelnica 1: *Rezultati detekcije dogodkov v časovnih vrstah.*

Stolpec *označeni* prikazuje koliko dogodkov je ročno označenih v časovnih vrstah, vrstici *zaznani* in *nezaznani* pa povedo, koliko od teh označenih dogodkov je bilo tudi zaznanih (ang. *true positive*) oziroma nezaznanih (ang. *false negative*). Stolpec *neoznačeni* navaja število dogodkov, ki jih je naša metoda zaznala, vendar jih strokovnjak ni označil (ang. *false positive*).

Iz teh rezultatov lahko izračunamo natančnost zaznavanja (ang. *precision*) dogodkov, ki je 0.73.

Ker je pravilno gručenje bistveno za točno klasificiranje dogodkov na pomembne in nepomembne, smo merili le kvaliteto dobljenih gruč v primerjavi z označevanjem, ki ga je izvedla oseba, ki ni domenski strokovnjak. Za mero kvalitete je bila uporabljena čistost (ang. *purity*), ki je definirana kot:

$$\textit{čistost} = \frac{1}{N}\sum_k \max_j |w_k \cap c_j|.$$

kjer je w_k iz množice gruč in c_j iz množice označb (razredov). Povprečna izmerjena čistost gručenja dogodkov na testnih primerih je 0.94. To pomeni, da le v 6 % primerov dogodek napačno pripišemo neki gruči, ki sicer vsebuje dogodke drugačne tipa.

7 ZAKLJUČEK

Predstavili smo učinkovito metodo za sprotno (ang. *real time*) detekcijo in klasifikacijo dogodkov v časovnih vrstah. Natančnost zaznave dogodka na testnih časovnih vrstah je 0.73 ter čistost gručenja je 0.94.

Pri nadaljnjem delu bomo poskušali izboljšati zaznavanje dogodkov, kar bo vodilo do še bolj učinkovitega preprečevanja pomembnih izrednih dogodkov.

8 LITERATURA

[1] Valery Guralnik, Jaideep Srivastava: *Event Detection from Time Series Data*, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999

[2] Yoshinobu Kawaharam, Masashi Sugiyama: *Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation*, Proceedings of SIAM International Conference on Data Mining, 2009

[3] Peter Sykacek, Stephen Roberts: *Bayesian Time Series Classification*, Advances in Neural Processing Systems 14, 2002

[4] Pierre Geurts: *Pattern extraction for time series classification*, Principles of Data Mining and Knowledge Discovery, 2001

[5] Li Wei, Nitin Kumar, Venkata Lolla, Eamonn Keogh , Stefano Lonardi, Chotirat Ann Ratanamahatana: *Assumption-free anomaly detection in time series*, Proceedings of the 17th International Conference on Scientific and Statistical Database Management, 2005

[6] Stephane G. Mallat: *A theory for multiresolution signal decomposition: the wavelet representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989

[7] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, Jenifer C. Ali: *Class-based n-gram models of natural language*, Computational Linguistics, 1992

[8] Peter J. Rousseeuw: *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics, 1987

BEHAVIOUR ANALYSIS BASED ON OBSERVED LOW-LEVEL ACTIONS: An overview

Violeta Mirchevska[1], Aleš Tavčar[2], Matjaž Gams[2]
[1]Result d.o.o., Bravničarjeva 11, 1000 Ljubljana, Slovenia
[2]Department of Intelligent Systems, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
e-mail: violeta.mircevska@result.si

ABSTRACT

The purpose of this paper is to give an overview of the field of agent behaviour analysis. General guidelines for approaching the problem are presented as well as techniques for extracting representative activity patterns based on observations of low-level actions and techniques for analysing preferences.

1 INTRODUCTION

The research and modelling of behaviour of agents is notably gaining in importance in modern information society. Behaviour models can be used in a variety of applications. Modelling the behaviour of the elderly in order to detect health problems based on their behavioural patterns and modelling behaviour patterns and strategies of movement of people in emergency cases in order to improve escape procedures are just two examples.

This paper presents an overview of methods for analysing behaviour of agents.

2 BEHAVIOUR ANALYSIS

In general, agent behaviour is purposeful. Agents perform actions in order to achieve their goals. For example, when traversing a road network, drivers take actions in order to reach their destination choosing roads that optimize their personal cost in terms of time, money, simplicity of reaching destination and similar. The task of analysing the behaviour of agents can be divided in two subareas:

1. extraction of representative activity patterns
2. analysing preferences

Extraction of representative activity patterns refers to the problem of determining the sequence of actions agents take in order to achieve a particular goal. Analysing preferences refers to the problem determining the personal cost function of a given agent i.e. determining what “effectiveness” is for a particular agent. In the previously specified example, extraction of representative activity patterns means determination of the most often used roads for traversing from one point to another. Analysing preferences, on the other hand, is related to the cost function of a particular agent traversing the road network determining which factors (e.g. time, money, simplicity of reaching final destination) influence agent’s decisions the most.

Section 2.1 presents general guidelines how the problem of behaviour analysis should be approached. Section 2.2 presents techniques for extraction of representative activity patterns. Techniques for analysing preferences are presented in Section 2.3.

2.1 Problem decomposition

Stone (1998) argues that the behaviour of agents acting in a limited communication, real-time, noisy environment with both teammates and adversaries is too complex to be addressed by a simple mapping from agents’ sensors to their actuators. Therefore, he introduces layered learning, a general-purpose machine learning paradigm for complex domains (Stone 1998, 2000). The key idea of this paradigm is that learned low-level behaviours can facilitate the learning of higher level behaviours. For example, when analysing collaborative and adversarial multi-agent domain, learning can begin with individual behaviours, which facilitate multi-agent collaborative behaviours, and eventually lead to the full-team collaborative or adversarial behaviours. The task decomposition in layered learning is not automated and is to be determined as a function of the specific domain at investigation.

Based on a widely acknowledged psychological hierarchy of human behaviour, Thurau (2006) proposes the behaviour analysis to be addressed on three levels:

- Strategic behaviour, for achievement of long term goals
- Tactical behaviour, for mid-term goal achievement
- Reactive behaviour, which refers to simple movements and solely sensor dependent actions

Thurau presents player analysis in the game QUAKE II. At the strategic level he analyses goal oriented movement, such as situation dependent item pickups and area control. At the tactical level he analyses smart localized behaviours, such as player ambushing an enemy player. At the reactive layer, analysis of behaviours that can be characterized as a direct functional mapping of game states onto agent actions, such as aiming, shooting and dodging projectiles, is performed.

2.2 Extraction of representative activity patterns

Extraction of representative activity patterns refers to the problem of determining the sequence of actions agents take in order to achieve a particular goal. This section presents techniques for extracting representative activity patterns.

MASDA

MASDA (Bežek 2006) is an algorithm for discovering strategic behaviour developed on the Department of Intelligent Systems at the Jožef Stefan Institute. By tracking low-level behaviour of a group of agents and using basic domain knowledge, it is able to discover common agent strategy. The algorithm incorporates an abstraction process, which enables it to separate agent activity patterns that are caused by the agent following its strategy, from activity that is caused by reactions to local environmental changes. The MASDA algorithm creates graphical and symbolic description of the discovered strategic behaviour. The graphical descriptions visually present multi-agent activity, while the symbolical descriptions present important characteristics of multi-agent interaction in a human-comprehensible way.

The algorithm has successfully been employed in the domain of robotic soccer. In this domain, it discovered relevant patterns of strategic behaviour for 11 vs. 11 player actual play. Moreover, it was able to generate executable models for playing 3 vs. 2 Keepaway.

The process of discovering strategic behaviour starts with transformation of a game trace to relevant agent actions based on domain knowledge. Each action is marked by the role of the agent performing the action.

Central concept of the MASDA algorithm is the *action graph* which is created once the agent actions and roles are reconstructed using domain knowledge. An action graph is a directed graph, where nodes represent state space at the start of an agent action and connections correspond to agent actions. An example action graph, obtained from actions in a RoboCup game, is presented in Figure 1.

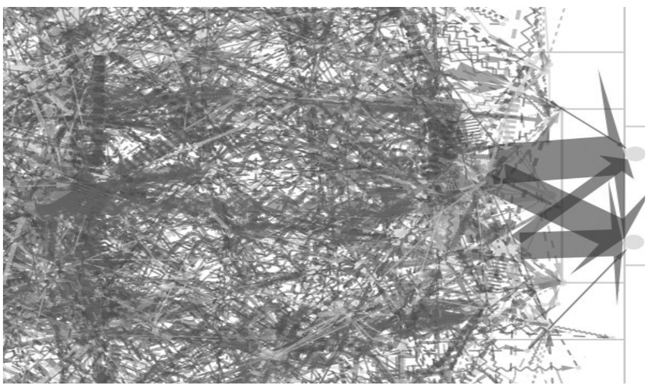


Figure 1 Action graph

The action graph obtained directly from the reconstructed agent actions and roles contains many nodes and, therefore, is difficult to be comprehended by humans. Therefore, its level of detail needs to be reduced, preserving at the same time the attained action concepts. MASDA accomplishes this with hierarchical clustering of graph nodes. The appropriate level of abstraction obtained by hierarchical clustering needs to be carefully determined since too little abstraction leads to models reflecting agents' reactions to local environment changes and too much abstraction may neglect important details of strategic behaviour. An action graph representing strategic activity in the RoboCup game is presented in Figure 2.

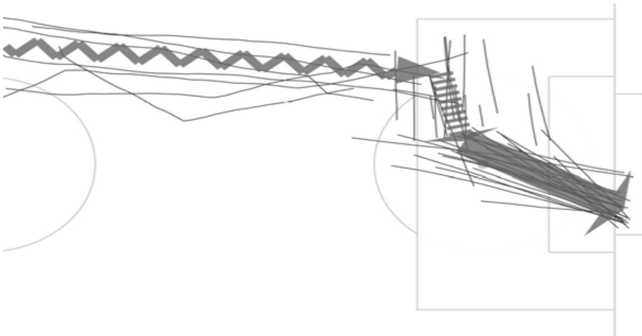


Figure 2 Graphical representation of strategic activity

Once strategic activity is discovered using action graphs, MASDA generates symbolic description of the observed behaviour patterns. As each node defines a unique action concept, it can be used to generate rules that describe this specific agent action. In particular, data for rule inducing algorithms is generated as follows: Positive examples are action instances in a target node and negative examples as instances in nearby nodes (i.e. near misses). For each instance, all pairs of agent role-domain feature are generated and the true ones are stored. Example of a symbolic description of learned strategy for successful goal attack in the RoboCup domain is: IF during opponent's goal attack the right back is left behind, THEN the most successful attack starts on the left side, then the ball is passed in the middle, then from the middle to the right where the attacker scores the goal.

Layered behaviour analysis

Kaminka et al. (2002) and later Horman et. al. (2004, 2005) presented an approach for learning sequential coordinated behaviour of teams. Their method translates observations of a dynamic environment into a time-series of recognized basic actions. This time-series is then analysed to find repeating subsequences characterizing each team. Two methods are used for extracting such characteristic sequences: frequency counts and statistical dependencies. Both approaches are able to extract meaningful sequences; however, the statistical dependency approach is able to correctly reject sequences that are frequent, but are due to random co-occurrence of behaviours. Learning sequential behaviour based on frequency of appearance is more suitable for low noise environments, where most of the observed actions are a result of the agent following a strategic plan, whereas statistical dependency is more suitable for higher noise environments.

Ramos et al (2008a, 2008b) present approach for discovering tactical plays (e.g. offensive plays) adopted by soccer-agents during a match within the context of formations. Formations are represented by planar topological graphs. Tactical plays are presented by the path of the ball occurring in particular context. The planar topological graph enriches this information with information concerning the players participating in the particular tactical play and the zones through which the play has taken place.

Aler et al. (2009) presents an approach for determining reactive behaviour in the domain of robotic soccer. Their purpose was to program RoboCup agents by imitating the actions of humans playing soccer. Therefore, humans were let control RoboCup agents in a soccer virtual environment. This way data concerning how humans act in soccer based on the current situation was collected. In particular, they examined the conditions under which human players perform the following five actions: turn, run slow, run fast, kick ball and kick to goal. The conditions were represented by information concerning the position of the agent, the relative position of the agent to the ball, two closest opponents and the opponents' goal. Reactive behaviour was represented by rules induced on the data obtained from humans controlling soccer agents.

Behaviour libraries

Recognition of agent behaviour may, also, be addressed by developing behaviour libraries which encode the complete repertoire of expected behaviour. This technique focuses on matching the observations of agent behaviour with library prototypes. This approach has been successfully used in domains such as intelligent user interfaces (Lesh et al. 1998) in virtual environment training (Tambe 1995) and games (Doirado et al. 2010). However, this approach is inappropriate for open or adversarial settings in which the behaviour repertoire of the observed agents is unknown at design time.

ISAAC system

ISAAC (Raines et al. 2000) is one of the first systems for analysing teams participating in the RoboCup competition. It enables automated post-hoc, off-line agent-team analysis. ISAAC`s novelty stems from a key design constraint that arises in team analysis: multiple types of models of team behaviour are necessary to analyse three different levels of granularity:

1. individual agent actions,
2. inter-agent interactions and
3. global team behaviour.

These models of team behaviour are automatically acquired via machine learning over teams' external behaviour traces. Additionally, ISAAC employs multiple presentation techniques that can aid human in understanding the generated models. As can be seen in Figure 3, acquiring the models involves a mix of data mining and inductive learning but the methods used are specific to the granularity of analysis being modelled. Analysis of an individual agent action (individual agent key event model) uses the C5.0 decision tree inductive learning algorithm, an extension to C4.5, to create rules of success or failure (Quinlan 1994). For analysis of agent interactions (multiple agent key interaction model), pre-defined patterns are matched to find prevalent patterns of success. Finally, to develop rules of team successes or failures (global team model), game level statistics are mined from all available previous games and again inductive learning is used to determine the causes for success and failure.

2.3 Analysing preferences

Analysing preferences refers to the problem determining the personal cost function of a given agent. This section presents a technique for analysing preferences.

Ziebart et al. (2009) present an approach for developing a

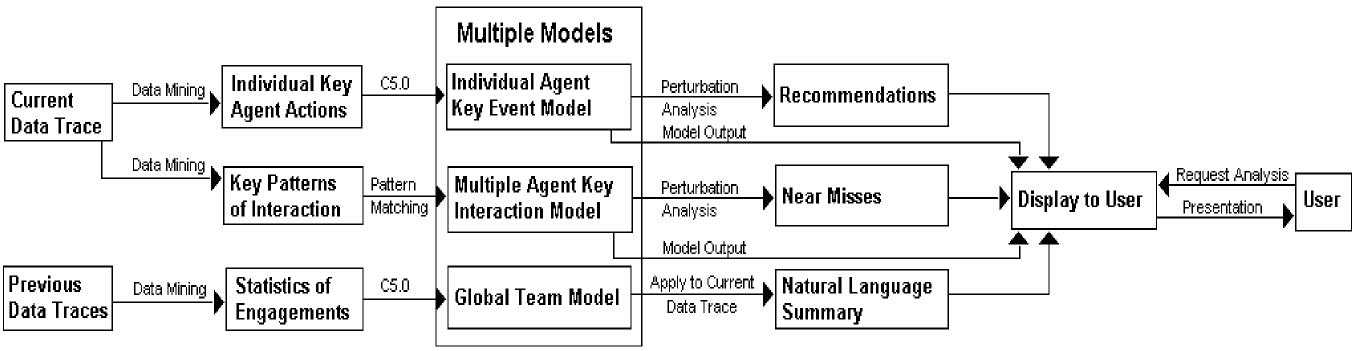


Figure 3 ISAAC model generation and analysis

conditional probabilistic model which predicts human decisions based on their preferences given a contextual situation. They employ the principle of maximum entropy within the Markov Decision Process framework, reducing the problem of human behaviour modelling to recovering a context-sensitive utility function that explains the observed behaviour by the conditional probabilistic model. Test domain for their research was the domain of learning driver route choices in which they modelled driver preferences with respect to time, safety, fuel costs, maintenance costs and similar. Having information concerning driver preferences, they were able to predict road congestions and predict drivers' destination given partial trajectory.

3 CONCLUSION

This paper presented an overview of the field of agent behaviour analysis. When analysing the behaviour of agents, two aspects should be considered: (1) extraction of representative activity patterns and (2) preference analysis. Extraction of representative activity patterns refers to the problem of determining the sequence of actions the agents take in order to achieve a particular goal. Analysing preferences refers to the problem of determining the personal cost function of a given agent i.e. determining what "effectiveness" is for a particular agent. Techniques for addressing both analysis aspects were presented.

The problem of analysing the behaviour of agents in limited communication, real-time, noisy environments with both teammates and adversaries is too complex to be addressed by a simple mapping from agents' sensors to their actuators. Therefore, layered learning approach needs to be used. Based on widely acknowledged psychological hierarchy of human behaviour, behaviour analysis on the following three levels is proposed: (1) strategic behaviour, (2) tactical behaviour, and (3) reactive behaviour. Strategic behaviour refers to actions an agent performs in order to achieve long-term goals. Tactical behaviour refers to actions an agent performs for mid-term goal achievement. Reactive behaviour encompasses simple movements and solely sensor dependent actions.

ACKNOWLEDGEMENTS

This work was partially supported by the European Union, European Social Found.

References

[1] Aler, R., Valls, J. M., Camacho, D., Lopez, A. (2009): Programming Robosoccer Agents by Modeling Human Behaviour, *Expert Systems with Applications* 36, pp. 1850 – 1859

[2] Bežek, A. (2006): Avtomatsko modeliranje večagentnih sistemov, *Doctoral dissertation*, University of Ljubljana, Faculty of Computer and Information Science, In Slovene

[3] Doirado, E. and Martinho, C. (2010): I Mean It! Detecting User Intentions to Create Believable Behaviour for Virtual Agents in Games, *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, pp. 83-90

[4] Horman, Y., Kaminka, G. A. (2004): Improving Sequence Recognition for Learning the Behaviour of Agents, *Proceedings of the Third Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, pp. 1332-1333

[5] Horman, Y., Kaminka, G. A. (2005): Removing Statistical Biases in Unsupervised Sequence Learning, *Proceedings of Intelligent Data Analysis (IDA 2005)*

[6] Kaminka, G. A., Fidanboylu, Cg, A., Veloso, M. M. (2002): Learning the Sequential Coordinated Behaviour of Teams from Observations, *Proceedings of the RoboCup-2002 Symposium*

[7] Lesh, N., Rich, C., Sidner, C. L. (1998): Using Plan Recognition in Human-Computer Collaboration, *Proceedings of the 7th International Conference on User Modelling*

[8] Quinlan, J. (1994): C4.5: Programs for Machine Learning, Morgan Kaufman, 1994

[9] Raines, T., Tambe, M., Marsella, S. (2000): Automated Assistants to Aid Humans in Understanding Team Behaviours, *Proceedings of the 4th International Conference on Autonomous Agents*, pp. 419 – 426

[10] Ramos, F., Ayanegui, H. (2008a): Discovering Tactical Behaviour Patterns Supported by Topological Structures in Soccer-Agent Domains, *Proceedings of the 7th International Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2008)*, pp. 1421-1424

[11] Ramos, F., Ayanegui, H. (2008b): Discovering Behaviour Patterns in Multi-agent Systems, *Proceedings of the 2nd KES International Conference on Agents and Multi-agent Systems: technologies and applications*

[12] Stone, P. (1998): Layered Learning in Multi-agent Systems, *Doctoral dissertation*, Carnegie Mellon University

[13] Stone, P. (2000): Layered Learning in Multi-agent Systems: A Winning Approach to Robotic Soccer, *MIT Press*

[14] Tambe, M. and Rosenbloom, P. (1995): RESC: An Approach for Real-time, Dynamic Agent Tracking, *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995)*, pp. 103 – 110

[15] Thurau, C. (2006): Behaviour Acquisition in Artificial Agents, *Doctoral dissertation*, Bielefeld University

[16] Ziebart, B. D., Maas, A., Bagnell, J. A., Dey, A. K. (2009): Human Behaviour Modelling with Maximum Entropy Inverse Optimal Control, *Proceedings of AAAI Spring Symposium on Human Behaviour Modelling*

Iskanje točnih in razumljivih hibridnih klasifikacijskih dreves

Rok Piltaver
Odsek za inteligentne sisteme
Institut "Jožef Stefan"
Jamova 39, 1000 Ljubljana, Slovenija
e-mail: rok.piltaver@ijs.si

POVZETEK

V prispevku je predstavljen algoritem, ki na osnovi učne množice, klasifikacijsega drevesa in dodatnega klasifikatorja BB poišče množico hibridnih dreves s čim večjo razumljivostjo in točnostjo hkrati. Algoritem generira hibridna drevesa tako, da v nekaterih listih danega klasifikacijskega drevesa zamenja klasifikatorje večinskega razreda za klasifikator BB. Algoritem vrne le Pareto množico hibridnih dreves.

Prispevek poleg algoritma opisuje tudi meri za razumljivost in točnost, postopek iskanja Pareto množice in dve optimizaiji osnovne različice algoritma.

1. Uvod

Na področju strojnega učenja algoritme za gradnjo klasifikatorjev iz množice učnih primerov in klasifikatorje same najpogosteje ocenjujemo po točnosti (zgrajenih) klasifikatorjev. Točnost ocenimo z merami kot so klasifikacijska točnost (delež pravilno klasificiranih primerov), površina pod ROC krivuljo, občutljivost, specifičnost, natančnost (ang. precision), priklic, F-mera, idr [3]. Kadar želimo oceniti točnost klasifikatorja in nimamo podatkov na katerih bomo uporabljali klasifikator, točnost ocenimo na delu učne množice, tako da jo razdelimo na učno in testno ali opravimo večkratno prečno preverjanje.

Poleg točnosti klasifikatorja je pomembna tudi njegova razumljivost. Uporabniki klasifikatorjem bolj zaupajo, kadar jih lahko vidijo v grafični obliki in razumejo njihovo delovanje ter naučeno znanje. Razumljivost klasifikatorja je ključnega pomena tudi, kadar želi uporabnik iz učnih podatkov izluščiti pravila oz. relacije, ki veljajo med napovedovanim in ostalimi atributi [3]. Med najbolj razumljive klasifikatorje spadajo klasifikacijska drevesa in odločitvena pravila ter klasifikatorji na osnovi naivnega Bayesa in logistične regresije, ki jih je mogoče prikazati z uporabo nomogramov.

Razumljivost in točnost klasifikatorja sta si nasprotujoči lastnosti. Znano je, da dosegajo kompleksnejši učni algoritmi z močnejšim jezikom za opis relacij v učnih podatkih (npr. meta klasifikatorji, SVM, umetne nevronske mreže) višjo točnost od razumljivih klasifikatorjev. Hkrati pa so taki (točnejši) klasifikatorji uporabniku videti kot črna škatla (ang. black-box classifier, BB), ki je sicer sposobna točno napovedovati vrednosti novih primerov, vendar ne ponuja razumljive razlage za odločitve in ne opisuje relacij v učnih podatkih na uporabniku razumljiv način.

Cilj algoritma, ki je opisan v tem prispevku, je uporabniku ponuditi pester nabor klasifikatorjev v razponu od najbolj razumljivih do najbolj točnih ter mu prepustiti odločitev o izbiri najprimernejšega klasifikatorja za njegove potrebe. Osnovna ideja je, da algoritem zgradi množico klasifikatorjev, ki so hibridi med razumljivim in točnim klasifikatorjem. Tudi kadar se uporabnik po pregledu ponujenih klasifikatorjev odloči, da bo namesto ponujenih hibridnih klasifikatorjev uporabil običajen klasifikator (maksimalno razumljiv ali točen), ima še vedno korist, saj je seznanjen tudi z ostalimi klasifikatorji z drugačnim razmerjem med razumljivostjo in točnostjo.

V naslednjem razdelku je predstavljen formalizem za predstavitev in shranjevanje hibridnih klasifikatorjev. Tretji razdelek govori o merah točnosti in razumljivosti, ki jih uporablja algoritem, in o načinu primerjanja hibridnih klasifikatorjev po obeh kriterijih hkrati. V četrtem razdelku je podan splošen opis algoritma. Peti razdelek opisuje enostavno implementacijo algoritma, medtem ko sta v šestem razdelku predstavljeni metodi za izboljšanje učinkovitosti algoritma. V zadnjem razdelku so podani zaključki in ideje za nadalnje raziskave.

2. Model za predstavitev znanja

Osnovno struktura, s katero je predstavljeno znanje hibridnega klasifikatorja, je klasifikacijsko drevo. Le-to smo izbrali, ker so metode za gradnjo dreves zelo znane, pogosto uporabljane in v praksi dajejo dobre rezultate, hkrati pa so klasifikacijska drevesa enostavna za prikaz in razumevanje. Vsako notranje vozlišče drevesa vsebuje pogoj, ki primer glede na vrednost določenega atributa posreduje določenemu nasledniku v drevesu. Vsak list drevesa določa vrednost razreda za primere, ki zaradi vrednosti atributov pristanejo v tem listu. Vsak list lahko obravnavamo kot klasifikator večinskega razreda (ang. majority class classifier), ki vse primere klasificira v razred, ki je v podmnožici učnih primerov, ki padejo v dani list, najpogostejši.

Formalizem za shranjevanje znanja hibridnih dreves, ki ga uporablja opisani algoritem, se od klasifikacijskega drevesa razlikuje po naslednji posplošitvi. Namesto klasifikatorja večinskega razreda se lahko v podmnožici listov drevesa pojavi nek drug klasifikator (npr. BB). Hibridno drevo klasificira primere, ki padejo v običajne liste, kot običajno drevo. Primere, ki padejo v liste z BB klasifikatorjem, klasificira z uporabo le-tega. V različici algoritma opisani v tem prispevku se v vseh hibridnih listih uporablja isti BB klasifikator.

3. Kvaliteta hibridnih dreves

V tem razdelku so opisane mere za ocenjevanje kvalitete hibridnih dreves, ki so rezultat obravnavanega algoritma. V prvem podrazdelku je opisana mera za točnost, v drugem za razumljivost in v tretjem za oba kriterija hkrati.

3.1. Točnost hibridnih dreves

Za oceno točnosti hibridnega drevesa k, ki jo označimo s t_k, uporabljamo število pravilno klasificiranih primerov izračunano na učni množici. Ta mera usmerja algoritem k iskanju hibridnih dreves, ki imajo čim boljše razmerje med številom pravilno klasificiranih primerov in številom vseh klasificiranih primerov v učni množici. Razliko v številu pravilnih klasifikacij v listu i med klasifikatorjem večinskega razreda in BB klasifikatorja označimo z $\delta_i = t_{i,\mathrm{BB}} - t_{i,\mathrm{drevo}}$.

Menjava običajnega lista drevesa za BB list je smiselna samo takrat, ko je točnost BB klasifikatorja $t_{i,\mathrm{BB}}$ na primerih, ki padejo v list i, večja od točnosti klasifikatorja večinskega razreda $t_{i,\mathrm{drevo}}$. Torej le za liste i, kjer velja $\delta_i > 0$.

Če v drevesu obstajajo listi, v katerih ima višjo točnost BB klasifikator ($\delta_i > 0$), in listi, v katerih ima višjo točnost klasifikator večinskega razreda ($\delta_j < 0$), lahko algoritem najde klasifikatorje, ki imajo višjo klasifikacijsko točnost od drevesa in od BB klasifikatorja hkrati.

3.2. Razumljivost hibridnih dreves

Mera za razumljivost hibridnega drevesa k, ki jo uporabljamo v tem prispevku in označujemo z r_k, je enostavna, kar nam omogoča uporabo optimizacij algoritma predstavljenih v razdelku 6 in izpeljavo dokazov o smiselnosti optimizacije. Razumljivost celotnega hibridnega drevesa ocenimo kot povprečno razumljivost primerov iz učne množice. Razumljivost primerov, ki padejo v običajne list drevesa je po definiciji 1. Razumljivost primerov, ki padejo v BB liste pa je 0. Običajno drevo ima po tej meri maksimalno razumljivost 1, medtem ko ima drevo z enim samim vozliščem, v katerem je BB klasifikator, minimalno razumljivost 0. Mera usmerja algoritem k iskanju čim bolj razumljivih hibridnih dreves. Mera ni primerna za ocenjevanje razumljivosti dreves različne velikost, ker je enaka 1 ne glede na velikost drevesa, čeprav je manjše drevo za uporabnika lažje razumljivo od večjega. Ker pa mero uporabljamo le za primerjavo hibridih dreves, ki so bila z zamenjavo listov dobljena iz enega samega osnovnega drevesa, je mera intuitivna, smiselna in praktično uporabna.

Razliko razumljivosti v listu i v katerem klasifikator večinskega razreda zamenjamo za BB klasifikatorja označimo z $-\Delta_i = N_i / N$. Razlika v razumljivosti je enaka razmerju med številom primerov, ki padejo v list i (N_i), in številom vseh primerov (N), krat -1, ker gre za zmanjšanje razumljivosti. Vsako hibridno drevo ima torej razumljivost večjo ali enako 0 in manjšo od 1. Najbolj točna hibridna drevesa imajo lahko razumljivost večjo od 0 in so od BB klasifikatorja boljša v točnosti in razumljivosti hkrati.

3.3. Množica najboljših hibridnih dreves

Ker je cilj algoritma poiskati čim bolj točna in hkrati čim bolj razumljiva hibridna drevesa, je rezultat množica hibridnih dreves in ne le eno samo hibridno drevo. Gre namreč za dvokriterijski optimizacijski problem [2] v katerem želimo maksimirati oba kriterija (razumljivost in točnost). Kadar primerjamo dve hibridni drevesi ju moramo primerjati po obeh kriterijih hkrati in ne po vsakem kriteriju posebej.

Če je hibridno drevo $S_i = (t_i, r_i)$ boljše ali enako hibridnemu drevesu $S_j = (t_j, r_j)$ po obeh kriterijih in je strogo boljše vsaj po enem, je S_i boljše od S_j. V takem primeru pravimo, da S_i dominira nad S_j in zapišemo $S_i \succ S_j \Leftrightarrow (t_i \geq t_j \wedge r_i > r_j) \vee (t_i > t_j \wedge r_i \geq r_j)$. Če je hibridno drevo S_i slabše ali enako od S_j po obeh kriterijih in vsaj po enem strogo slabše, je S_i slabše od S_j. V takem primeru pravimo, da je S_i dominirano in zapišemo $S_i \prec S_j \Leftrightarrow (t_i \leq t_j \wedge r_i < r_j) \vee (t_i < t_j \wedge r_i \leq r_j)$. Če je S_i po enem kriteriju boljše in po drugem slabše od S_j sta hibridni drevesi po kvaliteti neprimerljivi [2]. Med neprimerljivimi hibridnimi drevesi izbira le uporabnik, s stališča algoritma pa so vsa neprimerljiva hibridna drevesa enako dobra.

Za uporabnika je zanimiva le tako imenovan Pareto množica hibridnih dreves. Po definiciji so v Pareto množici S' množice $S = \{S_i\}$ vsi nedominirani elementi S_i množice S [2]. S formulo to zapišemo tako: $S' = \{S_i \in S; \nexists S_j \in S: S_j \succ S_i\}$. V našem primeru so v Pareto množici vsa hibridna drevesa, ki so med seboj neprimerljiva in za katere ne obstaja hibridno drevo, ki bi dominiralo nad nekim hibridnim drevesom v Pareto množici.

4. Uporaba algoritma

Cilj algoritma je poiskati in prikazati Pareto množico hibridnih dreves ter na zahtevo uporabnika podati izbrano hibridno drevo. Algoritem se osredotoča zgolj na kombiniranje danega klasifikacijskega drevesa z danim BB klasifikatorjem, ne pa z gradnjo izhodiščnega klasifikacijskega drevesa in BB klasifikatorja. Zato na vhodu sprejme učno množico, izhodiščno klasifikacijsko drevo in BB klasifikator, ki jih mora priskrbeti uporabnik. Pri tem je pomembno, da je dano klasifikacijsko drevo dovolj majhno, da bodo hibridna drevesa izpeljana iz njega za uporabnika dovolj razumljiva. Primerno velikost drevesa je mogoče doseči s pravilno nastavitvijo parametrov rezanja drevesa. Dani BB klasifikator mora biti kar se da točen. Če imajo vsi klasifikatorji večinskega razreda v listih danega drevesa višjo klasifikacijsko točnost od BB klasifikatorja, listov ni smiselno menjati za BB klasifikator in zato ni mogoče zgraditi hibridnih dreves, ki so vsaj po enem kriteriju boljša od izhodiščnega klasifikacijskega drevesa. Dovolj visoko klasifikacijsko točnost je mogoče doseči s primerno izbiro algoritma za gradnjo BB klasifikatorja in njegovih nastavitev.

Ker je algoritem za iskanje hibridnih dreves kompatibilen s knjižnico algoritmov za strojno učenje Weka [3], obsega zagotavljanje izhodiščnega klasifikacijskega drevesa in BB klasifikatorja le enostavno izbiro algoritmov in njihovih nastavitev iz Wekine knjižnice algoritmov.

5. Naivna implementacija

Algoritem deluje tako, da pregleda celotno množico hibridnih dreves, ki nastanejo iz izhodiščnega klasifikacijskega drevesa z menjavo podmnožice listov za BB klasifikator, in v njej poišče podmnožico nedominiranih hibridnih dreves (Pareto množico). Prostor preiskovanja S lahko predstavimo z množico $\{0,1\} \times \{0,1\} \times \ldots \times \{0,1\} = \{0,1\}^N$, kjer je N enak številu listov v izhodiščnem drevesu, vrednosti 0 ali 1 pa za vsak element kartezičnega produkta pomenita, ali je določen list hibridnega drevesa običajen ali BB list. Velikost preiskovanega prostora je zato enaka 2^N. Izhodiščno drevo lahko obravnavamo kot hibridno drevo in ga označimo z S_0 ter v izbrani notaciji zapišemo kot (0,0, … 0). Tudi BB klasifikator lahko obravnavamo kot hibridno drevo z oznako $S_{2^\wedge N}$ in zapisom (1,1, … 1).

Iz hibridnega drevesa S_i dobimo novo hibridno drevo S_j tako, da list l, ki je v S_i običajen list, zamenjamo za BB list. $S_j = (s_{j,1}, s_{j,2}, \ldots, s_{j,2^\wedge N})$, kjer je $s_{j,l} = 1$ in $s_{j,k} = s_{i,k}$ za $k \neq l$. Iz kvalitete hibridnega drevesa $S_i = (t_i, r_i)$ se kvaliteta hibridnega drevesa S_j izračuna po naslednji formuli $S_j = (t_i + \delta_l, r_i - \Delta_l)$. Zamenjava običajnega lista za BB list ima enak učinek na razumljivost in točnost ne glede na stanje izhodiščnega hibridnega drevesa (število in izbor BB listov).

Osnovni algoritem deluje v treh korakih. V prvem koraku s formulami iz razdelka 3.1 in 3.2 izračuna δ_l in Δ_l za vsak list izhodiščnega drevesa. V drugem koraku uporabi popolno preiskovanje, da najde vsa možna hibridna drevesa, in izračuna njihove razumljivosti in točnosti z uporabo rezultatov iz prvega koraka. V zadnjem koraku iz množice hibridnih dreves izloči dominirana hibridna drevesa.

Rekurzivni algoritem, ki se uporablja v drugem koraku, vodi dve množici: v prvi množici (O) so najdena in obdelana hibridna dreves, v drugi (T) pa najdena in še neobdelana hibridna dreves. Izvajanje se prične s prazno množico O, in množico T v kateri se nahaja le izhodiščno klasifikacijsko drevo S_0. Algoritem (slika 1) iterativno obdeluje množico neobdelanih hibridnih dreves: iz vsakega neobdelanega hibridnega drevesa S_i se zgradijo nova hibridna drevesa S_j, ki se jih doda v novo množico T (*Ttmp*), tako da se po en običajen list l zamenja za BB list. Po tem se hibridno drevo S_i doda novi množici O. Če je potrebno se požene nova iteracija. Skupaj s hibridnimi drevesi S_j se izračunavajo in shranjujejo tudi podatki o njihovi razumljivosti (r) in točnosti (t), ki se jih izračuna s formulo opisano v prejšnjem odstavku. Če v množici T ni več nobenega hibridnega drevesa, ki se ne nahaja v množici O, se izvajanje ustavi, saj množica O v tem primeru že vsebuje vseh 2^N možnih hibridnih dreves iz prostora $S = \{0,1\}^N$.

find(O, T) { *nasel* = false; *Ttmp* = { };
 for (S_i in T) { for (l: $s_{i,l} = 0$) {
  nasel = true; $S_j = S_i$; $s_{j,l} = 1$; $S_j.t = S_i.t + \delta_l$;
  $S_j.r = S_i.r - \Delta_l$; if (S_j not in *Ttmp*) *Ttmp*.add(S_j);
  } O.add(S_i);
 } if (*nasel*) return find(O, *Ttmp*) else return O; }

Slika 1: *Rekurzivna funkcija za iskanje hibridnih dreves*

V tretjem koraku algoritem poišče Pareto množico O'. Najenostavnejši način za iskanje Pareto množice S' je, da za vsak $S_i \in S$ preverimo ali v množici S obstaja kak element S_j, ki dominira nad S_i ($S_j \succ S_i$). Če takšen S_j ne obstaja, potem S_i pripada Pareto množici S', sicer pa ne. Učinkovitejši algoritem je prikazan na sliki 2 in prav tako kot prvi opisani algoritem deluje neglede na število kriterijev.

S' = {};
for (S_i in S) { iDominiran = false;
 for (S_j in S') { if ($S_i \succ S_j$) $S' = S' \setminus \{S_j\}$;
  else if ($S_j \succ S_i$) { iDominiran = true; break;}
 } if (iDominiran = false) $S' = S' \cup \{S_i\}$; }

Slika 2: *Algoritem za iskanje Pareto množice*

Za dvokriterijske probleme obstaja še en učinkovit algoritem (slika 3). Le-ta najprej sortira vse elemente množice S, za katero iščemo Pareto množico S', po prvem kriteriju od najboljših do najslabših, kjer je prvi kriterij enak upošteva še drugi kriterij. Nato v enem prehodu skozi sortiran seznam elementov množice S najde Pareto množico S' [1]. Med prehodom algoritem vodi spremenljivko, ki predstavlja trenutno vrednost drugega kriterija, v našem primeru razumljivost r_{tmp}. Ob začetku prehoda skozi sortiran seznam algoritem prvi element v seznamu $S_1 = (t_1, r_1)$ določi kot edini element Pareto množice S', in priredi $r_{tmp} = r_1$. Za vsak naslednji element S_i, ki izpolnjuje pogoj $r_i > r_{tmp}$, naredi naslednje: element S_i doda v Pareto množico S' in nastavi $r_{tmp} = r_i$. Algoritem je prikazan na sliki 3.

sort(S); $S' = \{S_1\}$; $r_{tmp} = r_1$;
for S_i in S_2 .. $S_{2^\wedge N}$ { if ($r_i > r_{tmp}$) { S'.add(S_i); $r_{tmp} = r_i$; }}

Slika 3: *Algoritem za iskanje Pareto množice (2 kriterija)*

6. Izboljšava osnovnega algoritma

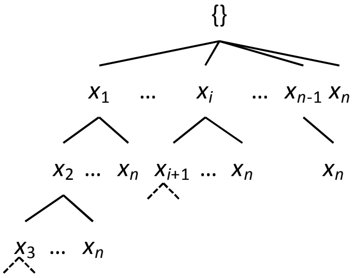
Časovna zahtevnost naivne implementacije algoritma narašča eksponentno z številom listov osnovnega drevesa in je enaka O(N!). Nekatera drevesa algoritem obravnava večkrat (pogoj if(S_j not in *Ttmp*) slika 1), kar je potratno. Kako se temu izgniti je opisano v razdelku 6.1. Prav tako se je mogoče izogniti pregledovanju hibridnih dreves, za katera je že pred izračunom njihove točnosti in razumljivosti jasno, da bodo dominirana. Ta izboljšava je opisana v razdelku 6.2.

6.1. Preprečevanje večkratnih obiskov istega dreves

V prvi iteraciji funkcije find({},{S_0}) se v množico *Ttmp* doda N elementov, ker je vseh N listov v S_0 običajnih listov in je zato vsakega mogoče zamenjati za BB list. V naslednjem koraku (rekurzivnem klicu) se za vsako od N hibridnih dreves s po enim BB listom zgradi $N - 1$ novih hibridnih dreves, saj je v vsakemu drevesu mogoče zamenjati $N - 1$ običajnih listov. V naslednjem klicu se za vsako drevo zgradi $N - 2$ hibridnih dreves in tako naprej, vse dokler se ne zgradi le še hibridno drevo $S_{2^\wedge N}$, ki ima v vseh listih BB klasifikator. Zato je število pregledanih dreves enako $N(N-1)(N-2)$ … (1) = N!, kar je za $N > 3$ več od števila vseh možnih hibridnih dreves (2^N). To kaže na neučinkovitost naivne implementacije algoritma.

Enostavno naštevanje podmnožic listov ni primerno, ker za učinkovitost izračuna točnosti in razumljivosti ter za pohitritev opisano v naslednjem podrazdelku potrebujemo zaporedje hibridnih dreves, ki je urejeno po številu BB listov. Naslednja zakonitost omogoča odpravo večkratnih obiskov hibridnih dreves in ohranja zahtevano urejenost.

Elemente množica $A = \{x_i\}$ lahko uredimo z relacijo $<(x_i, x_j) \Leftrightarrow i < j$, tako da jim pripišemo indekse. Z isto relacijo lahko uredimo tudi vsako podmnožico $B \subseteq A$. Vse podmnožice $B \subseteq A$ lahko predstavimo z drevesom prikazanim na sliki 4, v katerem vsaka pot od korena do nekega vozlišča predstavlja določeno podmnožico B z elementi, ki jih srečamo v vozliščih na poti. Nivo zadnjega vozlišča na poti je enak številu elementov v pripadajoči množici. Vsako vozlišče, ki vsebuje element x_i ima za naslednike le vozlišča z elementi x_j za katere velja $<(x_i, x_j)$.



Slika 4: *Del drevesa podmnožic množice* $A = \{x_i\}$

Elemente drevesa na sliki 4 lahko naštejemo s klicem rekurzivne funkcije findSubSet({}, *A*), ki je prikazana na sliki 5. Funkcija sprejme dve množici *used* in *unused*. Prva množica vsebuje vse elemente, ki so bili že uporabljeni višje v drevesu, druga pa elemente, ki jih še lahko uporabimo nižje v drevesu. Funkcija *e*.indeks() vrne indeks elementa *e*, *u*.max() vrne maksimalni indeks v množici *u*, *u*.clone() vren kopijo množice *u*, *u*.add(*e*) množici *u* doda element *e*, *u*.remove(*e*) pa iz množice *u* odstrani element *e*.

```
findSubSets(used, unused) { list = { } ;
  for (e in unused) if (e.indeks()>used.max()) {
    tmpList = findSubSets(used.clone().add(e),
      unused.clone().remove(e));
    for (s in tmpList)  list.add(s);
  } list.add(used); return list; }
```

Slika 5: *Algoritem za iskanje vseh podmnožic*

Algoritem iz slike 5 lahko priredimo za učinkovitejše iskanje hibridnih dreves. Listi izhodiščnega klasifikacijskega drevesa naj predstavljajo elemente množice *A*, in naj bodo oštevilčeni (urejeni z relacijo $<(x_i, x_j)$). S funkcijo findSubSet({}, *A*) poiščemo vse podmnožice množice *A*. Vsaka tako dobljena podmnožica predstavlja eno hibridno drevo, v katerem so listi izhodiščnega klasifikacijskega drevesa, ki so elementi množice, zamenjani za BB liste, ostali listi pa so listi običajnega drevesa.

6.2. Izogibanjem neperspektivnim drevesom

Tako kot lahko poiščemo Pareto množico hibridnih dreves lahko z uporabo iste relacije (dominira nad) poiščemo Pareto množico listov l izhodiščnega drevesa, ki jim za kriterijski vrednosti pripišemo razliki v točnosti δ_l in razumljivosti $-\Delta_l$. Če iz hibridnega drevesa $S_i = (t_i, r_i)$ naredimo novo drevo S_j, tako da list l spremenimo v BB list je kvaliteta novega lista $S_j = (t_i + \delta_l, r_i - \Delta_l)$. Ker sta t_i in r_i neodvisna od izbire lista l na kvaliteto novega hibridnega drevesa vpliva le izbor lista l. Kadar imamo na voljo več listov l_i, so zaradi naslednje relacije za menjavo zanimivi le nedominirani listi: $l_m \succ l_n \Leftrightarrow (t_i + \delta_m, \mathrm{r_i} - \Delta_m) \succ (t_i + \delta_n, r_i - \Delta_n)$. V iterativnem postopku gradnje hibridnih dreves v zaporedju, ki je urejeno po števili BB listov, zato ni potrebno zgraditi dreves, ki bi jih dobili z zamenjavo dominiranih listov. Dokaz sledi v naslednjem odstavku.

Predpostavimo, da smo hibridno drevo S_i dobili tako, da smo v izhodiščnem drevesu zamenjali neko strogo podmnožico nedominiranih listov in vsaj en dominiran list. Potem je S_i zagotovo dominirano, kar lahko dokažemo tako, da konstruiramo novo drevo S_j, ki dominira nad drevesom S_i. S_j konstruiramo tako, da v S_i zamenjmo nek dominiran list l_d nazaj v običajen list in nek nedominiran običajen list l_n v BB list. S_i in S_j se razlikujeta le v tem, da ima S_i poleg ostalih zamenjanih listov zamenjan tudi list l_d, medtem ko ima S_j poleg ostalih zamenjanih listov (istih kot S_i) zamenjan tudi list l_n. Ker velja $l_n \succ l_d \Rightarrow S_j \succ S_i$, smo dokazali, da je pri zamenjavah listov potrebno upoštevati le nedominirane liste.

Na tem mestu je potrebno opozoriti, da množica S in množica $S_{-e} = S \setminus \{e\}$, ki je enaka S brez nekega nedominiranega elementa e, nimata nujno iste Pareto množice. V Pareto množici S_{-e}' se ne pojavi odstranjeni element e, ki je element S'. Poleg tega se lahko v S_{-e}' pojavi kak nov elment iz S_{-e}, nad katirim dominira odstranjeni element e. Iz Pareto množice S' lahko hitro izračunamo Pareto množico S_{-e}', tako da uporabimo algoritem iz slike 2, ki ga inicaliziramo tako, da v S_{-e}', damo vse elemente S' razen e, in zunanjo zanko pošljemo le po elemntih množice S_{-e}', nad katerimi dominira e. Še učinkoviteje lahko množico S_{-e}' izračunamo z uporabo grafa tranzitivne redukcije relacije dominira nad ($\succ$) ali Hassejevega diagrama.

Množici hibridnih dreves, ki jo dobimo z opisanima izboljšavama, je sicer dosti manjša od množice, ki jo vrne osnovni algoritem, kljub temu pa še vedno lahko vsebuje dominirana hibridna drevesa. Dokaz bomo zaradi prostorske omejitve izpustili. Izboljšan algoritem kljub temu najde vsa Pareto optimalna hibridna drevesa. Časovna zahtevnost izboljšanega algoritma je eksponentna $O(2^N)$ vendar z majhno konstanto, kar omogoča hitro iskanje Pareto množice hibridnih dreves. Izboljšava pa ne omogoča iskanja Pareto množice, ki izhaja iz klasifikacijskega drevesa z bistveno večjim številom listov, kar pa ni omejujoče, saj so velika dreves nerazumljiva in zato za uporabnike opisanega algoritma nezanimiva.

7. Zaključek

V prispevku smo predstavili algoritem za iskanje Pareto množice hibridnih klasifikacijskih dreves po kriterijih točnosti in razumljivosti. Meri za točnost in razumljivost, ki smo jih definirali, vodita preiskovanje, dve izboljšavi osnovne različice algoritma, pa učinkovito omejujeta preiskovanje na perspektivni podprostor. V nadalnjem delu bomo izboljšali mero za razumljivost, uvedli hevristično preiskovanje in raziskali vpliv izhodiščnega klasifikacijskega drevesa in uporabe večih BB klasifikatorjev.

Reference

[1] S. Borzsonyi, D. Kossmann, K. Stocker. "The Skyline Operator", *Proceedings of the 17th International Conference on Data Engineering*, str. 421-430, 2001.

[2] K. Deb. Multi-Objective Optimisation using Evolutionary Algorithms, Wiley, 2001.

[3] I. H. Witten, E. Frank. Data Mining: Practical Machine Learning Tools and Techniques (2nd edition), Morgan Kaufmann, 2005.

EUSAS: European Urban Simulation for Asymmetric Scenarios

Aleš Tavčar[1], *Violeta Mirchevska*[2], *Matjaž Gams*[1]
[1] Department of Intelligent Systems, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Result d.o.o., Bravničarjeva 11, 1000 Ljubljana, Slovenia
e-mail: {ales.tavcar, matjaz.gams}@ijs.si, violeta.mircevska@result.si

ABSTRACT

The purpose of this paper is to give a short overview of the project EUSAS and the problems that need to be addressed. First, a short introduction explaining the problem for which the system will be developed is given. Next, a short overview of the system structure and the modules that compose the system is given. Finally, we will describe the scenarios selected to test and demonstrate the system.

1 INTRODUCTION

The aim of the EUSAS project is to develop, demonstrate and deliver a **new approach to mission analysis and training for low level units facing asymmetric threats in an urban environment**.

The need for such new capabilities comes from several factors. First, in recent years the military community has been increasingly involved in military operations on urban terrain in Kosovo, Iraq and Afghanistan. Since enemy military commanders strategically tend to concentrate combat operations within urban areas where their lower technical level is compensated by the difficulty for blue force to use their assets in such an environment. Second, the current military doctrines for urban operations are not mature. Indeed, since several decades military doctrines were built upon the fact that urban areas should be avoided for combat operations due to the risk of higher friendly casualties.

To increase efficiency of military operations on urban terrain, mission analysis simulations are needed to adapt the military doctrine in such difficult environment and new tools need to be developed in order to successfully train soldiers in a safe virtual environment.

2 SYSTEM OVERVIEW

The overall system can be coarsely divided into two parts: the training part and the planning part. The training platform is based upon the serious game component and will be mainly used by policemen and soldiers which will allow them to take part in virtual conflicts to gain experience needed in real world crisis situations. The planning part, on the other hand, is more suitable for squad leaders to prepare the rules of engagement and adequate strategies that could be used during the actual mission. It is important to emphasise that the system will not make any decisions about which strategy should be used, but it should be regarded only as an advisory tool.

By appropriate use of **data farming** for mission analysis and **serious game** technology for mission training, both using **highly realistic and reusable models of human behaviour,** the EUSAS project will provide the following enhancements up to a level 4 of technology readiness level (TRL):

- Mission analysis and training capabilities in a virtual environment with high level of fidelity (taking into account physical, emotional, social, cultural, environmental aspects) and large variety of behaviours modelled,
- Automatic learning to ease human behaviour modelling digitisation into applications and adaptation of existing models,
- Evaluation and optimization of rules of engagement and military procedures,
- Verification, validation and accreditation capability of human behaviour models through qualitative and quantitative evaluation.

Figure 1 illustrates the overall system structure and the information flow between components.

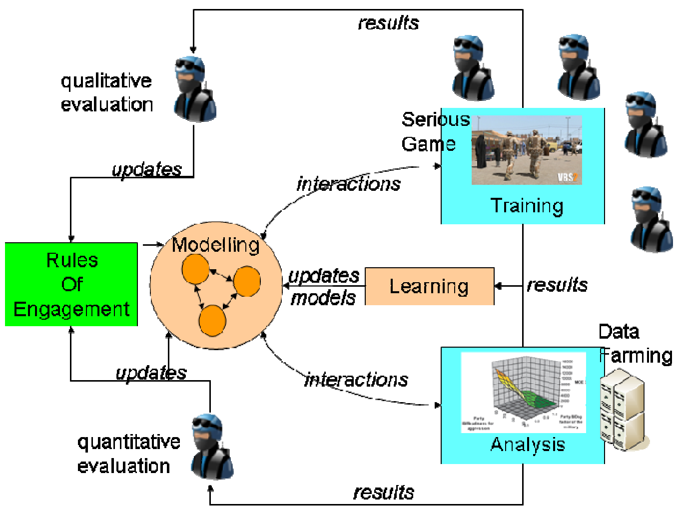


Figure 1: EUSAS system structure

In the next subsections a more detailed description of the modules, that are part of the EUSAS system, is given.

2.1 Serious game

The training part of the system will use a serious game approach which provides a 3D virtual environment based on a first person shooter game. The serious game platform should also provide advanced mission editing and gunnery simulation. The simulation environment selected for the training part is VBS2.

VBS2 is a simulation tool adapted to experiment when immersion in the situation is required. The VBS2 simulation architecture is similar to most modern multi-user first person shooter games and includes artificial intelligence, physic engine (based on Nvidia PhysX) and terrain procedural rendering.

The soldiers participating in the training part will play the defined scenario and during the game the system will collect relevant information about agent movements, actions, agent features, etc. Based on the features collected from the game the learning component will discover some behaviour patterns displayed by the soldiers during the game. A more detailed description of this component will be provided in one of the next subsections.

2.2 Data Farming

Data Farming is the process of using a high performance computer or computing grid to run a simulation thousands or millions of times across a large parameter and value space.
In EUSAS data-farming will be performed using a multi-agent simulation environment. The behaviour of soldier agents in the simulation will be defined in the behaviour models constructed during the learning part. The models describing the behaviour of civilians will have to support some sort of parameterization in order to provide a mechanism for changing the behaviour during the data farming part.
These type of models are necessary to execute multiple runs of the simulation models for different parameters in order to adequately cover the required parameter space. Such a large number of simulation can be distributed to a grid of computers to improve the overall simulation time proportionally to the number of machines available for the data farming task.
After the data farming part the system should produce outputs and useful information in the form of charts, rules or descriptions, which would help squad leaders to choose the best strategy to employ during the actual mission.

2.3 Learning module

The learning module will provide methods for the analysis of human (soldier) and agent behaviour from learning data.

What we want to achieve is the efficient extraction of relevant activity patterns. Extraction of representative activity patterns refers to the problem of determining the sequence of actions agents take in order to achieve a particular goal. For this task we have decided to use the algorithm MASDA

MASDA

Multi-Agent Strategy Discovering Algorithm (MASDA) (Bežek 2006) is an algorithm for discovering strategic behaviour developed at the Department of Intelligent Systems at the Jožef Stefan Institute. MASDA incorporates an abstraction process, which enables it to identify behaviour that is a result of an agent following a strategy and not just a sequence of actions the agent takes in response to local environment changes. The algorithm translates multi-agent action sequences and observations of a dynamic, complex and multivariate world state into a graph-based and rule-based strategic representation. By using hierarchically ordered domain knowledge the algorithm is able to generate strategic descriptions and corresponding rules at different level of abstraction.
The algorithm was applied to the RoboCup domain (Kitano et al. 1997) and in the next paragraph the main steps of the algorithm are described using screenshots obtained while inducing a team strategy in this domain.

The algorithm consists of three steps that are shown on Figure 2: data pre-processing (I.), graphical description creation (II.) and symbolic description learning (III.). Graphic descriptions visually present multi agent activity, while symbolic descriptions present important characteristics of multi-agent interaction in a human-comprehensible way.

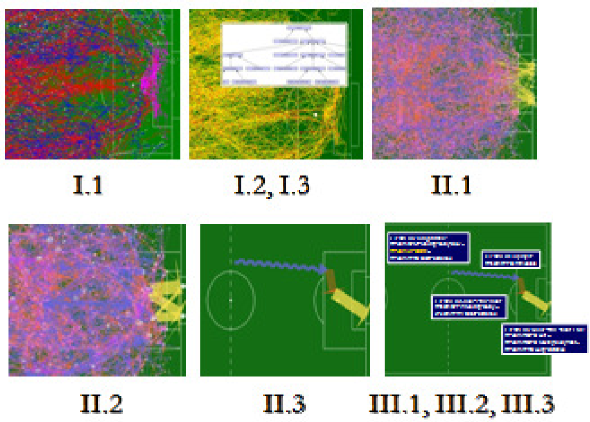


Figure 2: MASDA processing steps

MASDA was able to successfully create human comprehendible strategic descriptions in the form of graphical arrows and related strategic rules from basic observations of a team playing soccer (11 vs. 11). Moreover, the algorithm was able to construct a set of executable rules by observing agents play 3 vs. 2 Keepaway (soccer) and execute the play similarly to the observed team.

2.4 Behaviour Models

Behaviour models stored in this component will be used to simulate the behaviour of two types of agents: soldiers and civilians.
Behaviour of agents that control soldiers and policemen during data farming will be updated using information gained by observing real soldiers or policemen play the serious game. At some point MASDA will be triggered to generate an appropriate agent's behaviour.

The second type of models will be used to simulate behaviour of individuals and crowds.
The calibration (creation and update of behaviour models) of the agents who control civilians (during training and data farming) will be done experimentally employing the Data Farming methodology.

In different sociological theories, several models with different levels of abstraction exist in order to better explain complex human states and behaviours. The same principle will be used in EUSAS. Three different type of models that model human behaviour at different levels will most probably be used. Micro level models will concentrate on behaviour of individuals, meso and macro level models, on the other hand, will try to describe crowd's behaviour and capture crowd specific properties. However it is still to define if all these models will be used by the system.

The agent models that will be used in EUSAS will be based on the reference models MPECS (Schmidth 2005). These models provide capabilities for object-oriented model specification. Its application area is in the field of agent-based simulation, especially for models deriving from social studies.

3 SCENARIOS

The following section briefly describes the proposed scenarios and measures of effectiveness for crowd control. The scenarios represent a reference point for the definition of environment requirements, computer models and the actual simulations that will run during the data farming part. For this two different scenarios were proposed: a civilian and a military. The civilian scenario frame is based on a variety of records from the Gothenburg riots, taking place in Sweden in 2001. The second scenario, on the other hand, is not based on real events but was proposed during the expert workshop and it incorporates experiences and expertise from ISAF forces in Afghanistan.

We will start with briefly describing the civilian scenario. It is intended for police or other security forces operating in urban environments and dealing with civilians.

The hypothetical background for this scenario follows a long political dispute between two main opposing parties (A and B), each supported by extremist fractions. Party A organizes a celebration or meeting. After this event the supporters of party A continues to celebrate in a centrally located park during the evening. Violent supporters of party B are showing up later in the evening, sparking a large fight which turns against control forces in the area.

Figure 3 depicts member of each parties (dark yellow and cyan colour circles), as well as police officers (blue colour circles).
Figure 3: Graphical representation of the civilian scenario

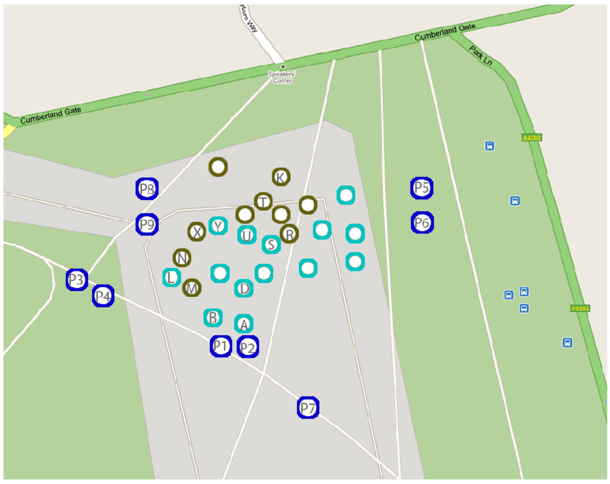


Figure 3: Graphical representation of the civilian scenario

The main chain of actions for this scenario is composed of five parts:

- The group is gathering in the park
- Confrontations are building-up between individuals
- Groups of people are starting to fight each other
- Full scale escalation demanding crowd containment

The military scenario is a generic scenario based on known circumstances and challenges in the current mission environment in Afghanistan.
The overall mission of the depicted coalition forces is to provide and maintain security in specified areas of responsibility which also includes facing insurgencies.
The insurgents are usually challenging the host nation government and the coalition by asymmetric threats.

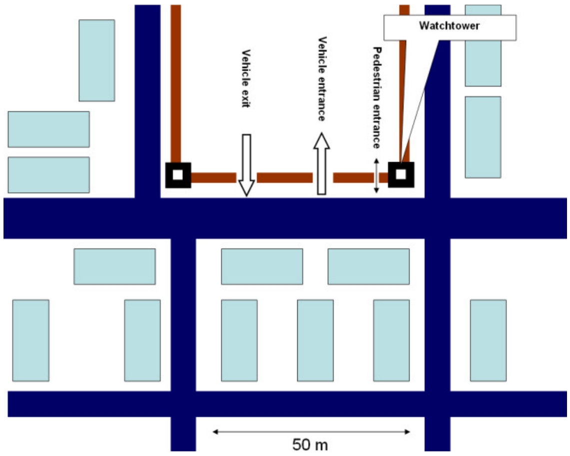


Figure 4: Detailed graphical representation of the military scenario

One of the possible vignettes for this scenario is a turmoil in front of the pedestrian entrance. A limited number (up to 20) of people is waiting for entrance to the operation base in order to apply for a job. At some point a confrontation between daytalers and employees of the base has started. Ending in a full-scale insurgency against the armed forces.

The described vignettes are just an example of what can happen in the given scenarios frameworks. These vignettes can be altered according to specific needs.

3 CONCLUSIONS

In this paper we have presented a short overview of the system that will be developed for the EUSAS project. The most important part of the system is the learning module that will provide methods for updating behaviour models by observing real soldiers play in a virtual environment. For this task the MASDA algorithm was selected as it provides functionalities for pattern extraction from observations of a multi-agent system.

MASDA is the only algorithm of this kind, as far as we know, that is in principle domain independent, uses no domain knowledge and is still able to deduce meaningful rules and is capable of behaviour cloning.

However, further research has to be done to determine, if MASDA will be able to learn executable models or just the descriptive ones. In addition, we need to determine if MASDA can be adapted to exploit behaviour libraries (i.e., collections of behaviour patterns that encode the complete repertoire of expected behaviour) and how can the observed agent behaviour be matched with library prototypes.

To conclude, the tools that EUSAS will provide will have a great impact on the way mission planning and analysis will be performed in the future. The methods that will be included in the learning part will most probably be close to the margins of what is feasible to achieve in behaviour modelling.

ACKNOWLEDGEMENTS

This work was partially supported by the European Union, European Social Found.

References

[1] Bežek, A. (2006): Avtomatsko modeliranje večagentnih sistemov, *Doctoral dissertation*, University of Ljubljana, Faculty of Computer and Information Science, In Slovene

[2] Bežek, A., Gams, M., Bratko, I. (2006): Multi-agent strategic modeling in a robotic soccer domain, *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems,* pp. 457 – 464

[3] Bežek, A., Gams, M. (2005): From Basic Agent Behavior to Strategic Patterns in a Robotic Soccer Domain, Informatica vol. 29, pp. 461-468

[4] Kitano, H., Tambe, M., Stone, P., Veloso, M. (1997): RoboCup synthetic agent's challenge, In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)

[5] Schmidt, B. (2005): Human Factors in Complex Systems

The Modelling of Human Behaviour, In Proceedings of the 19^{th} European conference on modeling and simulatoin

GOVORNI KORPUS ZA POTREBE IZVAJANJA RAZISKAV IN OVREDNOTENJA REZULTATOV NA PODROČJU RAZPOZNAVANJA GOVORCEV

Tomaž Šef, Robert Blatnik
Odsek za inteligentne sisteme, Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana, Slovenija
tomaz.sef@ijs.si, robert.blatnik@ijs.si

POVZETEK

Članek predstavlja postopek izgradnje specializiranega govornega korpusa za potrebe razvoja in ovrednotenja sistemov razpoznavanja govorcev. Posebna pozornost je namenjena raziskovanju vpliva različnih prenosnih kanalov na zanesljivost razpoznavanja govorcev.

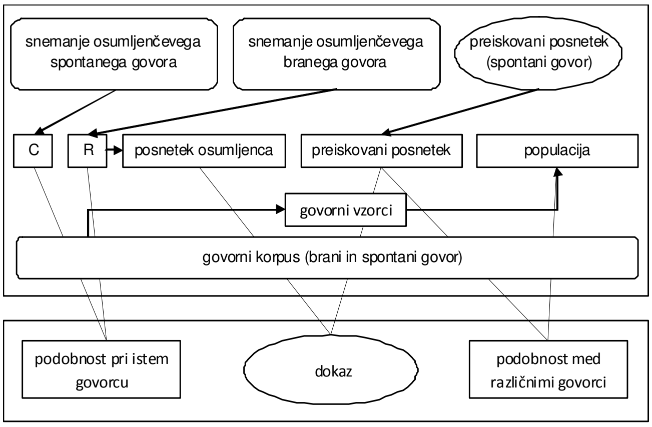
1 UVOD

Razpoznavanje govorcev je biometrični proces označevanja, katerega cilj je identifikacija/verifikacija ljudi na podlagi njihovega glasu. Analiza očesne šarenice in prstnega odtisa sta druga dobra primera biometričnih metod identifikacije oseb, kjer se testni vzorec neposredno primerja z znano predlogo. Na identifikacijo glasov pa lahko gledamo še iz drugega zornega kota. Podobno kot pri razpoznavanju obraza in grafološki analizi pisave imamo opraviti z variabilnostjo signala, kar vnaša v proces identifikacije dodatno stopnjo kompleksnosti [1].

Uspešno obvladovanje (stvarnega) komercialnega in forenzičnega razpoznavanja govorcev nujno vključuje **obravnavo variabilnosti v govoru**: zunanji pogoji se (običajno) hitro spreminjajo ali so močno degradirani, proces pridobivanja govornega signala ni vedno pod nadzorom itd., vključno z velikim obsegom virov variabilnosti govornega izražanja.

Za namene določanja, analize in meritve učinkov oz. posledic najpomembnejših virov variabilnosti, ki jih najdemo v (dejanskih) komercialnih in forenzičnih aplikacijah, in proučevanja njihovega vpliva na sisteme avtomatskega razpoznavanja, je bilo potrebno **razviti specifično govorno bazo in doseči nadzorovane pogoje za razpoznavanje govorcev v slovenskem jeziku**. Na njeni osnovi je mogoče ovrednotili razlike v snemalnih pogojih, ki nastopajo v praktičnih aplikacijah, kompenzirati njihove učinke ter kvantitativno opredelili negotovost oz nezanesljivost, ki je posledica teh dejavnikov. Snemalni pogoji lahko vplivajo na vse vrste metod razpoznavanja govorcev.

Podatkovno vodena Bayesova metoda za avtomatsko razpoznavanja govorcev zahteva poleg preiskovanega posnetka (oz. sledi) uporabo še treh baz izgovarjav (Slika 1): referenčno govorno bazo osumljenca (R), ki služi izdelavi statističnega modela njegovega glasu, kontrolno govorno bazo osumljenca (C), ki služi ocenjevanju notranje variabilnosti glasu osumljenca in govorno bazo potencialne populacije, ki vsebuje takšne posnetke glasov, da nobeden naključno izbran posnetek iz te baze ni izgovorjen s strani iste osebe, kot je preiskovani posnetek (oz. sled). Kakršno koli neujemanje govornih baz zaradi okoliščin pri prenosu govornega signala, vrste snemalne naprave, šuma, jezikoslovne vsebine in trajanja posnetkov lahko vpliva na zanesljivost dobljenih rezultatov [2].



Slika 1: Shematski prikaz uporabe različnih baz izgovarjav pri avtomatskem razpoznavanju govorcev.

2 RAZLIKE MED »STVARNIMI« IN »LABORATORIJSKIMI« POGOJI

Medtem ko pri nekaterih aplikacijah lahko ocenimo, privzamemo ali predvidimo delovne pogoje, pri večini komercialnih aplikacij razpoznavanja govorcev in skoraj vseh metodah razpoznavanja govorcev v forenzične namene to ni možno [3].

Pojem »stvarni pogoji« uporabljamo kot nasprotje »**laboratorijskim pogojem**«, ko ne moremo nadzirati, pričakovati ali predvidevati pogojev v katerih se bodo pridobili posamezni govorni posnetki. Celo več; obtoženec običajno ne želi korektno sodelovati in skuša ovirati ali preprečiti pridobitev kakršnihkoli zanj obremenilnih informacij.

Zaradi »stvarnih pogojev« pridobivanja posnetkov je govorni signal bolj spremenljiv oz. variabilen. Vire variabilnosti govornega signala lahko razvrstimo v naslednje kategorije [1]:

(i) *svojske variabilnosti govornih signalov istega govorca*: vrsta govora, staranje, časovni presledek med dvema posnetkoma, narečje, žargon, socialni status, čustveno stanje, vpliv omamnih sredstev itd.,

(ii) *izsiljene oz. umetne variabilnosti govornih signalov istega govorca*: »Lombardov« učinek, stres zaradi zunanjega vpliva, »cocktail-party« učinek itd.,

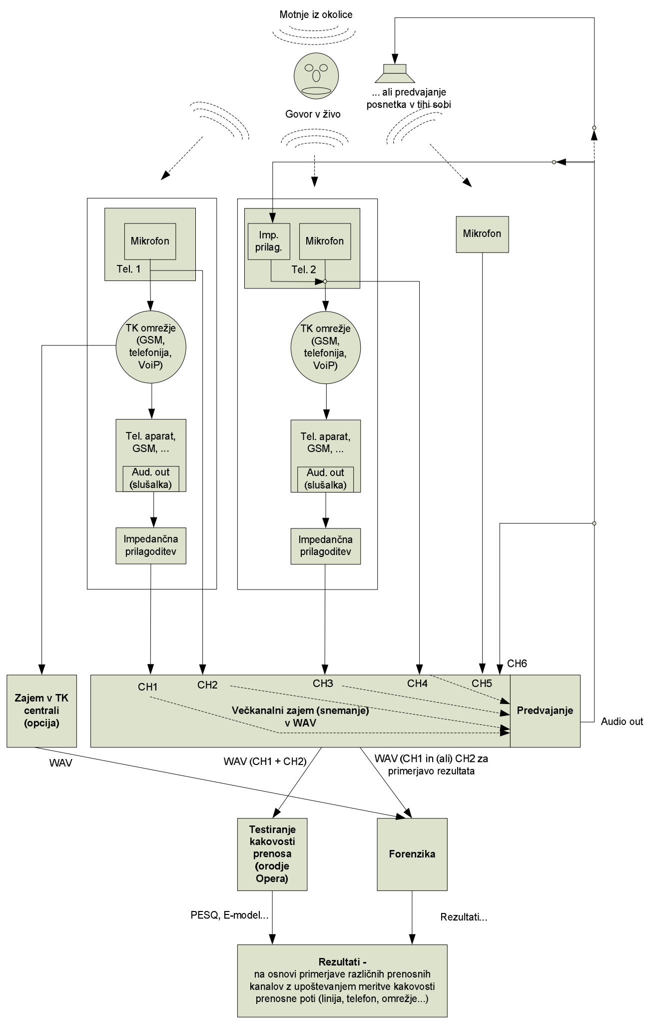
(iii) *zunanja variabilnost odvisna od kanala*: tip telefona ali mikrofona, fiksna/mobilna telefonija, komunikacijski kanal, pasovna širina, dinamični obseg oz. razpon, električni in akustični šum, odmev, popačenje itd.

Forenzični pogoji so doseženi, ko se dejavniki variabilnosti, ki predstavljajo t.i. »stvarne pogoje«, pojavljajo brez kakršnegakoli principa, pravila ali norme. Lahko so konstantni preko celotnega klica ali pa se hipoma pojavijo ali izginejo; na celoten proces vplivajo povsem nepredvidljivo [5, 6].

3 GOVORNI KORPUS V SLOVENSKEM JEZIKU

Zasnovali smo govorni korpus različnih govorcev (trenutno se v zbirki nahajajo glasovi preko 30 oseb moškega spola). Govorni signal smo zajemali paralelno preko več različnih kanalov (različni mikrofoni, različne telefonije: VoIP, PSTN, ISDN, GSM, VoWLAN). **Osnovne zahteve govornega korpusa so:**

1. **Snemanje v več delih**:
   - brani govor nasproti spontanemu govoru,
   - branje besedil z različnimi hitrostmi,
   - neposredno snemanje in snemanjem preko telefona,
   - uporaba različnih mikrofonov in telefonskih aparatov,
   - različna narečja govorcev (lahko se spreminja že pri istem govorcu glede na to ali besedilo bere ali pa spontano govori),
   - stalne izgovarjave besedil za vse govorce preko vseh delov posnetkov proti specifičnim izgovarjavam vsakega govorca znotraj posameznega posnetka.
2. **Fonološka in zlogovna uravnoteženost**:
   - govorna baza je fonološko in zlogovno uravnotežena tako, da posnema frekvenco pojavljanja fonemov in zlogov v govorjenem slovenskem jeziku.
   - »standardno« frekvenco pojavljanja posameznih enot smo izmerili na govornem korpusu s približno 15.000 besedami.
3. **Starostna porazdelitev**:
   - ustrezno starostno porazdelitev govorcev v podatkovni bazi smo določili ob upoštevanju socioloških dejavnikov povezanih z uporabo omenjenih tehnologij,
   - enakomerna starostna porazdelitev vedno ne odraža dejanske starostne porazdelitve uporabnikov v konkretni komercialni aplikaciji,
   - pri forenzičnih aplikacijah enakomerna starostna porazdelitev osumljencev storitve kaznivih dejanj ne ustreza dejanskemu stanju,
   - večjo utež smo dali starostni skupini med 25 in 45 leti.
4. **Časovni interval med posameznimi posnetki:**
   - variabilnost parametrov med posameznimi posnetki je pomemben dejavnik, ki ga je potrebno upoštevati pri izgradnji govorne baze namenjene razpoznavanju govorcev,
   - med posameznimi snemanji je bil vsaj 14 dnevni zamik.



Slika 2: Okolje za snemanje govornih posnetkov in izvajanje meritev kakovosti zvoka.

Zbiranje podatkov in snemanje govornih posnetkov:

- »neposredno« snemanje preko mikrofonov je potekalo v mirni sobi in je bilo ustrezno nadzorovano,
- pri vsakem »neposrednem« snemanju smo uporabljali (vsaj) dva simultana kanala: na enem kanalu smo ves čas snemanja uporabljali isti mikrofon, na drugem kanalu pa je bil pri vsakem snemanju drugačen mikrofon,
- uporabili smo neusmerjeni mikrofon (postavljen na mizo v razdalji okrog 30 cm od govorca), dinamični mikrofon (postavljen na mizo v razdalji okrog 10 cm od govorca) in dinamični mikrofon na slušalkah.
- pri vseh snemanjih so bili vsi mikrofoni in telefonske linije priključeni neposredno na ustrezen računalniški sistem, kjer se je govor zapisoval v digitalni obliki s frekvenco vzorčenja 44,1 kHz; vsaka stran telefonske povezave je posneta na svoj kanal,
- vsi posnetki so avtomatsko identificirani in označeni glede na jezik, spol, prenosni kanal ipd.,
- sestavni del govornega korpusa so posnetki preko PSTN, ISDN, GSM in VoIP telefonije,
- čas trajanja posameznih posnetkov je od ene do petih minut.

Večji del posnetkov so govorci hkrati govorili v dva namizna mikrofona, v prostoročni mikrofon VoIP telefona ter v GSM telefon in slušalko klasičnega analognega PSTN telefona. Na ta način smo isti govor posneli preko več sočasnih prenosnih kanalov, kar nam omogoča analize vplivov kanala na istem izvornem govornem signalu.

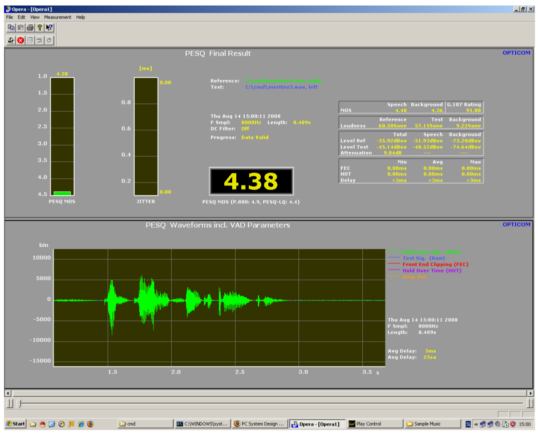
Govorci so pod nazorom operaterja snemanja govorili na tri načine: spontani govor, pogovor in branje. Vsak način govora smo posneli v dolžini najmanj dveh minut. Vsako snemanje smo pričeli z branjem teksta nekega članka, pri čemer se je govorec lahko vsaj približno privadil na snemalne naprave. Po branju smo v pogovoru, ki smo ga snemali, govorca pripravili na ustrezno temo, ki mu je blizu. Pri tem je bil na posnetkih slišen tudi govor sogovornika, ki je vodil pogovor in snemanje. Na ta način smo skušali zagotoviti čim bolj naraven in sproščen način govora. V nadaljevanju smo posneli še spontani govor, ki je v obliki monologa o določeni temi trajal prav tako 2 minuti. Izkazalo se je, da govorcu močno olajšamo spontani govor v obliki monologa, če se le ta smiselno in tematsko navezuje na pričeti pogovor v prejšnjem delu snemanja, saj se ljudje praviloma počutijo nelagodno, ko morajo pred določeno osebo več časa nepripravljeni govoriti o poljubni temi.

Za organizirano snemanje govora ljudi je potrebno veliko priprav in organizacijskih naporov. Nujni so vnaprej pripravljeni scenariji, po katerih se snemanje izvaja. Prav tako je potrebno ljudi ustrezno pripraviti na snemanje, saj prisotnost različnih snemalnih naprav in operaterja lahko znatno vpliva na naravnost govora.

4 MERJENJE KAKOVOSTI PRENOSA GOVORNEGA SIGNALA

Kakovost zvoka v mobilnih omrežjih (GSM, UMTS), brezžičnih lokalnih omrežjih (VoWLAN) in brezžični telefoniji (DECT) je močno odvisna od pogojev znotraj samega radijskega omrežja in od zunanjih pogojev, kot so šum iz ozadja in nelinearni akustični odmev. Za izboljšave kakovosti zvoka so v mobilnih in brezžičnih telefonih vgrajeni različni dušilci šuma in izločevalniki odmeva, ki pa se razlikujejo glede na procesno moč aparatov in kapaciteto baterije. Poleg tega na kakovost zvoka in zmanjševanje šuma močno vpliva uporaba različnih kodekov. Kakovost prenosa zvoka v mobilni in brezžični lokalni telefoniji je močno odvisna tudi od pokritosti signala oziroma števila baznih postaj (GSM, UMTS) oziroma dostopnih točk (WLAN, DECT).

Za potrebe analize zunanjih vplivov na kakovost prenosa govornega signala smo zgradili testno okolje za zajem govora in objektivno merjenje kakovosti prenosa govora (Slika 2). Izvedene so bile meritve kakovosti prenesenega govornega signala v različnih telefonijah in pri različnih pogojih delovanja omrežja (obremenitve omrežja, različni nivoji RF signala v WLAN itd.). Meritve smo izvajali po standardizirani metodi za objektivno ocenjevanje kakovosti govora PESQ (angl.: Perceptual Evaluation of Speech Quality) [4].



Slika 3: Rezultat meritve PESQ z orodjem Opera na dveh telefonih v fiksnem omrežju.

Ustreznost testnega okolja smo najprej preverili z meritvami prenosa zvoka med fiksnima telefonoma. Rezultat PESQ MOS (angl. Mean Option Score) je bil pričakovano dober in se je gibal od 4,24 do 4,38; odvisno od izbrane referenčne datoteke. Zakasnitev je bila prav tako minimalna in je znašala ca. 3 ms. Na sliki 3 je prikazan izpis meritve PESQ na dveh telefonih v fiksnem omrežju.

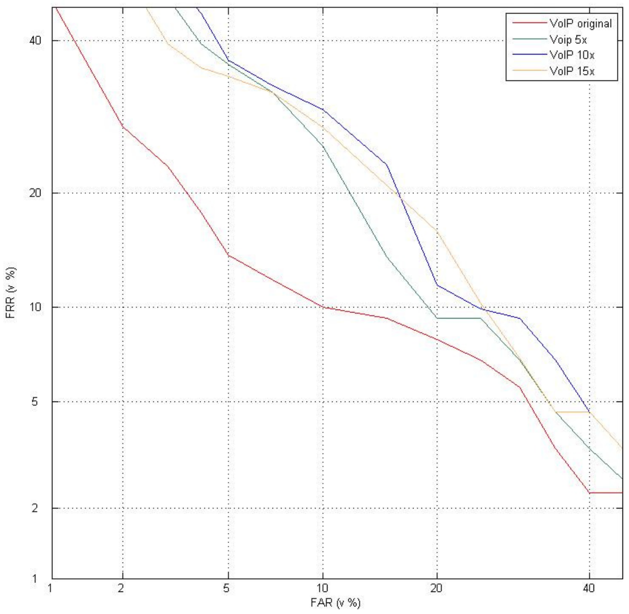
Nadalje smo izvedli meritve pri klicni povezavi iz fiksnega telefona v omrežju Centrex na GSM mobilni telefon. Pri tem smo naleteli na težavo, saj je na IJS uporabljen GSM vmesnik za klice iz fiksnega omrežja v mobilna omrežja, ki je v meritvah povzročal velik odmev in motnje na merjenem GSM telefonu, zato so bili rezultati kakovosti zelo slabi. Rezultat meritev PESQ je bil MOS 2,58, kar je povsem nezadovoljivo. Hkrati pa je bila izmerjena zakasnitev 206 ms za GSM omrežje pričakovana.

Sledile so izvedbe meritev pri povezavi dveh mobilnih telefonov v omrežju GSM. Pri zajemu datoteke smo naleteli na oscilacije na avdio vhodu telefona, kar je imelo za posledico nekoliko slabši rezultat MOS, ki je bil v našem primeru 3,18. Zakasnitev 200 ms pa je bila v mejah pričakovane.

Pri razvoju sistemov za razpoznavanje govorcev je v zadnjem času največji poudarek na zagotavljanju čim večje neobčutljivosti na vplive prenosnega kanala. Največji izzivi so v primerih, ko izvajamo učenje sistema na podatkih, ki so pridobljeni preko ene vrste telefonije, testiranje sistema pa se izvaja na podatkih, ki so posneti preko druge vrste telefonije oziroma neposredno preko mikrofona. Po dosedanjih izkušnjah se je izkazalo, da so rezultati v primerih mešanih pogojev znatno slabši od rezultatov, pridobljenih v enakih pogojih prenosnega kanala [2].

5 VPLIV KAKOVOSTI POSNETKOV NA SAMODEJNO RAZPOZNAVANJE GOVORCEV

Izvedli smo primerjavo uspešnosti sistema za samodejno razpoznavanje govorcev v odvisnosti od kakovosti prenosa govornega signala. Rezultati meritev kakovosti govora v telefonijah in uspešnosti razpoznavanja govorcev so pokazali, da je na osnovi objektivne ocene kakovosti prenosa govora v telefoniji mogoče predvidevati tudi uspešnost sistema za samodejno razpoznavanje govorca, ki za razpoznavanje uporablja govorni posnetek glasu, ki se prenaša preko določene telefonije.



Slika 4: Krivulje DET avtomatskega sistema za razpoznavanje govorcev pri različnem številu sočasnih VoIP kanalov (prenos govora v VoWLAN).

Pri prenosu govora v VoWLAN smo ugotovili, da se napaka razpoznavanja EER (angl. Equal Error Rate) sistema PerSay [7] za razpoznavanje govorcev povečuje s povečevanjem prenosa sočasnih VoIP kanalov, in sicer od 10% za eno VoIP povezavo na WLAN kanalu do skoraj 20% za 20 sočasnih VoIP povezav. To dejstvo pripisujemo občutnemu poslabševanju kakovosti prenosa govora v VoWLAN telefoniji, ki je dala rezultate MOS 3,5 za eno VoIP povezavo in do 2,5 za 20 sočasnih VoIP povezav na WLAN kanalu (Slika 4). Vendar pa se napaka razpoznavanja EER ne povečuje v nedogled ampak se pri 15 sočasnih VoIP povezavah na WLAN kanalu ustali.

6 SKLEP

Predstavljena je prva različica govornega korpusa za potrebe izvajanja raziskav in ovrednotenja rezultatov na področju razpoznavanja govorcev. Govorni korpus se še dograjuje; predvsem z novimi glasovi.

Pri snemanju govornega korpusa se je izkazalo za pomembno, da imamo pod nadzorom tudi kakovost prenosa govornega signala preko različnih prenosnih kanalov oz. telefonij, saj se le ta lahko s časom (odvisno od okoliščin) tudi spreminja.

Na primeru VoWLAN telefonije smo testirali vpliv kakovosti prenesenih posnetkov na samodejno razpoznavanje govorcev. Rezultati so skladni s pričakovanji.

V nadaljevanju nameravamo izvesti meritve še za druge telefonije in njihove medsebojne kombinacije ter zgraditi računalniški model, ki bo lahko napovedoval zanesljivost avtomatskega sistema razpoznavanja govorcev v odvisnosti od kvalitete prenosnih poti in vrste uporabljenih telefonij.

Literatura

[1] J. Ortega-Garcia, J. Gonzalez-Roidriguez, V. Marrero-Aguiar, AHUMADA: A large speech corpus in Spanish for speaker characterization and identification, Speech Communication 31, 2000, str. 255-264.

[2] A. Alexander, Forensic Automatic Speaker Recognition Using Bayesian Interpretation and Statistical Compensation for Mismatched Conditions, *doktorska disertacija*, Lausanne, EPFL, 2005.

[3] T. Šef, P. Baucon, Sodno izvedenstvo in razpoznavanje (identifikacija) govorcev v kazenskem postopku, *Pravosodni bilten*, Ljubljana, 2/2007.

[4] PESQ: An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech CODECs, ITU-T Recommendations, str..862, 2001.

[5] P. Rose, Technical forensic speaker recognition: evaluation, types and testing of evidence, *Computer Speech and Language*, 2005.

[6] P. Rose, *Forensic speaker Identification*, Taylor & Francis, 2002.

[7] http://www.persay.com/pdf/SPID_V6_DataSheet.pdf

SEARCHING FOR MEANINGFUL MODELS IN MACROECONOMIC DOMAIN

Vedrana Vidulin, Matjaž Gams
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: vedrana.vidulin@ijs.si

ABSTRACT

Models constructed with standard data mining and machine learning algorithms can contain irrelevant parts, consequently leading to wrong conclusions. This paper presents an approach to domain analysis based on human-computer interaction. The approach facilitates human examination of a model constructed by a computer, the detection of irrelevant parts and their correction. The method is tested on the macroeconomic problem of how research and development sector influences economic welfare of a country. The results show that with the help of our method improved models are easily discovered.

1 INTRODUCTION

Data mining (DM) and machine learning (ML) systems are mostly oriented towards automation and less on human understanding of how the algorithm came at the solution or why the solution is the best one [1]. Without such explanations user has incomplete information to make conclusions. On the other hand, if DM system is able to explain its reasoning, user can provide background knowledge, which can improve the results of further DM steps [2].

In certain cases, DM and ML algorithms produce models that contain errors. A typical example of an error is overfitting. This happens due to pathologies in induction algorithm [3], noise or missing values in data. To correct errors, different automatic approaches were suggested, e.g. post-pruning for decision trees [4] and correction by means of extreme value distribution for rules [5]. However, experiments showed that in spite of automatic corrections, as post-pruning, certain errors persisted (the candidate model in Figure 2). Therefore, we argue that user should have the key role in DM process since he/she can recognize problematic models based on experience and explanations, and correct the models with the help of Human-machine data mining (HMDM) method we propose. Through the analysis and corrections user finally constructs the correct mental model of a domain.

The method is empirically tested on a macroeconomic domain, through the analysis of relations between research and development (R&D) sector and economic welfare of a country. The goal of analysis is to find such relations for which are likely to be the promoters of economic welfare. With the help of HMDM method we were able to detect important relations which are acknowledged in economic literature [6].

2 RELATED WORK

To establish the successful interaction, system should first explain its reasoning in a human-understandable form. Literature covers two types of explanations. Static are mostly oriented towards explaining how an instance is classified by models constructed with statistically and mathematically oriented algorithms such as Naïve Bayes, SVM and kNN [7][8]. Interactive are conceived, not only to explain, but also to collect corrective feedback about the classification [2][9]. We adopted interactive explanations. However, they are generally oriented towards improving prediction without the need for user to understand the underlying domain model. In contrast, our approach is oriented towards domain analysis, where user should understand the model to construct the correct mental model of a domain.

Once user understands the system, he/she can provide the corrective feedback. Feedback can be incorporated by introducing new examples [10], labels of examples [11] or background knowledge. The advantage of the last is that system quickly converges towards better model. Such approaches are implemented as constraints, co-training of ML model and model representing user's view of data [2], and manual construction of a model with the help of data visualization techniques [12]. Our approach belongs to the last group. However, while the presented approaches focus on correction of a single model to improve prediction, we went beyond the single model for the purpose of domain analysis.

3 HUMAN-MACHINE DATA MINING

The basic idea of our approach is to construct a large number of models in a specific manner to extract those that are relevant, i.e. meaningful and correct. While DM methods are able to construct a large number of models, they do not always detect and correct errors within those models. On the other hand, humans alone cannot analyze any meaningful share of all generated models, but are successful at detecting errors. Therefore, by cooperating relevant models can be discovered with more success.

Our approach is based on two assumptions: first, only several models best represent the domain; second, DM method will discover relevant models in (sub)optimal form and they will be recognized by human. Human examines and evaluates the result of DM, and suggests the future direction of search in the way seemingly most promising from the previous search attempts.

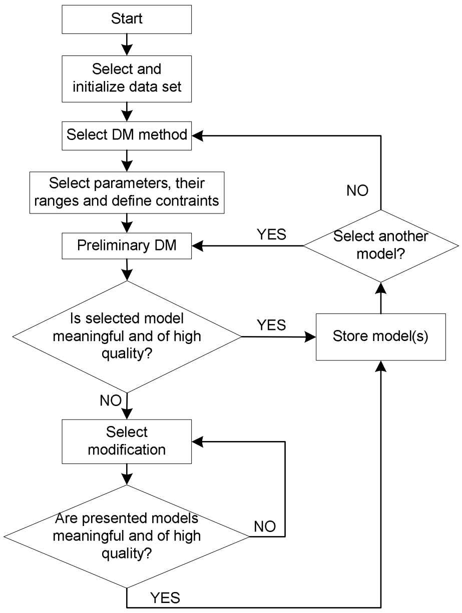


Figure 1: *The HMDM interaction model.*

The procedure is similar to the exploratory and confirmatory analysis within multivariate analysis [13]. The goal of the former is to find nonrandom patterns in data to form hypotheses interesting for further study. Similarly, the first heuristic we apply is to examine models constructed with different algorithm's parameters to detect possibly relevant models. The goal of the latter is to seek specific answers to the defined hypotheses. Similarly, when a candidate model is found, we apply several heuristics to cross-check the relevance of the observed and similar models.

The interaction model of HMDM method is presented in Figure 1. At the beginning user selects a data set and one or more DM methods. In this paper we used J48 algorithm from Weka [14], which constructs decision trees. Then, by selecting parameters and ranges of their values, user defines variations in constructed models. By defining constraints user further defines the type of models that are going to be excluded from the search. The role of the preliminary DM is to construct models specified through parameters and constraints, and to remove duplicate models. The models constructed by preliminary DM are presented to user in decreasing order of quality. User observes and selects one or several candidate models. If user is satisfy with the model, he/she can store it. In contrast, modifications are applied. There are three modifications: remove attributes, add attributes and expand relevance indicator.

Remove attributes. Since attributes are the basic building blocks of models, by meaningful manipulation with them we can make conclusions about the model's relevance. For example, when a relevant attribute is removed and the model reconstructed on the reduced data, the new model is of lower quality, and vice versa. When model contains irrelevant attributes, this is an indication that it should be corrected. To present the relevance of attributes we devised an explanation named *remove graph*. Each node in a graph (a relevance indicator) represents a removed attribute or a set of removed attributes. By clicking at the node, user can view the model obtained after removal, which is at the same time the proposal for the candidate model correction. The graph can be constructed automatically with the procedure similar to wrapper approach to attribute selection [15]. The difference is that the primary goal of wrapper approach is to find a single best attribute set to construct a model for prediction purposes, while our goal is to generate alternative models for the purpose of finding those that are relevant.

Add attributes. The relevance of an attribute or a set of attributes is isolated by the means of adding attributes to an empty set, this time constructing an *add graph*. When subsampling is used, as in the case of cross-validation, addition of an attribute can cause an increase in quality simply by appearing in a model constructed on one of the folds. Since we search for models that improve both in quality and meaning, dependent upon structure, we account only for those attributes that cause both change in quality and structure. The graph is constructed in reverse order to remove graph, with the difference that a relevance indicator is added when model constructed after adding an attribute is of higher quality that the parent model and the attribute is contained within the model.

Expand relevance indicator. This tool enables user to select an attribute or a set of attributes to add to the remove/add graph.

4 MACROECONOMIC DATA ANALYSIS

4.1 Evaluation

The quality of a model is evaluated with 10 fold cross-validation and expressed through accuracy (ACC), corrected class probability estimate (CCPE) and Kappa statistic. ACC denotes the percentage of examples correctly classified by the model. CCPE reflects the significance of the model in comparison to all possible models constructed on the same data. Kappa denotes how much the model is better that a random model. In all three cases higher number indicates better model.

4.2 Analysis

Select and initialize data set. We composed a data set by extracting 48 numeric attributes for 167 examples/ countries from UNESCO (http://www.uis.unesco.org) and WIPO (http://www.wipo.int) databases. Attributes describe the performance, personnel and financial resources of R&D sector. The class is GNI per capita. GNI stands for Gross National Income and represents the total value of goods and services produced by the residents of a country, at home, as well as abroad [16]. The attribute was collected in discrete form from the World Bank (http://www.worldbank.org) database representing the official classification of countries into three income levels: low – 745 US$ or less (50 countries), middle – 746-9,205 US$ (79 countries), and high – 9,206 US$ or more (38 countries).

Define parameters and constraints. Two parameters of J48 were selected: minimal number of instances per leaf (MNIL) with values ranged from 2 to 15 and reduced error pruning (REP) with on/off options. The constraint we posed was that minimal number of relations/branches in a model that was set to 2. In this manner we eliminated those models that return majority class.

Preliminary DM. The preliminary DM constructed 20 decision trees, from which we selected the tree in Figure 2 constructed with parameters MNIL 12 and REP. The motivation for modification was an appearance of "Researchers – Female (FTE)" attribute in the candidate tree. This attribute, representing the position of females in research community, looks more like a consequence of the certain level of economic welfare than a likely promoter of economic welfare.

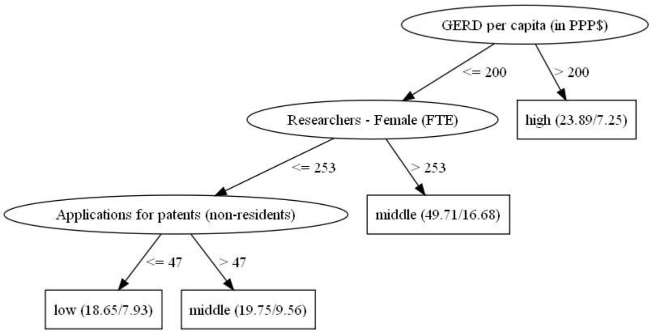


Figure 2: *The candidate tree.*

Remove attributes modification. Automatic construction of a remove graph resulted in the graph containing only two relevance indicators, showing that by not knowing "GERD per capita" and "Applications for patents (non-residents)", important information would be lost (the fall in all three quality estimates). GERD stands for Gross Domestic Expenditure on R&D, denoting expenditure on R&D performed on the national territory during the year [16]. PPP$ stands for purchasing power parity in American dollars. Statement of GERD in PPP$ allows fair comparisons between different countries. To reassure that "Researchers – Female (FTE) attribute is not important we removed it by expanding relevance indicator as can be seen from the first level of the remove graph in Figure 3. The first number within a relevance indicator is ACC, followed by CCPE and Kappa. The number in brackets that follow each quality estimate represents the difference in quality between the current indicator and its parent. The removal of female researchers attribute did not result in a considerable loss of information (ACC and Kappa did not change at all).

Next, we examined the trees within the relevance indicators. When any of the patent and female researchers attributes were removed the tree was obtained containing only the root node of the candidate tree having leaf "middle" instead of the right subtree. The tree indicates high importance of "GERD per capita" attribute. Further proofs of its importance are: a) when removed, it causes the highest fall in all estimates at the first level of the graph; b) when removed, another GERD attribute representing the same semantic category, only this time expressed as % of GDP (gross domestic product) takes its role; and c) when both GERD attributes are removed the significance falls considerably – fall in CCPE of 0.4032, i.e. 46% of the candidate tree's significance. Based on evidence, at this point we rank the important attributes according the level of their importance in following order: "GERD per capita", GERD as % of GDP" and patent attribute, and conclude that the candidate tree should be corrected.

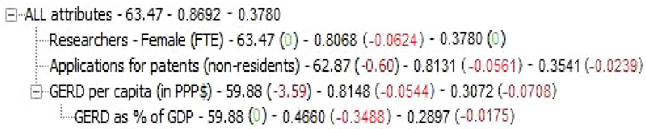


Figure 3: *The remove graph.*

Add attributes modification. The goal of this modification was to reexamine the importance of attributes from the previous step and to find suitable corrections of the candidate tree. An add graph was automatically constructed from the attributes in the candidate tree. Other relevance indicators were added during the analysis. In the add graph in Figure 4, underlined indicator indicates ACC higher than the candidate tree's ACC, while strikethrough indicator indicates that all attributes added are not present in the newly constructed tree. In general, the graph supported our conclusions about the important attributes. Further, two corrections were revealed. They were selected based on two criteria: first, at least two out of three quality estimates were better than those of the candidate tree; second, the total quality gain in comparison to the candidate tree, computed by summing the gains in all three estimates (ACC was divided by 100), was higher than -0.1000 (experimentally determined threshold). According these criteria, the tree could be slightly worse in quality if it is meaningful. However, high loss in quality indicates the tree unsupported by data. Although the tree might be meaningful, it is not acceptable under the quality criterion.

The first correction presented in Figure 5 shows the difference between the countries based solely on the level of investment into R&D. At the first glance, "GERD as % of GDP" subtree looks like an error. Analysis of the data, however, showed that the majority of "low" countries invest somewhere between 0.15 and 1.2% of their GDP into R&D, while the left side of the subtree only represents the special subgroups of "low" and "middle" countries with extremely low level of investment into R&D. This tree, together with other constructed trees, shows that countries from "high" group are best differentiated from the other two groups according the higher level of investment into R&D. In contrast, countries from “middle” group can be differentiated from “low” group according the a) higher level of investment into R&D, and b) higher number of applications for patents where the first applicant is a non-resident of a country (Figure 6).

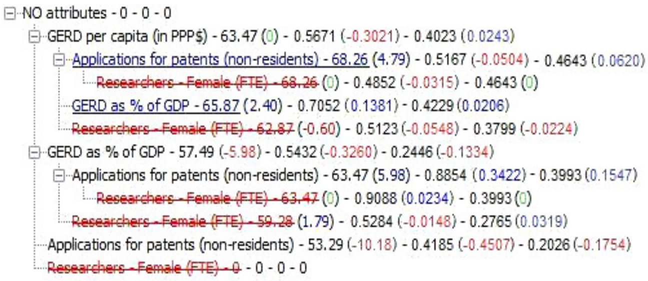


Figure 4: *The add graph.*

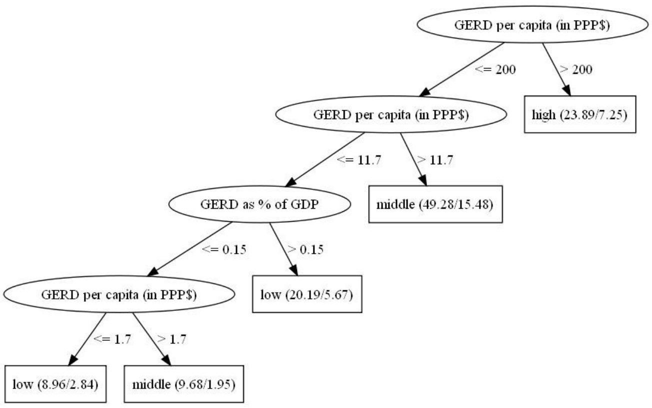


Figure 5: *The tree constructed on two GERD attributes.*

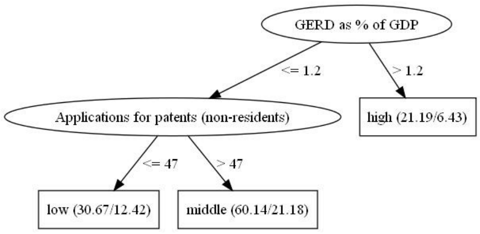


Figure 6: *The tree constructed on “GERD as % of GDP” and patent attributes.*

In conclusion, two relevant relations were discovered: first, higher level of investment into R&D leads to better economic welfare of a country; second, higher number of applications for patents leads to better economic welfare of a developing country (“middle” and “low” countries). The two factors are acknowledged in economic literature [6] showing that our method is able to discover valuable relations.

5 CONCLUSIONS

In this paper, we proposed HMDM method for detection of errors in models and their correction through the human-computer interaction. The method was applied on macroeconomic domain by analyzing the impact of R&D sector on economic welfare of a country. The results show that, with the help of our method, it is possible to find improved models.

The part of future work is to test the method on other domains and with other DM methods, such as regression trees.

References

[1] Y. Zhao and Y. Yao, On Interactive Data Mining, in *Encyclopedia of Data Warehousing and Mining*, ed. J. Wang (Idea Group, London, 2008), pp. 1085-1090.

[2] S. Stumpf, V. Rajaram, L. Li, W. Wong, M. Burnett, T. Dietterich, E. Sullivan and J. Herlocker, Interacting Meaningfully with Machine Learning Systems: Three Experiments, *Int. J. Hum-Comput. St.* **67**(8) (2009) 639-662.

[3] D. Jensen and P.R. Cohen, Multiple Comparisons in Induction Algorithms, *Mach. Learn.* **38** (2000) 309-338.

[4] J.R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, CA, 1993).

[5] M. Možina, J. Demšar, J. Žabkar and I. Bratko, Why is Rule Learning Optimistic and How to Correct It, in *Machine Learning: ECML 2006*, eds. J. Fürnkranz, T. Scheffer and M. Spiliopoulou (Springer, Berlin, 2006), pp. 330–340.

[6] N.C. Varsakelis, Education, Political Institutions and Innovative Activity: A Cross-country Empirical Investigation, *Res. Policy* **35** (2006) 1083-1090.

[7] M. Možina, J. Demšar, M. Kattan and B. Zupan, Nomograms for Visualization of Naive Bayesian Classifier, in: *Proc. of 8th Eur. Conf. on Principles and Practice of Knowledge Discovery in Databases* (Pisa, IT, 2004), pp. 337–348.

[8] E. Štrumbelj, I. Kononenko and M. Robnik Šikonja, Explaining Instance Classifications with Interactions of Subsets of Feature Values, *Data Knowl. Eng.* **68**(10) (2009) 886-904.

[9] T. Kulesza, W-K. Wong, S. Stumpf, S. Perona, R. White, M.M. Burnett, I. Oberst and A.J. Ko, Fixing the Program My Computer Learned: Barriers for End Users, Challenges for Machine, in *Proc. of Conf. on Intelligent User Interfaces (IUI)* (Sanibel Island, FL, 2009).

[10] J.A. Fails and D.R. Olsen Jr., Interactive Machine Learning, in *Proc. of the 8th Int. Conf. on Intelligent User Interfaces* (Miami, FL, 2003), pp. 39-45.

[11] S. Tong and D. Koller, Support Vector Machine Active Learning with Applications to Text Classification, *J. Mach. Learn. Res.* **2** (2002) 45-66.

[12] M. Ware, E. Frank, G. Holmes, M. Hall and I.H. Witten, Interactive Machine Learning: Letting Users Build Classifiers, *Int. J. Hum-Comput. St.* **55** (2001) 281-292.

[13] K.V. Mardia, J.T. Kent and J.M. Bibby, *Multivariate Analysis* (Academic Press, London, 1979).

[14] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, San Francisco, CA, 2005).

[15] R. Kohavi and G.H. John, Wrappers for Feature Subset Selection, *Artif. Intell.* **97** (1997) 273-324.

[16] J. Black, N. Hashimzade and G. Myles, *A Dictionary of Economics* (Oxford University Press, New York, 2009).

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Vzgoja in izobraževanje v informacijski družbi

Education in Information Society

Uredili / Edited by

Vladislav Rajkovič, Tanja Urbančič, Mojca Bernik

http://is.ijs.si
http://lopes1.fov.uni-mb.si

15. oktober 2010 / October 15th, 2010
Ljubljana, Slovenia

PREDGOVOR

Kdaj v šolo le z USB ključem? Pogosto slišimo to vprašanje. Tehnološke možnosti so. Je pa kopica ovir dejanskih in namišljenih, od proizvajalcev celuloze in šolskih torb, tudi s kolesci, do učiteljev in staršev, ki bi morali spremeniti svoje delovne vzorce.

Marsikdo se sprašuje, kaj bi dejansko s tem pridobili v pogledu vzgoje in izobraževanja. Res je, da je George Polja že leta 1944 izdal svojo knjigo »Kako rešujemo probleme« (prevedena tudi v slovenščino), ki je še danes aktualna, a so se računalniki takrat šele rojevali. In prav usposabljanje za reševanje problemov je eno izmed temeljnih poslanstev šole, ki ga je možno udejaniti tudi brez računalnika. Toda kaj potem pričakujemo od računalnika?

Splošna dosegljivost in priročnost papirja in svinčnika kot miselnih pripomočkov je omogočila razcvet v znanosti, tehniki, šolstvu in drugod. Vsakdo je lahko vizualiziral in modeliral del svojih misli. Z računalnikom se zgodba ponovi na ustrezno višji tehnološki ravni. Dobili smo univerzalni stroj za modeliranje in vizualizacijo podatkov, ki je v pomoč našim miselnim procesom in posledično reševanju problemov. Ta izziv že leta prenašamo tudi v procese vzgoje in izobraževanja. Tehnologija nam z rešitvami kot so priročni prenosni računalniki, tudi tablični, računalniška omrežja, USB ključi, ipd. stoji ob strani.

To pa žal ni dovolj. Sedanji izobraževalni sistem še vedno precej temelji na nekaterih predpostavkah, ki so ustrezale razmeram 19. in 20. stoletja. Tedaj je bila osnovna učna metoda verbalno prenašanje znanja, vloga učenca pa, da posluša in si učno gradivo čimbolj zapomni. Kljub temu da živimo v informacijski družbi, smo vse prevečkrat sužnji starega.

Osvoboditev iz starih spon ni možna le s predpisom, ampak z zavestnim iskanjem v smeri metod sodobnega učenja in poučevanja. Tako tudi na letošnji že 13. konferenci Vzgoja in izobraževanje v informacijski družbi razvijamo in raziskujemo e-izobraževanje, e-učenje, izobraževanje na daljavo, e-vsebine in druga področja za prenovo procesov vzgoje in izobraževanja, ki imajo za cilj razvijanje kompetentnega mišljenja in dela vseh generacij.

Vladislav Rajkovič, Mojca Bernik, Tanja Urbančič

PREFACE

When will students carry only a USB stick to school? We often hear this question. Technological means are available. Nevertheless, there is an abundance of obstacles, ranging from real to imaginary, encompassing different stakeholders from cellulose manufacturers and manufacturers of school bags, including those on wheels, to teachers and parents who would have to change their work styles.

Many wonder what is to be in fact gained in terms of education. It is true that George Polja already back in 1944 published a book on problem solving titled »How to solve it« which is still relevant today, however computers at that time were only in their embryonic phase. It is exactly the education for problem solving that is one of the fundamental missions of the school which can be carried out without a computer. What do we then expect from the computer?

General accessibility and handiness of the paper and the pencil as thinking tools enabled science, engineering, education and other fields to blossom. Everyone was able to visualize and model a fraction of their thoughts. With the emergence of the computer the story repeated itself at a higher technological level. We gained a universal machine for modeling and visualization of data. It assists our thinking processes and consequently problem solving. This challenge is being implemented in the processes of education for years already. The technology is there to support us with solutions such as laptop computers, also tablet computers, computer networks, USB sticks, etc.

This unfortunately is not enough. The contemporary educational system is still heavily based on several assumptions that suited the circumstances of the 19th and 20th centuries. The basic teaching method back then was verbal knowledge transfer, while the role of the student was to listen and to best memorize the learning materials. Despite living in the information age we are still too often restricted by the old ways.

No regulation can loosen these restrictions. This is only possible by a conscious search for new methods of modern education and teaching. With this in mind the 13th conference »Education in information society« develops and researches e-education, e-learning, distance learning, e-content and other fields aiming at reengineering of education processes with the goal of developing competent thinking and work of all generations.

Vladislav Rajkovič, Mojca Bernik, Tanja Urbančič

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Predsednik
Vladislav Rajkovič
Univerza v Mariboru, Fakulteta za organizacijske vede
in Institut Jožef Stefan

Alenka Adamič
Zavod MIRK

Vladimir Batagelj
Univerza v Ljubljani,
Fakulteta za matematiko in fiziko

Igor Bernik
Univerza v Mariboru,
Fakulteta za varnostne vede

Mojca Bernik
Univerza v Mariboru,
Fakulteta za organizacijske vede

Janez Bešter
Univerza v Ljubljani,
Fakulteta za elektrotehniko

Lea Bregar
Univerza v Ljubljani,
Ekonomska fakulteta

Borut Čampelj
Ministrstvo za šolstvo in šport RS

Dejan Dinevski
Univerza v Mariboru, Medicinska fakulteta

Saša Divjak
Univerza v Ljubljani,
Fakulteta za računalništvo in informatiko

Tomi Dolenc
ARNES

Marko Ferjan
Univerza v Mariboru,
Fakulteta za organizacijske vede

Ivan Gerlič
Univerza v Mariboru, Pedagoška fakulteta

József Györkös
Univerza v Mariboru,
Fakulteta za elektrotehniko, računalništvo in informatiko

Eva Jereb
Univerza v Mariboru,
Fakulteta za organizacijske vede

Mitja Jermol
Institut Jožef Stefan

Peter Kokol
Univerza v Mariboru,
Fakulteta za zdravstvene vede

Alenka Krapež
Gimnazija Vič

Nives Kreuh
Zavod RS za šolstvo

Tanja Logar
Center za poklicno izobraževanje

Robert Marinšek
Center za mobilnost in evropske programe izobraževanja in usposabljanja

Gregor Mohorčič
Zavod RS za šolstvo

Mojca Orel
Gimnazija Moste

Milan Pagon
University of Iowa, CIMBA

Jože Rugelj
Univerza v Ljubljani, Pedagoška fakulteta
in Institut Jožef Stefan

Niko Schlamberger
Slovensko društvo Informatika

Tomaž Skulj
Zveza za tehnično kulturo Slovenije

Branislav Šmitek
Univerza v Mariboru,
Fakulteta za organizacijske vede

Olga Šušteršič
Univerza v Ljubljani, Zdravstvena fakulteta

Tanja Urbančič
Univerza v Novi Gorici

Margareta Vrtačnik
Univerza v Ljubljani,
Naravoslovno tehniška fakulteta

Wechtersbach Rado
Zavod RS za šolstvo

Nemščina kot tuj jezik in jezik stroke (ekonomija): primer dobre prakse German as a Foreign and Business Language (Economics): Example of Good Practice

Jelka Bajželj
ESIC Kranj
jelka.bajzelj@guest.arnes.si

Povzetek

V prispevku avtorica predstavi primer dobre prakse pri poučevanju nemščine kot tujega jezika in jezika stroke (ekonomija). Uporaba IKT je več kot smiselna povsod tam, kjer do zastavljenih ciljev na ta način prihajamo lažje in hitreje, razvijanje IKT-kompetenc(e) dijakov in študentov je prav gotovo eden od (temeljnih) pogojev za njihovo kasnejše uspešno spoprijemanje z izzivi (tako poklicnimi kot drugimi) sodobnega sveta.

Ključne besede: nemščina kot tuj jezik in jezik stroke (ekonomija), informacijsko komunikacijska tehnologija, film, iskanje, selekcioniranje in analiziranje informacij, strategije uspešnega učenja

Abstract

In the article the authoress presents an example of good practice in teaching German as a foreign language and business language (economics). The Use of Information Communication Technology (ICT) is more than reasonable wherever it enables reaching the set goals more easily and more quickly. E- Competence development of the secondary and university students is one of the (basic) conditions which prepare the students to successfully tackle (the professional and other) challenges of the modern world.

Keywords: German as a foreign and business language (economics), Information Communication Technology, film, search, information selection and analysis, successful learning strategies

Zdravstvene težave v informacijski družbi Health Problems in the Information Society

Tomaž Bernard
Srednja šola Jesenice
Cesta talcev 2, Jesenice
tomaz.bernard@guest.arnes.si

Povzetek

Razvoj je v naš življenjski slog vnesel veliko pozitvnega obenem pa je nastavil tudi pasti za naše zdravje. Naše telo ni ustvarjeno za večurno sedeče mirovanje ob računalniku. To se odraža na zdravju današnje populacije, ki ji dandanes lahko rečemo kar sedeča populacija.. Uporabniki informacijske tehnologije se srečujejo s preobremenitvami vida, gibal ter stresom in splošnim slabim počutjem. Odgovor je v preventivi, ki predvideva pravilno razporeditev tipkovnice, zaslona, in drugih predmetov v delovnem okolju, pravilna uporaba tipkovnice in miške, vaje za sprostitev mišic ter zdrav življenjski slog tudi v prostem času.

Ključne besede: Informacijsko komunikacijska tehnologija, raztezanje, preobremenitve vida, preobremenitve gibal, prisilna drža

Abstract

Information communication technology has brought many positive changes into our lives but at the same time also set some traps for our health. Our body was not created for hours-long sitting by the computer, therefore every year there are more injuries that result from constant repeating of the same movements and maintaining the same position. Users of computers have to deal with overburdening of the eyes and extremities, as well as stress and generally not feeling well. It is especially important to introduce them to the prevention measures like correct position of the screen, keyboard and other objects on the work surface, the correct use of the mouse and keyboard, the exercises for muscle relaxation and a healthy way of life in their free time.
Keywords: Overburdening of the sight and mobility, coercive poise, information communication technology, streching

Razvoj strategij za kakovostno izvedbo višješolskega izobraževanja
Development Strategies for Quality Implementation of Vocabulary Education

Branka Balantič,
Tehniški šolski center Kranj,
Višja strokovna šola,
Kidričeva c. 55, SI-4000 Kranj
branka.balantic@guest.arnes.si

Branka Jarc Kovačič
Tehniški šolski center Kranj, Višja strokovna šola
Kidričeva c. 55, SI-4000 Kranj
branka.jarc-kovacic@guest.arnes.si

Zvone Balantič
Univerza Mariboru, Fakulteta za organizacijske vede
Kidričeva c. 55a, SI-4000 Kranj
zvone.balantic@fov.uni-mb.si

Povzetek

Korektno delo in razvoj višjega šolstva zahteva tudi ustrezen pristop h kakovosti. Kakovost je potrebno meriti s samoevalvacijo. V preteklosti je bilo samoocenjevanje šol pogosto nesistematično in zgolj intuitivno. Raziskava je usmerjena k pridobivanju objektivnih kazalnikov in izhodišč, ki jih šola lahko uporabi pri pripravi strateškega načrta višje šole.

Za oblikovanje celovite ocene kakovosti je primerna t.i. SWOT analiza (Strenghts, Weaknesses, Opportunities, Threats). SWOT analiza prinaša potrebne informacije, ki šoli zagotavljajo pridobivanje virov in kapacitet potrebnih za konkurenčnost na njenem področju delovanja. Metodologija poseže tudi po t.i. TOWS matriki, ki iz analize črpa točke za razvoj strategij, vključenih v rezultate.

TOWS matrika prikazuje posamezne strategije, ki sledijo priložnostim, strategije, ki bodo premagale slabosti, strategije za zmanjšanje ranljivosti in strategije, ki so osnova za obrambni plan.

Na višjih šolah je imenovana komisija za spremljanje in zagotavljanje kakovosti. Ustanovitev in delo komisije je opredeljeno v 15. členu Zakona o višjem strokovnem izobraževanju. Poleg znanih priporočil lahko s SWOT analizo in TOWS matriko prepoznamo medsebojne relacije, ki jih oblikujemo v ustrezno učinkovito in predvsem realno strategijo.

Ključne besede: višje šolstvo, kakovost, SWOT, TOWS

Abstract

Correct work in vocational school development demands appropriate quality approach. As quality needs to be measured self-assessment can be used. In the past self-assessment was non-systematic and only intuitive. The research is focused on determining objective indicators and preparation of vocational school strategic plan.

For comprehensive quality assessment SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis is appropriate. SWOT analysis delivers school needed information which enables it source acquisition and capacity needed for competitiveness on its working field.
TOWA matrix shows individual strategies which follow opportunities, strategies which will overcome weaknesses, strategies for vulnerability reduction and strategies which are basis for protection plan.

Commissions for monitoring and quality assurance are established on vocational schools. Commission establishment and role is defined in article 15 of Vocational education act. Beside known recommendations SWOT analysis and TOWOS matrix can recognize mutual relations which can be modeled into suitable, efficient and especially real strategy.

Keywords: Vocational school, quality, SWOT, TOWS

Cilji, naloge in pomen spletnih študijskih gradiv za ruščino kot tuj jezik pri sodobni uri ruščine

Aims, Tasks and Meaning of E-learning Contents for Present - Day Lessons of Russian as a Foreign Language

Maryna Bilash
Linguarus, d.o.o.
Dalmatinova 7, 1000 Ljubljana
marina.bilash@linguarus.si

Natalya Nikolenko
Linguarus, d.o.o.
Dalmatinova 7, 1000 Ljubljana
n-nikola@yandex.ru

Povzetek

V prispevku poudarjam vlogo aktivnega vključevanja učencev v izobraževalni proces, kar lahko dosežemo z uporabo e-gradiv za ruščino pri sodobnem pouku ruskega jezika kot tujega. Spletna učna gradiva za ruščino predstavlja 50 odstotkov gradiva, predvidenega z učnim načrtom za gimnazije in so sestavljena iz slovničnih, govornih ter komunikativnih vsebin, osvojitev katerih omogoča učencem pridobivanje osnove za komunikacijo v jeziku okolja ter jim dovoli opravljati osnovne komunikacijske potrebe pri sporazumevanju z nosilci jezika na družbenih in kulturnih področjih. V prispevku sem predstavila primere združevanja slovničnih in leksikalnih tem, kar omogoči učitelju ruščine zanimivo in kvalitetno pripravo potrebnega praktičnega gradiva za izvedbo pouka v razredu, pripravljanje interaktivnih vaj, organizacijo samostojnega dela učencev ter izvedbo nadzora nad preučevanjem aktualnih tem. Dvoletna priprava e-gradiv za ruščino pa nam je omogočila številne primere izvedbe pouka ruščine z uporabo spletnih učnih gradiv, ki kot primere dobrih praks vključujem v besedilo prispevka.

Ključne besede: e-gradiva, ruščina, aktivna vloga učencev v izobraževalnem procesu, sodobna priprava na pouk ruskega jezika

Abstract

The contribution emphasises on the active role of students in the educational process achieved by the use of e-learning contents for lessons of Russian as a foreign language. E-learning contents for Russian represent 50 percents of the materials foreseen for the course syllabus for gymnasiums, and include grammar, speaking and communicative tasks. The obtainment of these contents enables students to gain the basics for communication in the language of their surroundings and allows them to satisfy their basic communication needs when communicating with native speakers in the social and cultural field. In the contribution I presented examples of combining grammar and lexical topics which make it possible for teachers of Russian to prepare interesting and quality material needed for teaching in class, interactive exercises, tasks students have to do on their own and which grant the realization of control over the research of current topics. A two year long preparation of e-learning contents for Russian has enabled several realizations of Russian lessons with the use of e-learning contents which I mentioned as good examples in the contribution.

Keywords: e-learning contents, Russian, active role of students in the educational process, present-day preparation for Russian lessons.

Medpredmetno povezovanje v spletni učilnici Moodle
Cross - Curricular Learning in Moodle's Virtual Learning Environment

Sašo Bizant
TŠC Kranj
Kidričeva c. 55 Kranj
sasobizant@guest.arnes.si

Marija Šubic
TŠC Kranj
Kidričeva c. 55 Kranj
marija.subic@guest.arnes.si

Povezava na spletno učilnico Moodle:
http://ucilnica.tsckr.si/course/view.php?id=36

Povzetek

Prispevek prikazuje izvedbo medpredmetnega povezovanja med računalništvom, angleščino in matematiko s pomočjo spletne učilnice Moodle. Klasično poučevanje in učenje je bilo podprto in kvalitetno izboljšano z virtualnim učnim okoljem spletne učilnice. Tako kombinirano učenje za doseganje učnih ciljev se je izkazalo kot učinkovito, zanimivo, kreativno in vzpodbudno za nove izzive.

Ključne besede: medpredmetno povezovanje, kombinirano učenje, virtualno učno okolje - Moodle, načrtovanje pouka, sodelovanje

Abstract

The aim of this article is to present an example of cross-curricular learning in the case of Computing, English and Mathematics by using the Moodle e-learning platform. The traditional face-to-face learning and teaching were supported and augmented by Moodle's virtual learning environment. Such blended learning in order to attain the goals set in the learning process has proved to be effective, interesting and creative, and encourages new challenges.

Keywords: cross-curricular learning, blended learning, Virtual Learning Environment (VLE) - Moodle, lesson planning, collaboration

Uporaba orodja Plane Graphic Calculator pri obravnavi linearne funkcije
Using the Tool Plane Graphic Calculator when Considering the Linear Function

Klavdija Brecelj
OŠ Ivana Roba Šempeter pri Gorici
klavdijabrecelj@gmail.com

Povzetek

Informacijsko komunikacijska tehnologija je del življenje okoli nas in se zelo hitro razvija. Učitelji so tisti, ki učencem posredujejo določena znanja in veščine, zato je nujno, da IKT vključijo v proces poučevanja in s tem učencem nudijo različne oblike pridobivanja znanja. S pomočjo uporabe IKT postane izobraževalni proces za učence zanimivejši in s tem so bolj zainteresirani za pridobivanje novih znanj. V članku je prikazan primer uporabe internetnega orodja Plane Graphic Calculator pri učni temi linearna funkcija, v devetem razredu osnovne šole pri predmetu matematika. Učenci so se najprej seznanili s programom , potem pa so ga samostojno uporabljali v fazi ponovitve učne snovi in obravnave nove snovi.

Ključne besede: informacijsko komunikacijska tehnologija, poučevanje, linearna funkcija, internetno orodje Plane Graphic Calculator

Abstract

Information and Communication Technology is part of life around us and is growing rapidly. Teachers are those who provide students adequate knowledge and skills. It is essential to include ICT in the educational process in order to offer students different forms of teaching. With the use of ICT, the educational process becomes more interesting and students are more interested in acquiring new knowledge. The present abstract gives an example of the use of the Plane Graphic Calculator software tool in teaching linear function in the ninth grade of primary school mathematics course. Students familiarize with the program first, and then they use it independently through the phase of learning and revising of new topics.

Keywords: Information and Communication Technology, teaching, linear function, software tool Plane Graphic Calculator

Uvajanje računalniško podprtega laboratorijskega dela pri pouku biologije na Škofijski gimnaziji Antona Martina Slomška v Mariboru
Introduction of Computer Based Laboratory for Teaching Biology at the Episcopal High School Anton Martin Slomšek Maribor

Jožica Brecl
Škofijska gimnazija A. M. Slomška
Vrbanska cesta 30, 2000 Maribor
jozica.brecl@guest.arnes.si

Andrej Šorgo
Univerza v Mariboru, Fakulteta za naravoslovje in matematiko
Koroška cesta 160, 2000 Maribor
andrej.sorgo@uni-mb.si

Povzetek

V letu 2010 smo v sklopu izobraževanja bodočih učiteljev biologije izvedli teden dni trajajoč računalniško podprt eksperimentalni pouk biologije v vseh razredih Škofijske gimnazija A. M. Slomška, Maribor. Vsak dijak je izvedel računalniško podprte laboratorijske vaje, ob tem da so ure vodili študenti. Odziv dijakov in študentov na takšno delo je bil pozitiven, ponovno pa se je potrdila ugotovitev, da dijaki niso ovira pri vpeljavi takšnih metod dela v redni pouk.

Klučne besede: računalniško podprto laboratorijsko delo, IKT, kompetenca

Abstract

As a part of teacher trainings of prospective biology teachers we have performed a week of computer supported experimental work at the upper secondary school Škofijska gimnazija A. M. Slomška in the year 2010. Each student has performed computer supported laboratory exercises under the leadership of prospective teachers. Response of students and prospective teachers was positive, and once more was argued that students are not the obstacle in introduction of such methods into regular laboratory work.

Keywords: computer suppted laboratory work, ICT, competences

Pretvorba spletnega učnega gradiva v tiskano gradivo
Transformation of Web - Based Materials into Printed Materials

Lea Bregar
lea.bregar@ef.uni-lj.si

Margerita Zagmajster
Andragoški center Slovenije
Šmartinska 134a, 1000 Ljubljana
margerita.zagmajster@acs.si

Povzetek

S širjenjem e-izobraževanja in vse večjo ponudbo spletnih učnih gradiv narašča tudi število tiskanih gradiv, ki so nastala iz spletnih gradiv. V prispevku so na osnovi poznavanja načel ter izkušenj pri razvoju in uporabi spletnih ter tiskanih gradiv prikazani postopki, ki so potrebni z vsebinskega in oblikovnega vidika ter sloga pisanja za smiselno pretvorbo spletnega gradiva v tiskano gradivo. Nekatere posebne okoliščine spodbujajo pretvorbo spletnih gradiv v tiskana gradiva, v splošnem pa velja, da mora odločitev o pretvorbi spletnih učnih gradiv v tiskana temeljiti na skrbnem premisleku o stroških in koristih.

Ključne besede: e-izobraževanje, študij na daljavo, spletno učno gradivo, spletni izobraževalni program, tiskano gradivo

Abstract

In line with growth of e-learning and expanded supply of web-based learning materials increased number of printed materials, derived from web materials have been available. In the paper, principle procedures for turning content, design and writing style of web materials into printed ones are examined. These procedures are based on acknowledgement of basic principles and practical experiences related to web and printed learning materials development and delivery. Some specific circumstances encourage transformation of web-based materials into printed materials. However, generally the decision on this transformation has to rely on profound consideration of related costs and benefits.

Keywords: e-learning, distance education, web-based learning material, web-based course, printed material

Z razvojem IKT kompetenc do računalniškega znanja učencev OŠ ob tehniških dnevih
The Development of Interactive Whiteboard (IWB) Competences on Technical Days enables Primary School Pupils to acquire Computer Skills

Urška Brožič
Rafolče 22, 1225 Lukovica
urska_brozic@yahoo.com

Povzetek

V prispevku predstavljam izvedbo tehniških dni z računalniško vsebino v višjih razredih osnovne šole. Uporaba in delo z računalniškimi programi postaja nuja današnjega časa, zato si na naši šoli želimo, da ga osvojijo vsi učenci. Osnovno računalniško znanje učencev je pomanjkljivo. V učnem načrtu za osnovno šolo je računalniško znanje zajeto le pri izbirnih predmetih, ki pa niso obvezni za vse. Zato smo se odločili, da določene računalniške vsebine predstavimo učencem od 6. do 9. razreda na tehniških dnevih. V 6. razredu osnove programa Excel, v 7. razredu program Word, v 8. razredu programa PhotoFiltre in PowerPoint. V 9. razredu se učenci naučijo posneti kratek video z videokamero in ga obdelati v programu Studio Version 7. Izvedbo tehniških dni nam je v veliki meri olajšala lastna spletna učilnica, ki so se jo na tehniškem dnevu oz. pri pouku naučili uporabljati tudi učenci.

Ključne besede: tehniški dnevi z računalniško vsebino, Excel, Word, PhotoFiltre, PowerPoint, StudioVersion 7, spletna učilnica

Abstract

The article represents the realization of technical days at our school, where pupils of higher grades are introduced and encouraged to use computers. Using computers and having the basic knowledge of computer programs is nowadays almost a necessity, hence our school tries to provide that knowledge to all of our pupils. We have noticed the lack of basic knowledge regarding computers and unfortunately the curriculum includes this subject merely as an optional subject that does not include all of our pupils. That is why, we have decided to introduce it to our 6th to 9th grade pupils in the context of technical days. Therefore, our 6th grade pupils are introduced to the basics of Excel, the basics of Word are introduced to our 7th grade pupils, while pupils in the 8th grade work with programs, such as PhotoFiltre and PowerPoint. The oldest pupils learn how to make a short video with the help of a video camera, and later on also how to process that video with the Studio Version 7 program. Our school has and promotes the usage of E-Classroom which has proven to be of a great help in the realization of technical days. Pupils are encouraged to acquire the knowledge of computer skills also by using E-Classroom throughout their classes.

Keywords: technical days entailing the use of computers, Excel, Word, PhotoFiltre, PowerPoint, StudioVersion 7, E-Classroom

Pedagoške razsežnosti projekta Vizualizacije glasbe z barvami (VGzB)
Possibilities of Application of the Project "Colour Visualization of Music" in Teaching

Peter Ciuha
peter.ciuha@guest.arnes.si

Bojan Klemenc
bojan.klemenc@fri.uni-lj.si

Franc Solina
franc.solina@fri.uni-lj.si

Povzetek

VGzB temelji na sistemu barvnih znakov, povezanih z glasbenimi toni. Znaki sorodnih barv se nahajajo na tonih, ki so med seboj v harmoniji ali sozvočju. V uvodu so predstavljene osnove, na katerih je osnovan sistem barvnih znakov – matematični model harmonije. Sledi razmislek o vidnem prikazu različnih izraznih elementov glasbe – melodiji, kompoziciji, ritmu in harmoniji. Te sorodnosti omogočajo razvoj računalniškega programa, ki uporabi te elemente za vizualizacijo ali prikaz barvnih in oblikovnih struktur glasbe. Program posnema človeško zaznavo, v kateri so posamezni sestavni deli določeni s prepoznavanjem celote. Omogoča tudi razvoj orodij, ki lahko povečajo razumevanje med poslušanjem ali igranjem glasbe. Igranje glasbe pa lahko pridobi novo kvaliteto z uporabo interaktivnih barvnih glasbil, ki v povezavi s programom igralcu v vsakem trenutku z barvami prikažejo raznolike možnosti tvorjenja zvočnih harmonij. S tem spremenijo komponiranje glasbe v igro in privlačno barvno-zvočno potovanje. Tu se pojavi velik izziv za pedagogiko in metodiko, da osmisli in uporabi prihajajoča multimedijska barvna orodja. Ta bi lahko v veliki meri približala proces učenja proti igri, ki je otrokova najbolj naravna oblika učenja in delovanja. V umetniškem ustvarjanju pa lahko ponovno vzpostavimo ravnovesje med našo logično in intuitivno-kreativno naravo.

Abstract

Colour visualization of music is based on a system of colour signs, which are connected with musical tones. Tones which are in harmonic relationships are represented by related colours. In the introduction we outline the foundations on which the system of colour signs is based – the mathematical model of harmony. We discuss possibilities of visual representation of expressive elements of music – melody, composition, rhythm and harmony. These relationships enabled us to develop a computer program that uses these elements for visualization. The program mimics human perception in which the parts are determined by perception of the whole. Furthermore the program enables the development of tools that can enhance the understanding during listening to music or playing. Playing of music can acquire a new quality with the use of interactive coloured musical instruments, which by using colours show the performer different possibilities for forming musical harmonies and thereby change the composing of music into a game and attractive colour-aural journey. Here we stumble upon a challenge for educational science and methodology: how to use such upcoming multimedial tools. These tools would bring the processes of learning and playing a game closer together, as game is a child most natural form of functioning. Furthermore in the area of artistic creation we can once again establish a balance between our logic and intuitive nature.

Keywords: visualization, music, colours, learning, creativity

Odločitveni model za izbiro najboljšega kandidata za vodjo podružnice
Decision Making Model for Choosing the most Appropriate Candidate as the Head of a Branch Unit

Nevenka Colja
OŠ Miren, Slovenija
nevenka.colja@gmail.com

Povzetek

V prispevku je predstavljen večparametrski odločitveni model, ki služi ravnatelju v osnovni šoli kot pomoč pri izbiri kandidata za vodjo podružnice. Pri delu je bila uporabljena sodobna informacijsko-komunikacijska tehnologija s poudarkom na lupini ekspertnega sistema za večparametrsko odločanje DEXI. Ta nam omogoča izgradnjo baze znanja, ocenjevanje variant, razlago dobljenih rezultatov in kaj-če analizo. Razvili smo odločitveni model po vseh fazah odločanja in ga preizkusili na praktičnem primeru. Na podlagi večparametrskega modela so bili kandidati-variante ocenjeni po več lastnostih. Podrobneje je analizirana končna odločitev o izbiri najprimernejšega kandidata za vodjo podružnice, ki je z uporabo modela nepristranska in dovolj transparentna tako za odločevalca, kot tudi za kandidate.
Na podlagi SWOT analize smo v zaključku ugotavljali prednosti in pomanjkljivosti izdelanega modela in podali njegovo končno oceno ter možnosti za širšo uporabo v praksi.

Ključne besede: ravnatelj, vodja podružnice, DEXI, odločanje

Abstract

The article is about a multi-attribute decision making model, which helps the headmaster of an elementary school to choose the most appropriate candidate as the head of their branch unit. The modern information and communication technology with emphasis on an expert system shell for multi-attribute decision making, named DEXI, was used. This system facilitates the creation of knowledge database, evaluation of options, explanation of the evaluation results and what-if analysis. A decision making model for all decision making phases was developed and used in practice. Based on the multi-attribute model the candidates/options were evaluated regarding several characteristics. What is more, the final decision on choosing the most appropriate candidate as the head of a branch unit was analysed in detail. The decision based on this model is therefore impartial and sufficiently transparent not only for the decision maker but also for the candidates.
At the end the SWOT analysis was used to establish the advantages and disadvantages of our model, to pass the assessment and to present the possibilities for its wider application.

Keywords: headmaster, head of branch unit, DEXI, decision support

Model samoocenjevanja stopnje informatizacije šole

Borut Čampelj
Ministrstvo za šolstvo in šport
borut.campelj@gov.si

Povzetek

V Sloveniji šole nimajo veliko primerov praks ali strokovne literature, ki bi jim pomagale pokazati nadaljnjo pot pri procesu informatizacije. Ena izmed poti je preko indikatorjev - kazalnikov, ki posamezno šolo umestijo glede na nivo trenutne informatizacije šole. V prispevku bomo nakazali možnosti za večparametrski hierarhični model samoevalvacije informatizacije posamezne šole, ki ga je možno učinkovito izvesti tudi s pomočjo programa Dexi. Tri glavne skupine so: šola in okolje, učitelji in skupnosti, učenci in polje bivanja. V predlaganem odločitvenem drevesu naj bi s kazalniki izmerili stanje na šolah ter nakazali smernice za nadaljnji razvoj in potrebne spremembe. Smiselnost razvoja in uporabe modela za vsako skupino je v tem, da se z IKT resnično na šoli spremenijo stvari na bolje v skladu z novimi pristopi, predlaganimi standardi, pričakovanji… IKT je lahko v teh procesih le orodje, ki lahko po eni strani poenostavi administrativni del, ga naredi bolj preglednejšega in uporabnega, še posebej pa spremeni potek poučevanja in učenja, tj. spremeni se vloga učitelja v motivatorja, moderatorja, vzpodbujevalca, tistega, ki pomaga, usmerja in vzgaja pri vrednotenju informacij, ki jih mladi neprestano prejemajo z različnih virov. Učenec pa naj bo v šoli veliko bolj ozaveščeno aktiven, da se v izven šolskem času lahko posveti zdravemu življenju (hobiji, prosti čas, …).

Ključne besede: šola, IKT, stopnja informatizacije, evalvacija, samoevalvacija, vodenje

Abstract

Slovenian schools don't have many cases of practice or special literature, which would help them to show the right way of school informatisation process. Selfevaluation with indicators that place individual school considering the level of informatization is one of them. In the paper is proposed the multiple attribute decision making model, which could be implemented by Dexi software. Three main areas of school informatisation are: school and environment, teachers and community, pupils and living area. With the model we would like to measure existing situation and show the way for further development and necessary changes. The viability of developing model is that the evaluation should cause real changes with new approach standards, expectations, … ICT is just a tool, but it could simplify a lot the school administration and mostly could change the teaching and learning where the teacher becomes a motivator, moderator, promoter, and person, who helps, guides and raises pupils in the assessment of new information from everywhere. And the pupil is in the school more consciously active, so there is more time for healthy life (hobbies, spare time,…).

Keywords: school, ICT, level informatization, evaluation, selfevaluation, leadership

Stanje in trendi uporabe informacijsko komunikacijska tehnologija v slovenskem izobraževalnem sistemu
Present State and Trends of Using Information and Communication Technology in Slovene Education System

Ivan Gerlič
Univerza v Mariboru, Fakulteta za naravoslovje in matematiko
ivan.gerlic@uni-mb.si

Povzetek

V referatu je predstavljen krajši pregled rezultatov raziskave »Stanje in trendi uporabe IKT v slovenskih osnovnih in srednjih šolah«, ki jo avtor izvaja vsaki dve leti za potrebe Ministrstva za šolstvo in šport ter nacionalni projekt »Informatizacija slovenskega šolstva«. Predstavljen je povzetek rezultatov s področja uporabe strojne in programske opreme v slovenskem šolstvu, kvalitete pedagoških kadrov, specialno - didaktične problematike pouka Informatike in Računalništva in širše uporabe računalniške in sodobne informacijske tehnologije v izobraževanju.

Ključne besede: strojna oprema, programska oprema, kadri, didaktična problematika, raziskava.

Abstract

The following article analyses some results of investigation “Present state and trends of using ICT in Slovenian primary and secondary schools”. This research the author performs biannual for Ministry of education and sport and national project “Informatization of Slovenian education system”. The article presents some problems on educational hardware, software, pedagogic staff and some didactic criteria of learning Informatics and Computer sciences and computer applications in the educational process.

Keywords: hardware, software, pedagogic staff, didactic problems, research.

Uporaba e-gradiva pri naravoslovju
Using E-learning Materials Science

Leonida Gregorič
Osnovna šola Dornberk
Gregorčičeva 30 a, Dornberk
gbogdan@siol.net

Povzetek

V četrtem razredu osnovne šole imajo učenci jeseni naravoslovni dan gozd. V gozdu opazujejo rastje in podrast kjer vsako leto opazijo veliko različnih vrst gob, ki jih naberejo, poimenujejo in spoznajo njihov pomen, zgradbo, razmnoževanje, življenjski ritem, užitne in strupene gobe.

Računalnik je pri delu tako učitelju, kot tudi učencem v veliko pomoč, saj računalniški programi dajo veliko možnosti za usvajanje nove snovi ter utrjevanje in poglabljanje novih in drugih znanj. Učenci so nadgradili znanje gob tako, da so s pomočjo e-gradiv spoznali gobe.

Ključne besede: naravoslovje, internet, računalniški program, e-gradiva, gobe.

Abstract

In the fourth grade of elementary school science students have the day in the autumn woods. In the woods and undergrowth vegetation observe where each year observed many different types of mushrooms, which are accumulated named and understand their meaning, structure, reproduction, life rhythm, edible and poisonous mushrooms.

The computer is at work as a teacher, as well as students of little help because the software is made great potential for acquiring new materials and the consolidation and deepening of new and other skills. Pupils build knowledge of mushrooms so that they are using e-materials realized mushrooms.

Keywords: science, internet, software, e-materials, mushrooms.

Razvoj računalniških didaktičnih programov s poudarkom na medpredmetnem povezovanju
Developement of Didactic Programms Focuaed on Iterdisciplinary Curriculum

Marina Ivanič
OŠ Rudolfa Ukoviča Podgrad, Podgrad 99b, 6244 Podgrad
marina.ivanic@gmail.com

Andreja Istenič Starčič
Univerza na Primorskem, Pedagoška fakulteta Koper,
Cankarjeva 5, 6000 Koper
andreja.starcic@siol.net

Povzetek

V prispevku avtorici obravnavata tematiko medpredmetnega povezovanja, katera je bila ponovno izpostavljena pri prenovi učnih načrtov (2006). Predstavljata prednosti in zahteve kakovostnega medpredmetnega povezovanja, ki daje temelje celostnemu poučevanju. Analizirata priporočene medpredmetne povezave v veljavnem učnem načrtu za matematiko in predlagata nove, ciljno usmerjene povezave. Predstavita tudi začetno izobraževanje učiteljev matematike in računalništva za uporabo računalnika v izobraževanju. Študente, bodoče učitelje, je potrebno na to vlogo pripraviti in jim omogočati pogoje za razvoj kompetenc, ki jim omogočajo uporabo IKT-ja pri pouku. V ta namen študentje pri predmetu URI razvijajo svoje zamisli in povezanost IKT z medpredmetnim povezovanjem. V prispevku je prikazano, katere aplikacije študentje najpogosteje izbirajo za uresničevanje medpredmetnih povezav.

Ključne besede: medpredmetno povezovanje, osnovna šola, matematika, računalniško podprto učenje, izobraževanje učiteljev

Abstract

In the contribution the authors deal with the interdisciplinary curriculum that was highlighted in the renovation of the curricula (2006) again. They present the advantages and requirements of the development of an interdisciplinary curriculum which is underlying for the integral teaching. They analyse the recommended interdisciplinary correlations in the current curriculum for mathematics, and propose additional objective-based correlations. Initial training for the teachers of mathematics and computer science on computer-based teaching at school is presented. Students of teacher-training study courses - the future teachers have to be prepared for this role and given the support for the development of competence for the use of ICT in teaching. For this purpose, the students involved in the subject URI develop their ideas and the correlation of the ICT with the interdisciplinary curriculum. The paper shows which applications are most frequently chosen by the students for the implementation of interdisciplinary correlations.

Keywords: interdisciplinary curriculum, primary school, mathematics, computer aided learning, teachers education

Uporaba spletne učilnice Moodle na srednji šoli
Using E-classroom Moodle in Secondary School

Jana Jemec
ESIC Kranj, Ekonomsko-trgovska šola Kranj
jana.jemec@siol.net

Povzetek

Na Ekonomsko-trgovski šoli Kranj smo v šolskem letu 2007/08 postavili spletno učilnico Ekonomc s pomočjo aplikacije Moodle. Sledilo je postopno izobraževanje učiteljev za delo s spletno učilnico. V praksi so se pri delu s spletno učilnico pokazale številne prednosti in priložnosti, pa tudi nekaj pomanjkljivosti in slabosti, ki jih ima tak način dela.
Rezultati analize ankete o uporabi spletne učilnice pri učiteljih na naši šoli nam kažejo, da se Moodle z viri in dejavnostmi, vse bolj uveljavlja kot uspešno orodje za izobraževanje in je izziv današnjega časa za e-šolstvo.
Omogoča učinkovito in dejavno poučevanje, učenje in nadgradnjo teoretičnega znanja. Spletna učilnica omogoča in zahteva drugačne organizacijske oblike in načine izvedbe izobraževalnega procesa ter učitelja postavlja vedno bolj v vlogo mentorja in koordinatorja dela.

Ključne besede: Moodle, IKT, e-gradiva, spletna učilnica

Abstract

E-classroom Ekonomc was set in the school year 2007/08 with the aplication Moodle in our secondary school. We started with progressive education for teachers to learn the activities of e-classroom. Practical work with e-classroom brings us many advantages and opportunities, but also deficiencies and weaknesses of such teaching.
The resaults of survey about using e-classroom shows us that Moodle activities and resources are very successful tool for education and the challange of our time on e-education.
Provides effective and active teaching, learning and upgrading of theoretical knowledge. E-classroom allows and requires different organizational forms and methods of carrying out the educational process and teachers are increasingly put in the role of the supervisor and coordinator of work.

Keywords: Moodle, ICT Technology, e-materials, e-classroom

Poučevanje praktičnega pouka - učnega podjetja na daljavo
Long Distance Teaching of Practical Work - Practice Firm

Ana Jurgele
ESIC, Ekonomsko trgovska šola Kranj
Cesta Staneta Žagarja 33, Kranj
ana.jurgele@gmail.com

Povzetek

Namen prispevka je predstaviti uporabo spletne učilnice pri poučevanju praktičnega pouka. Spletno učilnico lahko uporabimo v vseh fazah učnega procesa, od začetne motivacije do zaključnega ocenjevanja dijakovega znanja. Izobraževanje je trajen proces, ki se ne konča ob koncu pouka. Učili naj bi se tudi takrat, ko vstop v šolo ni mogoč iz različnih razlogov. Lahko so to dijaki s posebnimi potrebami, vrhunski športniki, dijaki, ki zaradi osebnih in družinskih okoliščin ne morejo redno obiskovati pouka ali opravljati določenih učnih obveznosti. Zanje pomeni delo na daljavo učinkovito komunikacijsko bližnjico med njimi in ostalimi udeleženci izobraževanja. Olajšano jim je sporazumevanje s sošolci, z vrstniki in učitelji kakor tudi samo pridobivanje znanja, saj jim delo v e-učilnici oz. Moodlu zagotavlja in omogoča oboje.

Ključne besede: e-učilnica, Moodle, pouk na daljavo, praktični pouk

Abstract

The aim of the article is to present the use of e-classroom during practical work lessons. e- classroom can be used at all phases of the teaching and learning from the initial motivation to the concluding grading of students. Learning as a process does not stop at the end of school lessons but it is also possible when students cannot attend school for different reasons. There can be students with special needs, students athletes, students that cannot come to school regularly for family or personal reasons or they are not able to meet their learning commitments. More suitable and efficient for them is long distance learning because of communication shortcuts between students and other participants of the learning process. The use of e-classroom/Moodle enables not only communication among school friends, mates and teachers, but also it assists to acquire knowledge.

Keywords: e-classroom, Moodle, long distance learning, practical work lessons

Integriranje spletne aplikacije Bubbl v vzgojno – učni proces
Introduction of Bubble Web Application into Educational Process

Ana Jurgele
ESIC, Ekonomsko-trgovska šola Kranj
Cesta Staneta Žagarja 33, 4000 Kranj
ana.jurgele@gmail.com

Povzetek

Namen prispevka je predstaviti program za izdelavo miselnih vzorcev za vsa predmetna področja in na vseh nivojih izobraževanja tako za dijake kot za učitelje. Miselni vzorci so zaradi svoje enostavnosti in privlačnosti zelo priporočljivi, saj naredijo učenje učinkovitejše, urijo možgane, tako da bodo dijaki sposobni logičnega razmišljanja. Ta brezplačna spletna aplikacija se v učnem procesu uporablja za sestavljanje nalog, za utrjevanje znanja, za medsebojno sodelovanje in primerjavo pridobljenega znanja. Miselni vzorci omogočajo in spodbujajo ustvarjalno in kritično razmišljanje, samostojno ali skupinsko raziskovanje, odgovorno predstavljanje in izmenjavanje informacij, komunikacijo in sodelovanje. Možnosti uporabe miselnih vzorcev ostajajo v prihodnosti široko odprte.

Ključne besede: miselni vzorec, spletno orodje, motivacija.

Abstract

The purpose of this article is to present Mind Maps for school subjects at all levels of education, for students and teachers respectively. Mind Mapping is useful and advisable in successful learning as it is a visual way to brainstorm ideas; it is transparent, simple and encourages logical thinking of the students. The important point is that Mind Maps at Bubbl.us website are free and they can be used in setting tasks for the students, in revision of lessons as well as in promoting cooperation among students and comparing the knowledge obtained in the process. Mind Maps encourage creative and critical thinking, individual or team research, the exchange of information, communication and cooperation. This shows that the use of Mind Maps is a real prospect in education in the future.

Keywords: mind map, internet tools/software, motivation

Tehnološko podprta uporaba koncepta učnih izidov v e-izobraževanju
Technology Supported use of the Learning Outcomes Concept in E-education

Tomaž Klobučar
Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
klobucar@e5.ijs.si

Povzetek

Evropsko ogrodje kvalifikacij (EOK) definira učne izide kot ugotovitve o tem, kaj udeleženec ob zaključku učnega procesa zna, razume in je sposoben opraviti. Pri tem so učni izidi opredeljeni v smislu znanja, spretnosti in kompetenc. Prispevek obravnava vlogo in uporabo učnih izidov v e-izobraževalnih sistemih in nekatere probleme, na primer določanje in enotno opisovanje učnih izidov, načrtovanje in opisovanje učnega procesa, pripravo in iskanje najustreznejše učne vsebine glede na pričakovani učni izid ter tehnološko podprto preverjanje in ocenjevanje pridobljenih spretnosti in kompetenc. V prispevku so na kratko prikazani standardi na tem področju in orodja, ki učitelju, učencu in drugim subjektom v e-izobraževanju olajšajo delo z učnimi izidi. Predstavljeni rezultati izhajajo iz mednarodnega projekta ICOPER (Interoperable Content for Performance in a Competency-driven Society; http://www.icoper.org) iz programa eContentplus, v katerem sodeluje Institut »Jožef Stefan«.

Ključne besede: učni izid, e-izobraževanje, tehnološko podprto učenje, visoko šolstvo, znanje, spretnost, kompetenca, izobraževalno orodje

Abstract

European Qualifications Framework (EQF) defines learning outcomes as statements of what a learner knows, understands and is able to do on completion of a learning process. Here, learning outcomes are defined in terms of knowledge, skills and competences. This paper discusses the role and usage of learning outcomes in e-learning systems, as well as some relevant problems such as specification and unified description of learning outcomes, design and description of a learning process that enables obtainment of the outcomes, preparation and search for the most useful learning opportunities according to the intended learning outcome, and technology enhanced assessment of obtained skills and competences. We present standards in the field and technical solutions that help teachers, learners and other stakeholders in e-education to use the concept of learning outcomes. Presented work is a result of an international ICOPER (Interoperable Content for Performance in a Competency-driven Society; http://www.icoper.org) project from the eContentplus programme, where Jožef Stefan Institute is one of the partners.

Keywords: learning outcome, e-education, technology enhanced learning, higher education, knowledge, skill, competence, educational tool

Uporaba računalniških programov pri pouku družbe v 5. razredu osnovne šole
The use of Computer Programs at Lesson of Sociaty in the 5th Class of Primary School

Slavka Kozel
OŠ Trzin
Mengeška 7b, 1236 Trzin
slavka.kozel@gmail.com

Povzetek

V prispevku je predstavljen pomen uporabe IKT v vzgojno-izobraževalnem procesu, varna uporaba interneta in elektronske pošte ter možnosti uporabe računalniških programov Microsoft Power Point in Slikar pri pouku v drugem vzgojno-izobraževalnem obdobju. Predstavljen je primer sodelovalnega učenja med petošolci in vključitev didaktične sestavljanke, ki jo učenci oblikujejo s pomočjo programa na spletni strani in jo v reševanje po elektronski pošti pošljejo sošolcu.
Ob sodelovalnem učenju usvajajo nove pojme, razvijajo dobre medsebojne odnose, oblikujejo pozitivno samopodobo in nadgrajujejo znanje s področja računalništva.

Ključne besede: Microsoft Power Point, IKT tehnologija, e-mail, sodelovalno učenje, sestavljanka, kmečke hiše na Slovenskem.

Abstract

The article presents the importance of using information and communication technology in educational process, the safe use of Internet and e-mail and the possibility of using the computer programme Microsoft Power Point in the second triennium of primary school. An example of collaborative learning between pupils of the fifth class is presented. The pupils form a didactic puzzle through the programme on the website and they send it by e-mail to schoolmates to solve it.
Through cooperative learning the students assimilate new concepts, develop good relationships, develop positive self-esteem and build up skills in computer science.

Keywords: Microsoft Power Point, Information and Communication Technology, e-mail, cooperative learning, puzzles, farmhouses in Slovenia.

Interaktivnost brez interaktivne table
Interactivity without an Interactive Whiteboard

Alenka Kralj
Osnovna šola Šmartno, Šmartno pri Litiji
Pungrt 9, 1275 Šmartno pri Litiji
Podružnična šola Štangarske Poljane
alenka.kralj@guest.arnes.si

Povzetek

Sodoben način poučevanja uvaja v šole uporabo računalnika, projektorja, zadnjih nekaj let pa je v marsikateri učilnici prisotna interaktivna tabla.
V uvodu je izpostavljena pomembnost didaktičnega gradiva in učiteljeva vloga pri oblikovanju in uporabi tovrstnega gradiva.
V nadaljevanju je opisana možnost uporabe gradiva brez interaktivne table in predstavljeno gradivo, izdelano s pomočjo programa za delo z interaktivno tablo.

Ključne besede: didaktično gradivo, interaktivna tabla, glasbena vzgoja, zvok, slika, spoznavni stili

Abstract

Contemporary way of teaching at school introduces the use of computer, projector, and in the last few years an interactive whiteboard was introduced to many classrooms.
In the introduction the importance of didactic material and a teacher's role at forming and using that kind of material is pointed out. Further on, the option of using material without an interactive whiteboard is described and the material elaborated with the help of a programme for working with an interactive whiteboard is presented.

Keywords: didactic material, interactive whiteboard, Music, sound, picture, cognitive styles

Spletna učilnica kot pripomoček za shranjevanje in dostop do didaktičnega gradiva
Virtual Classroom as a Tool for Storage and Access to Teaching Materials

Alenka Kralj
Osnovna šola Šmartno, Šmartno pri Litiji
Pungrt 9, 1275 Šmartno pri Litiji
Podružnična šola Štangarske Poljane
alenka.kralj@guest.arnes.si

Povzetek

V uvodnem delu članka se bralec na kratko seznani z nekaterimi vzroki, ki učitelja usmerijo v oblikovanje spletne učilnice. Osrednji del članka predstavlja spletno učilnico s prosto dostopnim didaktičnim gradivom ter opisuje nekatera gradiva in dejavnosti, ki jih uporablja učitelj – oblikovalec spletne učilnice pri delu z učenci v prvih letih otrokovega šolanja.

Ključne besede: spletna učilnica, kombinirani pouk, didaktično gradivo, moodle, indirekten pouk

Abstract

In the introductory part of the article the reader gets acquainted with particular causes which direct a teacher towards shaping a virtual classroom. The central part of the article presents a virtual classroom with a free-accessible didactic material and describes some materials and activities used by the teacher – the former of the virtual classroom at working with pupils in the first years of their schooling.

Keywords: virtual classroom, combined lessons, didactic material, moodle, indirect lesson

Uporaba programa Art Rage pri likovni vzgoji v prvi triadi OŠ
Using Art Rage in Art Education in the First Three Year Cycle in Primary School

Irena Kresevič
OŠ Rudolfa Ukoviča Podgrad
Podgrad 99b, 6244 Podgrad
ikresevic@gmail.com

Povzetek

V prispevku je predstavljen program Art Rage. To je brezplačen računalniški program, primeren za vse starosti učencev. S pomočjo omenjenega programa učenci spoznavajo strojno in programsko oprema računalnika ter usvajajo cilje in znanja z različnih predmetnih področij. Prikazane so možnosti uporabe omenjenega programa pri realizaciji ciljev likovne vzgoje v prvi triadi. Avtorica ugotavlja, da so bili učenci za delo s tem programom zelo motivirani, saj so ga pričeli uporabljali tudi pri drugih dejavnostih in v prostem času.

Ključne besede: Art Rage, likovna vzgoja, risanje, slikanje, učni načrt

Abstract

This contributionr presents a program Art Rage. This is a free computer program suitable for all age students. With the help of the program pupils learn about hardware and software equipment, acquire computer skils and goals in different subject areas. It shows the potential uses of the program in the realization of the objectives of art education in the first three years. The author notes that pupils have been working with this program are very motivated, since they started using it for other activities and leisure.

Keywords: Art Rage, art education, drawing, painting, curriculum

Sodobna e-gradiva – skupina NAUK
Modern E-learning Resources – NAUK Group

Matija Lokar	**Boris Horvat**	**Iztok Kavkler**	**Primož Lukšič**	**Alen Orbanić**
Univerza v Ljubljani, Fakulteta za matematiko in fiziko Inštitut za matematiko, fiziko in mehaniko Jadranska 19, 1000 Ljubljana matija.lokar@fmf.uni-lj.si	Univerza v Ljubljani, Fakulteta za matematiko in fiziko Inštitut za matematiko, fiziko in mehaniko Jadranska 19, 1000 Ljubljana boris.horvat@fmf.uni-lj.si	Univerza v Ljubljani, Fakulteta za matematiko in fiziko Inštitut za matematiko, fiziko in mehaniko Jadranska 19, 1000 Ljubljana iztok.kavkler@fmf.uni-lj.si	Univerza v Ljubljani, Fakulteta za matematiko in fiziko Inštitut za matematiko, fiziko in mehaniko Jadranska 19, 1000 Ljubljana primoz.luksic@fmf.uni-lj.si	Univerza v Ljubljani, Fakulteta za matematiko in fiziko Inštitut za matematiko, fiziko in mehaniko Jadranska 19, 1000 Ljubljana alen.orbanic@fmf.uni-lj.si

Povzetek

Cilji skupine NAUK (Napredne Učne Kocke) je razvoj tako teoretičnih kot praktičnih izhodišč na področju uporabe IKT v izobraževalnem procesu na vseh nivojih. Pri tem se ukvarjamo predvsem s problematiko gradnje in uporabe e-gradiv, razvijamo orodja, potrebna za izdelavo, uporabo in prilagajanje e-gradiv ter seveda tudi izdelujemo tudi sama e-gradiva. V prispevku so prikazana e-gradiva, ki so nastala v zadnjem letu in pokrivajo področja računalništva, matematike, fizike in logike. Dosegljiva so bodisi prek osrednjega portala http://www.nauk.si bodisi neposredno prek repozitorija http://gradiva.nauk.si.

Ključne besede: e-izobraževanje, učni gradniki, e-gradiva

Abstract

The objectives of the NAUK group (Advanced Learning Cubes) are the development of both theoretical and practical concepts at all levels of educational ICT. In so doing, we explore the problems of construction and use of e-learning resources. Furthermore, we are also developing the tools necessary to manufacture, use and adapt the resources and, at the same time, we develop the resources. The article presents e-learning resources that were prepared in the last year, covering the areas of computer science, mathematics, physics and logic. They are accessible either through a central portal http://www.nauk.si or directly through the repository at http://gradiva.nauk.si.

Keywords: e-learning, learning blocks, e-learning resources creation

Izhodišča za pripravo e-učbenikov
Baselines for the Preparation of Electronic Textbooks

Matija Lokar
Univerza v Ljubljani, Fakulteta za matematiko in fiziko
Inštitut za matematiko, fiziko in mehaniko
Jadranska 19, 1000 Ljubljana
matija.lokar@fmf.uni-lj.si

Primož Lukšič
Univerza v Ljubljani, Fakulteta za matematiko in fiziko
Inštitut za matematiko, fiziko in mehaniko
Jadranska 19, 1000 Ljubljana
primoz.luksic@fmf.uni-lj.si

Boris Horvat
Univerza v Ljubljani, Fakulteta za matematiko in fiziko
Inštitut za matematiko, fiziko in mehaniko
Jadranska 19, 1000 Ljubljana
boris.horvat@fmf.uni-lj.si

Povzetek

V okviru skupine NAUK (Napredne Učne Kocke) – http://www.nauk.si se ukvarjamo z razvoj tako teoretičnih kot praktičnih izhodišč na področju uporabe IKT v izobraževalnem procesu na vseh nivojih. Pri tem je v zadnjem času zelo v ospredju problematika e-učbenikov. Kakšen naj bo e-učbenik, kakšne naj bodo njegove značilnosti, v čem naj se razlikuje od klasičnega učbenika, itd. V prispevku bomo poskusili predstaviti nekaj priporočil, za katere menimo, da naj jih upoštevajo avtorji e-učbenikov.

Ključne besede: e-izobraževanje, e-učbeniki

Abstract

The NAUK group (Advanced Learning Cubes) - http://www.nauk.si is engaged in the development of theoretical and practical concepts of ICT use in all levels of education. Recently, the proposed introduction of electronic textbooks has become a major topic nationwide. There are multiple dilemmas to be solved, e.g. what is the e-textbook, what should it include, what should its differences be from the conventional textbook, etc. This paper attempts to make some recommendations, which we believe should be taken into account by the authors of e-textbooks.

Keywords: e-learning, electronic textbooks

Priporočila za delo z glasovalnimi napravami pri slovenščini
Tips for using Interactive Response System in Slovene Lessons

Tatjana Lotrič Komac
OŠ Naklo
Glavna cesta 47, 4202 Naklo
tatjana.lotric@guest.arnes.si

Tina Žagar Pernar
OŠ Naklo
Glavna cesta 47, 4202 Naklo
tina.zagar1@guest.arnes.si

Povzetek

Za učitelja so povratne informacije o znanju učencev ključnega pomena, saj s tem lažje in uspešneje nadzoruje osvojene funkcionalne in izobraževalne cilje ter nadgrajuje učne vsebine. Tradicionalne oblike pridobivanja tovrstnih povratnih informacij od učencev so omejene na dvigovanje rok oz. izpostavljanje učencev, ki jih učitelj izbere sam. Sodobne oblike zbiranja podatkov o znanju in mnenjih učencev z glasovalnimi napravami omogočajo pregled nad večjim delom učne populacije, saj lahko svoje mnenje oz. odgovor sočasno poda večje število učečih se, ki ob tem dobijo tudi takojšnjo povratno informacijo o svojih dosežkih.
Prispevek tako po eni strani prinaša priporočila za delo z glasovalnimi napravami, temelječa na izkušnjah avtoric, po drugi pa tudi ideje, kako glasovalne naprave izrabiti na več načinov.

Ključne besede: interaktivna tabla, glasovalne naprave, slovenščina, interaktivna učna gradiva

Abstract

Information about student understanding is key element for teacher as it provides easier and more successful supervision over already achieved learning goals. Traditionally teachers were relying on hand-raising, volunteering and selecting students. Modern forms of collecting information about knowledge and opinion (interactive response system) enable closer look into data base of all students' achievements, as they can all simultaneously answer the questions or give opinion and furthermore response system gives them feedback.
On one hand paper gives tips for using interactive response system, based on authors' experiences, on the other hand it points out some ideas how to increase creativity in using interactive response system.

Keywords: interactive whiteboard, interactive response system, Slovene language, interactive learning resources

Modeli za razvijanje in izboljševanje ravni različnih vrst pismenosti s pomočjo IKT
Models for Developing and Upgrading Various Levels of Literacy by Using ICT Tools

Marija Lubšina Novak
marija.lupsina@guest.arnes.si

Povzetek

Spremembe v informacijski družbi zahtevajo od posameznika drugačna znanja in spretnosti, kot jih ponuja klasična, tradicionalna šola. Zato v sodobni šoli iščemo drugačne pristope, da bi kar najbolj zagotovili celosten razvoj vseh veščin in spretnosti ter kompetenc mladega človeka. V projektu Berem in ustvarjam – se učim, ki ga podpirata Evropski socialni sklad in Ministrstvo za šolstvo in šport RS, smo zastavili in preverjali učinkovitost modelov za razvijanje in izboljšanje ravni bralne pismenosti za otroke in mladostnike. Na Osnovni šoli Bizeljsko smo po modelu 4 v Mitološki bralni zbirki skušali doseči zastavljene cilje.

Ključne besede: branje, ustvarjanje, IKT, opismenjevanje, modeli, medpredmetno povezovanje

ABSTRACT

Modifications in information society demand different knowledge and skills from individuals than the ones offered by classical, traditional school. Contemporary schools are searching for different approaches to ensure integrated development of all skills and competences of a young adult. One of the goals of "I am reading and creating – I am learning" project, which is supported by the European Social Fund and the Ministry of Education and Sport, was to check the efficiency of the models for developing and upgrading various levels of reading literacy among children and adolescents. According to Model 4 in the Mythological Reading Collection in Bizeljsko Primary School we have tried to achieve previously set goals.

Key words: reading, creating, ICT, elimination of illiteracy, models, connecting subjects

Učna ura s pomočjo e-gradiv
A Lesson with E-materials

Jožica Mlakar Broder
OŠ Matije Valjavca Preddvor
Šolska ulica 9, Preddvor
jozica.mlakar-broder@guest.arnes.si

Povzetek

V 6. razredu smo pri pouku matematike poskusno uporabljali e-gradiva in ugotovili, da so zelo uporabna za utrjevanje in preverjanje znanja, za pridobivanje dodatnih informacij in za samostojno učenje. Učenci jih lahko uporabljajo tudi kot samostojno domače delo. Na roditeljskem sestanku sem starše povabila v računalniško učilnico in jim prikazala učno uro s pomočjo e-gradiv. Ponovili smo množenje in deljenje decimalnih števil, pridobljeno znanje pa nato preverili še s preverjanjem in kontrolno nalogo. Starši so bili nad takim načinom učenja navdušeni, saj so ugotovili, da je matematika lahko tudi zanimiva. Obljubili so, da bodo svoje otroke spodbujali k takemu načinu dela tudi doma – tako bodo čas, ki ga vsakodnevno preživijo pred računalnikom, koristno uporabili.

Ključne besede: e-gradiva, samostojno učenje, domače delo

Abstract

E-materials have been experimentally used in the 6th grade for learning and revision of maths, for gaining new information as well as for autonomous learning. The students can use e-materials as autonomous homework.
Learning with e-materials was introduced to parents on parent meeting. They were invited to the computer room where they revised multiplying and dividing decimal numbers, later they checked their knowledge with revision and test. Parents were impressed by such learning method. They found maths interesting to learn. They promised to encourage their children to use e-materials at home. Thus, the time spent on a computer would be more useful.

Keywords: e-materials, autonomous learning, homework

Uporaba e-gradiv pri pouku geografije
Teaching Geography with E-materials

Mirjam Martinuč Bernard
Gimnazija Jesenice
mirjam.martinuc-bernard@guest.arnes.si

Marjana Kralj
Srednja gostinska in turistična šola Radovljica

Povzetek

Pri delu s pričujočimi e-gradivi »Obča geografija in izbrani problemi sodobnega sveta«, se dijaki seznanjajo s teorijo, pridobijo sposobnost samostojne uporabe elektronskih medijev za pridobivanje podatkov in komunikacijo z součečimi. Učijo se dela s podatki in vrednotijo rezultate. Pričujoča e-gradiva odlikuje postopnost, ki jo omogočajo večstopenjski multimedijski gradniki od klasičnih fotografij do tematskih zemljevidov, video posnetkov, animacij in simulacij. Učenje in poučevanje z aktualnimi e-gradivi med drugim poteka na Gimnaziji Jesenice ter Srednji gostinski in turistični šoli Radovljica. Ugotovitve so sledeče: a) Obča geografija je med projekti e-gradiv, ki so aktualna preko podjetja Videofon d.o.o. nadpovprečno obiskana; tako po številu klikov učečih, kot po številu mentorskih skupin; b) dojemljivost dijakov za tovrstno metodo dela je visoka; c) uporabnost e-gradiv je posebno visoka v športnih oddelkih, pri dijakih s statusom kulturnika in dijakih, ki so zaradi zdravstvenih težav dalj časa odsotni od pouka; d) e-gradiva se uporablja pri ponavljanju in utrjevanju znanja za maturo.

Ključne besede: e-gradiva, geografija, srednja šola, multimedija, elektronski mediji

Abstract

The presented e-material »General geography and the problems of modern world" provide students the theory, the ability to start independently using electronic media to gather information, as well as a method to communicate with their classmates. Students learn how to work with the data and evaluate the results. The highlight of the presented e-materials is a design, based on multimedia elements designed with several stages and presented as classical photographs, thematic maps, videos, animations and simulations. Learning and teaching with the presented e-materials is performed at Gimnazija Jesenice and Srednja gostinska in turistična šola Radovljica. The findings are the following: a) among e-material projects the General geography from Videofon d.o.o. is attended above average measured by click numbers of students and by number of mentor groups; b) methodology is easily understood by students; c) e-materials proved especially useful for sport classes, students with status of artist, and students missing classes due to health problems; d) e-materials are used in preparation for general examination before leaving secondary school.

Keywords: e-material, geography, secondary school, multimedia, electronic media

Implementacija IKT v projektu Multikulturna glasbena vzgoja
The Implementation in the Project Multicultural Music Education

Lorena Mihelač
Šolski center Novo mesto, Srednja šola Metlika
Partizanski trg 4, 8330 Metlika
lorena.mihelac@guest.arnes.si

Povzetek

Članek predstavlja izvedbo projekta »Multikulturna glasbena vzgoja«, ki poteka že peto leto na Šolskem centru Novo mesto, v izobraževalni enoti Srednja šola Metlika. Projekt sloni na e-gradivu »Glasbeno popotovanje« istega avtorja, ki je objavljeno in javno dostopno tudi na internetnih straneh (glej: http://www.unisvet.si/index/index/activityld/82). Članek prikazuje različne faze projekta, od začetnega dela in postavitve samega projekta, izdelave seminarske naloge, PPT-predstavitve do praktične realizacije projekta v razredu. Pojasnjuje vlogo dijakov in učitelja ter pomembnost dobre koordinacije in interakcije za uspešno realizacijo projekta. Izpostavlja vlogo IKT-ja v projektu, kako preiti od začetne ideje uporabe IKT-ja do konkretne realizacije. Članek pojasnjuje tudi vlogo spletne učilnice, brez katere bi bila sama realizacija projekta precej bolj zapletena in bistveno manj učinkovita. Izpostavlja možnosti medpredmetne povezave (npr. glasba in tuji jezik) glede na to, da so določene informacije dostopne samo na tujem jeziku. Članek prikazuje tudi, katere dejavnosti je možno nadaljevati ob zaključku projekta ter kako pridobiti skupni izdelek nekega razreda, ki bi lahko koristil vsem dijakom tega razreda v njihovi bodoči stroki.

Ključne besede: multikulturna glasbena vzgoja, IKT

Abstract

The article is representing the realization of the project »Multicultural music education«, which is already the fifth year a part of the music subject in the School center Novo mesto, Srednja šola Metlika. The project is based on the e-pedagogical material »Travelling with music« from the same author, and is available also on the internet sites (look at: http://www.unisvet.si/index/index/activityld/82). The article is showing different stages of work in this project, from the very first beginning till the realization of the seminar excercise, PPT presentation and the practical performance in the classroom. It explains the role of the students and the teacher as also the impact of a well coordinated work and interaction on the succesful realization of the project. The article points out also the role of ICT, how to implement the idea of using ICT into a real process, and it tries to explain the role of the »internet classroom«, which is crucial in the realization of the project. Some interdisciplinar connections are exposed in the project as also the possibilities how to upgrade the project after its realization, and how to create materials connected with this project, which could help the students in their professional development.

Keywords: multicultural music education, ICT

Model ocenjevanja institucij za izobraževanje zaposlenih
Assessment Model for Employee Educational Institutions

Modrijan Urška
Suhadolčanova 86, 1000 Ljubljana
modrijan.urska@gmail.com

Vladislav Rajkovič
Univerza v Mariboru, Fakulteta za organizacijske vede
Kidričeva 55a, 4000 Kranj, Slovenija
vladislav.rajkovic@fov.uni-mb.si

Povzetek

Namen raziskave je bilo ugotoviti kako izbrati najprimernejšo izobraževalno institucijo, ki bo izobraževala zaposlene. Pri izboru najprimernejše pa se ponavadi pojavi problem, ko želimo izbrati tisto, ki bo izobraževanja izvajala. Na trgu je veliko število izobraževalnih institucij, ki ponujajo izobraževanja za zaposlene.
Raziskava je bila narejena s pomočjo programa Dex-i, v katerem je bil oblikovan odločitveni model za izobraževanje zaposlenih. Na podlagi različnih dejavnikov, kateri so pomembni za izobraževanje zaposlenih, so bile ocenjene izobraževalne institucije. V modelu je bila obravnavana ponudba izobraževanj, ugled institucije in ostala ponudba, ki vpliva na kvaliteto. Podatki so bili nato prenešeni v program Vredana, s pomočjo katerega so bile narejene tudi številne primerjave in analize variant (analiza tipa "kaj-če").
V uvodu magistrskega dela so predstavljeni razlogi, zakaj danes tudi v naši družbi že namenjamo več pozornosti izobraževanju odraslih, kakšne so razmere na področju izobraževanja odraslih in kako naprej ter zavedanje vrednosti intelektualnega kapitala v podjetju. V poglavju o sprejemanju odločitev in o modelih odločanja je naveden postopek za reševanje problema za izbor najprimernejše izobraževalne institucije. V modelu so opisane štiri (hipotetične) izobraževalne institucije za izobraževanje zaposlenih. Model je bil oblikovan v programu DEX-i in prenešen v program Vredana, s pomočjo katerega so bile narejene tudi številne primerjave in analize variant. Tako je v zaključku predstavljena najprimernejša izobraževalna institucija, ki bo izobraževala zaposlene. Podjetje bo imelo bolje izobražen kader za delo, katerega zaposleni v podjetju opravljajo.

Ključne besede: Izobraževanje, vseživljenjsko izobraževanje, Izobraževalna institucija, znanje

Abstract

The goal of this research was to find the way, how to choose the most appropriate educational institution, which will educate the employees. But when choosing the right one to execute the education usually a problem appears. There are many institutions which offer education for employees.
The research was made using a programme called Dex-I, which was used to form a decision model for educating employees. Based on different factors important for educating people the institutions were assessed. This model discussed education offer, institutions reputation and other offers influencing the quality. The data was loaded in a program named Vredana, which was used to make numerous comparisons and variants analysis' (what –if analysis).
The introduction presented the reasons why we focus more and more attention to educating adults, what is the adult educating situation and what is the awareness of the value of potential capital. Decision making section indicated the procedure for solving the problem of choosing the most appropriate institution. 4 different hypothetical educational institutions were described. The model was formed in DEX-i program and was loaded to Vredana which enabled making numerous comparisons and analysis'. So the last part of the thesis was used to present the most appropriate institution to educate adults. A company will have better educated employees to execute their work.

Keywords: Education, Educational institution, Potential capital, Individual evolution, Success

Uporaba programa Slikar in Microsoft Word v učnem procesu drugega vzgojno-izobraževalnega obdobja v osnovni šoli
ICT Literacy by using Computer Programmes Microsoft Word and Paint with Children in the Second Trienniun of Primary School

Božena Opara
Osnovna šola Trzin
Mengeška 7b, 1236 Trzin
opara.bozena@gmail.com

Povzetek

V članku predstavljam izvedbo projektnega učnega dela v okviru medpredmetnih povezav slovenskega jezika, likovne vzgoje in družbe. Predstavitev predstavlja dejavnosti, ki ustrezajo učnim ciljem 4. razreda osnovne šole. Cilj projekta je bil z uporabo računalniške tehnologije napisati fantazijsko zgodbo in jo opremiti s fantazijsko sliko. Učenci so po branju in obravnavi zgodbe V. Pečjaka Drejček in trije marsovčki s pomočjo računalniškega programa Slikar narisali vesoljsko vozilo, ga kopirali v program Microsoft Word in opremili z fantazijsko zgodbo, ki je ustrezala naslovu Drejček in marsovčki potujejo po vesolju. Izdelke so kopirali, javno predstavili, ovrednotili in izdelali razstavo. Vsi učenci so dosegli zastavljene učne cilje.

Ključne besede: Projektno učno delo, računalniška programa Slikar, Microsoft Word, ikona, datoteka, shrani, kopiraj, prilepi, literarno delo, otroški fantazijski svet, poustvarjalno pisanje, likovno ustvarjalna slika, predstavitev in vrednotenje izdelka, razstava, medpredmetna povezava.

Summary

The article presents the realization of project work in context of cross-curricular links of Slovenian language, art education and society education. The presentation represents activities which correspond to the learning objectives of the forth grade of primary school. The project aim was using ICT technology to write a fantasy story and draw a fantasy picture. After reading and hearing stories by Vid Pečjak "Drejček in trije marsovčki" the students have used a computer program Paint and have drawn a spacecraft, they have copied it into Microsoft Word and equipped it with a fantasy story, corresponding to a story "Drejček in marsovčki potujejo po vesolju" (Drejček and the Martians travel in space). The products were copied, publicly presented, evaluated and exhibited. All students have achieved their learning goals.

Keywords: project work, computer programmes Paint and Microsoft Word, an icon, file, save, copy, paste, literary work, children's fantasy world, reproducing writing, creative art image, presentation and evaluation of the product, exhibition, cross-curricular links.

Medpredmetno povezovanje mehatronika-informatika
Interdisciplinary Integration Mechatronics – Information Science

Tanja Pirih
tanja.pirih@gmail.com

Povzetek

Pri obravnavi učne snovi o električni energiji pri predmetu mehatronika smo se odločili, da raziščemo porabo električne energije na naši šoli.
Želeli smo ugotoviti, koliko energije lahko privarčujemo že z enostavnimi ukrepi, kot so ugašanje luči, ko le te niso več potrebne (prostori niso zasedeni ali je dovolj naravne svetlobe) in z ugašanjem naprav, kadar jih ne potrebujemo.
Da bi bil projekt bolj nazoren in zanimiv za dijake smo se povezali s predmetom Informatike. Pri delu smo uporabili naslednje programe: v Pro/Engineer, AutoCAD, PhotoFiltre, Word, Excel in PowerPoint.
Ugotovili smo, da smo z enostavnimi ukrepi kot so ugašanje luči, ko le te niso več potrebne (prostori niso zasedeni ali je dovolj naravne svetlobe) in ugašanjem naprav prihranili: 26,45 % pri VT (visoki tarifi) in 31,18 % pri MT (mali tarifi).
Dijaki so bili nad medpredmetnim povezovanjem navdušeni, kar dokazujejo tudi številni dobri izdelki.

Ključne besede: varčevanje, razsvetljava, energija, osvetljenost, luxmeter, Pro/Engineer, AutoCAD, PhotoFiltre, Word, Excel, PowerPoint

Summary

Concerning the electrical energy at the Mechatronics class we decided to evaluate the use of electrical energy of our school.
We want to find out how much energy we can save with simple measures such as switching off the lights when they are not needed (when classrooms are either not used or there is enough natural light) and by turning off all the devices when not in use.
To make the class more interesting for the pupils we joined with Information Science class. Doing the work we used the following computer programmes: Pro/Engineer, AutoCAD, PhotoFiltre, Word, Excel and PowerPoint.
We found out that simple measures, such as turning off the lights when the classrooms are not in use (e.g. all breaks) and when there is enough natural light as well as turning off the devices when not in use, saved 26,45 % of high price energy and 31,18 % of low price energy.
Pupils were of interdisciplinary integration very impressed as evidenced by a number of good products.

Keywords: saving, lighting, energy, illumination, luxmeter, Pro/Engineer, AutoCAD, PhotoFiltre, Word, Excel, PowerPoint

Kodeks uporabe IKT v izobraževanju
Code of Conduct for using ICT in Education

Franci Pivec
IZUM Maribor
Franci.Pivec@izum.si

Povzetek

Kodeksi ravnanja (in etični kodeksi) so način zagotavljanja prevlade pozitivnih vplivov v skupnosti. Visokošolska okolja, ki so standardizirala ravnanja z IKT, izkazujejo višjo kakovost delovanja od tistih, ki tega niso storila. Značilna je tudi njihova večja pripravljenost za spremembe. Univerza je prostor znanstvene komunikacije, zato ji IKT in še posebej internet predstavlja vstop v novo razvojno fazo in temu se najboljše univerze močno posvečajo. S tem se tudi etika vrača v jedro poslanstva visokošolskih institucij. Številne države po svetu sprejemajo standarde uporabe izobraževalne tehnologije (NETS), ki so nastali in se stalno dopolnjujejo v okviru združenja ISTE in predstavljajo kodeks ravnanja študentov, učiteljev, administratorjev in drugih nosilcev kakovostnega študija. Ob standardih pa je potrebno zagotoviti tudi bistvene pogoje za njihovo izvajanje.

Ključne besede: kodeks ravnanja, etični kodeks, etična zrelost, ideja univerze, standardi izobraževalne tehnologije (NETS), ISTE

Summary

Codes of conduct (and ethical codes) are a way of ensuring that positive impact in the community prevails. Higher education environments that have standardized ICT management show higher quality of performance if compared to those who have not. Also characteristic for these environments is their strong willingness for change. A university is a place of scientific communication and, thus, ICT and especially the Internet represent access to a new development phase to which the best universities are strongly dedicated. In this way, also ethics returns to the core of the mission undertaken by higher education institutions. Many countries around the world are adopting National Educational Technology Standards (NETS), which have been developed and are continuously updated within the ISTE Association and at the same time represent a code of conduct for students, teachers, administrators and all others involved in high-quality study. Along with the standards it is necessary to provide the essential conditions for their implementation.

Keywords: code of conduct, ethical code, ethical maturity, idea of university, National Educational Technology Standards (NETS), ISTE

Od spletne oglasne deske do spletne učilnice
From the E- notice Board to the Online Classroom

Stanislava Polajžer
ŠCRM Kamnik
Novi trg 41 a, Kamnik
stanislava.polajzer@guest.arnes.si

Povzetek

Pri svojem delu z udeleženci Izobraževanja odraslih sem ugotovila, da ne zadoščajo več le gradiva, dostopna na spletni oglasni deski, ampak da se pri številnih udeležencih kaže potreba po bolj individualnem komuniciranju s profesorjem, s pomočjo kakšnega drugega ustreznega medija in ne le z osebnim kontaktom v šoli.
Po pregledu in presoji obstoječih možnosti in po posvetu z e-strokovnjakinjo na naši šoli sem se odločila, da si delo in učenje predmetov, ki jih poučujem, olajšam s kakovostno komunikacijo preko spletne učilnice.
V prispevku je opisano, kako je potekalo delo prek spletne oglasne deske in uvajanje spletne učilnice v Izobraževanje odraslih v okviru maturitetnega tečaja pri pripravah na maturo iz geografije ter kako spletno učilnico sprejemajo udeleženci in drugi vpleteni v Izobraževanje odraslih.

Ključne besede: spletna oglasna deska, komuniciranje v spletni učilnici, spletna učilnica, spletna geografija.

Abstract

Working with Adult education students I realised that resources available on the e-notice board are no longer sufficient. Many students expressed their needs for an individual communication with the professor by means of other more appropriate media as the personal contact at school is no longer enough.
After having checked and evaluated the possibilities with the school e-expert, I decided to make my work and teaching easier with the good quality communication over the online classroom.
In this article I describe how the online classroom work went on, how the online classroom was introduced to the Adult education, particularly to the course called »maturitetni tečaj« at geography and how the online classroom was accepted by the Adult education students and other participants.

Keywords: e-notice board, communicating via the online classroom, the online classroom, online geography

IKT pri pouku fizike
ICT at Physics Lessons

Betka Potočnik
Osnovna šola Križe
Cesta Kokrškega odreda 16, Križe
betipot67@gmail.com

Povzetek

Razvoj računalniške tehnologije z uporabo različnih računalniških orodij in medijev omogoča vrsto novih možnosti za popestritev in dopolnitev pouka fizike. Učitelj jih uporablja v celotnem učnem procesu in tako povečuje nazornost pouka, motiviranost učencev in aktivnost učencev pri pouku. Ob tem pa se učitelj sooča tudi s povsem novimi pogoji poučevanja. Hiter razvoj informacijske tehnologije je povzročil, da se učenci učijo na drugačen način. V prispevku je opisan primer uporabe informacijsko-komunikacijske tehnologije pri izvedbi opazovalne naloge pri predmetu fizika.

Kjučne besede: informacijsko komunikacijska tehnologija, učni proces, fizika, aktivne metode učenja, opazovalna naloga

Summary

The development of computer technology with the use of different computer tools and media enables various possibilities of how to make the subject of physics more interactive and interesting. The technology can be used in the entire learning process which contributes to clarity of the lessons, pupils' motivation and it gets children to be more involved in the learning process. Along with this the teachers are faced with new teaching conditions. The rapid technology development paved the way to pupils learning differently nowadays. The article deals with the use of information communication technology at the observational task at physics' lessons.

Keywords: Information communication technology, learning process, physics, active learning methods, observational task

IKT in domače naloge
Doing Homework by using E-materials

Andreja Prezelj
OŠ Koroška Bela Jesenice
C. talcev 2, 4270 Jesenice
andreja.prezelj@telesat.si

Povzetek

V članku posvečam pozornost domačim nalogam, ki so obveznost in odgovornost vsakega učenca. Glede na to, da le te večini učencev pomenijo neke vrste popoldansko nadlogo, ki je ne izpolnijo, sem se odločila, da jim ponudim drugačen način domačega dela – domače naloge, ki jih opravljajo s pomočjo IKT (uporaba računalnika, e – gradiv). Pripravila sem aktualne in zanimive vsebine, z namenom, da jih pritegnem k opravljanju domačega dela, ki vodi k utrjevanju in poglabljanju znanja, pridobivanju učnih navad in usvajanju novih tehnik učenja. Takšno delo vodi tudi v odkrivanje zanimivih vsebin, h katerim se lahko učenci vračajo v svojem prostem času in čas ob računalniku kvalitetno preživljajo.
Takšna oblika dela od učitelja zahteva skrbno načrtovanje in posredovanje domačih nalog. V članku sem predstavila nekaj primerov domačega dela z uporabo računalnika. Praksa je pokazala, da se učenci takšnih nalog lotevajo z veseljem in so za takšno delo precej bolj motivirani. To utemeljujem z rezultati ankete, ki sem jo izvedla med učenci in s primerjavo števila ustrezno opravljenih domačih nalog. Za učitelja pa je takšno delo tudi naporno, saj zahteva veliko inovativnosti ter predvsem dodatnega dela in časa. Takšne naloge lahko postanejo tudi oblika dela z nadarjenimi učenci.

Ključne besede: domača naloga, uporaba računalnika, e – gradiva, evalvacija

Summary

In my article I am dealing with the pupils' homework at Geography lessons since homework is every single pupil's responsibility and obligation. During my teaching I have come to the conclusion that homework is usually 'a pain in the neck' and that pupils usually avoid it. Therefore I offered different kinds of homework – homework based on ICT that includes computer learning and use of e-learning materials. I prepared interesting and up-to-date topics in order to revise and improve their knoledge as well as to develope their learning strategies and techniques. E-learning materials can help pupils to gain new knowledge and they can also spend some quality time using computer.
This method reqiures a very careful planning. In my artilce there are some examples of good practice included. Pupils were willing to do their homework because e-materials raise their motivation and I can confirm that statement with the results of my survey made among pupils. From the teacher's point of view such homework certainly demands inovation and extra working time. Homework dealing with e-materials is suitable also as a method of teaching and learning for gifted children.

Keywords: homework, use of computer, e-learning materials, evaluation

Priprava in analiza govornega nastopa z uporabo IKT
Preparation and Analysis of Oral Presentation by using ICT

Stanka Rakar
OŠ Koroška Bela Jesenice
Cesta talcev 2, 4270 Jesenice
stanka.zupancic@telesat.si

Marija Mežnarec
OŠ Koroška Bela Jesenice
Cesta talcev 2, 4270 Jesenice
marija.meznarec@guest.arnes.si

Povzetek

Otroci so doma pripravili govorni nastop, pri katerem so na različne načine iskali informacije, ki so jih predstavili v govornem nastopu. V šoli sva jim predstavili na kakšen način vse lahko pridobivajo informacije: iz knjig, revij, preko interneta. V raziskavi, ki sva jo izvedli po končani nalogi, sva ugotavljali, katere vire so uporabili. Ugotovili sva, da je večina otrok, s pomočjo staršev, informacije in slikovno gradivo za predstavitev dane naloge, iskala in našla s pomočjo računalnika in interneta.

Govorni nastop sva v šoli posneli in ga kasneje prikazali učencem. Pri delu sva uporabili kamero, računalnik in projektor. Posnetke sva uporabili za ogled govornega nastopa in za analizo le tega. Otrokom sva predstavili kriterije ocenjevanja, nato pa smo skupaj analizirali in ocenjevali nastope.

Otroci so bili nad uporabo IKT navdušeni in si želijo še več takega pouka.

Za to metodo dela sva se odločili, ker otroci v tem obdobju že radi uporabljajo računalnik, predvsem za igranje raznih igric. Želeli sva jim približati računalnik tudi kot vir raznih informacij. V tem primeru so iskali razne podatke in sličice na dano temo govornega nastopa. Poleg tega sva želeli učencem prikazati uporabo kamere. Na koncu so si svoj nastop lahko še ogledali. Učenci so s pomočjo posnetka videli sebe pri govornem nastopu in zraven analizirali, kaj jim je šlo dobro, kaj pa bi lahko še izboljšali.

Ključne besede: govorni nastop, pridobivanje informacij, IKT

Abstract

Pupils prepared their oral presentation at home using different sources: books, magazines, internet. At school we informed them how to gain different pieces of information. After pupils' oral presentations we gave a questionnaire to all the pupils in order to find out what kind of sources they used. The majority of pupils found the information and pictures/photos on the internet. Parents often helped them.

We filmed all oral presentations and later pupils watched themselves on the video. We needed a camera, a computer and a projector. Videos were used for self-analysis of every single pupil. Based on the assessment criteria pupils could analyse and assess their presentations.

Pupils liked ICT a lot and they would like to use it even more often.

We used this method because pupils at that age start to use computer mostly for playing computer games. We wanted to present them the computer as a source of different information. They had to find different data and pictures/photos in order to prepare their oral presentation. We also used a camera and we filmed pupils because they could analyse themselves and they could find new ways to improve their presentations in the future.

Keywords: oral presentation, acquiring information, Information and Communication Technologies (ICT)

Mehatronizacija tehnoloških procesov z didaktičnih vidikov
Mechatronisation of Technological Processes from Didactic Aspects

Gorazd Rakovec
Cegelnica 25a, 4202 Naklo
gorazd.rakovec@siol.net

Povzetek

Klasičen način poučevanja pričenja najprej s podajanjem temeljnih zakonov in se nadaljuje z njihovo sintezo v teorijo obravnavanega področja – v našem primeru mehatronizacijo tehnoloških procesov. Pristop, ki poudarja strokovne vidike, pa pričenja poučevanje kompetenc z analizo praktičnega primera, ki ga razčlenjuje in abstraktno poglablja vse do ciljne teorije, ki jo želi obravnavati. Študent pride do teorije po obratni poti – z analizo primera - namesto s sintezo osnovnih abstraktnih zakonov. Njegove pridobljene kompetence zato ne temeljijo na abstraktnih neoprijemljivih pojmih, temveč na zaznavah v fizičnem svetu. Študent spozna isto teorijo, le da so njeni temelji na trdnih tleh fizičnega sveta, zato teorijo tudi laže uporablja v praksi. Na ta način se tudi skrajša uvajalno obdobje študentov, ki pridejo iz šole na delo v industrijo.

Ključne besede: poučevanje kompetenc, mehatronika, tehnološki procesi, nadzorni sistemi, didaktični sistem, analiza primerov

Abstract

Classical way of lecturing begins with teaching of basical laws continuing with their synthesis to theory of basic field – mechatronisation of technological processes in our case. The approach from the professional side begins lecturing with the analysis of practical case which is analysed to the final theory. The student learn the theory with the case analyse instead of synthesis of abstract basic laws. His competences are based on physical cases instead of abstract ideas. The student learns the same theory, but the fundamentals of knowledge are of physical nature and the theory is usable in the practice. The probation time of students which finished the study and start working in industry is much shorter.

Keywords: teaching competences, mechatronics, technological processes, monitoring systems, didactical system, case study

Timsko poučevanje slovenščine preko omrežja
Team Teaching of Slovene by the Help of the Internet

Irena Rakovec Žumer
ESIC Kranj, Gimnazija
Kidričeva cesta 65, 4000 Kranj
irena.rakovec-zumer@guest.arnes.si

Povzetek

V š. l. 2008/2009 smo v gimnazijah začeli poučevati po posodobljenih učnih načrtih. Vpeljane so bile didaktične novosti, kot so timsko poučevanje in uporaba novih tehnologij. Pri dijakih na ta način razvijamo različne kompetence: sporazumevalno, medkulturno in digitalno. IKT nam je v veliko pomoč pri timskem poučevanju. Uporabljali smo internetna gradiva, pesmi in odlomke poslušali v različnih jezikih, pogledali videospote, naredili blog, delali v šolski spletni učilnici ...
Pri poučevanju športnikov pa je IKT še prav posebej nepogrešljiva. Dijaki so veliko odsotni in le ob pomoči IKT lahko opravijo vse obveznosti, zato avtorica v prispevku predstavlja projekt Timsko poučevanje dijakov športnikov preko omrežja, pri katerem je v š. l. 2009/10 sodelovalo Ministrstvo za šolstvo in šport skupaj z Zavodom za šolstvo, CPI, Olimpijskim komitejem Slovenije, Združenjem športnih zvez in Telekomom Slovenije. V š. l. 2009/10 je Gimnazija ESIC Kranj na področju timskega poučevanja športnikov preko omrežja skupaj s še dvema šolama (II. Gimnazijo Maribor in ŠCPET) orala ledino. Projekt pa se bo v š. l. 2010/11 nadaljeval v obliki enoletnega poskusa Uvajanje izobraževanja na daljavo za dijake športnike. Vodil ga bo ZRSŠ, poleg naše šole bodo sodelovale še štiri splošne gimnazije.

Ključne besede: timsko poučevanje, IKT, projekt Timsko poučevanje dijakov športnikov preko omrežja, projekt Uvajanje izobraževanja na daljavo za dijake športnike

Abstract

In the school year 2008/2009 the renewed updated curricula were launched in the Gimnazija schools. Along with this some didactic novelties were introduced, like team work and the use of modern technologies; as in this way the students will be able to develop various competences simultaneuously: communicational, intercultural and digital. ICT is very helpful at any team teaching. We took the advantage of internet materials, we listened to poems and various literary abstract in different languages, we had a look at the related video spots, further on a blog was set up and a lot of activites carried out in the web classroom.
ICT is indispensable at teaching students – sportspeople, as they are often absent from classes, so ICT enables them to meet all the school requirements – thanks to the Slovene Ministry of Education and Sport (MSŠ), The National Education Institute (ZRSŠ), the Slovene Institute for Vocational Education and Training (CPI), the National Olympic Commitee, the Union of Slovene Sports Associations and Telecom Slovenia, which have decided to support the project Team teaching of the students sportspeople with internet. And our school centre ESIC Kranj is glad to to be able to participate in this project. In the school year 2008/09 the Gimnazija school ESIC Kranj with two other schools started with team teaching of students -sportpeople through internet network (2nd Gimnazija school Maribor and ŠCPET). The project will continue experimentally in the school year 2010/11 under the name: Introducing virtual training of the students – sportspeople. The project will be lead by the ZRSŠ. Our shool with another four will participate in the project.

Keywords: team teaching, ICT, project Team teaching of the students – sportspeople with internet, Introducing virtual training of the students – sportspeople

Videokonferenca – nove možnosti in priložnosti v izobraževanju
Videoconference – New Possibilities and Opportunities in Education

Janja Razgoršek
Šolski center Slovenj Gradec, Višja strokovna šola
Koroška 11, 2380 Slovenj Gradec,
janja.razgorsek@sc-sg.net

Zdenko Potočar
Ekonomska šola Novo mesto, Višja strokovna šola
Ulica talcev 3a, 8000 Novo mesto
zdenko.potocar@guest.arnes.si

Povzetek

Avtorja v prispevku obravnavava nov, sodoben način izobraževanja, in sicer s pomočjo videokonferenc. Videokonference omogočajo dvosmerno komunikacijo in aktivno sodelovanje študentov v izobraževalnem procesu. Opisala sva, kaj so videokonference, kakšne so njihove prednosti v izobraževanju, pa tudi kakšne pomanjkljivosti prinašajo s seboj in kako se jim lahko izognemo. V empiričnem delu predstavljava, kako sta potekala vpeljava in usposabljanje za uporabo videokonferenc pri poučevanju na Višji strokovni šoli Slovenj Gradec in Višji strokovni šoli Novo mesto. Prikazala sva, kako so potekale priprave na posamezne videokonference in kako je potekala njihova izvedbo. Skušala sva ugotoviti, koliko so bile videokonference učinkovite, in kako doseči še večjo učinkovitost le-teh v prihodnje. Prispevek sva zaključila z napotki za izvedbo učinkovitih videokonferenc v izobraževanju.

Ključne besede: videokonference, videokonferenčni prenos, učinkovite videokonference v izobraževanju

Abstract

In the contribution we deal with a new, modern way of education, education with the aid of videoconferences. Videoconferences allow a two-way communication and an active participation of students in the educational process.We describe what videoconferences are and what their advantages are in education, but also what flaws they carry with them and how to avoid these. In the empirical work we present how the introduction of and training on the use of videoconferences for teaching purposes took place on the Vocational College Slovenj Gradec and the Vocational College Novo mesto. We demonstrate the course of our preparations for individual videoconferences and how they were executed. We try to determine how effective the videoconferences were, and how to achieve an even greater effectiveness in the future. The contribution concludes with directives for an effective execution of videoconferences in education.

Keywords: videoconferences, videoconference transfer, effective videoconferences in education

Z informacijsko tehnologijo od tihega k »glasnemu« znanju
From Tacit to Noticeable Knowledge with Information Technology

Vanda Rebolj
vanda.rebolj@siol.net

Povzetek

V članku najprej predstavimo znanje in argumentiramo uporabo informacijske tehnologije za njegovo merjenje. Osredotočimo se na znanje, ki ga pridobimo v neformalnem izobraževanju in priložnostno, saj se zdi, da je merjenje formalno pridobljenega znanja bolj razvito. Iščemo rešitev, da bi bilo merjenje znanja, ne glede na izvor, neovirano dostopno in množično. Predstavljamo načrtovanje, izdelavo in evalvacijo avtomatiziranega e-testa. Strokovno ga obravnavamo z dveh strani: tehnološko in humanistično, to je vezano na védenje o človeku. Zanima nas veljavnost testov in sicer kako dobro izmerijo, kar želimo meriti? O e-testiranjih razmišljamo ekonomično, zato nas zanima, kaj prinašajo h klasičnim oblikam merjenja znanja.
Opozarjamo na pomen priprave testiranca na testiranje, še posebej, ker u merimo tiho znanje, ki je konceptualno neizoblikovano. Posebej predstavljamo, kako izdelamo ponovljive teste in kako znanje certificiramo. Dokazujemo dodano vrednost takih testov, zato jih je smiselno razvijati, jih razširiti iz ekspertnega v vsako okolje in dalje raziskovati.

Ključne besede: e-test, veljavnost testa, neformalno pridobljeno znanje, tiho znanje, testni program, testno orodje, merjenje znanja

Abstract

The article first presents knowledge and arguments for the use of information technology for its measurement. It focuses on non-formally and informally acquired knowledge since measurement of formally acquired knowledge seems to be more developed. The article seeks solutions for mass access to measurement of knowledge regardless to its origin. Planning, producing and evaluation of an automatic e-test are presented. Two expert viewpoints are discussed: technological and humanistic which is bound to knowledge about humans. Validity of tests and their measurement of what is wished to be measured are the article's interest. E-testing is thought about in an economical way, so it is of interest what it brings to classical forms of knowledge measurement.
Preparation of the testee for the testing is pointed out, especially because of measurement of salient knowledge, which is conceptually undefined. Every test inevitably includes recognizable structures and concepts. Producing repeatable tests and certification of knowledge are also presented. Since there is proof of added value of such tests, there are indications for development, transfer from expert to non-expert environments and further research.

Keywords: e-test, test validity, unformal knowledge, tacit knowledge, testing software, testing tool, knowledge measurement

Priprava slikovnega gradiva in stavnice s programom SMART Notebook
Preparing Images and beting with the Program SMART Notebook

Darja Rijavec
Branik 33 P, 5295 Branik
darja.rijavec@gmail.com

Povzetek

Interaktivna tabla s programon SMART Notebook (v slovenščini SMART beležnica), omogoča številne dejavnosti, s pomočjo katerih učenci osvajajo novo znanje na veliko bolj nazoren in zanimiv način. Članek predstavlja možnost uporabe slikovnega gradiva pri pouku v prvi triadi in natančen opis prenosa slik iz programa Tux Paint, Art Rage , Photo Filtre in interneta v SMART beležnico. Avtorica predstavi tudi izdelavo in uporabo stavnice pri začetnem opismenjevanju.

Ključne besede: SMART Notebook, slikovno gradivo, stavnica

Abstract

Interactive whiteboard with SMART Notebook programon (in Slovenian SMART notebook), provides a series of activities to help students acquire new knowledge in a much more obvious and interesting way. This paper presents the possibility of using images in teaching in the first three and the precise transfer of images from the program Tux Paint, Art Rage, Photo Filtre and Internet in SMART notebook. The author presents the development and use of betting in early literacy.

Keywords: SMART Notebook, illustrations, Gambling

Uporaba interaktivne table oziroma i-table pri pouku zgodovine
The use of an Interactive Whiteboard in a History Classroom

Vesna Robnik
OŠ Mislinja
Šentilj pod Turjakom 1, 2382 Mislinja
vesna.robnik@gmail.com

Povzetek

Interaktivne table so v svetu že zelo uveljavljene kot kvalitetno učno sredstvo v izobraževanju. Pri nas pa se učitelji z njimi v večjem številu srečujemo šele v zadnjem času. Poučevanje z interaktivno tablo oziroma i-tablo je zagotovo zanimivejše, nazornejše in kvalitetnejše.
Učenci ob pravilni uporabi i-table niso samo opazovalci dogajanja na tabli, ampak tudi njegovi soustvarjalci. Delo z i-tablo je zasnovano tako, da zahteva tako od učitelja kot od učencev aktivno vlogo. Osnovni namen je, da učence spodbudimo k samostojni, ustvarjalni in poglobljeni izgradnji in razumevanju učne snovi ter bolj poglobljenemu in povezanemu znanju.

Ključne besede: Aktivna vloga učitelja in učenca, i-tabla, učno sredstvo, poučevanje, zanimivejše, nazornejše, poglobljeno in povezano znanje.

Abstract

Interactive boards have become established teaching tools in today's education. However, teachers in Slovenia have only started using them in recent years. Teaching with the help of interactive boards is definitely more interesting, demonstrative and qualitative.
With the right use of interactive boards learners are not just observers but also become co-creators. Working with interactive boards demands active teacher and learner roles. The main purpose of interactive boards is to encourage learners to be more independent and creative in the process of learning. It also helps them strengthen und connect different facts.

Key words: Active teacher and learner role, interactive board, teaching tool, teaching, more interesting, clearer and more connected knowledge.

Muzejska razstava skozi objektiv kamere
Museum Exhibition as seen through the Camcorder

Vesna Robnik
Šolski center Slovenj Gradec
Srednja ekonomska šola Slovenj Gradec
Koroška ulica 11, 2380 Slovenj Gradec
vesna.robnik@gmail.com

Maja Kumprej Gorjanc
Koroški pokrajinski muzej
Glavni trg 24, 2380 Slovenj Gradec
maja.kumprej.gorjanc@gmail.com

Marcelj Furman
Šolski center Slovenj Gradec
Srednja ekonomska šola Slovenj Gradec
Koroška ulica 11, 2380 Slovenj Gradec
furman.marcelj@gmail.com

Tine Pajk
OŠ Mislinja
Šentilj pod Turjakom 1, 2382 Mislinja
tine.pajk@gmail.com

Povzetek

Živimo v času, ko naš vsakdan obdajajo različne nove tehnologije. Uporaba teh tehnologij pri pouku pa je učencem danes blizu, saj je internet že vsakdanji način sporazumevanja z ljudmi in nepogrešljiv vir pri iskanju informacij mladih. Zaradi tega se je nam učiteljem Srednje ekonomske šole Slovenj Gradec porodila ideja o videokonferenčni učni uri pri pouku zgodovine. Za izvedbo učne ure smo za sodelovanje zaprosili Koroški pokrajinski muzej. Videokonferenca je komunikacija dveh ali več ljudi na daljavo. Pri videokonferenci se prenašata zvok in slika, lahko pa se prenašajo tudi drugi podatki. Za vzpostavitev videokonference potrebujemo zaslon, spletno ali drugo kamero, mikrofon, zvočnik ali slušalke, internet in ustrezno programsko opremo (Skype, Cu- See Me, …). Za videokonferenco smo uporabili program Skype. Potrebovali smo tudi zaslon, spletno kamero, mikrofon, slušalke in internetno povezavo.
Pri izvedbi učne ure smo uresničili cilje, ki smo si jih zadali:

- aktivna vloga učitelja in učenca
- raziskovalno učenje
- razvijanje zgodovinskega mišljenja
- sposobnost analiziranja in razvijanja kritičnega mišljenja.

Ključne besede: zgodovina, komunikacija na daljavo, videokonferenca, kritično mišljenje, kustos

Abstract

We live in a time when different sorts of technology surround us. Nowadays, students are familiar with the use of technology in school lessons. The Internet has become a way of communication between people as well as an indispensable tool for searching for information. It is because of this fact that we came up with the idea of using videoconference in a History lesson. Our partner for Co-operation was Carinthian Regional Museum. Videoconference is communication between two or more people that do not have direct physical contact but are at different locations. When talking about videoconferences, we talk about the transmission of sound and picture. However, other data can also be transmitted. To set up a videoconference one needs to have a monitor, a webcam, a microphone, speakers or headsets, the Internet and the appropriate software (Skype, Cu- See Me, …). We used the programme Skype to set up the videoconference. We also needed a monitor, a webcam, a microphone, headsets and an Internet connection.
We realized the following lesson aims:

- active role of the teacher and students,
- research learning,
- the development of historical thinking,
- the ability to analyze and develop critical thinking.

Keywords: lesson, History, long distance communication, videoconference, critical thinking, curator

Odločitveni model kot podpora pri izbiri učnega pripomočka v visokem šolstvu
Decision Support Model for Selecting a Learning Tool in Higher Education

Helena Erika Rojc
Univerza v Ljubljani, Biotehniška fakulteta
Jamnikarjeva 101, 1000 Ljubljana
helenaerika.rojc@bf.uni-lj.si

Petra Grošelj
Univerza v Ljubljani, Biotehniška fakulteta
Jamnikarjeva 101, 1000 Ljubljana
petra.groselj@bf.uni-lj.si

Lidija Zadnik Stirn
Univerza v Ljubljani, Biotehniška fakulteta
Jamnikarjeva 101, 1000 Ljubljana
lidija.zadnik@bf.uni-lj.si

Povzetek

Razvoj informacijske tehnologije je prinesel novosti tudi v šolstvu. S pojavom novega učnega pripomočka – interaktivne table se je začel spreminjati tudi način poučevanja. Učitelji se skušajo preko interaktivnih vsebin približati današnji generaciji mladih, ki informacijsko tehnologijo uporablja vsakodnevno, ne samo za razvedrilo in pridobivanje informacij, ampak tudi za učenje. Ker več osnovnih in srednjih šol uporablja interaktivne table, je smiselno razmišljati o njihovi vpeljavi v visoko šolstvo. V prispevku smo zgradili matematični model, ki je v pomoč pri odločanju, kakšna tabla bi največ doprinesla k spremembam in napredku poučevanja. S pomočjo metode analitičnega mrežnega procesa (ANP), smo glede na izbrane kriterije in podkriterije med seboj primerjali tri vrste šolskih tabel: interaktivno, belo in zeleno. Uteži za izbrane kriterije in podkriterije smo izračunali s pomočjo računalniškega programa Super Decisions. Rezultati so kot najboljšo izbiro pokazali interaktivno tablo, ki predavatelju omogoča večjo dinamičnost, trajnost zapisa in možnost sprememb.

Ključne besede: večkriterijsko odločanje, analitični mrežni proces (ANP), šolska tabla

Abstract

The development of information technology has brought innovations in education. The way of teaching started to change with the emergence of a new learning tool - an interactive whiteboard. Teachers are trying to approach interactive content to today's generation of young people who daily use information technology not only for entertainment and information retrieval, but also for learning. A growing number of primary and secondary schools use interactive whiteboards, so it is reasonable to think about their introduction into higher education. In this paper we have built a mathematical model that helps in deciding which kind of school boards could most contribute to changes and progress in teaching. The analytic network process (ANP) was used for selecting the most suitable school board among an interactive, white and green board, according to selected criteria and subcriteria. Preferences for the selected alternatives were calculated by the computer program Super Decisions. Results have shown that the best choice is an interactive whiteboard, which allows the lecturer to become more dynamic, enables records durability and the possibility of changes.

Keywords: multiple criteria decision-making (MCDM), analytic network process (ANP), school board

Priprava e-gradiv z matematično nadarjenimi učenci
Creating E-materials with Students gifted in Mathematics

Magda Slokar Čevdek
OŠ Solkan, Slovenija
magda.slokar@guest.arnes.si

Povzetek

Zakon o osnovni šoli uvršča nadarjene učence v skupino učencev s posebnimi potrebami in nalaga šolam, da tem učencem prilagodijo metode in oblike dela. V strokovni literaturi ni enotne definicije nadarjenosti. Razlog je v tem, da nadarjeni niso neka homogena skupina, ampak se nadarjenost kaže v različnih oblikah in obsegih. Poleg običajnih učnih programov potrebujejo nadarjeni za optimalen razvoji svojih sposobnosti tudi prilagojen pouk in dejavnosti. V konceptu Odkrivanje in delo z nadarjenimi učenci v osnovni šoli so navedene številne oblike dela, med njimi tudi vzporedni programi (pull-out) in individualni programi za delo z nadarjenimi. Kot rezultat takšnega načina dela so nastala e-gradiva, ki so predstavljena v prispevku.
Uvodoma najprej spregovorimo o nadarjenih učencih in osvetlimo kompleksnost definicije nadarjenosti. Sledi predstavitev značilnosti matematično nadarjenih učencev. V nadaljevanju podrobneje opišemo pripravo e-gradiv za matematiko z nadarjenimi učenci in predstavimo dva primera takšnih gradiv.

Ključne besede: nadarjeni učenci, matematika, e-gradiva, interaktivne naloge, algebrski izrazi, vzporedni programi

Abstract

Gifted students are placed into the groups of students with special needs according to the Basic School Act which demands schools to adapt the methods and the ways of teaching.
There is no unified definition of a talent in professional literature. The gifted children are not a homogeneous group because the gift is shown in different shapes and sizes.
Besides general tuitional programmes they need adapted lessons and activities for the optional development of their talents.
In the concept Detecting the Gifted Students and Activities with them in Primary and Junior High School numerous activities are given such as parallel (pull-out) and individual programmes. E-materials presented in my contribution are the result of that way of teaching.
In the first part of my contribution we discuss the gifted students and enlighten the complexity of the definition. While in the second part the characteristics of the students gifted in mathematics are presented. In the following part creating e-materials with the students are described and two examples of such materials are presented.

Keywords: gifted students, mathematics, e-materials, interactive exercises, algebraic expressions, parallel programmes

Academic Teachers' Perceptions about ICT in Education, a Survey

Solza Sopova Grceva
FON University, Faculty for ICT
Bul. Vojvodina bb, Skopje, R.Macedonia
solza.grceva@fon.edu.mk

Abstract

In this paper we present the analysis, conclusions derived from a universities/faculties survey, concerning the use and perception of ICT in education. It has been conducted during the third year of implementation of the EU project "Integrating E-Learning across Teachers Curriculum".
Our overall objectives were to understand the relationship the current faculty members have with educational technology, and to determine how educational technology is affecting their teaching practices. Our belief is that this information will be useful for the universities/faculties to build up their future policies.
More specifically, we tried to evaluate the following issues:

- Faculty perceptions and priorities with respect to educational technology
- Faculty experiences with and problems using educational technologies
- Faculty preferences regarding technology support and education

Keywords: teachers survey, ICT in education, ICT perceptions

Integration of ICT across the Educational System: General Guidelines

Solza Sopova Grceva
FON University, Faculty for ICT
Bul. Vojvodina bb, Skopje, R.Macedonia
solza.grceva@fon.edu.mk

Ivan Kraljevski
FON University, Faculty for ICT
Bul. Vojvodina bb, Skopje, R.Macedonia
Ivan.kraljevski@fon.edu.mk

Abstract

ICT supported or technologies enhanced teaching includes a whole range of necessary resources, policies and skilled teachers. This issue involves infrastructure, equipment and facilities, educational management, teacher training and support, capacity building of educational institutions, educational content, and above all educational policy.
Teachers' faculties are faced with the challenge of preparing a new generation of teachers for effective use of new learning tools in their teaching practices. For many teacher education programs, this daunting task requires the acquisition of new resources, expertise and careful planning. Above all, it is highly necessary to integrate ICT or even further, E-Learning across teachers' curriculum.
This paper is designed to provide a guide to help teacher educators, schools, administrators and policy-makers infuse, integrate, or embed ICT into teachers' education. The main factors and indicators of success are determined. The strategy and resources presented in this paper were developed by an international group of experts with extensive experience in the integration of ICT into teacher education. It is a result of a successful 3-years European project with a title "Integrating E-Learning across the teachers' curriculum", (IETC), implemented in R. Macedonia. Four teachers' training faculties and four pilot schools (three primaries and one secondary) from R. Macedonia, together with the Universities of Edinburgh, Groningen and Leuven were involved.

Keywords: ICT in education, ICT curriculum, teaching resources

Odločitveni model za vključevanje osnovnih šol v mednarodne projekte za vseživljenjsko učenje
The Decision Making Model for Intergration of Primary Schools in International Lifelong Learning Projects

Maruška Spasovski
maruska.spasovski@gmail.si

Povzetek

V devetletni osnovni šoli si čedalje več učiteljev prizadeva popestriti učne ure, učencem približati učno snov in spodbuditi učence k samostojnemu delu in učenju. Z vključevanjem v razne projekte, tako učitelji kot učenci na zanimiv način pridobivajo nova znanja in spoznanja.Mednarodni projekti in sodelovanje v njih predstavljajo velik izziv za šolo. Sodelovanje v projektu in uspešna izvedba projekta sta zelo zapletena, zato je potrebno natančno preučiti kateri projekt ustreza določenim zahtevam.V članku je predstavljen postopek vrednotenja mednarodnih projektov za osnovne šole. V prvem delu je opredeljen problem, metodologija ter informacije o posameznih mednarodnih projektih.Pomen opisnih kriterijev pri ocenjevanju projekta je podan v drugem delu. S pomočjo računalniškega programa Dexi pa je, ob upoštevanju kriterijev, predstavljen odločitveni model, katerega rezultat je končna ocena projekta.S pomočjo modela ocenjevanja je predstavljen nov, preprost in natančen način izbire najbolj primernega projekta, ki bo olajšal delu učitelju, ki se spopada s takimi problemi.

Ključne besede: vseživljenjsko učenje, vrednotenje, mednarodni projekti, samostojno delo, program Dexi

Abstract

In the 9-year programme of primary school education teachers give more and more effort to diversify school lessons. They are also trying to bring educational material and to encourage pupils to work and learn independently.By participating in various projects, both teachers and pupils are on an interesting way acquiring new knowledge and insights.International projects represent a major challenge for the school. Participation and the successful implementation of the project are very complex. It is necessary to carefully examine which project fits for certain requirements.The article presents procedure of evaluating international projects for elementary schools.The first part defines the problem, methodology and information on various international projects.The second part discusses the significance of written reports. The model and his criteria are introduced with the help of the Dexi programme. The result is the final assessment of the project.The model presents a new, simple and accurate way of choosing the most suitable project. The model will also facilitate the work of teacher who is faced with such tasks and problems.

Keywords: lifelong learning, evaluation, international projects, independent work, Dexi programme

Nove zmožnosti spletne učilnice Moodle 2.0
New Features of Online Learning - Moodle 2.0

Sašo Stanojev
ESIC, Ekonomsko – trgovska šola
Cesta Staneta Žagarja 33, 4000 Kranj
saso.stanojev@guest.arnes.si

Povzetek

Platforma za spletne učilnice Moodle nam v svoji novi različici prinaša precej novosti, tako na skrbniškem, kot na uporabniškem nivoju. V prispevku bodo predstavljene nekater ključne posodobitve in pridobitve, ki bodo učiteljem omogočile sodobnejši pristop k organizaciji in vodenju spletne učilnice prek izboljšanega uporabniškega vmesnika in z uporabo storitev Web 2.0.

Ključne besede: Moodle 2.0, novosti, Splet 2.0, odprta koda

Abstract

Moodle learning management in his new version delivers significant new features at the administration and user level. The article presents some key updates and additions, which will enable teachers to more modern approach to organizing and managing virtual classroom through improved user interface and the use of Web 2.0 services.

Keywords: Moodle 2.0, features, Web 2.0, opensource

Medpredmetno povezovanje angleškega jezika in računalniške stroke
Inter-subject Connections of English Language and Computer Science

Gašper Strniša
Tehniški šolski center Kranj
Kidričeva 55, 4000 Kranj
gasper.strnisa@gmail.com

Povzetek

Računalništvo in angleški jezik sta neposredno povezana v vseh pogledih, zato je medpredmetno povezovanje med temi predmeti praktično nujno. V prispevku so predstavljena teoretična izhodišča in prednosti medpredmetnih povezav, ter praktični primer izvedbe učne ure z naslovom: Uvod v vizualno programiranje. Rezultati ankete so pokazali, da je takšen način poučevanja dijakom zelo všeč in da si želijo več takšnih ur. Učiteljem takšno delo predstavlja veliko dodatnega dela, a se je zaradi zadovoljstva dijakov vredno potruditi.

Ključne besede: medpredmetna povezava, računalništvo, angleški jezik, dijaki, tehnična gimnazija.

Abstract

Computing and the English language are directly connected in every aspect and that's why inter-subject connections between those two school subjects are so important. In this contribution the theoretical starting points and the advantages of inter-subject connections are presented. There is also a practical case of a lesson: Introduction to visual programming. The results of a survey showed that students enjoy this way of teaching and that they want more of such lessons. Such a way of teaching presents teachers with a lot of additional work but it is worth trying because of the students' satisfaction.

Keywords: inter-subject connections, computer science, the English language, students, technical high school.

Uporaba IKT v šoli
Application of ICT in Schools

Tatjana Šček Prebil
BIC Ljubljana, Živilska šola
tatjana.scek@guest.arnes.si

Povzetek

Ritem in način življenja v 21. stoletju sta neposredno povezana z informacijsko- komunikacijsko tehnologijo (IKT), ki je bistven element za nadaljnji razvoj gospodarstva in družbe. V svetu globalizacije in hitrih sprememb je IKT v šolah pomemben dejavnik družbenega razvoja. Klasične oblike poučevanja in vodenja šole se prepletajo s sodobnimi oblikami. Vseživljenjsko učenje postaja realnost. Koliko in kako zaposleni uporabljajo IKT pri svojem delu, bomo prikazali v spodnjem prispevku.

Ključne besede: šola, učitelji, dijaki, učenje, znanje, IKT

Abstract

The rhythm and lifestyle in the 21st century are closely connected with information and communication technology (ICT) which remains the essential element for further development of economy and society. In the world of globalization and constant changes the ICT application in schools has become meaningful and valuable. The classic methods of teaching as well as managing a school interweave with the modern ones. The lifelong learning has become a reality. The extent of ICT application is demonstrated in the following article.

Keywords: school, teachers, students, learning, knowledge, ICT

Uporaba IKT pri pouku zgodovine v OŠ:
Holokavst skozi oči otroka v času 2. svetovne vojne
Holocaust through the Eyes of a Child

Marjeta Šifrer
OŠ Naklo, Naklo
marjeta.sifrer@guest.arnes.si

Matjaž Jeruc
OŠ Naklo, Naklo
matjaz.jeruc@guest.arnes.si

Povzetek

V prispevku je predstavljen primer dobre prakse pri poučevanju zgodovine v osnovni šoli. Tema Holokavst in usoda Židov v času 2. svetovne vojne, je predstavljena s pomočjo življenjske zgodbe deklice Hane. Za aktivno in ustvarjalno poučevanje zgodovine si pomagamo tudi z uporabo informacijskih in komunikacijskih tehnologij. Učence navajamo na aktivno in koristno uporabo računalnika ter interneta, pri iskanju informacij ter pošiljanju sporočil.

Ključne besede: informacijsko komunikacijska tehnologija, Žid, Jud, holokavst, genocid, nacisti, Davidova zvezda, koncentracijska taborišča

Abstract

This article describes a practical example of teaching history in primary school. Holokaust and the fate of Jews during the 2nd World War is represented through the life story of a girl named Hana. Information and communication technology is used to teach history in a more active and creative way and also to encourage pupils in their quest of finding a more useful and active way to search the information on the World Wide Web and to communicate via computer.

Keywords: David star, holokaust, information and communication technology, genocide, nazis, concentration camps, Jews

Razvijanje digitalne kompetence v procesu posodabljanja gimnazij
Development of Digital Competence in the Process of High School Updating

Tomaž Šmid
Valjavčeva 4, 4000 Kranj
tomaz.smid@guest.arnes.si

Povzetek

V prispevku kot primer razvijanja predvsem digitalne kompetence v procesu uvajanja prenovljenih učnih načrtov in posodabljanja gimnazije navajam poseben projekt, ki smo ga razvili na Gimnaziji Ledina in je imel za osnovo medpredmetno povezovanje, timsko učenje in uporabo IKT. Osnovni cilj je bil pri dijakih razvijati kompetenco medkulturnosti in digitalno kompetenco.Pri obravnavi štirih besedil iz sklopa slovenskega povojnega pripovedništva so bila le-ta predstavljena s pomočjo digitalnih tehnologij z njihovega slovenističnega, sociološkega, zgodovinskega, filozofskega in psihološkega vidika.
Dijaki so sodelovali v spletni komunikaciji na forumih, klepetalnici, izdelovali wiki in pisali preverjanje v obliki internetnega testa. Izvedena je bila tudi videokonferenca med dijaki Gimnazije Ledina ter Gimnazije Novo mesto.S takim načinom dela so dijaki razvijali različne spretnosti in kompetence: učne, sodelovalne, interakcijske, komunikacijske ter miselne strategije, kot so ustvarjalnost, kritičnost, reševanje problemov.

Ključne besede: digitalne tehnologije, digitalna kompetenca, videokonferenca, wiki, posodabljanje gimnazij timsko poučevanje

Abstract

In this article I primarily speak about developing digital competence in the process of introducing revised curriculum and updating of high schools giving an example of a project realized at Gimnazija Ledina. This was a project in many ways based on cross-curricular coordination and cooperation, team work and the usage of ICT. The major objective is to develop the idea of intercultural and digital competences.
There were selected four texts from the Slovene literature from the period after the Second World War, and analyzed from a different point of view at the following subjects: Slovene, Sociology, History and Psychology. The students cooperated in various forums, chat rooms, the making of Wiki; their knowledge was tested on the Internet. We also managed to organize a video-conference between high schools Ledina and Novo mesto.By taking part at this project, the students developed various skills and competences: learning, cooperation, interaction, communication, and not the least how to think creatively, critically, and with the ability to focus on problem-solving tasks.

Keywords: digital technologies, digital competence, video-conference, Wiki, grammar school update, tandem teaching

Kako z manj narediti več? (Uporaba programske opreme za IKT na naši šoli) How to make More with Less? (The usage of ICT Software at our School)

Jožica Štrajhar
Ulica Toma Brejca 17, 1241 KAMNIK
jozica.strajhar@amis.net

Povzetek

Varčevanje je tudi v šoli pomembna vrednota: s financami, z učiteljevim časom. Za nakup računalnikov še nekako zberemo finančna sredstva (v šoli in doma), pri nakupu strojne opreme pa zmanjkuje sredstev. Ker moramo delovati vzgojno, brez »piratskih« verzij, rešujemo problem z boljšo izrabo programov, ki jih kupimo, in z uporabo brezplačnih programov. Opisujem, kako na naši šoli poskušamo z manj plačljive programske opreme narediti več. Varčujemo tudi z učiteljevim časom – s pripravo kratkih izobraževanj za uporabo računalniških programov in objavo gradiv v spletni učilnici.

Ključne besede: programska oprema, Publisher, PhotoFiltre, PhotoMix, ExeLearning, Notebook, Edison, CiciCad

Abstract

At our school a lot of attention is being paid on economizing since this is of big worth in schools: saving with our teachers' time, with finances. We somehow manage to gather means (at school and at home) for the purchase of computing machines, however always seems to be a kind of reduction for the purchase of software. Since we have to act educatively, without »pirate« versions, we try to solve this problem with the better exploitation of software, which was been purchased as well as the usage of free software programmes. We want to show our effort at our school to do more with less payable software programmes. We also economize with our teacher's time – we arrange short courses for our teachers to get them acquainted with the usage of computer programmes and we publish materials on our websites.

Keywords: software, Publisher, PhotoFiltre, PhotoMix, ExeLearning, Notebook, Edison, CiciCad

Digitalna enciklopedija naravne in kulturne dediščine na slovenskem ter njen pomen za vzgojo in izobraževanje v informacijski družbi Digital Encyclopaedia of the Slovenian Natural and Cultural Heritage and Its Significance for Education in Information Society

Gregor Torkar
ZRC SAZU
Geografski inštitut Antona Melika
Gosposka ulica 13, Ljubljana
gregor.torkar@guest.arnes.si

Boris Horvat
IMFM & Hruška d.o.o.
Jadranska 19, 1000 Ljubljana
boris.horvat@fmf.uni-lj.si

Mateja Šmid Hribar
ZRC SAZU
Geografski inštitut Antona Melika
Gosposka ulica 13, Ljubljana
mateja.smid@zrc-sazu.si

Daniel Vladušič
XLAB d.o.o.
Pot za Brdom 100, 1000 Ljubljana
daniel.vladusic@xlab.si

Povzetek

V prispevku je predstavljena Digitalna enciklopedija naravne in kulturne dediščine na Slovenskem (DEDI). Gre za prvi poskus večmedijske digitalne predstavitve slovenske naravne in kulturne dediščine. DEDI na enem mestu predstavlja 4 vrste dediščine (nepremično, premično in živo kulturno ter naravno dediščino) v 3 okoljih: v sklopu digitalne enciklopedije, v interaktivnem spletnem atlasu Geopedija ter v tri-razsežnostnem geografskem informacijskem sistemu Gaea+. V prispevku so predstavljeni različni načini iskanja in uporabe informacij enciklopedije DEDI ter njihova vloga v vzgojno-izobraževalnem procesu. Prispevek je dober vir informacij in navodil za vse učitelje, ki želijo uporabljati enciklopedijo DEDI v razredu.

Ključne besede: digitalna enciklopedija, dediščina, vzgoja in izobraževanje, e-učenje

Abstract

The paper presents Digital encyclopaedia of the Slovenian natural and cultural heritage (DEDI) which is the first attempt to multimedia digital presentation of Slovenian natural and cultural heritage. DEDI presents four different types of heritage (immovable, movable and live cultural and natural heritage) in 3 areas: in the context of the digital encyclopaedia, an interactive online atlas Geopedija and three-dimensional geographic information system Gaea +. The paper explains different ways of searching and using of information in DEDI and its importance for education in information society. It is a good source of information and instructions for teachers, who want to start using DEDI in the classroom.

Keywords: digital encyclopaedia, heritage, education, e-learning

Učenje programske opreme s pomočjo video vsebin
Software Learning with the Help of Video Content

Marko Urh
Univerza v Mariboru, Fakulteta za organizacijske vede
Kidričeva 55a, 4000 Kranj, Slovenija
marko.urh@fov.uni-mb.si

Povzetek

V prispevku so predstavljene lastnosti, trenutno stanje na trgu in značilnosti učenja programske opreme s pomočjo video vsebin. Zahtevnejše stranke, vse hujša konkurenca, hiter tehnološki razvoj in druge značilnosti današnjega sveta so vzroki, ki narekujejo drugačno funkcioniranje posameznika, tako v poslovnem kot zasebnem življenju. Učenje in izobraževanje pri tem ni izjema. Prikazane so tehnološke, finančne, pedagoške in ostale lastnosti, ki jih prinaša učenje programske opreme s pomočjo video vsebin. Med drugim so podane prednosti in slabosti klasičnega učenja in učenja programske opreme s pomočjo video vsebin. Prikazani so nekateri pozitivni trendi, ki nakazujejo pomembnosti učenja programske opreme. Predstavljene so značilnosti največjih svetovnih ponudnikov učenja programske opreme s pomočjo video vsebin in njihov pogled na celovito oskrbo udeležencev v procesu izobraževanja. Zaključek prispevka podaja predloge in mnenja, ki jih lahko posamezniki in organizacije uporabijo pri izdelavi vsebin namenjenim učenju s pomočjo računalnika.

Ključne besede: e-učenje, učenje programske opreme, video vsebine

Abstract

In this article we get to know the features and characteristics of a current market situation and learning through video content. The customers now days are becoming more and more demanding, there’s a fierce competition on the market, progress in technology is developing rapidly and that all leads and dictates a different functioning in the life’s of individual's, business and private. Learning and education is no exception. Article presents the technological, financial, educational and other features that learning through video content's brings. We also get to know the advantages and disadvantages of both classical learning and learning about software through video content's and some of the positive trends which indicate the importance of learning about software. Some of the world’s largest providers that offer how to learn about software through video content are also shown in the article, their opinions and looks on the comprehensive care of the participants in the learning process. In the end of the article I conclude with some suggestions and opinions that both individuals and organizations can use when making content’s that are used for software learning.

Keywords: e-learning, software training, video courses

Raziskava o uporabi in sevanju mobilnih telefonov med dijaki
Research about Using and Radiation of Mobile Telephones between Students

Barbi Vidmar
Srednja šola Zagorje
Cesta zmage 5, 1410 Zagorje ob Savi
barbi_vidmar@yahoo.com

Povzetek

V naši raziskovalni nalogi smo poskušali ugotoviti, koliko sevanja oddajajo naši mobilni telefoni, koliko časa uporabljajo mobilne telefone naši dijaki in kaj sploh vedo o sevanju. Postavili smo dve hipotezi:

1. Mobilni telefoni oddajajo sevanje v okolico.
2. Srednješolci se premalo zavedajo sevanja mobilnih telefonov in drugih tehnologij in ne razmišljajo o posledicah.

Prvo hipotezo smo potrdili z merjenjem sevanja mobilnih telefonov z osebnim dozimetrom v razredu, v kleti in zunaj. Mobilni telefoni sevajo, vendar še v okviru dovoljenih mej.
Drugo hipotezo smo potrdili z anketo, ki so jo izpolnili vsi dijaki naše šole: o sevanju vedo zelo malo ali nič. Pomembno bi bilo vse srednješolce osveščati o tem, da vsa tehnologija oddaja določeno sevanje in jih seznaniti z ukrepi, kako zmanjšati vpliv sevanja na posameznika. Predlagali smo nekaj ukrepov.

Ključne besede: mobilni telefon, sevanje, tehnologija, bazna postaja.

Abstract

The subject of this research is trying to find out how much radiation emit our mobile telephones, how much time use mobile telephones our secondary school students and how much they know about radiation. We created two hypotheses:
Mobile telephones emit radiation in the environment.
Secondary school students know very little or nothing about radiation of mobile telephones and other technology. They don’t think about consequences.
First hypothesis was confirmed with measurement radiation of mobile telephones in the class, in the cellar and outside. Mobile telephones emit radiation, but in allowed borders.
The second hypothesis was confirmed with questionnaire.
It is importuned to elucidate our secondary school students, that all technology emit radiation. We suggested some steps about that.

Keywords: mobile telephone, radiation, technology.

Zahtevana znanja diplomiranih informatikov
Knowledge and Skills Requirements for MIS Graduates

Borut Werber
Univerza v Mariboru, Fakulteta za organizacijske vede
Kidričeva 55a 4000 Kranj
borut.werber@fov.uni-mb.si

Daša Jakše
Univerza v Mariboru, Fakulteta za organizacijske vede
Kidričeva 55a 4000 Kranj
dasa.jakse@fov.uni-mb.si

Povzetek

V prispevku so predstavljeni in analizirani delni rezultati predštudije, ki je bila izvedena v času od decembra do januarja 2010. Cilj raziskave je bil zbrati in analizirati mnenja vodilnih informatikov na informacijskem področju iz večjih slovenskih podjetij o zahtevanih znanjih diplomiranega informatika, ki jih potrebuje danes in jih bo potreboval v prihodnosti. Pokazali smo razvrstitev zahtevanih znanj po skupinah in po pomembnosti, jih primerjali s predhodnima raziskavama, ki sta bili izvedeni leta 1993 in 1997 in nekaterimi tujimi raziskavami. Ugotovili smo, da so po tej pomembnosti na prvem mestu kompetence iz organizacijskih spretnosti, kar velja danes in bo tudi v bodoče. Rezultati predštudije so primerljivi z rezultati podobnih raziskav v svetu.

Ključne besede: izobraževanje, znanja, spretnosti, kompetence, informacijski sistem, informacijska tehnologija, vodilni informatiki

Abstract

The paper presents the results of a preliminary pilot study, conducted during the period from December to January 2010. The aim of this study was to collect and analyze the MIS managers' perceptions from major Slovenian companies on the needed knowledge and skills required by MIS graduates now and in the future. We showed the classification of required skills by groups, by their importance and comparison of results with previous surveys which were carried out in 1993 and 1997. Findings of investigations are compared and confront with the outcome of certain studies done in other countries. We found out that as most important were organizational skills, as is the case today and will be in the future. Outcomes of our study are comparable with the results of similar studies in the world.

Keywords: education, knowledge, skills, competences, information systems, information technology, empirical study, MIS managers

IKT kot ovira in možnost za posodobitev slovenskega gimnazijskega izobraževanja
ICT as an Obstacle and Opportunity for Modernization of Slovene High School Education

Srečo Zakrajšek
IAM (Inštitut in akademija za multimedije, Ljubljana
sreco.zakrajsek@iam.si

Povzetek

Sodobnega izobraževanja ni mogoče izvajati brez kakovostne opreme in usposobljenih kadrov. Rezultati raziskave, ki smo jo izvedli leta 2009 in 2010 v slovenskih gimnazijah, so pokazali, da le-te ne dosegajo spodnjih zahtev za sodobno opremljene šole v razvitih evropskih državah, zato gimnazije tudi ne uporabljajo vrste tehnologij in možnosti, ki jih daje sodobno, z IKT podprto izobraževanje, in posledično tudi gimnazijski maturanti nimajo vrste kompetenc, ki jih pridobijo v srednješolskem izobraževanju njihovi vrstniki v razvitejših državah.
Na realno zelo slabo stanje na področju IKT v gimnazijah in kasneje na višjih nivojih izobraževanja vpliva vrsta dejavnikov, med katerimi je na prvem mestu (predvsem zaradi neustreznega financiranja) premajhno vlaganje v tehnološko posodabljanje šol in dosedanje nevključevanje potrebe po IKT v izvajanju izobraževalnega procesa. Kot zelo pozivno dejstvo gre sprejeti, da je Strokovni svet RS za splošno izobraževanje leta 2008 sprejel prenovljene gimnazijske učne načrte, ki eksplicitno zahtevajo uporabo sodobne IKT, prav tako kompetenčni pristop v skladu z osmimi ključnimi kompetencami, izmed katerih je četrta prav digitalna kompetenca.
Neustrezno so za sodobno poučevanje izobraženi in usposobljeni učitelji, ki pa imajo na žalost podporo tudi v zastareli pedagoški stroki in sindikatih, ki ne dopuščajo nobenih strukturnih sprememb. Tako je nastal zaprt krog, iz katerega bo brez radikalnih ukrepov težko najti hitre in pozitivne izhode.
V prispevku so prikazani rezultati anket in analiz na področju IKT v gimnazijah ter predlogi za spremembo stanja.

Ključne besede: gimnazija, IKT, tehnološka posodobitev

Abstract

Modern education cannot be performed without quality equipment and qualified cadres. Results of the research, performed in 2009 and 2010 in Slovene high schools showed that they do not reach the minimum requirements for modern equipped schools in developed European countries; consequently high schools do not use a large number of technologies and possibilities, given by modern, ICT supported education and consequently, high school graduates do not have a number of competencies, acquired in high school education by their contemporaries in more developed countries.
Many factors influence the current poor condition regarding ICT in high-schools and in higher levels of education; the most important of them (especially because of inappropriate financing) is the lack of investments for technological modernization of schools and previous lack of need to include ICT into the educational process. It is a very positive fact that in 2008 the Council of Experts of the Republic of Slovenia for General Education adopted renovated high-school curricula which explicitly demand the use of modern ICT, as well as competent approach in accordance with eight key competencies, where the fourth is the digital competence.
Teachers are unsuitably educated and qualified for modern teaching and unfortunately, they have the support in the ancient pedagogical profession and in syndicates who do not allow any structural changes. A closed circle was created from which (without radical measures) it will be hard to find quick and efficient exits.
In this article are shown the results of surveys and analyses in the field of ICT in high schools, with suggestions for changes.

Keywords: high school, ICT, technological modernization

Integracija obstoječih e-gradiv v programsko opremo i-table pri slovenščini
Integration of E-learning Resources into Interactive Whiteboard in Slovene Lessons

Tina Žagar Pernar
OŠ Naklo
Glavna cesta 47, 4202 Naklo
tina.zagar1@guest.arnes.si

Povzetek

Vse hitrejši tehnološki razvoj, ki vpliva na vključevanje sodobnih učnih pristopov, temelječih na individualizaciji in diferenciaciji, od učitelja zahteva nenehno prilagajanje, posodabljanje in nadgrajevanje učnih pripomočkov in gradiv. Zaradi želje po ustvarjanju kakovostnih gradiv za pouk in nenehne tekme s časom se učitelji vse pogosteje soočajo s preobremenjenostjo in stresom. Avtorica prispevka opaža, da so učitelji v praksi še vedno preveč zaposleni z obvladovanjem tehnologije kot take (v zadnjem času predvsem rabe i-table), ne zavedajo pa se korelacij med različnimi tehnološkimi možnostmi in raznovrstnimi v spletu razpršenimi e-gradivi. Izkušnje avtorice kažejo, da bi z različnimi oblikami povezovanja, sodelovanja in iskanja novih poti lažje prišli do znanja in razvoja ter ne nazadnje dosegli boljše rezultate s čim manj stresa in izgorevanja. V ta namen prispevek nakazuje možnosti nadgradnje že obstoječih e-gradiv z interaktivno tablo pri slovenščini, in sicer pri pouku književnosti v 8. razredu.

Ključne besede: interaktivna tabla, slovenščina, obstoječa interaktivna učna gradiva

Abstract

Technological development with its influence on modern teaching approaches (which include individualisation and differentiation) demand from teacher much adaption and improvement of learning instruments and resources. Due to strong teachers' wishes to create quality learning materials and due to lack of time teachers confront overtaxing and stress. Author of this paper notices that teachers put much effort into handling with ICT on one hand but on the other hand they are still unaware of correlations between technical possibilities and on-line resources. According to her own experiences the author stresses that collaboration and new ways would lead to better teaching results with less stress and without burning out. In order to point that out this article presents some tips how to integrate already existing interactive learning resources into teaching with interactive whiteboards in Slovene lessons in the 8th grade.

Keywords: interactive whiteboard, Slovene language, existing interactive learning resources

Umetnostna zgodovina: Učimo se umetnost (MOTIV) v programu Power Point
Art History: Learning art (MOTIVE) in Programme Power Point

Metka Žnidar
Srednja šola Jesenica
Ulica bratov Rupar 2, 4270 Jesenice
metka.znidar@gmail.com

Povzetek

V svoji dolgoletni praksi šele dve leti uporabljam pri predmetu umetnost računalnik in moram priznati, da mi olajša delo in občutno izboljša kakovost pouka. Dijake lahko bolje motiviram in njihovo znanje je boljše
Pri pouku umetnost je ključno, da dijaki snov sprejemajo preko računalnika in projektorja,
Zato je potrebno v učno uro vključiti veliko slikovnega gradiva in razpredelnic.
Opisala sem urejanje slik v power pointu za predstavitev učne ure na temo MOTIV.
Slike za v power point sem našla večinoma na internetu, zato sem jih morala tudi urediti.

Ključne besede: motiv, power point, oblika, pike

Abstract

I use the computer for subject matter only two years, it is a great help to me and to students. The motivation for learning is higher. The knowledge of students improves. I show you the steps how I prepared a lesson for art with pictures from internet in power point.

Keywords: motive, power point, photoshop, image size, piksel, inch

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Izkopavanje znanja in podatkovna skladišča

Data Mining and Data Warehouses

Uredila / Edited by

Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

12. oktober 2010 / October 12th, 2010
Ljubljana, Slovenia

Preface / Predgovor

Data Mining and Data Warehouses (SiKDD 2010)

Data handling technologies have significantly progressed in the 90's. The first phases mainly dealing with storing and efficiently accessing the data, resulted in the development of industry delivering tools for handling large databases, standardization of related processes, queering languages, etc. When the data storage was not a primary problem any more the need for improving the database organization resulted in the databases supporting not only transactions but also analytical views of the data. At this point data warehousing with On-Line-Analytical-Processing entered as a usual part of a company information system, requiring from the user to set well defined questions which is not always easy and possible. On the other hand, Data Mining offers automatic data analysis trying to obtain some new information from the existing data and enabling the user some new insights in the data. The Slovenian KDD conference covers a broad area including Statistical Data Analysis, Data, Text and Multimedia Mining, Semantic Technologies, Link Detection and Link Analysis, Social Network Analysis, Data Warehouses.

Odkrivanje znanja in podatkovna skladišča

Tehnologije, ki se ukvarjajo s podatki so v devetdesetih letih močno napredovale. Iz prve faze, kjer je šlo predvsem zato kako podatke shraniti in kako do njih učinkovito dostopati, se je razvila industrija za izdelavo orodij za delo s podatkovnimi bazami, prišlo je do standardizacije procesov, povpraševalnih jezikov itd. Ko shranjevanje podatkov ni bil več poseben problem se je pojavila potreba po bolj urejenih podatkovnih bazah, ki bi služile ne le transakcijskem procesiranju ampak tudi bolj analitskim pogledom v podatke – pojavilo se je skladiščenje podatkov (data warehousing), ki postaja vse bolj standarden del informacijskih sistemov v podjetjih. Paradigma OLAP (On-Line-Analytical-Processing) zahteva od uporabnika, da še vedno sam postavlja sistemu vprašanja in dobiva nanje odgovore in na vizualen način preverja in išče izstopajoče situacije. Ker seveda to vedno ni mogoče, se je pojavila potreba po avtomatski analizi teh podatkov oz. z drugimi besedami to, da tehnologija sama pove, kaj bi utegnilo biti zanimivo za človeka – to prinašajo tehnike odkrivanja znanja (data mining), ki iz obstoječih podatkov skušajo pridobiti novo znanje in tako uporabniku ponudi novo razumevanje dogajanj zajetih v podatkih.

Slovenska KDD konferenca pokriva vsebine, ki se ukvarjajo z analizo podatkov in odkrivanjem zakonistosti v podatkih: pristope, orodja, probleme in rešitve.

Editors and Program Chairs / Urednika

- Marko Grobelnik
- Dunja Mladenić

TWO PASS K-MEANS ALGORITHM FOR FINDING SIFT CLUSTERS IN AN IMAGE

Nenad Tomašev, Dunja Mladenić
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
e-mail: nenad.tomasev@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

This paper explores the ways to represent images as bags of SIFT feature clusters. SIFT features themselves are widely used in image analysis because of their properties of scale and rotation invariance. The usual way to group them is to segment the image into regions and then assign features to the corresponding image parts. When images themselves are not available for privacy reasons, this is not possible. We created a hybrid clustering algorithm which offers more flexibility than simple spatial k-means clustering. The algorithm parameters were optimized by a stochastic procedure. The impact of different elements of local representation on final clustering quality is discussed.

1 INTRODUCTION

Due to the constant increase in the amount of multimedia data, automatic image processing is becoming more and more important. For any sort of machine learning algorithms to be employed, images first need to be represented in a way which allows for easy manipulation and calculation. There exists a variety of local or global image features, capturing aspects of objects present in an image. One of the most frequently used local features are SIFT features. They are quite robust and invariant to translation, scaling and rotation [1].

Each SIFT feature has the following structure: (x,y) location in an image where it was found, scale at which it was found, orientation of the direction of the change it describes, as well as the most important part – the feature descriptor. The descriptor is a set of orientation histograms in the vicinity of the keypoint. Usually, a 16x16 pixel neighborhood is sampled, observing 16 histograms of dimensions 4x4, each having 8 bins. This results in the descriptor length of 128 values, so the descriptor is in fact quite high-dimensional. The keypoints themselves are usually found as scale-space extrema of difference of Gaussian image convolutions [2].

SIFT features can be used in various ways. For representing an entire image, it is customary to first quantize the most typical feature vectors and assign each feature on an image to its closest typical vector, thereby representing an image as a histogram of frequencies of these so-called *codebook* vectors. They are obtained as centroids of clusters of features from a stratified sample on a collection of images. On the other hand, we might want to observe an image as a set of objects in a scene. Each object would then be represented as a group of spatially close SIFT features corresponding to the region in the image where the object is located [1]. In order to achieve this, features have to be grouped in a certain way. The ways of performing the grouping vary according to the specific image task. Sometimes it is beneficial to group features according to the descriptors and to observe an object as a small group of appearance clusters of features describing different textures [3]. More usual way to do is to find regions of interest by first segmenting the image at some level of detail, and then just assigning features to image segments based on their location. A simple K-means clustering based feature location in an image is also an option, though certainly not the most flexible choice.

Sometimes, there are certain privacy issues involved and it is not possible to obtain the images themselves, just some sort of feature representation. This means that there are times when it is not possible to rely on image segmentation for SIFT feature grouping. It is also noteworthy to mention that image segmentation does not necessarily lead to the most convenient image decomposition for purposes of SIFT-based object representation. Even if it does, one must still be careful with threshold parameters for segment merging/splitting, in order to get the desired level of detail.

The goal of this project was to make an algorithm for SIFT feature clustering in a single image, which works just with locally extracted information and which would be flexible enough to meet various preferential requirements for determining feature groups.

2 VIEWS ON SIFT FEATURE CLUSTERING

As previously mentioned, sometimes it is beneficial to group SIFT features before proceeding with the analysis. These feature clusters can then be represented either by their centroids in a simplistic manner, or mapped to corresponding codebook vectors to obtain histograms.

It would be ideal if the so-constructed clusters would match either whole objects or object parts perfectly. Naturally, this is not possible in general case. However, it should be our aim to try and fit those clusters as close as possible to the ideal case.

The first approach that comes to mind is to simply use one of the existing clustering algorithms, for instance k-means, and just define the distance measure between features conveniently. This metric should be defined so as to include information about the difference in keypoint location, feature descriptor and scale. We could also include information about colors in the vicinity of the points by calculating color histograms in some small pixel neighborhoods, if the images are provided. We could form a linear combination of these terms by weighting them according to their relative importance for the image mining task at hand. However, things are not all that simple.

Consider the image shown in Figure 1. There are two cars present in the image, which are exactly the same. If we think of how the SIFT clusters might look like in the end, it is clear that by using the above described approach, there will be big differences between the small and the big car. This is because the spatial distance is weighted the same in both cases and parts of the bigger car are further away from each other than in case of the smaller car. Hence, we will obtain different clusters for those two objects just because they are of different size.

In this particular example, normalizing the distance between each pair of keypoints by some function of the scales at which the respective SIFT features were detected could definitely improve the result. This is also not ideal, because some objects contain features of different scales. If the distribution of scales in the object's features has a big variance, any sort of direct distance normalization based on scale would induce a clustering on an object where features would most likely be broken up into two or more potentially spatially inconsistent groups.

Figure 1: *An image of a car, containing a small copy of itself in the upper left corner.*

One might consider the above example slightly artificial because both cars were the same, but the same argument goes for having similar objects of the same type in several places in an image, not all of them having the same size. In Section 4 we will propose one solution for this problem.

3 RELATION BETWEEN IMAGE SEGMENTS AND SIFT FEATURE CLUSTERS

Image segmentation can be a useful way to obtain SIFT feature clusters after decomposing the original image into subsegments. However, this can also sometimes lead to redundant feature clusters and more complex representations of the original image. Observe the images in Figure 2. Even though it is possible to interpret every single leaf as an individual object for further analysis, it is hardly necessary. It would be quite acceptable having a few of the leaves clustered together, the more the better. They bear nearly identical SIFT features and even have similar colors. However, the shown segmentation has produced 64 different segments in the image. This is much more than what would be required for further processing.

Figure 2: *An image of an autumn tree branches and leaves and the corresponding segmentation obtained by SRM segmentation algorithm.*

Since image segments usually represent either objects or parts of objects, it would be reasonable to assume that the features contained within a single segment should be relatively homogenous with respect to their cluster labels. This doesn't mean that all the features within a single segment must come from a single SIFT feature cluster. In the case of the leaves depicted above, it would also be acceptable to have some slight overlap of clusters within segments. Also, image segments tend to be of quite irregular shapes, sometimes extending along the edge of the entire image. Since distance between keypoints plays a role in the clustering, it is difficult to produce such irregular clusters, especially if the density of keypoints within is low, meaning that the region is not highly textured. Hence, it is also acceptable to have several feature clusters within the segment and vice versa. What is best depends on the image in question and there is no universal answer to cover all the possible cases.

4 PROPOSED ALGORITHM

There are as many ways to approach the problem as there are clustering algorithms in general. In this particular case, we would like to propose a simple extension to the k-means algorithm, which is usually used to quickly find spatial feature clusters when more time-consuming methods are not applicable. This is the procedure we will be considering in the rest of the paper.

- INITIALIZATION:
  - Perform K-means based on x,y coordinates of the features in order to place initial centroids at spatially reasonable locations
  - Calculate local color histograms for neighborhoods of all keypoints
- LOOP:
  - Calculate local color histograms for all centroids
  - Calculate separately distances in coordinates, scale, color histograms and feature descriptors between all keypoints and all centroids
  - For each keypoint:
    - Rank centroids separately according to distance from the keypoint in either of the metrics (in an ascending way, the closest centroid having rank of 1). Mark these ranks $r_{xy}(c_i)$, $r_c(c_i)$, $r_s(c_i)$, $r_d(c_i)$ for centroid c_i, respectively
    - Calculate distance from the keypoint to the centroid as a linear combination of these ranks: $d(X_i, c_i) = r_{xy}(c_i) + \alpha r_c(c_i) + \beta r_s(c_i) + \gamma r_d(c_i)$
    - Assign keypoint to the closest centroid
  - If there were no reassignments or the error change is below threshold, END
  - Calculate new centroids and go back to start of the loop

The use of ranks instead of explicit distances addresses the issue raised in Section 2 of dealing with similar objects on various scales. This way, there is no need to try explicitly combining individual distances which might behave differently in different situations. Certainly, some of the information is lost this way, but hopefully it's compensated enough by overcoming some shortcomings of the direct approach. It should also be noted that instead of combining continuous variables, here there is only a discrete set of rank combinations for a centroid, which could also be ordered in a more general way.

Since the initial spatial K-means is done in the initialization step, this second K-means pass with the modified rank-based metric is there as sort of cluster refinement.

Parameters allow for setting the algorithm to favor either of the aspects of a keypoint when deciding on cluster assignment. It is clear that the choice of parameters should reflect the context in which the clustering is performed. If it is known that the images in the dataset are not very colorful, then local color histograms definitely won't contribute, they could even worsen the end configuration. The three parameters α, β, γ represent relative importance of color, scale and feature descriptor compared to significance of keypoint (x,y) coordinates when determining feature clusters. Hence, the influence of coordinate distance is assumed to be non-zero, naturally.

5 EXPERIMENTAL SETUP

When dealing with a specific task, it is necessary to set the rank weights well. We mentioned image segmentation as a way of grouping SIFT features, which has its own advantages and disadvantages. Here we check which setup of the rank weights seems to be most consistent with respect to image segmentation. We approach this as a stochastic optimization problem.

A small set of 50 images was handpicked for evaluation from the IMAGENET challenge dataset [4]. The images were chosen from several different categories, namely plants, animals, landscapes, people, cars, gears, etc. We chose those images which had neither too simple a segmentation, nor too complex one. The aim wasn't to try solving the generalized problem, but just to gain some insight into which setup could be expected to produce clusters which follow the structure of image segments.

For segmentation, statistical region merging method was used [6]. SIFT features were extracted by SiftWin application developed by David Lowe [7].

Optimization was performed by simulated annealing [5]. It is a method which gives a trade-off between exploration and exploitation in the search space, controlled by the temperature parameter. The non-negative probability of choosing the worse solution during the search helps escaping local optima and is given by the following formula:

$$p(s_{i+1} = o_i \mid Err(o_i) > Err(s_i)) = e^{(Err(s_i) - Err(o_i))/Td}$$

In the above equation, the current solutions are denoted by *s*, the configurations being currently checked by *o*, *T* is the temperature parameter and *d* the average change caused by mutations to the solutions, which is calculated in the first few steps.

For this particular case, it was assumed that the greatest influence should be exhibited by the coordinate distance, in order to get spatially consistent clusters. Therefore, the bounds were set on all three parameters α, β, γ to fall into the interval [0,1].

Fitness of configurations was estimated as the average within-segment entropy of cluster labels. Minimizing this entropy would lead to more homogenous solutions.

Two different optimization runs were performed, one by setting a fixed number of clusters (K=20) on all images, the other by stating that for every image the number of clusters is equal to number of segments on that image. 300 iterations were performed in each of these runs.

Images in the dataset had a varying number of segments, ranging from 13 to 76, with an average of a little over 34 segments per image. The number of SIFT features on an image ranged from 487 to 3062 and the average was around 1410 features per single image.

6 RESULTS

It turns out there was little difference between the two optimization runs in terms of the final suggested configurations.

	Descriptor rank weight	Color rank weight	Scale rank weight
$n_c = n_{segments}$	0	0.06	0.62
$n_c = 20$	0	0.04	0.63

Table 1: *Weighting schemes that were most aligned with SRM segmentation*

It was clear beforehand that descriptor similarity should not be given much weight if the goal was to achieve segmentation-like clustering of the keypoints, because similar features can be located anywhere in an image, so a high weight of feature similarity would certainly have disrupted the desired grouping. However, it still came as a surprise to see that in this particular case of statistical region merging segmentation, any use of the descriptor information whatsoever led to clustering configurations which were less aligned with the observed segmentations.

It is also intuitive that color should be given some significance, but not too much since there can be many regions in an image having same or similar colors. High scale weight suggests that features of similar scales tend to be grouped together in same segments often enough for it to become important.

Even with the described optimal parameter setting, the produced clustering of SIFT features differs greatly from given segmentation. Average within-segment entropy is roughly 50% of the maximum possible entropy for the used number of clusters. This can be explained by the fact that some image segments do not contain SIFT keypoints at all, so the entropy naturally gets a bit higher when more clusters are used than is actually needed. However, it is not entirely clear how one should go about guessing the proper number of clusters beforehand in the k-means setting. It is of course possible to run clustering for a range of values of k and pick the best one. Performing all these runs would have taken much more time, so we limited the experiments to this less-than-ideal case.

The parameters obtained via the described optimization process do not extend to the general case. The choice of parameters is context-dependent, and sometimes it is even advisable to give greatest weight to feature descriptor.

Both the weighting scheme and the final entropy estimates relate only to comparison with SRM segmentation.

7 CONCLUSION

We explored the possibilities of using a two-pass k-means approach when clustering SIFT features on an image, first performing spatial clustering for initialization and then refining the clustering in the second pass by using as distance a linear combination of centroid ranks obtained with respect to some selected individual distance measures. We performed stochastic optimization on parameters of the algorithm to see which configuration leads to best alignment with image segmentation. Most important was naturally the coordinate distance, followed by scale and color, while the feature descriptors proved to be of no importance in this particular case. Achieving segmentation-like clustering of keypoints is only one possible approach, so different weighting schemes could be chosen for other tasks.

8 ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC PASCAL2 (IST-NoE-216886).

References

[1] Z. Zhang and R. Zhang *Multimedia Data Mining*. Chapman & Hall, 2009.

[2] D. Lowe. Object recognition from local scale-invariant features, *Proceedings of the International Conference on Computer Vision*. pp. 1150-1157. 1999.

[3] K. Mikolajczyk, B. Leibe, B. Schiele. Local Features for Object Class Recognition. *In Proc. ICCV'05, Tenth IEEE International Conference on Computer Vision.* Beijing. China. 2005.

[4] IMAGENET. http://www.image-net.org/

[5] Franco Busetti. Simulated Annealing overview. www.geocities.com/francoburseti/saweb.pdf

[6] R. Nock, F. Nielsen. Statistical Region Merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1452-1458. 2004.

[7] D. Lowe. SIFT Keypoint Detector. http://www.cs.ubc.ca/~lowe/keypoints/

Multi-View Canonical Correlation Analysis

Jan Rupnik (1), John Shawe-Taylor (2)
(1) Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
(2) University College London, Gower Street, London WC1E 6BT, United Kingdom
e-mail: jan.rupnik@ijs.si

ABSTRACT

Canonical correlation analysis (CCA) is a method for finding linear relations between two multidimensional random variables. This paper presents a generalization of the method to more than two variables. The approach is highly scalable, since it scales linearly with respect to the number of training examples and number of views (standard CCA implementations yield cubic complexity). The method is also extended to handle non-linear relations via kernel trick (this increases the complexity to quadratic complexity). The scalability is demonstrated on a large scale cross-lingual information retrieval task.

1 INTRODUCTION

Principal Component Analysis is a very popular approach to dimensionality reduction in the field of statistics and machine learning. When observations arrive from two sources that share some mutual information a related approach called the Canonical Component Analysis was developed [3].

This paper presents an efficient method that generalizes CCA to more than two views, Multi-view Canonical Correlation Analysis (MCCA). Defining a measure of cross-correlation for more than two random variables is not straightforward and many possible measures have been proposed [4]. Typical approaches define cross-correlation as a function of pairwise correlations between variables (for example the sum, product or sum of squares). Sum of correlations problem formulation, SUMCOR, was first studied in [2], where the optimization problem was formulated and a method to solve it was proposed (a generalization of the power method for standard eigenvalue problem which has been proved to converge in [1]). We will adopt and extend this approach since it is closely related to a known linear algebra problem which can be solved efficiently. The method proposed by Horst was designed to find a one-dimensional common representation.

The paper is structured in the following way: section 2 introduces the canonical correlation analysis, section 3 describes the multi-view CCA method, section 4 involves evaluation, followed by conclusions in section 5.

2 CANONICAL CORRELATION ANALYSIS

Canonical Correlation Analysis (CCA) is a dimensionality reduction technique similar to Principal Component Analysis (PCA), with an additional assumption that the data consists of feature vectors that arose from two sources (two views) that share some information. Examples include documents written in two different languages, textual information paired with images, a set of feature vectors computed from audio information and a set of feature vectors computed from the frames in a video recording, etc. Instead of looking for linear combinations of features that maximize the variance (PCA) we look for a linear combination of feature vectors from the first view and a linear combination for the second view, that are maximally correlated.

Formally, let $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be the set of n sample points (pairs of observation vectors) where $x_i \in \boldsymbol{R}^p$ and $y_i \in \boldsymbol{R}^q$ represent feature vectors from p (or q)-dimensional vector spaces. Let $\boldsymbol{X} = [x_1, \dots, x_n]$ and let $\boldsymbol{Y} = [x_1, \dots, x_n]$ be the matrices with observation vectors as columns, which are viewed as two samples of observations of two random vectors (X and Y). The idea is to find two linear functional (row vectors) $\alpha \in \boldsymbol{R}^p$ and $\beta \in \boldsymbol{R}^q$ so that the random variables αX and βY are maximally correlated (α and β map the random vectors to random variables, by computing weighted sums of vector components). By using the sample matrix notation $\boldsymbol{X}$ and $\boldsymbol{Y}$ this problem can be formulated as the following optimization problem:

$$\max_{\alpha \in R^p, \beta \in R^q} \alpha \boldsymbol{X} \boldsymbol{Y}' \beta'$$
$$s.t.$$
$$\alpha \boldsymbol{X} \boldsymbol{X}' \alpha' = 1$$
$$\beta \boldsymbol{Y} \boldsymbol{Y}' \beta' = 1$$

The optimization problem can be reduced to an eigenvalue problem and includes inverting the variance matrices ***XX'*** and ***YY'***. If they are not invertible one uses a regularization technique by replacing them with $(1-\kappa)$ $\boldsymbol{XX'} + \kappa \boldsymbol{I}$, where $\kappa \in \boldsymbol{R}$ and $\boldsymbol{I}$ is the identity matrix.

A single canonical variable is usually inadequate in representing the original random vector, that is why one looks for k-1 other projection pairs $(\alpha_2, \beta_2), \dots, (\alpha_k, \beta_k)$, so that α_i, and β_i are highly correlated and each α_i is uncorrelated to α_j for $j \neq i$ (analogously for β).

The method was extended to handle nonlinear relations between two random vectors in [6]. The approach is based on the observation that computing the canonical correlation vectors can be done by using solely the inner product information between sample vectors and that one can omit directly using any vector features. This enables the use of the dual problem formulation and application of the kernel trick [7]. For a given choice of kernel function with a corresponding feature map, this is equivalent to first nonlinearly mapping both sets of sample vectors to a separate higher dimensional Hilbert spaces (the dimensions can be even infinite, for example when one uses a Gaussian kernel function) and look for linear relations in between the samples in those spaces. This usually makes the problem underdetermined – a high, possibly infinite, number of features and a smaller set of examples. To avoid overfitting, one needs to apply a regularization technique.

A typical regularization approach transforms the problem [7] into finding well cross-correlated projection vectors that have a high covariance as well. This enforces that the patterns discovered are not only well correlated across views but also well represented in the data.

3 MULTI-VIEW CANONICAL CORRELATION ANALYSIS

Consider a set of vectors $w_i \in \mathbb{R}^{n_i}, i = 1 \dots m$. For each random vector $\mathcal{X}_i$, with dimension n_i, we can define a univariate random variable $\mathcal{Z}_i$ as a linear combination random components: $\mathcal{Z}_i = w_i' \mathcal{X}_i$.We can now compute pairwise correlation coefficients for each pair of the variables $\mathcal{Z}_i$. The goal is to find the vectors w_i so that the sum of all pairwise correlations is the highest. One can prove that the optimization can be written as:

$$\max_{w_1,\dots,w_m} \sum_{i<j} w_i' X^i X^{j'} w_j$$

s.t.

$$w_i' X^i X^{i'} w_i = 1, \forall i,$$

where $X^i \in \mathbb{R}^{n_i \times n}$ are centered matrices of observations of random vectors X_i, containing n columns of sample vectors. Notice that every matrix X_i has the same number of columns – this corresponds to aligned sample assumption (column k of matrix X_i and column k of matrix X_j are aligned samples in two views).

We will reformulate the problem in dual form to make the problem feasible in the case of high dimensional data (e.g. text mining, where the number of features is the number of words encountered in the corpus) and with the use of the kernel trick make the solution more flexible than the linear model [9]. To express the problem in dual form we introduce new variables (we will also refer to them as dual variables), $\beta_i \in \mathbb{R}^n$, so$w_i = X^i \beta_i$. Let K_i be the kernel matrix computed on data X_i, which means that the element in the k-th row and l-th column of K_i is equal to:

$$< \phi_i(X_k^i), \phi_i(X_l^i) >$$

for some mapping $\phi_i : \mathbb{R}^{n_i} \to \mathcal{H}_i$ to some Hilbert space $\mathcal{H}_i$. The bracket denotes the inner product. The kernelized dual formulation of the problem is then:

$$\max_{\beta_1,\dots,\beta_m} \sum_{i<j} \beta_i' K^i K^{j'} \beta_j$$

s.t.

$$\beta_i' K^i K^{i'} \beta_i = 1, \forall i,$$

The constraints force the univariate random variables (linear combinations of the components of the original random vectors) to have unit variance.

If one of the kernel matrices is singular or is ill-conditioned the problem becomes numerically intractable. To remedy this problem one usually adds a low positive number on the diagonal elements of each kernel matrix in the variance equation (not the optimization criterion function).

By using Lagrangian multiplier techniques one can transform the constrained optimization problem to a generalized multivariate eigenvalue problem of the form:

$$\begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mm} \end{bmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} = \begin{pmatrix} \lambda_1 \beta_1 \\ \vdots \\ \lambda_m \beta_m \end{pmatrix},$$

Where A_{ij} are block matrices of dimension $n \times n$, β_i, are n-dimensional canonical vectors and λ_i are the generalized eigenvalues. Canonical vectors and the generalized eigenvalues are unknown and must be computed. The transformation of the kernelized dual to the multivariate eigenvalue problem can be conducted so that the λ_i become interpretable: their sum is directly proportional to the sum of correlations when one uses canonical projection vectors β_i to obtain univariate random variables from random vectors $\mathcal{X}_i$.

The solution (see [2]) to the multivariate generalized eigenvalue problem presented above can be found with a method similar to finding an eigenvector-eigenvalue pair in a square matrix by using power iteration method. The algorithm that finds the canonical projection vector requires a starting set of vectors which iteratively converge to a local optimum of the problem (several restarts with different starting vectors can prove useful). One must choose the number of iterations, denoted as *maxiter*, in advance or implement a stopping criterion.

So far we have discussed how to find a single canonical projection vector in each view. This is typically insufficient since too much information is discarded that way (In text mining for example, describing a document by a single number that represents the similarity of the document to the discovered latent vector). We denote the canonical vectors $\beta_1, \dots, \beta_m$ that we found as $\beta_1^1, \dots, \beta_m^1$ and try to find another set of concept vectors $\beta_1^2, \dots, \beta_m^2$ for which the sum of pairwise correlations is maximal with an additional constraint that they must be “different” from the first set. We can express this as a set of additional constraints:

$$\beta_i^{1\prime} K^i K^{i'} \beta_i^2 = 1, \forall i.$$

This forces the new set of vectors to be uncorrelated to the first. One can extend the problem to any number of sets of canonical projection vectors (each new set must be uncorrelated with all that have been discovered so far).

We can prove that the resulting optimization problem can be posed as a generalized multivariate eigenvalue problem and that it still satisfies the local convergence guarantees.

Algorithm 1: Horst algorithm

Input:matrices A_{ij}, initial vectors, $\alpha_i^0, where\ i, j: 1 \dots m$
Output: $\alpha_1^{maxiter}, \dots, \alpha_m^{maxiter}$
$for\ i = 1\ to\ maxiter\ do$:
  $for\ j = 1\ to\ m\ do$:
    $\alpha_j^i \leftarrow \sum_k A_{j,k}\alpha_k^{i-1}$
    $\alpha_j^i \leftarrow \frac{\alpha_j^i}{\sqrt{\alpha_j^{i\prime}\alpha_j^i}}$
  $end\ for$
$end\ for$

4 EXPERIMENTS

The following section includes information retrieval experiments on the European Parliament corpus. We computed the semantic space for documents from ten different languages and compared the retrieval performance with two alternative approaches, namely Cross-lingual LSI and k-means clustering.

Subsection 4.1 details the experimental setup, subsection 4.2 describes the evaluation measure, subsection 4.3 describes the other alternative cross-lingual methods and subsection 4.4 offers an insight into the latent concept vectors discovered by MCCA.

4.1 DATA SET AND PREPROCESSING

Experiments were conducted on the EuroParl, Release v3 [8] data set and include Danish, German, English, Spanish, Italian, Dutch, Portuguese, Swedish, Finish and French language. We first removed all documents that had one translation or more missing. We split the corpus in an aligned set of documents, each representing a speech in the parliament. Cleaning the set resulted in 107.873 documents per language. We kept the first 100.000 for training and remaining 7.873 for testing or for testing. We then extracted the bag of words model for each language, where we kept all unigrams, bigrams and trigrams that occurred more than thirty times. This resulted in roughly 200.000-dimensional feature spaces for each language. Finally we computed the tf-idf weighting and normalized every document.

4.2 MATE RETRIEVAL

We used the aligned test set to measure the quality of the latent space representation. Given a test document q (view X) and its aligned document q' (view Y) and test set S' (view Y) we compute the window10 mate retrieval score in the following way: project q and S into the common semantic space, compute the similarities between projections of q and S and assign score 1 if q' is one of the top 10 most similar documents to q.

4.3 COMPARING TO CL-LSI AND K-MEANS CLUSTERING

We will compare our method with Cross-Lingual Latent Semantic Indexing and k-means clustering [11]. CL-LSI is an adaptation of LSI [10] for more than one view. The idea is to merge all document matrices into a single matrix Y by concatenating the aligned feature vectors. The matrix Y can then be used as the input for clustering or LSI. The final step when comparing to MCCA is to split the concept vectors into shorter concept vectors for each view in concordance with how the views were merged.

Language	k-means	LSI	MCCA
EN	0.7486	0.9129	0.9883
SP	0.745	0.2907	0.9855
GE	0.5927	0.8545	0.9778
IT	0.7448	0.9022	0.9836
DU	0.7136	0.9021	0.9835
DA	0.5357	0.854	0.9874
SW	0.5312	0.8623	0.988
PT	0.7511	0.9	0.9874
FR	0.7334	0.9116	0.9888
FI	0.4402	0.7737	0.983

Table 1 *Mate retrieval window 10*

We tested the performance of the three methods on mate retrieval with window10 on the 100-dimensional subspaces that the methods produced. For each source language we used all remaining nine languages, and averaged each score over all languages. Results imply that the concepts detected by MCCA in Table 1 are of higher quality than that of LSI and clustering. One way to explain this result is that MCCA takes into account that data come from several sources that share some mutual information, whereas the clustering and LSI approaches discard that information (after the views are concatenated we perform standard LSI which is "unaware" that features come from different views). LSI and CCA both find new latent features that are more informative (can detect synonyms), whereas the clustering approach uses the original features and thus performs worse than the other two methods.

4.4 CONCEPT VECTORS

The multivariate random variables from MCCA in our experiments correspond to document-vectors (in the bag of words representation) in different languages. We will now consider sets of words that are correlated between the two or more languages (sets of words that have a correlated pattern of appearance across the aligned corpus). We will assume that such sets approximate the notion of 'concepts' in each language, and that such concepts are the translation of each other. To illustrate the conceptual representation we have printed few of the most probable (most typical) words in each language for the first few components found from the EuroParl (Figure 1). The words are sorted by their weights in the concept vectors.

5 CONCLUSIONS

We have presented an algorithm that can detect similar patterns across multiple domains. A straightforward approach would yield a cubic complexity in the number of samples whereas our implementation reduces the complexity to linear (quadratic if kernel methods are applied). We demonstrated the scalability and effectiveness on a large data set.

6 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886).

References

[1] Chu, M .T. and Watterson, L. J. (1993). On a Multivariate Eigenvalue Problem, Part I: Algebraic Theory and a Power Method. SIAM Journal on Scientific Computing, Vol.14 NO.5, 1089–1106.

[2] Horst, P. (1961). Relations among m sets of measures Psychometrika, 26, 129–149.

[3] Hotelling, H. (1935). The most predictable criterion Journal of Educational Psychology, 26, 139–142.

[4] Kettenring, J. R. (1971). Canonical analysis of several sets of variables Biometrika, 58, 433–451.

[5] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990) Indexing by latent semantic analysis Journal of the Society for Information Science, 41(6), 391–407.

[6] Francis R. Bach and Michael I. Jordan (2001). Kernel Independent Component Analysis Report No. UCB/CSD-01-1166.

[7] Shawe-Taylor, J. and Cristianini, N. (2004). Kernel Methods for Pattern Analysis Cambridge University Press.

[8] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation

[9] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[10] D. Lewis, Y. Yang, T. Rose, F. Li, *Rcv1: A new benchmark collection for text categorization research*. Journal of Machine Learning Research (JMLR), 5:361–397, 2004.

[11] Dobsa J., Dalbelo Basic B. (2004). Comparison of Information retrieval techniques: Latent semantic indexing and concept indexing, Journal of Information and Organizational Sciences, 28(1-2), 1-15

DA	menneskerettighederne, menneskerettigheder, forretningsordenen, rusland, ndringsforslag, ã ndringsforslag
DE	menschenrechte, russland, posselt, menschenrechtsverletzungen, zusammenarbeit, nderungsantrag, verfahrensantrag
EN	amendment, amendments, russia, human rights, cooperation, resolution, of order
ES	enmienda, enmiendas, rusia, n de orden, de orden, reglamento, posselt
FI	ihmisoikeuksien, ihmisoikeuksia, tyã jã, tarkistuksen, tarkistusta, tarkistus, tarkistuksia
FR	amendement, amendements, posselt, russie, rã solution, russe, l amendement
IT	emendamenti, emendamento, risoluzione, russia, regolamento, cooperazione, bielorussia
NL	amendement, mensenrechten, amendementen, rusland, van orde, resolutie, samenwerking
PT	ponto de ordem, de ordem, alteraã, alteraã ã, direitos humanos, directiva, regimento
SV	resolutionen, ryssland, ordningsfrã, posselt, arbetsordningen, samarbete, ryska
DA	omdelt, dagsordenen, tak, er omdelt, protokollen fra, protokollen, strukturfondene
DE	tagesordnung, der tagesordnung, das protokoll der, kommissar, wurde verteilt, wurde verteilt gibt, haushalt
EN	commissioner, president commissioner, agenda, budget, commissioner the debate, commissioner the, item is
ES	comisario, distribuido, gracias, acta de la, comisaria, presupuesto, comisario el debate
FI	esityslistalla, kiitoksia, kiitos, esityslistalle, esityslistalla on, lissabonin, esityslistan
FR	merci, commissaire, jour appelle, du jour appelle, tã distribuã, ã tã distribuã, jeudi
IT	commissario, grazie, ringrazio, sono osservazioni, commissario la discussione, vi sono osservazioni, giovedã
NL	rondgedeeld, zijn rondgedeeld, de orde is, orde is, commissaris, orde is het, begroting
PT	obrigado, presidente senhor, conselho, acta, hã alguma, hã alguma observaã, comissã rio estã
SV	tack, r jag fã, rã det, kommissionsledamot, budget, har delats ut, kommissionã

Figure 1 *Two sets of latent vector*

BUILDING A CONCEPT SHELL: ONTOLOGY POPULATION WITH FACTS FROM WWW

Inna Novalija, Dunja Mladenić
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773144
e-mail: inna.koval@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

This paper addresses the process of the ontology population with facts for a selected domain of interest extracted from the Web documents. We suggest an information extraction methodology based on the ontology structural and lexical features. The preliminary evaluation is performed in the financial domain using Cyc ontology.

1 INTRODUCTION

This paper presents an approach to ontology population with facts extracted from the Web documents. The usage of the extended ontology for textual information analysis constitutes the primary motivation for our research.

The suggested method is used for inserting the new financial knowledge into Cyc [7], which maintains one of the most extensive common-sense knowledge bases worldwide.

In [14] we have presented a methodology for semi-automatic ontology extension using ontology content and ontology structure information, where ontology content of a particular ontology concept represents the available textual background of the referred concept and ontology structure includes neighborhood concepts involved in the hierarchical and non-hierarchical relations with a referred concept. We have also defined ontology extension [14] as a process allowing for adding new concepts to the existing ontology or, augmentation of the existing textual representation of the relevant concepts with new available textual information.

Following a goal of inserting new knowledge into the existing ontology, in this paper we address a process of ontology-based fact extraction, incorporating previously developed methodology for ontology extension into the ontology population process.

We define a **Concept Shell** as a building block of the ontology extension process. Concept shell aggregates all available information about a candidate ontology concept. The information is described at two layers – specification layer and instantiation layer. While structural information is defined by finding existing related ontology concepts and their relationships, factual information is obtained during ontology population.

For example, Cyc concept *CommercialOrganization* represents a subclass of concept *Organization*, whose primary goal is to generate a profit for its owners, usually through buying and selling of goods or services. A specification layer of *CommercialOrganization* includes a number of related Cyc concepts, such as:

- *OrganizationTypeByProfitMotive,*
- *BusinessRelatedThing,*
- *TelecommunicationsCompany,*
- *BankingOrFinanceCompany,*
- *ManufacturingOrganization et al.,*

and a number of Cyc relations, such as:

- *companyIsInIndustry,*
- *executiveVicePresident,*
- *enterpriseValue,*
- *mainBusinessActivityOfOrgOccursAt,*
- *organizationGrantsFranchisesOfType,*
- *companyHasGeneralCounsel et.al.*

The instantiation layer for CommercialOrganization includes instances: *FileMaker-CommercialOrganization, Symantec-CommercialOrganization, EpinionsDotCom, Thunderstone-theCompany, Adaptec-CommercialOrganization, WiredDigital-CommercialOrganization, Snap-CommercialOrganization, Amiga-CommercialOrganization, LycosInc, ElectronicArtsInc, Cross-jones, Tascon, StJudeMedical, PanasonicInc, EgyptTrans-GasCompany, Jodco-Japan, DeloreanTheCompany, PeugeotTheCompany, BMWTheCompany, Mercedes-BenzTheCompany, LamborghiniTheCompany, SubaruTheCompany, ChevroletTheCompany et.al., etc.*

The relation instances at the instantiation layer of the *CommercialOrganization* are the following:

- *(companyIsInIndustry MicrosoftInc* **(***IndustryOfRegionFn ComputerHardwareIndustry UnitedStatesOfAmerica***)**
- *(companyIsInIndustry MicrosoftInc* **(***IndustryOfRegionFn SoftwareIndustry UnitedStatesOfAmerica***)***)*
- *(mainBusinessActivityOfOrgOccursAt KrispyKremeCorporation UnitedStatesOfAmerica)*
- *(mainBusinessActivityOfOrgOccursAt CVSCorp UnitedStatesOfAmerica)*

- *Time Interval : (TimeIntervalInclusiveFn (MonthFn November (YearFn 2008)) Now) Time Parameter : TimePoint (executiveVicePresident Nokia EskoAho)et al.*

The paper is structured as follows: Section 2 presents the related work; the methodology for ontology population is discussed in Section 3, Sections 4 describes the preliminary experiments and the results, the conclusion is covered in Section 5.

2 RELATED WORK

A number of approaches for automatic ontology population have proven themselves as effective tools of information extraction.

Natural language processing and unsupervised text mining are notably used for extending ontologies [15]. Extension of the existing ontology by automatically extending its relations was addressed by several researchers. The approaches include learning taxonomic [5]/non-taxonomic relations [12].

Described first by Hearst [10], the pattern based approach for instance and hyponym extraction uses a defined set of patterns while analyzing textual sources.

Etzioni et al. [8] developed a KnowItAll system for named entity classification. The approach performs pattern learning and can iteratively obtain new rules and new seeds.

The Open Information Extraction (OIE) paradigm for relation extraction from text was introduced by Banko et al. [3] and implemented in the TextRunner system. TextRunner performs self-supervised learning of a reliability classifier, single-pass extraction of tuples for all possible relations and redundancy-based assessing of probability for each trustworthy tuple.

Lexico-syntactic pattern-based ontology learning is handled by Text2Onto [6], a framework for ontology learning and data-driven change discovery.

SPRAT [13] is a tool for automatic semantic pattern-based ontology population. SPRAT system combines the name entity recognition, ontology-based information extraction and relation extraction in order to define patterns for the identification of a variety of entity types and relations between them.

Carlson et al. [4] present a method of coupling the semi supervised learning of category and relation instance extractors for ontology population with category and relation instances.

Several methods of the automatic ontology extension and population operate with enlarging of Cyc Knowledge Base (Cyc KB) [16], [17].

In our approach the available lexical and structural information of the large common-sense ontology is exploited for information extraction and validation. In particular, ontology-based patterns are used for extraction of relation instances from Web.

3 METHODOLOGY

In [14] we have proposed a new methodology for semi-automatic ontology extension, which combines text mining methods with user-oriented approach and supports the extension of multi-domain ontologies. Moreover, we have adapted the methodology in order to obtain an exhaustive specific methodology for Cyc knowledge base extension.

Experiments have been conducted in two domains – finances and fisheries & aquaculture. For the financial domain, we have used the Harvey [9] financial glossary, which contains around 6000 hyperlinked financial terms. The fisheries & aquaculture domain has been represented by the ASFA thesaurus [2], containing around 9900 terms involving several types of relationships: equivalence relationships, hierarchical relationships, associative relationships and notes.

As a part of the current research, we propose a method for ontology population with concept instances and relation instances extracted from the Web.

Basically, each ontology concept is represented as an information unit with all available information about it. The relationships the concept is involved in, the related existing ontology concepts, the concept instances and relation instances are combined in the concept shell.

The following steps are needed in order to define a concept specification and instantiation layers, which combine a concept shell:

1. *Identification of the Related Concepts*. When a new concept is added to the ontology, the methodology for semi-automatic ontology extension based on content and structure information [14] is used to find its hierarchically and non-hierarchically related concepts.

2. *Structural Inheritance.* From the superclasses the concept inherits its potential relationships.

3. *Concept Instances Search.* A number of patterns is used to search the Web for potential concept instances:

   "**c** such as **I**"
   "such **c** as **I**"
   "**c** including **I**"
   "**c**, especially **I**"
   "**c** like **I**"
   "**c** called **I**"
   "**I** is a **c**"
   "**I**, a **c**"

4. *Concept Instance Validation and Insertion.* Suggested concept instances are ontologically validated. In case no controversies are found, the new concept instances are inserted into the ontology.

5. *Relation Instances Search.* Patterns formed from the lexical and ontological information available about concept relations and their arguments are used to search the Web for potential relation instances.
   In our approach we assume that each ontological relation is represented by the relation denotation (in natural language) and argument types.
   For the methodology testing with Cyc we selected a list of fact raw types taking to the account BBN's proposed answer categories for question answering [1] and assuming that relation arguments are related to one of these types:

   - Quantity
   - Date&Time
   - Location
   - Product
   - Money
   - Event
   - ConceptualWork
   - Rate
   - Agent

   Using the ontology structure, it is possible to identify the relevant raw type for each argument of the particular relation, and therefore, apply a more efficient search procedure to find the relation instances.
   We form a relation instance pattern taking the relation denotation and extracted concept instance name. Sentences from text found on the Web, in which both relation denotation and a particular concept instance occur, are then extracted and checked for facts occurrence. For each argument type a specified technique is used. For instance, for Agent type we search for the relation name in the text and analyze the preceding and subsequent words.

6. *Relation Instances Validation and Insertion.* The extracted arguments are ontologically validated for each particular relation. The relation argument types are compared to the extracted argument types and validated relation instances are inserted into the ontology.

4 EXPERIMENTS & RESULTS

In order to evaluate the proposed methodology we conducted a simple fact extraction experiment in the financial domain using Cyc ontology.

Since Cyc knowledge base contains common sense knowledge [11], we assume that Cyc KB includes some financial knowledge - a financial knowledge base (Cyc FKB).

Figure 1 presents an extract from the concept shell for the new concept *EmergingMarket* added to Cyc. The light ovals represent the existing Cyc concepts, related to the new Cyc concept *EmergingMarket*. The hexagons are the relations (Cyc predicates). The structural ontological relations (such as superclass-subclass relation), that can be obtained in the ontology extension process, relate the new ontology concept to the existing ontology concepts. The relations, which operate with the instances of a particular concept, can be inherited by the new concept through the concept hierarchy. The dark ovals represent the new concept instances and the dark diamonds are the relation instances of the new Cyc concept.

Using the content and structure methodology for ontology extension, we identify that the new concept can be a subclass of the existing Cyc concept *Market*. Hence, the candidate concept *EmergingMarket* inherits a number of Cyc relations from its superclass concept. The binary predicate *hasMonopolyInMarket* relates instances of the *CommercialOrganization* with the market in which it has a monopoly.

With Hearst patterns [10] a number of potential instances for a concept *EmergingMarket* are extracted. One potential instance is *ChineseMarket.*

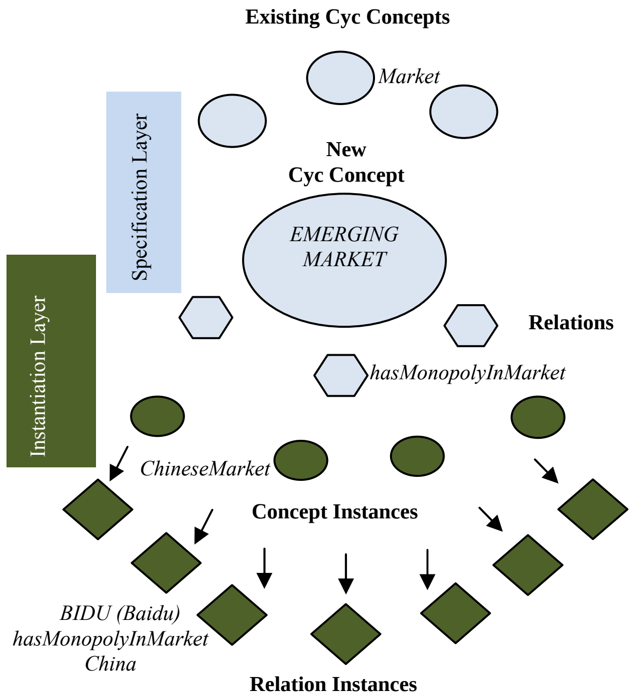


Figure 1: *Extract from Concept Shell for Concept "Emerging Market"*

With available information about the predicate *hasMonopolyInMarket*, we can form a relation instance pattern using lexical phrase *monopolies* from Cyc and

instance name *Chinese market*. Looking at the argument types of the *hasMonopolyInMarket* predicate, we automatically define that we have to search for *CommercialOrganization* which is mapped to the raw type Agent.
As a result, we automatically extract and validate a relation instance: *BIDU (Baidu) hasMonopolyInMarket* in *Chinese Market.*
The results of the experiment confirm the applicability of the suggested methodology for ontology population to Cyc Knowledge Base augmentation.

6 CONCLUSION

This paper addresses the process of the ontology population with extracted facts for a selected domain of interest. We suggest an information extraction methodology based on the ontology structural and lexical features. The preliminary evaluation is performed in the financial domain using Cyc ontology.
The future work should include further extension and population of Cyc Knowledge Base and using it for sophisticated text analysis. Furthermore, the proposed methodology for ontology population should be tested on other domains.

6 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and ACTIVE (IST-2008-215040).

References

[1] Annotation guidelines for answer types: http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html
[2] ASFA thesaurus, http://www4.fao.org/asfa/asfa.htm
[3] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni. Open information extraction from the web. *Proc. of the 20th international joint conference on Artifical intelligence*, pages 2670–2676. 2007.
[4] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, T. M. Mitchell. Coupled Semi-Supervised Learning for Information Extraction. *Proc. of the Third ACM International Conference on Web Search and Data Mining (WSDM)*, volume 2, page 110. 2010.
[5] P. Cimiano, A. Pivk, L. Schmidt-Thieme, S. Staab. Learning Taxonomic Relations from Heterogeneous Evidence. *Proc. of ECAI Workshop on Ontology Learning and Population.* 2004.
[6] P. Cimiano, J. Völker. Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery. *Proc. of NLDB 2005*, pp.227-238. 2005.
[7] Cycorp, Inc., http://www.cyc.com
[8] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91-134. 2005.
[9] C. R. Harvey. Yahoo Financial Glossary, Fuqua School of Business, Duke University. 2003.
[10] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. *Proc of COLING*. 1992.
[11] D. Lenat. Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communic. of the ACM 38 (11).* 1995.
[12] A. Maedche, S. Staab. Discovering conceptual relations from text. *Proc. of ECAI 2000*.
[13] D. Maynard, A. Funk, W. Peters. SPRAT: a tool for automatic semantic pattern-based ontology population. *Proc. of International Conference for Digital Libraries and the Semantic Web*, Trento, Italy. 2009.
[14] I. Novalija, D. Mladenić. Content and Structure in the Aspect of Semi-Automatic Ontology Extension. *Proc. of the 32nd International Conference on Information Technology Interfaces.* 2010.
[15] T. Sabrina, A. Rosni, T. Enyakong. Extending Ontology Tree Using NLP Technique. *Proc. of National Conference on Research & Development in Computer Science REDECS.* 2001.
[16] P. Shah, D. Schneider, C. Matuszek, R. C. Kahlert, B. Aldag, D. Baxter, J. Cabral, M. Witbrock, J. Curtis. Automated population of Cyc: Extracting information about named-entities from the web. *Proc. of the Nineteenth International FLAIRS Conference*. 2006.
[17] M. Witbrock, D. Baxter, J. Curtis, D. Schneider, R. Kahlert, P. Miraglia, P. Wagner, K. Panton, G. Matthews, A. Vizedom. An Interactive Dialogue System for Knowledge Acquisition in Cyc. *Proc. of the Eighteenth International Joint Conference on Artificial Intelligence.* 2003.

TOWARDS SEMANTIC DATA MINING WITH g-SEGS

Petra Kralj Novak[1], *Anže Vavpetič*[1], *Igor Trajkovski*[2], *Nada Lavrač*[1,3]
[1] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
[2] Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, 1000 Skopje, Macedonia
[3] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
e-mail: Petra.Kralj.Novak@ijs.si

ABSTRACT

This paper introduces the term semantic data mining to denote a data mining approach where domain ontologies are used as background knowledge for data mining. It is motivated by successful applications of SEGS (search for enriched gene sets), a system that uses biological ontologies as background knowledge to construct descriptions of interesting gene sets in experimental microarray data. We generalized this domain-specific system to perform subgroup discovery on arbitrary data, annotated by ontologies. We present a prototype of the new semantic data mining system named g-SEGS, implemented in the Orange4WS environment, and an illustrative example showing the application potential of semantic data mining.

1 INTRODUCTION

The most common setting in knowledge discovery is that we are given some data and a data mining task. The data is first manually preprocessed, then a data mining algorithm is applied and the ending result is a model or a set of patterns that can be further interpreted and visualized. It is generally recognized that the quality of the end model depends crucially on the quality of the data collection and preparation process.

Data by itself does not carry any meaning; it needs to be interpreted to convey information. Standard data mining algorithms do not 'understand' the data: the data is treated as meaningless numbers and statistics are calculated on them to build models, and the interpretation of the results is left to human domain experts. An example of an everyday data mining challenge is using the reference to time when the data was collected. Unless time is the main interest of investigation (as is the case in time series analysis), time should be treated just like one of the attributes. However, as standard data mining algorithms do not have specialized mechanisms to deal with time, it is the role of the domain expert to adequately preprocess the time entry.

This paper introduces the term *semantic data mining* to denote a data mining approach where domain ontologies are used as background knowledge for data mining (schematically presented in Figure 1). Our research is motivated by the SEGS [12] system which successfully uses biological ontologies as semantically annotated background knowledge to find descriptions of differentially expressed gene sets. We realized that some features of SEGS could be useful not only in functional genomics but also in other domains, and decided to generalize SEGS to become domain independent.

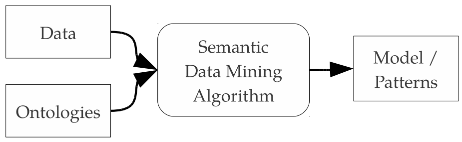


Figure 1: The proposed semantic data mining methodology schema.

We present a prototype semantic data mining system g-SEGS, a generalization of the SEGS system. g-SEGS uses as input: (1) data annotated by ontologies and (2) ontologies in the OWL format. The latter are used for efficient search and pruning of the pattern search space to generate patterns in the form of conjunctions of ontology terms, and uses the Fischer exact test and permutation testing to statistically validate the discovered patterns. As such, g-SEGS is a successful proof-of-concept semantic data mining system.

This paper is organized as follows: Section 2 presents the related work. Section 3 presents the new g-SEGS system and Section 4 provides an illustrative example. In Section 5, we conclude and give some directions for further work.

2 RELATED WORK

The idea of using hierarchies as background knowledge to generalize terms in knowledge discovery has been proposed already in early machine learning by Michalski [7]. More recent usage of ontologies in knowledge discovery includes [1, 11, 2] as well as domain specific systems that use ontologies as background knowledge for knowledge discovery [4, 12].

In [1], background knowledge is in the form of standard inheritance network notation and the algorithm KBRL—based on the RL learning program (Clearwater & Provost, 1990)—performs a general-to-specific heuristic search for a set of conjunctive rules that satisfy user-defined rule evaluation criteria. In [11], ontology-enhanced association mining is discussed and four stages of the (4ft-Miner-based) KDD process are identified that are likely to benefit from ontology application: data understanding, task design, result interpretation and result dissemination over the semantic web. The work of [2] first focuses on pre-processing steps of business and data understanding in order to build an ontology driven information system (ODIS), and then the knowledge base is used for the post-processing step of model interpretation.

An ontology driven approach for knowledge discovery in biomedicine is described in [4], where efforts to bridge knowledge discovery in biomedicine and ontology learning for successful data mining in large databases are presented.

A domain specific system that uses ontologies as background knowledge for data mining is SEGS [12]. The SEGS system finds groups of genes—the so-called gene sets—that are enriched. A gene set is enriched if the genes that are members of that gene set are statistically significantly differentially expressed compared to the rest of the genes. Compared to earlier work [10, 5], the novelty of the SEGS method proposed by Trajkovski et al. (2008) [12] is that it does not only test existing gene sets for differential expression but it also generates new gene sets that represent novel biological hypotheses. The SEGS method has four main components: the background knowledge, the hypothesis language, the hypothesis generation procedure and the hypothesis evaluation procedure.

3 GENERALIZED SEGS: G-SEGS

Motivated by successful applications of SEGS [6, 8], we decided to generalize it to become domain independent and named it g-SEGS. From the four main components of SEGS, only the SEGS hypothesis language and the generation and pruning procedure are general enough to be used unchanged in the new semantic data mining system g-SEGS. System g-SEGS inputs ontologies in the OWL format and data in the Orange [3] format, uses the hierarchical structure of the of the 'is-a' relation in ontologies for efficient search and pruning of the pattern search space, generates patterns in the form of conjunctions of terms from different ontologies, and uses the Fischer exact test and permutation testing to statistically validate the discovered patterns.

Interesting subgroups are constructed by conjunction of terms from the ontologies. All possible descriptions (by making all possible conjunctions) could be generated and evaluated for small ontologies. In case of very large ontologies, however, we need to prune the search space. In this case, we use the hierarchical property of the is-a relations of the ontologies. For example, if we constructed a subgroup with the following description $X \wedge Y \wedge Z$ ($X \in Ont1, Y \in Ont2, Z \in Ont3$), which covers m objects from class A, assuming a threshold of N ($N > m$) as the minimum number of objects that must be covered with the description, then we do not need to construct (and evaluate) all the intersections $x \wedge y \wedge z$, where $x \preceq X, y \preceq Y, z \preceq Z$ ($\preceq$ denotes more specific). This significantly reduces the search space of feasible descriptions.

g-SEGS is implemented in the Orange4WS data mining platform [9], which upgrades the freely available Orange data mining framework with several additional features: simple creation of new widgets from distributed web services, composition of workflows from both local and distributed data processing/mining algorithms and data sources, and implementation of a toolkit for creating new web services. By using these tools, we were able to give g-SEGS a user-friendly interface and the ability to be executed remotely as a web service. We defined the g-SEGS web service using WSDL (Web Service Definition Language). Using the created web service definition and the set of tools provided by Orange4WS, we created a web service for our system. Finally, also using Orange4WS, we imported the web service into the Orange environment, thus allowing g-SEGS to be used in various workflows together with other Orange widgets.

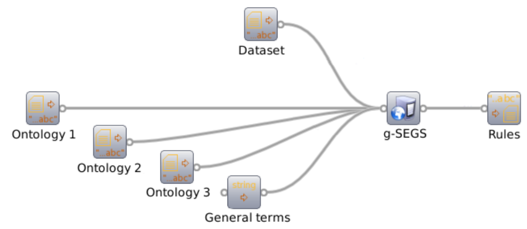


Figure 2: An Orange4WS workflow with g-SEGS.

A screenshot of an Orange4WS workflow with g-SEGS is presented in Figure 2. The workflow is composed of one widget for loading the data (Dataset), three widgets for loading the three ontologies, and one widget for specifying top-level ontology terms that are too general to appear in the final rules. These five widgets act as the input to the g-SEGS widget, which generates rules, displayed in the Rules widget.

g-SEGS inherited some limitations of SEGS, which include the limitation to four input ontologies, using a hierarchical structure (directed acyclic graphs only), which in practice means 'is-a' relations only, and cannot use attributes that are not annotated by ontologies.

4 AN ILLUSTRATIVE EXAMPLE

As a proof-of-concept of semantic data mining, we present the following example. Consider a bank which has the following data about its customers: place of living, employment, bank services used, which includes the account type, possible credits and insurance policies and so on. The bank also annotated the clients as 'big spenders' or not and wants to find patterns describing big spenders. Table 1 presents the example data.

An application of a 'standard' data mining algorithm (we chose the Orange [3] implementation of CN2) to these data produces the result presented in Table 2. These rules are very specific, due to the specificity of the attribute-values the data is described by. In classical data mining, such data should be manually preprocessed and attribute-values generalized to obtain more general rules and therefore more valuable results.

In addition to the data of Table 1, we propose to use three ontologies (depicted in Figure 3) to bring semantics into the knowledge discovery process. The result of applying g-SEGS to the data from Table 1 and ontologies from Figure 3 is presented in Table 3.

Table 1: A table of bank customers described by different attributes and a class 'big spender'.

id	occupation	location	account	loan	deposit	investment_fund	insurance	big_spender
1	Doctor	Milan	Classic	No	No	TechnologyShare	Family	YES
2	Doctor	Krakow	Gold	Car	ShortTerm	No	No	YES
3	Military	Munich	Gold	No	No	No	Regular	YES
4	Doctor	Catanzaro	Classic	Car	LongTerm	TechnologyShare	Senior	YES
5	Energy	Poznan	Gold	No	No	No	No	YES
6	Doctor	Rome	Gold	Apartment	No	No	Regular	YES
7	Finance	Bavaria	Gold	No	ShortTerm	GlobalShare	No	YES
8	Health-care	Frankfurt	Classic	Car	No	GlobalShare	Family	YES
9	Military	Warsaw	Gold	No	ShortTerm	No	Regular	YES
10	Education	Latina	Gold	Apartment	No	No	Family	YES
11	Health-care	Karlsruhe	Classic	Apartment	No	EuropeShare	No	YES
12	Retail	Munich	Classic	Car	LongTerm	TechnologyShare	Regular	YES
13	Education	Catanzaro	Gold	Car	No	No	No	YES
14	Doctor	Milan	Classic	No	No	EuropeShare	No	YES
15	Police	Munich	Gold	Apartment	No	No	No	YES
16	Retail	Stuttgart	Classic	Car	LongTerm	TechnologyShare	No	NO
17	Finance	Brescia	Gold	Apartment	No	EuropeShare	Regular	NO
18	Administration	Tarnow	Classic	Car	No	No	Senior	NO
19	Materials	Freiburg	Gold	Apartment	ShortTerm	GlobalShare	No	NO
20	Doctor	Poznan	Classic	Personal	ShortTerm	EuropeShare	Regular	NO
21	Administration	Cosenza	Classic	Car	No	No	No	NO
22	Unemployed	Munich	Classic	Car	No	No	No	NO
23	Military	Kalisz	Classic	Apartment	ShortTerm	EuropeShare	Regular	NO
24	Manufacturing	Cosenza	Gold	Apartment	LongTerm	No	No	NO
25	Transportation	Cosenza	Classic	Car	ShortTerm	No	Family	NO
26	Police	Tarnow	Gold	Apartment	No	No	No	NO
27	Nurse	Radom	Classic	No	No	No	Senior	NO
28	Education	Catanzaro	Classic	Apartment	No	No	No	NO
29	Transportation	Warsaw	Gold	Car	ShortTerm	TechnologyShare	Regular	NO
30	Police	Cosenza	Classic	Car	No	No	No	NO

Table 2: Rules generated by CN2 for data from Table 1. Coverage and confidence were computed in postprocessing.

Rules for class big_spender='YES'	coverage	conf.
loan='No' & account='Gold'	13.33%	100.00%
occupation='Doctor' & deposit='No'	10.00%	100.00%
occupation='Health-care'	6.67%	100.00%
occupation='Doctor'	16.67%	83.33%
occupation='Education' & account='Gold'	6.67%	100.00%

Table 3: Rules generated by g-SEGS from Table 1 data and ontologies from Figure 3.

Rules for class big_spender='YES'	coverage	conf.
Occupation(Public) & BankingService(Gold)	26.67%	87.50%
Occupation(Doctor)	20.00%	83.33%
BankingService(Gold)	46.67%	64.29%
Location(Germany) & Occupation(Service) & BankingService(InvestmentFund)	16.67%	80.00%
Location(Bavaria)	16.67%	80.00%

Characteristics of using g-SEGS (semantic data mining) are the following:

- more general rules compared to CN2 or other non-semantic data mining algorithms
- automated and therefore repeatable preprocessing - not prone to errors like human preprocessing
- g-SEGS rules have ontology terms with ontology names as conjuncts, while CN2 rules have attribute-value pairs.

5 CONCLUSIONS

This paper introduced the term *semantic data mining* which denotes a data mining approach where domain ontologies are used as background knowledge for data mining. We generalized a domain-specific system SEGS to perform semantic data mining on arbitrary ontology-annotated data.

There are many possible fields of application of semantic data mining. It can be directly applied to domains where data are characterized by sparsity and taxonomies are available, like market basket analysis, to give an example. Despite its current limitations, the new semantic data mining system g-SEGS shows major advantages compared to non-semantic systems, which include more general rules and automated data preprocessing. Hence, g-SEGS is a significant step towards practical semantic data mining.

Acknowledgments

The research presented in this paper was supported by the Slovenian Ministry of Higher Education, Science and Technology (grant no. P-103) and the EU-FP7 projects e-LICO and BISON.

References

[1] J. M. Aronis, F. J. Provost, and B. G. Buchanan. Exploiting background knowledge in automated discovery. In *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 355–358, 1996.

[2] L. Brisson and M. Collard. How to semantically enhance a data mining process? In J. Filipe and J. Cordeiro, editors, *ICEIS*, volume 19 of *Lecture Notes in Business Information Processing*, pages 103–116. Springer, 2008.

[3] J. Demšar, B. Zupan, and G. Leban. Orange: From experimental machine learning to interactive data mining, white paper (www.ailab.si/orange). Faculty of Computer and Information Science, University of Ljubljana, 2004.

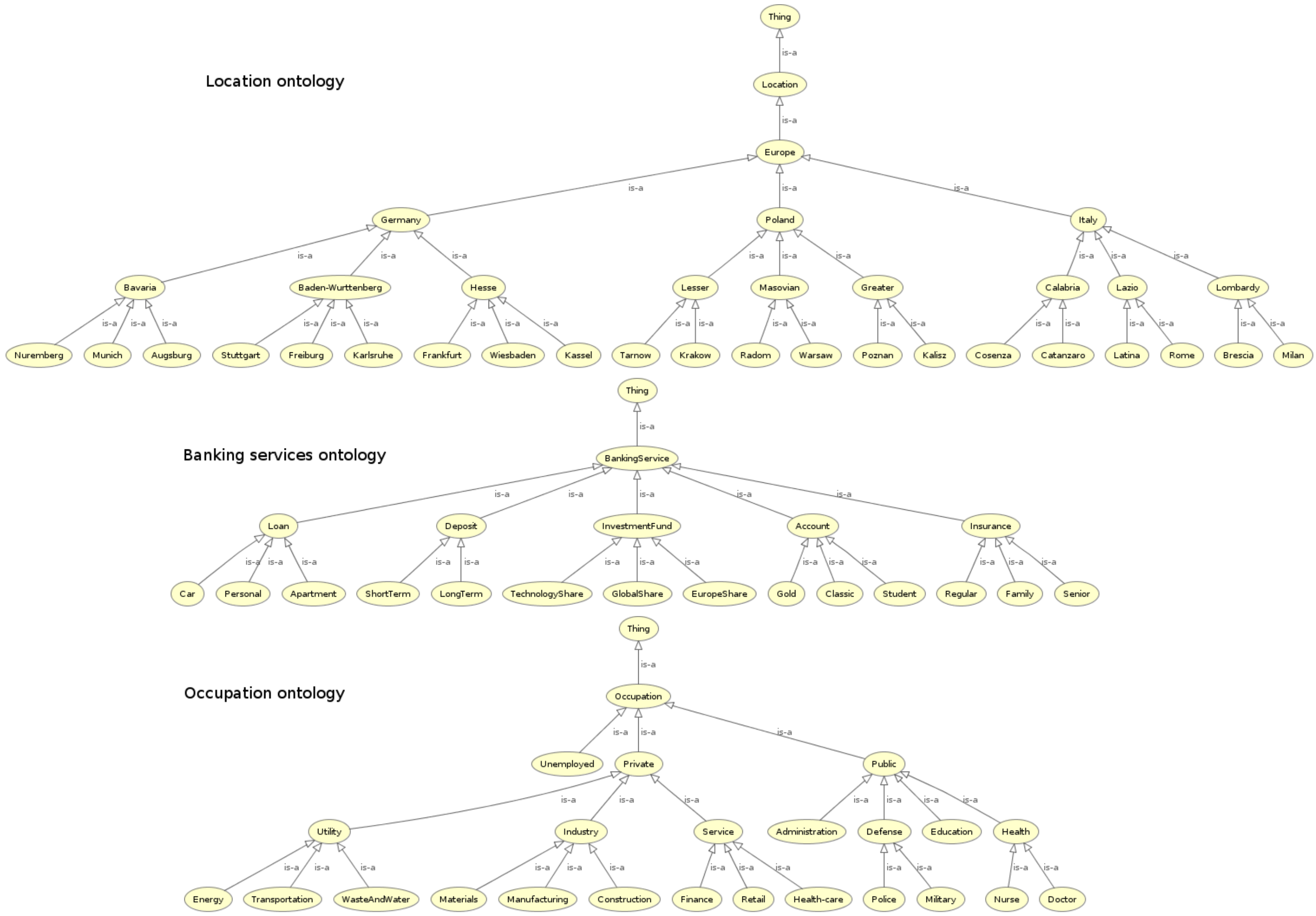


Figure 3: Three ontologies for data in Table 1.

[4] P. Gottgtroy, N. Kasabov, and S. MacDonell. An ontology driven approach for knowledge discovery in biomedicine. In *Proceedings of the VIII Pacific Rim International Conferences on Artificial Intelligence (PRICAI)*, 2004.

[5] S.Y. Kim and D.J. Volsky. Page: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(144), 2005.

[6] N. Lavrač, P. Kralj Novak, I. Mozetič, V. Podpečan, H. Motaln, M. Petek, and K. Gruden. Semantic subgroup discovery: Using ontologies in microarray data analysis. In *Proc. 31st Annual Intl. Conf. of the IEEE EMBS*, pages 5613–5616, 2009.

[7] R. S. Michalski. A theory and methodology of inductive learning. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, *Machine Learning: An artificial intelligence approach*, pages 83–129. Palo Alto: Tioga Publishing Company, 1983.

[8] I. Mozetič, N. Lavrač, V. Podpečan, P. Kralj Novak, H. Motaln, M. Petek, K. Gruden, H. Toivonen, and K. Kulovesi. Bisociative knowledge discovery for microarray data analysis. In *Proc. First Intl. Conf. on Computational Creativity*, pages 190–199, 2010.

[9] V. Podpečan, M. Juršič, M. Žakova, and N. Lavrač. Towards a service-oriented knowledge discovery platform. In V. Podpečan and N. Lavrač, editors, *Third-generation data mining: towards service-oriented knowledge discovery*, pages 25–36, 2009.

[10] P. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, and M.A. Gillette. Gene set enrichment analysis: A knowledge based approach for interpreting genome-wide expression profiles. *Proc. of the National Academy of Science, USA*, 102(43):15545–15550, 2005.

[11] V. Svátek, J. Rauch, and M. Ralbovský. Ontology-enhanced association mining. In *Semantics, Web and Mining, Joint International Workshops, EWMF 2005 and KDO 2005*, pages 163–179, 2005.

[12] I. Trajkovski, N. Lavrač, and Jakub Tolar. SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*, 41(4):588–601, 2008.

OPTIMIZING EMAIL-RELATED PRODUCTIVITY USING CONTEXTIFY

Gregor Leban, Marko Grobelnik
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773127
e-mail: gregor.leban@ijs.si, marko.grobelnik@ijs.si

ABSTRACT

Contextify is a tool for maximizing user productivity by showing email-related contextual information. The contextual information is determined based on the currently selected email and includes related emails, people, attachments and web links. This content is displayed in a sidebar in Microsoft Outlook and in a special dialog that can display an extended context.

1 INTRODUCTION

Despite the popularity of numerous web services, emails still play a crucial role in today's information exchange. It is very common for people to receive tens or even hundreds of emails per day. In this sea of information it is often difficult to stay organized and to find the right information when one needs it.

To help people perform mail-related tasks faster and with greater ease we developed Contextify. Contextify is an add-on for Microsoft Outlook. Its goal is to maximize user's productivity by unobtrusively finding and showing the relevant contextual information for the currently selected email. The main functions of the add-on are displayed in two windows which will be described next.

2 CONTEXTIFY SIDEBAR

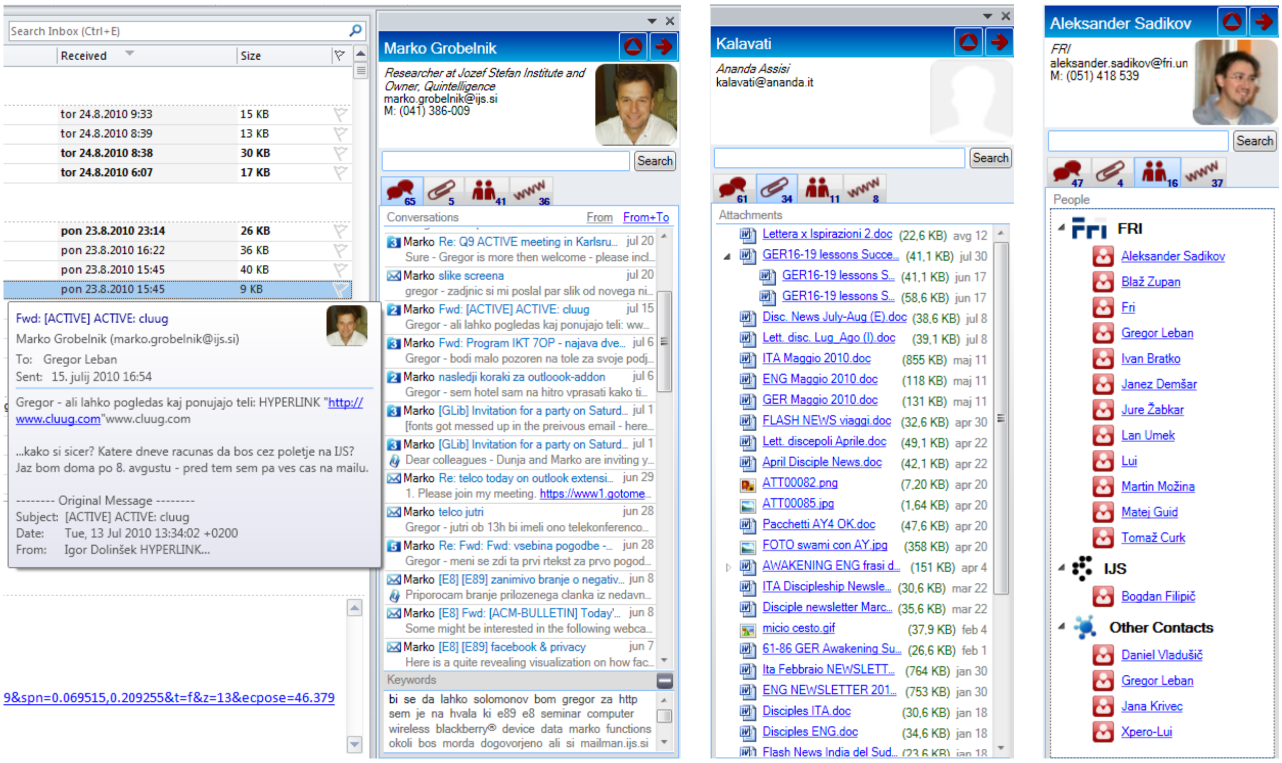
One way how Contextify displays the relevant contextual information is in a sidebar of Microsoft Outlook. The goal of the sidebar is to show important information related to the sender of an email being currently selected and shown within Outlook. This information is mainly gathered from past emails from this person but also from various online services such as Facebook, LinkedIn and Twitter.

An example of the sidebar is shown in Figure 1. The top of the sidebar displays the person's photo together with the relevant personal information collected from Outlook contacts and online social services (e.g. Facebook, LinkedIn). Other contextual information is organized into several tabs. In the first tab we show the list of recent emails from this person together with several email details (Figure 1.a). For emails that are a part of a thread, the whole thread can be brought into view by clicking on the email. Below the list of emails there is also a tag cloud of keywords that best describe the content of these emails. The second tab displays the list of exchanged attachments together with various details (Figure 1.b). Attachments that were exchanged several times, probably because being updated, are grouped together. They can be opened by clicking on the filename. The third tab lists the people who are participating in these emails (Figure 1.c) while the fourth tab displays the web links that were exchanged in the emails. The sidebar also provides searching capabilities where the context is determined based on the search terms and not the sender of the selected email. The found search terms are also highlighted for easier recognition. In the future we'll add a tab with a summary with person's appearance on various social services.

3 CONTEXTIFY DIALOG

An expanded context for the selected email can be displayed in the Contextify dialog. The goal of the dialog is to provide a visual display of the contextual information together with additional highlighting and filtering options. The context in this dialog consists of those emails where the participants (the sender and all the recipients) sufficiently match the participants in the currently selected email. The context is defined here as all emails sent to a similar social group. Such a definition of a context helps the user to see emails that are related to a particular aspect of his/her life.

An example of this dialog is shown in Figure 2. All emails that belong to the context are grouped into threads and displayed in the top left part of the dialog. Below the topic of each thread is a list of keywords that best describe the content of the thread. The bottom left part of the dialog shows a visualization of email activity for the computed context over time – each bar shows the number of emails



(a) (b) (c)

Figure 1. Information displayed in different tabs of the Contextify sidebar. The Email tab (a) shows recent emails from the selected contact, the Attachments tab (b) shows the exchanged files, and the People tab (c) shows the people included in the conversations.

that were received in a particular time period. By selecting or unselecting specific bars the users can display only emails from a specific time period. The right side of the dialog shows the social network for the participants in the context. There is a directed edge between persons A and B if there is at least one email sent from A to B. The font size for nodes depends on how often the person is present in the emails – this helps quickly identifying the most relevant participants. Selecting a person in the graph also highlights emails sent by this person in the list of emails.

4 CONTACT MANAGEMENT

Accessing and modifying information about contacts can be done in the Contact management dialog. Here, we can manually group different email addresses that represent the same person. In this way, when showing contextual information, emails from all person's email addresses will be displayed. By clicking the "Import contact information" button additional information about contacts can be automatically imported from Facebook, LinkedIn and Outlook contacts. The "Cleanup contacts" button provides a fast way to tidy-up contact names and to merge contacts which are likely to represent the same person. An example of the Contact management dialog is shown in Figure 3.

5 CONCLUSIONS AND FUTURE WORK

The goal of Contextify is to determine and show the relevant contextual information that should help the user to more efficiently perform his email-related tasks. We are currently working on text-mining algorithms to provide additional functionality such as email summarization, automatic categorization of emails into folders, improved extraction of information from email contents and more advanced contact management.

A video demonstration of the add-on can be seen at http://www.youtube.com/watch?v=hYpxhUvYM10 in lower resolution and in higher resolution at http://www.screencast.com/t/ODg5MzM1YWEt.

6 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency, the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and ACTIVE (IST-2008-215040).

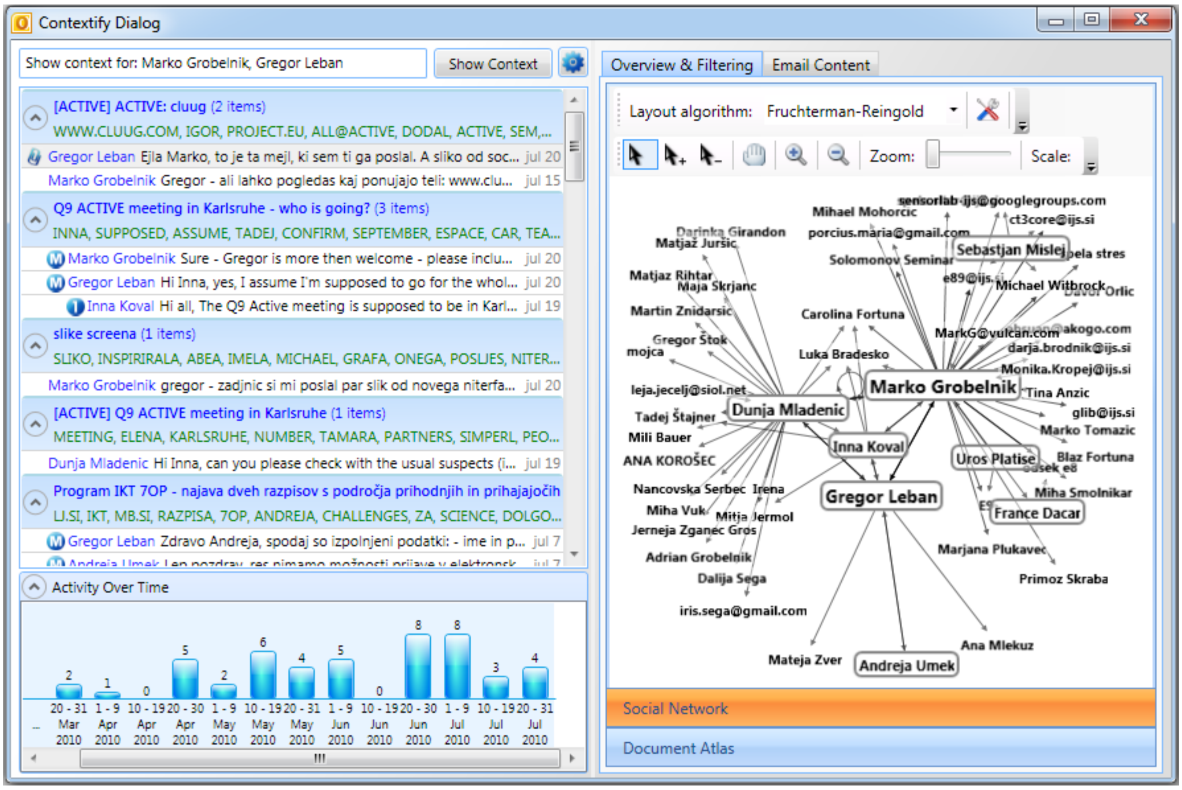


Figure 2. Contextify dialog displays an expanded version of the context.

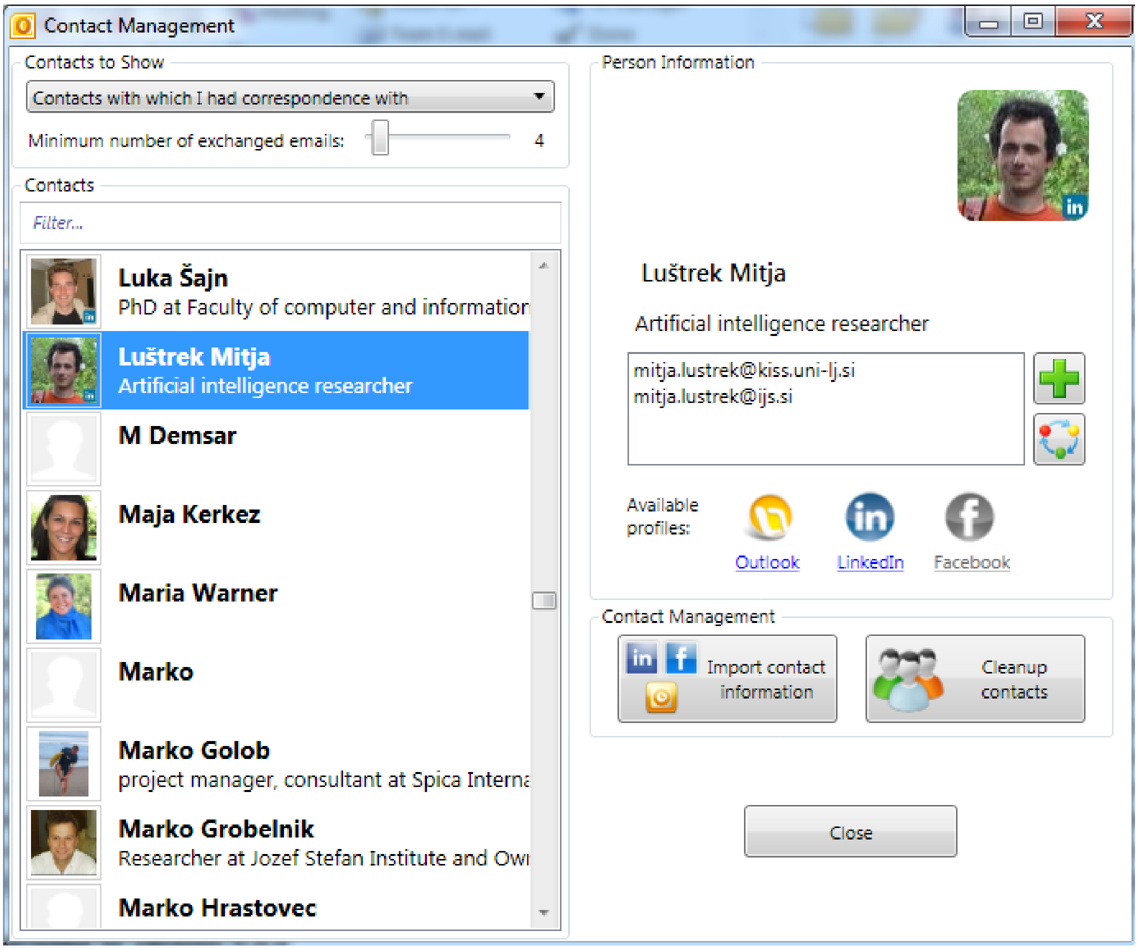


Figure 3. Contact management dialog displays information about the contacts collected from emails. Additional information about the contacts can be imported from social web-services such as Facebook and LinkedIn.

KNOWLEDGE PROCESS MINING AND OPTIMIZATION

Tadej Štajner, Carolina Fortuna, Dunja Mladenić, Marko Grobelnik
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773900
e-mail: {firstname.secondname}@ijs.si

ABSTRACT

This paper addresses the problem of optimizing knowledge processes that have been automatically identified from a data stream. We are proposing extension of TaskMiner tool for process mining and visualization in collaboration processes. The proposal is to enable TaskMiner graphical interface for process mining to supports a larger scope of the process analyst's workflow: process discovery, visualization and measurement, providing decision support for process refactoring and follow-up measurements of implemented optimizations.

1 INTRODUCTION

Knowledge processes, as seen in this paper, involve knowledge workers in an enterprise who are usually involved in several projects that require accessing different data sources, exchanging messages, browsing the Web etc. With the wide usage of computers in enterprises, one can expect that each knowledge worker has access to a personal computer, where a program can be installed to record activity on the level of complex events, such as, at time *T* a person *P* has accessed a document *D*. We assume that each event is associated to a context (e.g., a project) and that it is possible to cluster the events so that we automatically identify which events belong to the same context. Each context has data collections associated to it and possibly interconnected with some kind of relation. In our scenario, knowledge workers switch from working in one context to the other on weekly, daily or maybe hourly bases.

This paper presents bottom-up approach to optimization of *knowledge processes*, where knowledge processes are seen as loosely coupled sets of activities occurring in some context. The developed approach is semi-automatic, implemented as an extension of TaskMiner tool. TaskMiner enables displaying and reporting the state of the knowledge process which exposes the metrics needed for optimization. The initial version of TaskMiner has already been described in [2]. In this paper, we have extended TaskMiner with reporting capabilities that enable measurement of objective metrics, needed for process optimization.

The paper is structured as follows: Section 2 briefly presents the TaskMiner tool. The proposed approach to semi-automatic knowledge process optimization is presented in Section 3. Sections 4 gives conclusions and some directions for future work.

2 PROCESS MINING USING TASKMINER

Process Mining that is used as a starting point of our work is based on the bottom-up approach using data mining techniques to obtain a probabilistic process model. The setting is as follows: given a database, describing events in a business setting, executed by actors on resources, construct a probabilistic temporal model that best describes the action patterns appearing in the event. The model construction is performed by action mining followed by process mining [3]. Since the data is provided in the form of a graph, composed of multiple different node types, action mining is addressed as an example of a multi-relational clustering problem.

Process mining was performed by using Markov Models for finding frequent sequences of actions in the data, as one of the standard algorithms applied for process mining with the extension of pruning the obtained models by selecting only the statistically significant transitions [5]. It is developed for process mining on TNT (text, network, time) data proposed in [1]. Results from the developed prototype were shown on real-world datasets. The real-world data captured in TNT events reflects the situation of knowledge workers in a real-world setting, including interruptions and context switches, noise from different sources, under-defined tasks and contexts.

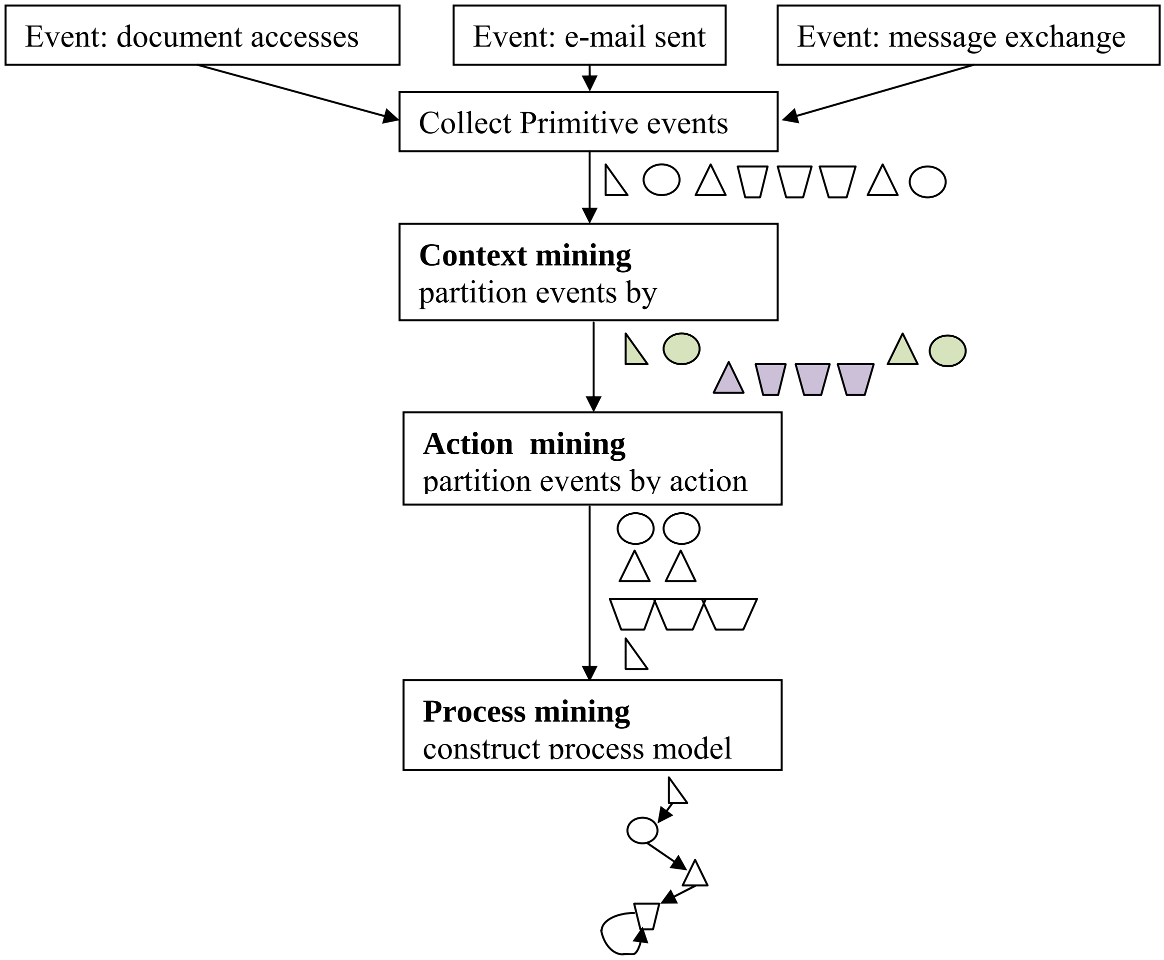


Figure 1: *Architecture of probabilistic process mining. Icons represent individual events, their shapes represent actions, and their colors represent contexts.*

Figure 1 shows the architecture of probabilistic process mining as implemented in TaskMiner. First, events are collected, such as, document access, e-mail send or receive, exchange of short messages between the users, etc. We then execute the steps of context and action mining.

In the case of context mining, context correspond to clusters of events using people and terms that appear in the events' contents, since these are the features which knowledge workers tend to consider as distinguishing for contexts.

On the other hand, action mining is a similar clustering procedure, but executed on a different feature set of the same event stream. In action mining, the features consist of content terms (without named entities), the type of the event and, an abstraction of social properties such as the roles of participants (i.e. is it a private conversation or a group, or does it span a single or multiple organizations). This view of features gives us a context-free representation of events. The purpose of clustering events into actions is to construct a process model from those actions.

Context mining is applied in order to identify contexts and partition events by context. Action mining is applied to partition events by actions. In process mining, a process model is constructed by finding sequences of actions based on the data. We assume that a process model contains actions belonging to the same context.

3 APPROACH DESCRIPTION

TaskMiner offers a graphical front end for the process mining service (see an example of visualization in Figure 2). It enables the process analyst to explore process models using different process mining views and parameters: viewing per-user or per-context model with varying degree of granularity. Its core process mining functionality, prototyped in [3] is also used for process-based prediction [2] and represents an implementation of a probabilistic model for the activities of knowledge workers [4].

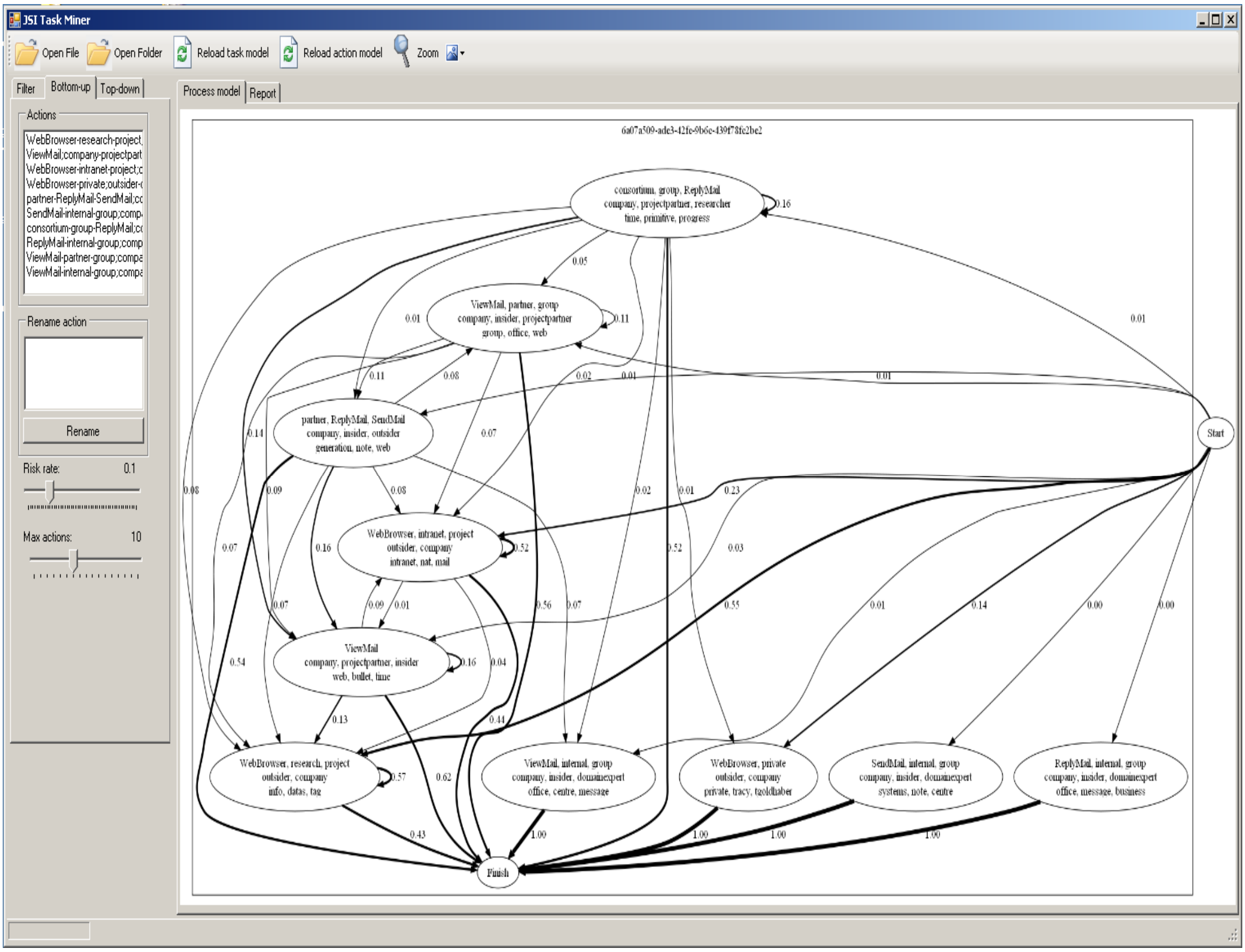


Figure 2: *TaskMiner process visualization view. Thicker lines represent more frequent transitions.*

Imagine a process analyst exploring the process visualizations, as seen in Figure 2, with the purpose of refactoring the process afterwards. Compared to other tasks, process optimization has another key component: objective metrics that one is able to optimize. With the data that we are using as input for process mining, we are able to measure metrics from the perspective of mined actions and their distribution over time:

- Total time spent executing some action
- Average time spent executing action
- Number of events representing action
- Percentage of time spent executing action

Although one may be able to derive some of these metrics from the total time and number of events per action alone, the average time spent per action is more easily interpretable as an objective metric. For instance, one is able to isolate the effects of a particular optimization approach either on average execution time (faster actions) or a lower number of required executions (less actions). Also, some optimization approaches affect only some actions - in those cases, one is still able to measure the effect on that particular type of action alone.

Since each individual event is only associated with its time stamp, we approximate its duration using the time difference from the previous event, as long as the last event occurred in the last 60 minutes with the assumption that individual actions are not longer than 60 minutes for this particular domain. The metrics reporting functionality are implemented as another tab in the TaskMiner interface, showing a data grid of metrics for each action.

As a prototype study, we have measured these metrics on the real-world log dataset from a major telecommunication provider. The process mining metric report in Figure 3 shows that all e-mail related actions tend to have very similar duration of around two to two and a half minutes, while all of the different browsing- related actions tend to average around 45 seconds. In other words, reading an e-mail takes on average the same effort as writing or responding to one.

6 CONCLUSION

We have described a bottom-up approach to optimization of Knowledge Processes, based on extending TaskMiner tool for process mining and visualization. The TaskMiner graphical interface for process mining now supports a larger scope of the process analyst's workflow: process discovery, visualization and measurement, providing decision support for process refactoring and follow-up measurements of implemented optimizations.

Future work involves using the implemented reporting to test various process optimization approaches for various common knowledge worker tasks, such as, applications for optimizing e-mail management or process-based resource prediction for document management.

7 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and ACTIVE (IST-2008-215040).

References

[1] M. Grobelnik, D. Mladenić, J. Ferlež; Probabilistic temporal process model for knowledge processes: handling a stream of linked text, Proc. of the 12th international multiconference on Information society IS-2009, SiKDD, pages 222–227. 2009.

[2] T. Štajner, C. Fortuna, D. Mladenić, M. Grobelnik; Prototype for forecasting future behavior and actions in collaboration process; ACTIVE Deliverable, http://wiki.active-project.eu/wiki/D3.5.1, 2010

[3] T. Štajner, C. Fortuna, D. Mladenić, M. Grobelnik; Scalable learning of probabilistic temporal models from data (prototype and evaluation); ACTIVE Deliverable, http://wiki.active-project.eu/wiki/D2.2.2, 2010.

[4] T. Štajner, D. Mladenić: Modeling knowledge worker activity; Workshop on Applications of Pattern Analysis - WAPA, Cumberland Lodge, 2010

[5] S. Jacquemont, F. Jacquenet, and M. Sebban. Mining probabilistic automata: a statistical view of sequential pattern mining. Machine Learning, 75(1):91–127, 2009.

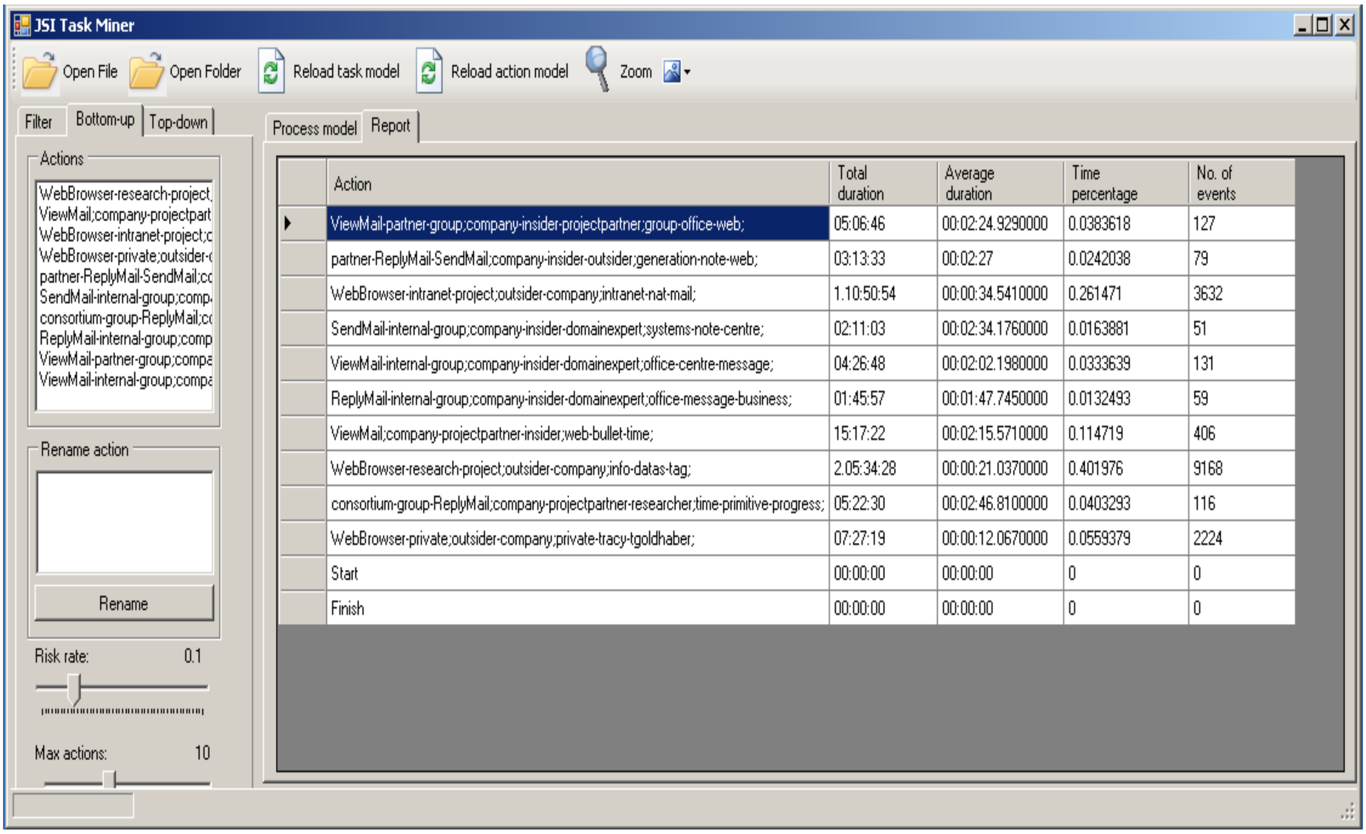


Figure 3: *Example process mining report in the extended TaskMiner*

USING ADVANCED BUSINESS INTELLIGENCE METHODS IN BUSINESS PROCESS MANAGEMENT

Andrej Kocbek[1], Matjaz B. Juric[2]
[1]Intera d.o.o.
Črtkova ulica 18, 2250 Ptuj, Slovenija
[2]Univerza v Mariboru
Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova ulica 17, 2000 Maribor, Slovenija
E-pošta: [1]andrej.kocbek@uni-mb.si, [2]matjaz.juric@uni-mb.si

ABSTRACT

Business intelligence refers to computer-based techniques used in discovering, processing and analyzing business data to present statistics of business process to managers and analytics. Is a set of applications and techniques used for gathering and analyzing data stored in data warehouses and data marts. Integrating advanced business intelligence methods, such as data mining, predictive analytics and complex analysis, into business process management becomes essential for business. In this way analysts and managers of business processes with the successful use of mentioned methods have the opportunity for effective business and this reduces total cost of ownership.
A major challenge is: How to establish business intelligence methods in process management using service oriented architecture? In the article we will discuss the issue of integrating business intelligence with business process management and use of advanced business intelligence methods to create complex reports and analysis.

1 INTRODUCTION

One of the most important steps in the efficient business is understanding business process flow and its proper managing and monitoring. This can be achieved by using business process monitor managements systems (E.g. Business Activity Monitoring) which graphically displays the efficiency of key performance indicators. Thus managers and analysts in a quick and effective way obtain information about critical business process activities – this leads to fast response and problem solution, business can be more efficient by using advanced business intelligence methods. In the article we will discuss the issue of using advanced business intelligence methods for creating complex reports and analysis and fill the gap integration between business intelligence and business process management.

2 BUSINESS INTELLIGENCE

In the business, most of the analysts and managers need access to accurate and timely information which helps them to reach their business goals. To understand the past behavior of their business, they must have access to preliminary data. But today the focus is on getting access to current transactions and business events in order to react quickly to new demands, market pressures, competitor movements, and other business challenges.
Business intelligence (BI) is the process where users obtain accurate and consistent business data from the enterprise data warehousing environment (e.g. data warehouses, data marts). With obtained data from different business context, users can analyze and identify trends and anomalies, execute simulations and obtain insights about business problems. [3]

2.1. Benefits of business intelligence

Recently the concept of business intelligence has gained huge significance in every organization. But if we want to understand this significance, we need to know and understand the most common benefits of business intelligence. We also need to be aware that benefits of using business intelligence depend on the size and structure of your business and will vary. However, there are a number of benefits that are common to all businesses that deploy BI. [3]

Common benefits of business intelligence are listed below:

- Helps align the organization towards its key objectives.
- Empower analysts and managers with a single version of the truth.
- Enables faster and fact-based decision making.
- Meet or exceed your customer expectations - based on factual information.
- Find latent problems by building a picture of the information we cannot see.
- Get instantaneous answers to next business question. The next business question is the question that you did not know existed until you began looking at data.
- Quickly recognize what are competitive advantages.
- Combines multiple sources of data for decision making
- Efficient collection and distribution of essential data and statistics.

2.2. Life cycle of business intelligence

The successful introduction of business intelligence requires four steps, which are with each other connected cyclically. Figure 1 show the life cycle of business intelligence.

- **Getting accurate data** – Getting large amount of data from data warehouses, data marts, databases and other sources.
- **Analyzing the data obtained through BI operations** – Complex elements are extracted into smaller segments for better understanding and creating new knowledge. Business problems are solved by indentifying business needs. Reporting components (annual reports, analysis, dashboards, charts, etc.) are created.
- **Indentifying trends, changes and incorrectness** – Trend forecasting by using predictive analytics. Identify threats and opportunities within the business by using complex mathematical methods and algorithms
- **Simulations and gaining new knowledge** about business problems and opportunities.

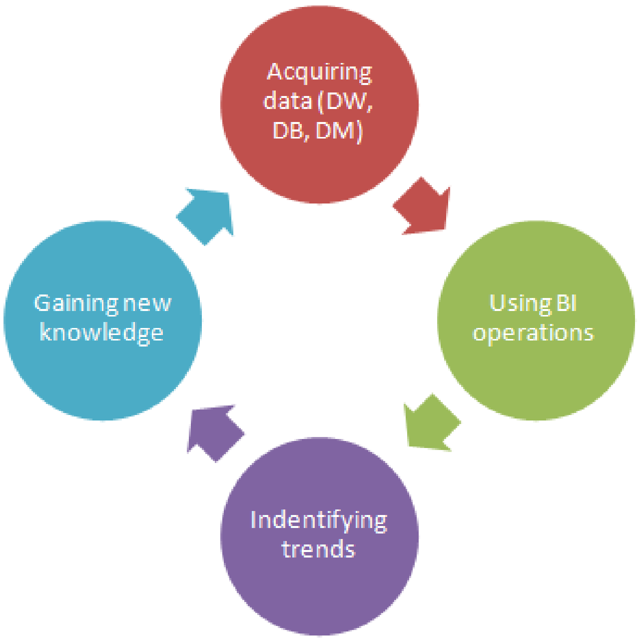


Figure 1*: Life cycle of business intelligence.*

3 ADVANCED BUSINESS INTELIGENCE METHODS

In the next section we will discuss using advance business intelligence methods in business process management to create complex reports and analysis. [4]

3.1 Complex analysis

Online Analytical Processing (OLAP) is an important component of business intelligence. Using a data structure (OLAP cube) enables fast data analysis and is defined as the ability to manipulate and analyze data from different perspectives. OLAP is a tool which allows fast creation of data inquiries and overcomes the limitations of rational databases. Its purpose is to enable fast and user-friendly queries for data. Data warehouses configured for OLAP use a multidimensional data model which allows complex analytical and ad-hoc queries with a rapid execution time. The typical OLAP applications are used in business reporting for marketing, sales, business process management, forecasting, financial reporting and etc.

With the support of OLAP the analyst navigate through the database, data warehouse or data mart and search for a particular subset of data, changing the data's orientation and defining analytical calculations. Common operations of OLAP are: [4]

- **Slice and Dice** – Slicing refers to the ability to combine and re-combine the dimensions to see different slices of the information. The dice operation is to slice a data cube on more than two dimensions. Analysis across several dimensions and many categories of data items are used to uncover business behavior and rules. Analysts use this operation to answer your next business question.
- **Drill Down/Up** – Specific technique where analyst navigate among different levels of data.
    - Up mean to move to less detailed information. We see less detailed information or the parent of the child that we are currently viewing.
    - Down mean to move to more detailed information.
- **Pivot** – This technique changes dimensional orientation of report. E.g. changing the dimensions from columns in row and vice versa of the table.
- **Nesting** – Display data dimension one dimension within another.

3.2 Predictive analytics

Nowadays, business is all about placing bets and knowing if the odds are in our favor. Business success depends on being capable to understand and analyze the future and take correct actions as soon as possible. Business analytics and managers must be able to predict the future scenarios well enough to prepare plans to seize opportunities, neutralize threads and mitigate risks. Predictive analytics here plays an essential role and helps to plan day-to-day operations in business. Predictive analytics, using obtained information about data from data warehouses with help of mathematical algorithm, predicts the trend of business processes. The results of predictive analytics are used to identify patterns, trends and forecast business activity flow. The accuracy of forecast depends on the complexity of data interdependence. By increasing the number of variables we get less accuracy forecast. [4]

Methods and algorithms for predictive trends are:

- Regression techniques (Linear regression model).
- Discrete choice models (Logistic regression, Probit regression, Multinomial logistic regression).
- Time series models.

- Survival or duration analysis.
- Classification and regression tress.
- Multivariate adaptive regression spines.
- Machine learning techniques (Neural networks, Radial basis functions, Naïve Bayes, K-nearest neighbors).

Predictive analytics model

Business predictive analytics model analyze historical events and data with goal to find patterns of subtle data in relation to surround which uses data on dependent and independent way. Prediction models often perform real-time transaction and calculations for the purpose of introduction data-driven decisions which lead to efficient and effective business. [6]

3.3 Data mining

Data mining is a process of obtaining and detecting patterns in data. Mainly is used in different practical situations such as marketing, supervision and fraud detecting. Is a process which detects important correlations, patterns and trends with analysis of large amount of data stored in data warehouses. Data mining uses tools and statistical and mathematical techniques to detect patterns in data. Knowledge of these hidden relations provides valuable insight into key business drivers. [3]

We know two basic types of data mining:

- **Hypothesis validation** –Used when we have an idea or a hunch about important relations between data elements.
- **Knowledge discovery** – This type of data mining is used when we may have unknown statistically significant relations between data elements which are difficult to recognized.

The most common tasks which can be done using data mining are:

- **Classification** – Indicates the characteristics of a particular group. A characteristic of each group is used to design a model, which allows the classification of new, unknown samples to one of the groups.
- **Clustering** – Indentifies groups of things that have certain common characteristics.
- **Association** – Finds the relations between events which happen at a given time.
- **Sequencing** – Is similar to association. The different is that here we are looking for relations according to the time period.
- **Regression** – Is a form of appraisement which is used to display information according to predictive value. Linear and nonlinear techniques are used.
- **Forecasting** – Predicting future values based on patterns and large data sets.

4 INTEGRATING BI WITH BPM

Organizations which want to understand and optimize their core processes and operations need to establish integration between business intelligence and business process management (BPM). With this integration users can directly invoke additional business processes as a result of business insight. Majority of medium and large organizations use the special software for managing and tracking critical business functions. Generally, the integration of complexity and heterogeneity of IT infrastructure and applications represents a major challenge. [5][7]

4.1 Integration problems

Most organizations have failed in process of establishing a successful integration between business intelligence and business process management. One of the major reasons is that the solutions and tools for business process management are offered by different vendors as tools which can be established BI, thus, the lower the change of integration significantly. Even more difficult is integrating that data management entirely on a different area organization such as the development of business processes. Solution that can solve these problems is use of service oriented architecture. SOA enables easy use of business processes with business intelligence services through enterprise service bus. [5]
Problems often occur with less-structured and more far-flung data environments where we have frequently changing system source. One of the problems is that most enterprise integration and BI is centralized, while the data is distributed. The solution is to move to a more distributed integration model, where data is ready and available as a service to the individual. The individual is already performing much of the innovation, exploration and integration, and this distributed model facilitates the process.

4.2 Data level integration

Recent approaches to integration between business intelligence and business process management have been very limited and simply reduced to the problem of technical integration between BPM in BI systems. Without doubt the technical integration between those systems is very complex. But when we solve this problem, efficiencies and opportunities in solving real-time problems in organization, increase. [1]

In this section we will discuss integrating BI and BPM. There are several levels (e.g. data level integration, user view integration, etc.) of integration of those two areas. Bellow we will take a look to the data level of BI and BPM integration.

As shown in Figure 2 first step represents the execution of processes in which new events are generated. Processes run on process server. Through Enterprise Service Bus (ESB) different web services communicate and one of them collects new generated information about events. ESB provides fundamental services for complex architectures via an event-driven and standards-based messaging-engine. The main purpose of the ESB is to interrupt links between services and their users. Later all collected events are filtered, merged and transformed. Structural data is then stored in specific data warehousing environment where they are collected and wait to be used when it is necessary. Collected data is now prepared to be use in BI and Business Activity Monitoring (BAM) systems. On these manner users have possibilities to generate complex reports, analysis and dashboards about business processes. [1] [2]

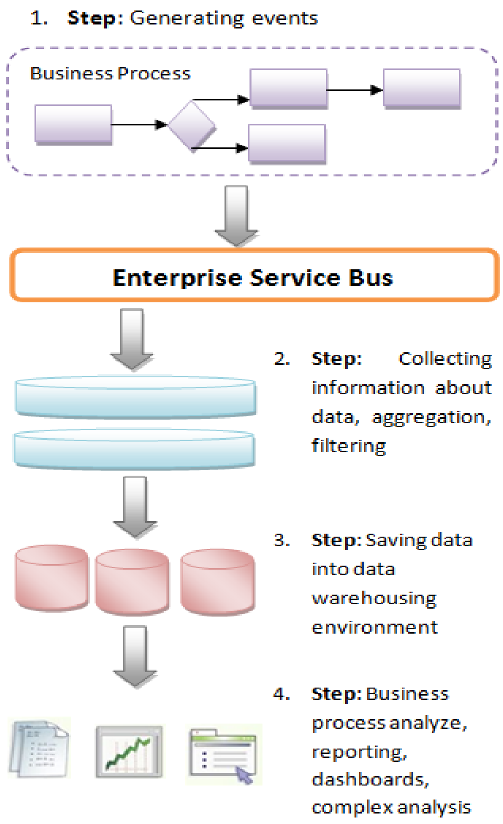


Figure 2: *Data level of BI and BPM integration*.

4 CONCLUSION

In this paper we discussed integration between business intelligence and service oriented architecture and meaning of using advanced business intelligence functions in the workflow management. Experience from past indicated that successful business requires business process monitoring using different systems and tools (E.g. BAM). This way business analysts and managers gain important information about activity of organization and with the help of reports, analysis, and dashboards they get insight into what happened in their own organization. But business results are more efficient if they are able to answer following question: *Why at some point the effectiveness of key performance indicators is under the critical level?* Answer to this question is using business intelligence in correlation with business process management and service oriented architecture. We described advanced business intelligence methods which help to create complex analysis in reports. If we want to use those methods with business processes successfully we need to establish integration between BI and BPM which is not always an easy task.

References

[1] Chuck Ballard, Ahmed Abdel-Hamid, Robert Frankus: *Improving Business Performance Insight … with Business Intelligence and Business Process Management*, RedBooks, 2006, pp. 67-123

[2] Eckerson, W., *Performance Dashboard: Measuring, Monitoring and Managing Your Business*. New York, NY: Wiley, 2006

[3] Jaklic Jurij: *Poslovna vrednost poslovne inteligence*, Ekonomska Fakulteta, Ljubljana, 18.10.2005

[4] McGraw Hill: *IBM Cognos 8: Business Intelligence – The Official Guide*, IBM, 2008, May

[5] Steven Rogers: *Oracle BIEE Integrates With SOA*, Oracle, 2007

[6] Indart, B., *Operationalizing Business Intelligence – Turning Insight Into Action*, Business Intelligence Journal, Second Quarter, 11, 2 pg. 35.

[7] Imhoff, C. (2005), *Streaming Processes for Front- Line Workers: Adding Business Intelligence for Operations, Operational Business Intelligence*, White Paper, Intelligent Solutions Inc., December 2005.

Operation part financed by the European Union, European Social Fund. Operation implemented in the framework of the Operational Program for Human Resources Development for the Period 2007-2013, Priority axis 1: Promoting entrepreneurship and adaptability, Main type of activity 1.1.: Experts and researchers for competitive enterprises.

STREAM MINING ON ENVRONMENTAL DATA

Maja Škrjanc, Dunja Mladenić
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773900
e-mail: {firstname.secondname}@ijs.si

ABSTRACT

This paper addresses a problem of environmental data analysis, where the data is obtained from different data sources including sensor measurements. We propose a framework for stream mining on environmental data and illustrate its applicability on real-world datasets. Moreover, in collaboration with domain experts two scenarios of data analysis are identified as potentially useful in environmental stream mining.

1 INTRODUCTION

Environmental phenomena share a dynamic nature of chance with the nature itself. With modern technologies we can observe, trace and analyse more and more of dynamic data describing the environmental changes. A sub-field of Knowledge Discovery called Stream Mining [2, 3] addresses the issue of rapidly changing data. The idea is to be able to deal with the stream of incoming data quickly enough to be able to simultaneously update the corresponding models, as the amount of data is too large to be stored: new evidence from the incoming data is incorporated into the model without storing the data. The underlying methods are based on the machine learning methods of on-line learning, where the model is built from the initially available data and updated regularly as more data becomes available.

This paper proposes a framework for stream mining on environmental data based on the identified needs of domain experts working on analysis of environmental data. Two specific scenarios are addressed, one supporting definition of alarm triggers and the other supporting analytic browsing in the context of historic data.

The paper is structured as follows: Section 2 presents the proposed framework to environmental stream mining. Application of the framework on real-world datasets is described in Section 3. Sections 4 gives conclusions and some directions for future work.

2 PROPOSED FRAMEWORK

The proposed framework for environmental stream mining named *EnStreaM* enables the users to analyze structural, i.e., more static data and stream data obtained from sensors. In addition, it supports usage of semantic annotations of the original data if available as part of the input data.

We assume that most of the environmental domain related tasks which include observation and monitoring of environmental phenomena use three main data inputs:

- structured data, which is more or less static during the monitored period of time
- data streams, which are usually sensor data related to the observed phenomena. Sensor data in general is collected from different locations and sensor platforms. They can have different representations of the sensor measurements and can be compliant to different standards (e.g., [4, 7])
- data annotations, annotating the structured data or providing semantic annotations of the sensor preferably using some standard form, (e.g., [5])

The data is describing environmental phenomena and each event is analyzed and monitored through time. To be able to calibrate and re-use appropriate prediction models, end-users need adequate analytical platform to combine structural data and data streams provided as in time series. The proposed framework enables the users to browse through history of observed phenomena by different categories, analyse and define data patterns and have the ability to identify rules for alarm triggering.

We use a general scenario of environmental stream mining for developing the framework. Figure 1 shows the proposed data model for the general scenario. On the top of that based on the discussions with domain experts, two specific scenarios are proposed: definition of alarm trigger and analysis of observed environmental phenomena in a historical context. Architecture of the proposed framework is shown in Figure 2. The stream mining is designed as a Web service that can be easily incorporated in a larger portal for environmental services [6]. The rest of this Section provides a more detailed description of the scenarios.

2.1 Definition of Alarm Trigger

Stream mining methods can be applied in the phase of monitoring and validation of a prediction model, used in a process of monitoring the environmental phenomena. The results enable the user to upgrade the monitoring process with alarm triggering rules. The EnStreaM framework enables monitoring of the model performance and possibly auto-refresh the model when necessary.

When the specified environmental phenomena i.e., event occurs, the expert user runs the existing process of analysis, which, among other results, also provides prediction of how the observed event will evolve. Based on prediction, the experts can select the appropriate response to the phenomena. There are special situations where the expert user needs to re-run the modelling phase of the prediction model. There are two typical situations when the user should refresh the prediction model [1]: the phenomenon is in reality evolving significantly different then the prediction model anticipated (changed event status); the input data for prediction model has significantly changed (e.g. weather forecast is significantly differs from the reality).

In both cases the framework enables the expert user to analyse and monitor the observed event through time and define appropriate actions accordingly. When some of the observed data used for prediction or event status is significantly changed, the user can re-create or update the alarm triggering rules accordingly. The definition of "significant changes" is defined by the user. We anticipated two different settings, which are both captured in proposed framework.

(1) Automatic monitoring the quality of the prediction model. One of the possibilities is that the quality of the prediction models is automatically monitored by comparing the predicted event status with the event status in reality (e.g., satellite image) and if the prediction quality is significantly dropped, the prediction model is updated – re-created with updated input data (e.g., from sensor stream of weather data).

(2) Automatic monitoring the input data streams. The input weather data is monitored and compared to weather forecast data. If the current weather conditions are significantly changed from the weather forecast data used for modelling phase, the prediction model is automatically updated with the fresh input weather data. The threshold for updates of the model depends on the input weather parameters.

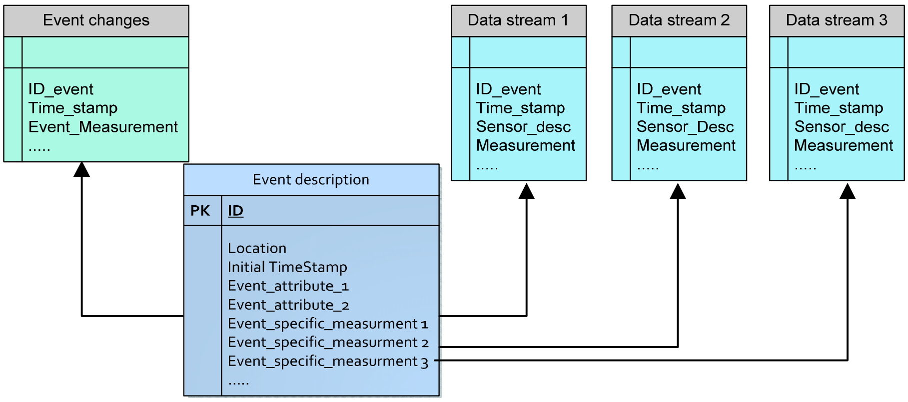


Figure 1: *Data model of general environmental data for the proposed stream mining framework.*

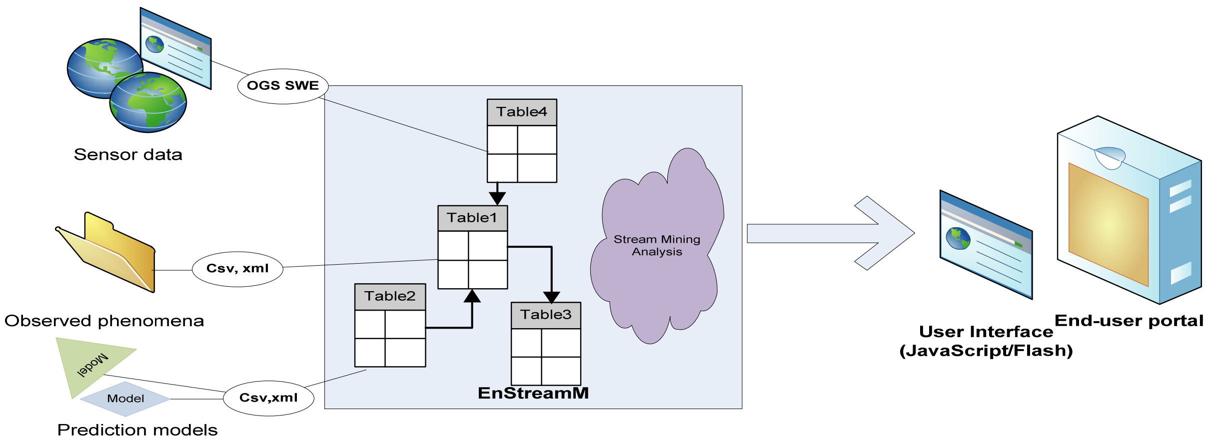


Figure 3: *Architecture of EnStreaM framework.*

2.2 Observed phenomena in a historical context

The proposed second scenario is focused on the empowerment of the user with additional knowledge extracted from advanced analytical browsing of past situations. EnStreaM enables complex data analysis, combining structural data and data streams from different sensors, locations and time. This enables the users to analyse, compare and search for certain phenomena from the past, while at the same time the data related to the observed phenomena will be updated online with incoming sensor data, e.g., from available web services.

In certain situations, the end-user define or calibrate additional input parameters, which are crucial for the quality of the prediction model. To be successful at the modelling phase, the end-user needs to gain expertise by analysing the past situation and appropriate solutions. Data about past situations could include information about predicted models and used user defined parameters. EnStreaM framework enables advanced analytical browsing of similar past situations, which empower the user with additional expertise. Moreover, the user can define what type of similarity is the most important for a given case in a specific situation. When determining the similarity factors, the user is able to search the database of past phenomena and retrieve a set of similar situations from the past. One additional possibility is that at the same time the stream mining capabilities enable the comparison of weather conditions from the online web services with historical values from the similar situations. The available database also includes different types of available predictive models and their parameters.

3 APPLICATION TO REAL-WORLD DATA

The proposed framework was developed by taking into account real-world scenarios as described in Section 2 and two case studies [1]. In this Section we describe the current status of the real-world case studies from two viewpoints:

- process of usage - description of how the user is currently exploiting the workflow process and,
- data availability- data sources that are currently used in case studies' process.

3.1 Oil spill case study

The main goal in the oil spill case study as described in [1] is to predict how the oil spill drifts and its consequences affect the environment. This prediction is the basis for decision making strategies and tactics of responding to oil spills in the sea. Historical data on past situations including weather conditions represents one of the key factors for a successful learning process for experts to create prediction models.

The most common process of predicting oil spills drifts is as follows. Domain expert gets request to create prediction model with the following data: location of the spill (latitude, longitude), amount or rate, oil type, type of spill (on the water, bellow he water), prediction window (in days) – for how many days ahead the customer needs prediction – time period. Domain expert collects the available data: bathymetry - grid of the see ground, grid of the coast line, weather forecast for selected time period (winds, current,etc) (see Figure 2). If the weather data is not available, experts can replace missing data with historical data from the same period of the year (one or several years ago) for the specific location or by a summarized data - calculated averages (e.g., based on sparse historical data) for the specified area. Alternatively, the expert can run the prediction model, which provides among other results 3D time-stamped representation of how the oil is expected to drift and changes of mass balance.

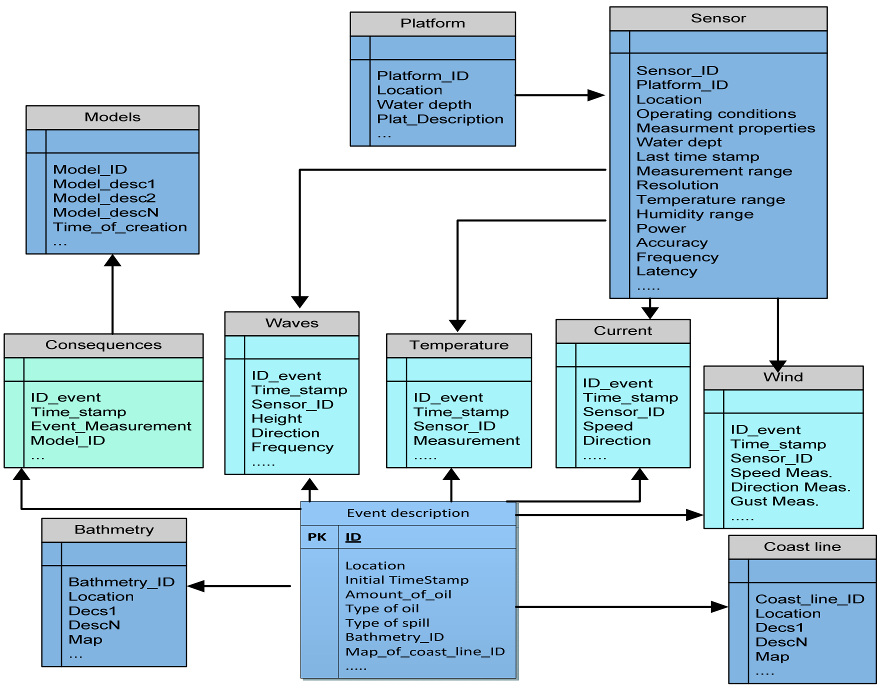


Figure 2: *Oil spill Data Model - adaptation of the general data model (Figure 1).*

Observed events and the corresponding data include initial characteristics of the oils spills, weather forecast data and to some extent dynamic data stream. Location(longitude, latitude), user input (amount, type of spill, oil type), bathymetry - map of the see-ground, coast line map,weather forecast for selected time frame, model outputs. Historical data about past oil spills was not systematically collected and stored, so therefore this data has limited availability. Currently available data does also not include information about the progress of the spill, e.g. data about the drift of the oil spill (at least not in a form required for stream mining analysis, where the data about prediction and "real" situation must be available for several selected time-stamps). This means that it is not possible to compare the "real situation" to predicted situations (predicted oil drift) in a form required for stream analysis.

3.2 Landslide case study

The main goal in the landslide case study is to predict if the selected road (in Guadeloupe) has to be closed for the traffic due to an upcoming landslide. The event is defined as a question: how high is the probability that the landslide will damage the road? The process of calculating the alarm for road closure is composed of sequential composition of five prediction models [1] – the workflow. For all the predicting models the user must specify special input parameters related to certain conditions, such as type of soil, characteristic of the landslide which determines the speed of possible slide propagation or other influential factors for the predicting model in question. By using the existing modelling services the user can obtain a prediction whether the landslide will damage the road.

From the data perspective, the expert user has to define: input data correlated to location, static data sources (elevation model, the geological map, borehole data – data about geological structure, map of geotechnical formations, map of the selected road), dynamic data source correlated to the location and with respect to a specified period of time (precipitation data from the meteorological sensor systems). The user can then create and calibrate prediction models via additional input parameters. Experts empirically determine the suitable parameters, related to certain conditions such as the type of soil, or other influential factors for the predicting model in question. The workflow also provides additional data as intermediate results (prediction models, which are created during the workflow process and are also input for next prediction model): Geological model, Groundwater map model, Landslide probability model, Landslide hazard map, Risk map.

The data sources are divided into input data and output data – which represents results of prediction models in format of raster or vector maps, available also in txt formats (for each measurement unit, appropriate probability is calculated). Some historical information about landslides data is available. The additional available data includes past landslides in Guadeloupe area (more than 50 landslides), described by time stamp, type of slide, how big is the slide impact, damage, etc. Additional data source include the weather sensor data from Guadeloupe area.

We have adjusted the general data model provided in Figure 1 to the specifics of this case study in a similar way as for the Oil spill (omitted due to space restrictions).

6 CONCLUSION AND FUTURE WORK

We have proposed a general framework for environmental stream mining and describe its adaptation to two real-world environmental case studies. Furthermore, we propose two possible scenarios of the framework usage which can be applied in various environmental situations: definition of alarm trigger and analysis of the observed phenomena in historical context. The domain experts found the scenarios potentially useful for the two case studies.

Future work involves implementation and testing of the proposed framework on publicly available environmental data, as well as on real-world data from the two case studies.

7 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and ENVISION (IST-2009-249120).

References

[1] N.R. Bodsberg, U. Bronner, H. Kobayashi, J. Langlois, D. Roman, G. Athanasopoulos; Envision Deliverable 1.1: Report presenting definition of case studies.

[2] J. Gama, M. M. Gaber (eds.), Learning from Data Streams: Processing Techniques in Sensor Networks, Springer Verlag, 2007.

[3] D. Mladenić, M. Grobelnik. 2005, Knowledge Discovery for Ontology construction, In Davies, Studer, Warren (eds.), Semantic web technologies: trends and research in ontology-based systems, Chichester: John Wiley & Sons, 2006, 9-27.

[4] Open Geospatial Consortium Inc. (OGC): OpenGIS Sensor Observation Service (SOS) Implementation Specification. A. Na, and M. Priest (eds.). Available at http://portal.opengeospatial.org/files/?artifact_id=26667

[5] Open Geospatial Consortium Inc. (OGC): OpenGIS Sensor Model Language (SensorML) Implementation Specification, M. Botts and A. Robin (eds.). Available at http://portal.opengeospatial.org/files/?artifact_id=21273

[6] D. Roman, S. Schade, A. J. Berre, N. R. Bodsberg, J. Langlois, 2009, Environmental Services Infrastructure with Ontologies –A Decision Support Framework, In Proceedings of EnviroInfo Conference 2009, Berlin, Germany.

[7] World Wide Web Consortium (W3C): Web Services Description Language (WSDL). E. Christensen, F. Curbera, G. Meredith, S. Weerawarana (eds.).

USER PROFILING BASED ON MOUSE MOVEMENT

Lorand Dali
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 477 3144;
e-mail: lorand.dali@ijs.si

ABSTRACT

The paper presents an approach to user profiling based on the user's mouse activity. The hypothesis which we try to verify in this work is that everybody uses the mouse in a specific way, and therefore a user model can be learned from the mouse activity. The aim of the user model is to recognize who a user is, given the way he uses the mouse. The data, collected from 10 users, consists of Windows events which were fired as a result of mouse activity. The described user profiling could be applied in security systems and for personalization.

1 INTRODUCTION

Authentication is a very important service for the security of a computer system. Many authentication methods such as passwords, fingerprints, iris recognition, face recognition, voice recognition have been used. We propose an authentication method based on mouse activity. The work builds on the assumption that usage of the mouse is specific to individual users. The advantage in using the mouse for authentication is that data is plentiful and cheap to collect and analyse. Also, mouse movement is harder to fake than a password.

The remainder of the paper is organized as follows. Section 2 describes the data collection and preprocessing, Section 3 talks about the experiments and the evaluation, and Section 4 draws the conclusions.

2 DATA COLLECTION AND PREPROCESSING

The data analyzed consists of events triggered by mouse usage on a Windows system. To help collect this data, ten users agreed to track their mouse over the time period of about a week. The users (4 female and 6 male) will be refered to with the fictional names of: Ana, Brian, Claudio, Dorina, Elsa, Flavia, Gerard, Holger, Iain and Jeffrey. During the mouse tracking, each event triggered by the mouse was recorded together with the following attributes:

- **Event Type.** Possible event types are: Move, LeftButtonUp, LeftButtonDown, RightButtonUp, RightButtonDown and MouseWheel.
- **Mouse Position.** X and Y coordinates of the mouse position on the screen at the moment when the event was triggered
- **Timestamp.** The time (in milliseconds) when the mouse event occurred.

Thus the raw data of about 800 000 events per user was collected. In what follows, the preprocessing steps applied to this data will be described.

2.1 Dividing into Gestures

By gesture I mean a sequence of events which happen close to eachother in time. A gesture ends when the user makes a break longer than one second between two successive mouse events. For each user we obtain a number of gestures somewhere between 5000 and 10 000.

2.2 Annotation of High Level Events

After segmenting the data into gestures, we annotate each gesture with higher level events such as: left click, right click, double click, movement, scroll, drag and drop. The higher level events are semantically more meaningful. The annotation is done based on a few simple rules.

Table 1 Annotation Rules

Annotation	Rule
LeftClick	LeftButtonDown → LeftButtonUp
RightClick	RightButtonDown → RightButtonUp
DoubleClick	LeftClick → LeftClick
Movement	Move → Move → … → Move
Scroll	MouseWheel → … → MouseWheel
DragNDrop	LeftButtonDown → Movement → LeftButtonUp

2.3 Approximating the Path of Mouse Movement by Line Segments

Each Movement as well as DragNDrop event is composed of a sequence of mouse moves (i.e. a sequence of points on the screen). The movement path made of a sequence of points is approximated by a line segments. An important observation is that linear regression, the usual way of fitting a line to a set of points cannot be used in this case because we have the points as a sequence in time, not as a set, and because of this the direction of the line we fit is important. In absence of a standard method to approximate the path by line segments, a simple algorithm was implemented.

```
function APPROX-PATH(points)
returns: lines
lines ← []
for i in [1,3,5,…n] do
   line ← MakeLine(points[i], points[i+1])
   lines.push(line)
end for
while not done do
   forall angle in angles(lines) do
       if angle < 30° do
           MakeLine(angle.prevLine,
                    angle.nextLine)
       end if
   end forall
end while
```

First between each pair of successive points we draw a line, then as long as there are angles smaller than 30° we join the two lines together into a single line. Figure 1 shows an example of a segment approximation of a path. The red dots show mouse positions on the path.

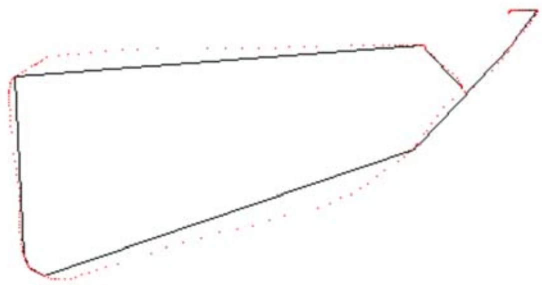


Figure 1 Line segment approximation of a path

Every line segment thus obtained can be described by three parameters: the length, the direction (i.e the angle between the segment and the OX axis), and the speed with which the user moved the mouse on that segment. Only the lines with length at least 20 pixels are taken into account.

2.4 Discretizing Line Segment Parameters

Each of the segment parameters (length, angle and speed) is a continuous value which is discretized. The angle is discretized into 12 values, each of the 12 values covering an angle of 30°, thus we obtain angles of (0° - 30°, 30° - 60° etc.). The segment lengths are discretized into 5 values (very short, short, medium, long, very long). The speed values are also discretized into 5 (very slow, slow, medium, long, very long). The thresholds for discretizing length and speed values are obtained by looking at the entire training data. For instance the threshold for very short is the length of the line which is longer than 20% of the lines.

3 EXPERIMENTS

The experiments try to find out ways of recognizing the users. This means that given some mouse events, can we determine to which user these events belong? We look at this problem as a multi-class classification problem, in our case we have 10 classes (the 10 users).

The data is divided into training data and test data. The training data consists of 70% of the gestures from each user. The test data is made of the remainder of 30% gestures from each user. The gestures are taken ordered by time and the ones from the test data are the gestures generated last.

3.1 Mouse Activity Maps

Because for each event the position is known where the mouse was when the event was triggered, we can draw activity maps to discover the areas with most activity.

Figure 2 and Figure 3 show move maps of two users, Flavia and Ana. The differences can be noticed quite easily. Flavia has a lot of movement on the right of the screen, while Ana uses the bottom part more and the right almost at all. Interesting features like the task bar, window title bar, start button, minimize and close buttons can be recognized. In a discussion with Ana she explained that she has widgets which she very rarely uses on the right part of the screen; this is why this part appears more white than the rest. Flavia also confirmed that her task bar and start button are on the right instead of at the bottom.

3.2 Mouse Movement Models

As described in the section about data preprocessing, the mouse movement is approximated by a sequence of segments. Each segment has a length, an angle (direction) and a speed. Based on these three features and of the time-dependency of the segments, user models can be built.

A user model is a vector whose entries are probabilities of the segment attributes taking certain values. For example, an entry could be the probability that for the user Flavia a segment has the length 'very short' and the speed 'fast'. Another entry could be the probability that the angle of a segment is between 30° and 60°. The complete list of features whose probabilities are computed is:

- length
- angle
- speed
- length + angle
- length + speed
- angle + speed
- length + angle + speed

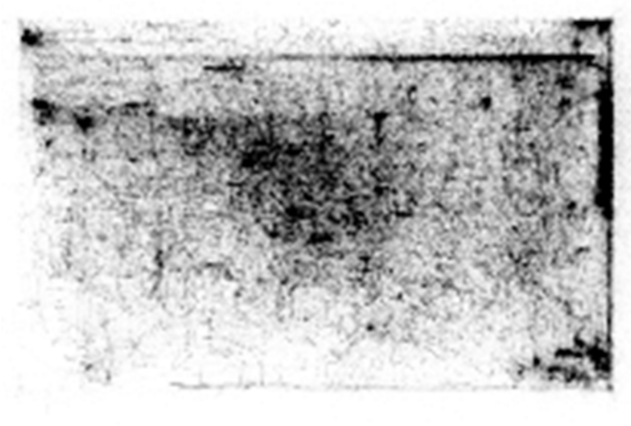


Figure 2 Flavia's Move map

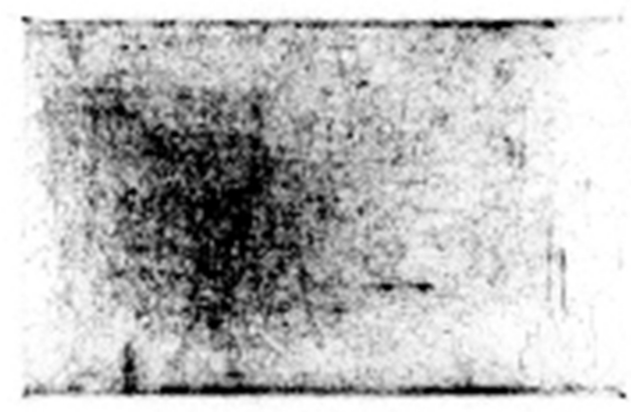


Figure 3 Ana's Move map

Moreover, the values of the previous segments are also taken into account. For instance we could have as a feature the probability that the current segment length is 'long' and the previous segment length is 'short'. For length and for speed the values of up to 5 previus segments are taken into account, for angle up to 3, for length+angle, length+speed and for angle+speed only the previous 2 and for length+angle+speed no previous segment is taken into account as that would make the feature vectors very sparse.

Having a user model expressed as a vector, we can compute the distance to other users. By finding the closest user to each user the directed graph in Figure 4 is obtained. There are two connected components, one of which has mostly male users (white nodes). In the other connected component the ratio between male users and female users is equal. An observation to make is that if user B is the closest to user A it is not true in general that also user A is closest to user B.

The experiments consist of computing a model from the training data for each user. Then, from the test data of each user we compute several test models. The test model is classified by finding the training model closest to it. From the test data of each user 100 sequences of segments are sampled. For each of these samples is classified and then the accuracy is computed.

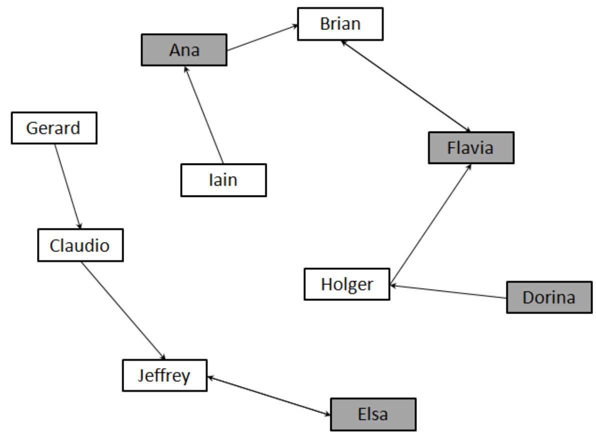


Figure 4 All users and the smallest distances between them

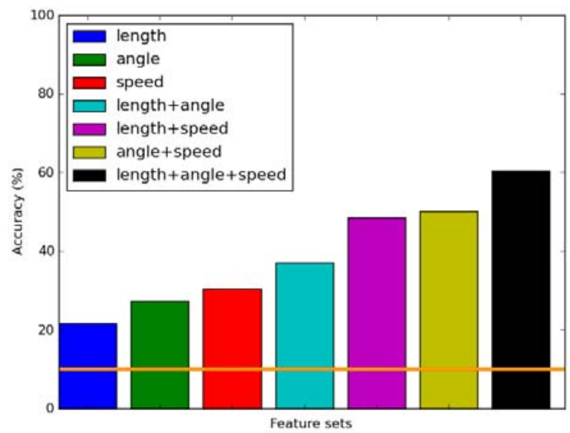


Figure 5 Accuracy of different feature sets

We try to find out which features are most helpful for the classification. In Figure 5 the accuracy of models based on seven different time independent feature sets are shown. The model which takes into account only the length of a single segment performs worst. It has an accuracy of about 20%. The best accuracy of about 60% can be obtained by considering the joint probability of length, angle and speed. For each of the test samples a sequence of 100 segments was used.

Having found that all three attributes of a segment have to be used for accurate classification, two important questions remain still open. How many segments per sample are enough? Can we improve the accuracy by taking time dependency into account?

To find the answers to the first question we have varied the number of sequences in a sample from 10 to 1000. Two feature sets are considered: one of length+angle+speed without N-grams, and the other taking into account all features with N-grams. Figure 6 shows that the model which does not take N-grams into account has an accuracy

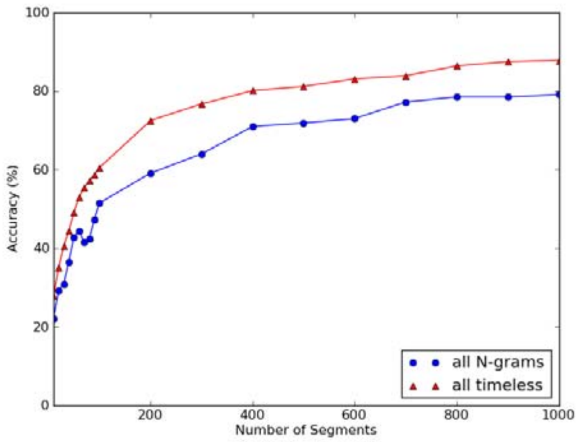


Figure 6 Increase of accuracy with the number of segments in the test models

of about about 10% better than the other. Another thing we notice is that the accuracy increases as the number of segments increase. Until around 200 the accuracy increases fast after which it increases at quite a small rate.

We have also noticed that the classification accuracy varies a lot from one user to another. For instance the accuracy of classifying data from Gerard at 400 segments is 93% while for Flavia only 70%.

4 CONCLUSIONS

We have presented a couple of methods for analyzing data obtained from mouse events produced by the activity of 10 users. We have foused mainly on move events. The experimental results show that the user which produced given mouse data can be determined with high accuracy. For this, all parameters of segments (length, angle, speed) should be taken into account and at least 200 segments are necessary to determine the correct user reliably. Surprisingly time features did not help in the classification but ‘confused’ it instead.

For the future, we plan to extend the user models with other features aside from movement. Also segments of smaller length could prove to be important, and a finer grained discretisation of length, angle and speed values might be necessary.

HAND GEOMETRY PERSONAL AUTHENTICATION USING MACHINE LEARNING

Mario Karlovčec
Faculty of Organization and Informatics, University of Zagreb, Croatia
Tel: +385 91 1505811
e-mail: mario.karlovcec@gmail.com

ABSTRACT

Personal authentication system based on hand geometry is proposed. We propose to perform identification and verification based on machine learning methods, in particular K-Nearest-Neighbor. Experimental results on verification show advantages of using machine learning over using the distance function

1 INTRODUCTION

This paper present biometric system based on hand geometry. Hand geometry includes measuring parts of hand like length and widths of fingers, width of palm, etc. Hand geometry is a non-invasive biometry, meaning that it is suitable for middle and low security requirements, because it's suitable for verifying someone's identity, but with a big database it cannot identify he's identity. [1]

The papers shows hand based biometry with two different authentication method: (1) based on machine learning, and (2) distance functions. Taking into account definitions form (1) and [2], we propose the following definition: Biometric is a science which includes procedures and technologies for automated person authentication based on physical, behavioral and chemical characteristics. Complete biometric system based on hand geometry is developed so different authentication methods can be properly tested. Biometric system is: "Multiple individual components (such as sensor, matching algorithm, and result display) that combine to make a fully operational system. A biometric system is an automated system capable of:

1. Capturing a biometric sample from an end user
2. Extracting and processing the biometric data from that sample
3. Storing the extracted information in a database
4. Comparing the biometric data with data contained in one or more reference references
5. Deciding how well they match and indicating whether or not an identification or verification of identity has been achieved." [3]

The sensor in the proposed system is standard document scanner. There is an application developed which uses algorithm for extracting features from acquired image of the hand, as well as matching algorithms for making the decisions about authentication.

2 PROPOSED APPROACH

The proposed system takes hand geometry sample using standard document scanner. To enroll a person to the database, 4 samples of hand geometry of that person's right hand must be acquired. This biometric template is stored in to the database together with personal information, like first and last name, including the persons age and sex, for possible future analysis. The hand geometry sample stored in database is a feature vector composed of 48 ordered pairs which give information for location of point on two dimensional plane:

$$F = ((x_0, y_0), (x_1, y_1), \dots, (x_{47}, y_{47}))$$

Locations of these points are extracted using the feature extraction algorithm described in the next section. When person wants to verify its identity, new scanning of the hand is required, so there is a sample of hand geometry which is compared with samples from the database. Verification (where "live" taken samples is compared with samples from database which are assigned to claimed identity) and identification (where all samples from database are taken into account), are performed with two different algorithms described in section 4.

Figure 1*: Biometric system based on hand geometry.*

The complete process is shown with scheme on figure 2. With image acquisition we get hand image as an input. Acquired image is subjected to preprocessing, and then final processing where features are extracted, resulting with feature vector. Feature vector is an input for last step – decision making, which is performed with two alternative methods: distance functions and machine learning. Decision making produces result of verification, i.e. identification of person whose hand image was acquired as an input.

3 FEATURE EXTRACTION

Image of hand can be acquired using different devices. Acquired image is first preprocessed, and then processed with extracted features as a result.

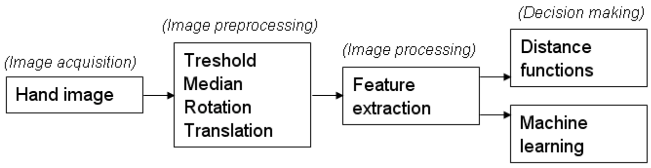


Figure 2*: Schematic view of proposed biometric system.*

3.1 Image acquisition

Methods for image acquisition are differing by type of used device and by required restrictions in placing hand on the device. Most often used devices for prototypes are CCD digital cameras and document scanner. Some systems for image acquisition use pegs for positioning the hand (like the one in [4]), other don't include the pegs for positioning (one of the examples is shown in [5]). The systems without the pegs can be completely unrestricted, allowing user to place the hand in any way on the scanning surface. More often there are some restrictions, like: fingers have to be separated, wrist has to be on the scanning device, and orientation of the placed hand has to be as vertical as possible. Advantage of system with the peg is that after scanning, the location of hand is known what makes the following image processing steps easier. The disadvantages are lower acceptance rate from users (pegs can cause difficulties in placing smaller hands), deformations of hand caused by pegs, the pages have to be removed from the image, the rate of misplaced hands is higher that with systems without pegs [1].
In proposed system image acquisition device is standard document scanner (figure 1). There are no pegs for positioning, but there is a requirement that the fingers are separated while scanning.

3.2 Image preprocessing

Threshold is applied on input image to gain binary silhouette of the hand. Noise is removed from the image with median filter.

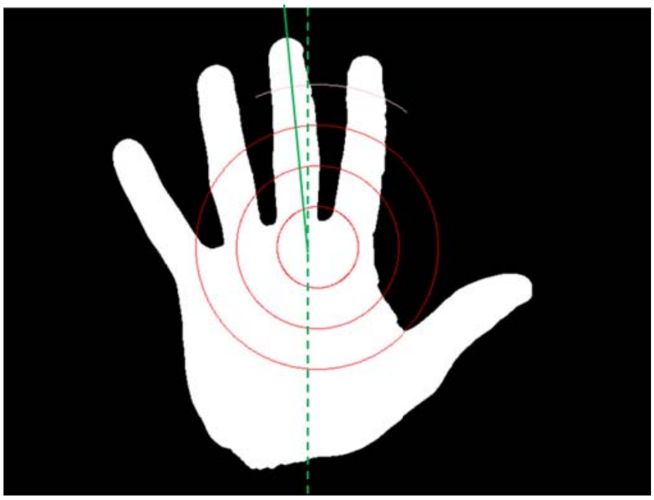


Figure 3*: Rotation of hand silhouette.*

As no pegs where used while acquiring image, we have to identify the location and orientation of the hand. The center of the image, concentric circles are grown until the middle finger and the pointer finger are identified (figure 3). When position of these two fingers is known, the whole hand is rotated so that the orientation of middle finger (which is used as a reference for orientation of whole the hand) is vertical. Using horizontal translation, hand is centered in the middle of the image.

3.3 Points extraction

Preprocessed image and identified locations of points on the edges of middle finger and pointer finger are used to identify all other points of the hand. This is done using basic geometric functions combined with previously know information. Figure 4 shows an example how points on ring finger are identified using previously known points on middle finger.

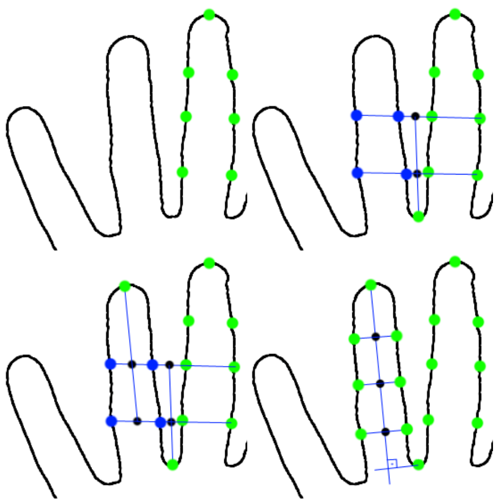


Figure 4: *Extraction of points on edges of ring finger.*

After the points extraction is done there are 48 points identified on the hand. Locations of these points make a feature vector which is stored in the database and used in latter steps for verification and identification.

4 AUTHENTICATION

There are two main categories of personal authentication, these are: identification and verification of person's identity. Identification is problem of establishing an individual's identity, so it answers the question: "Who is he?". Verification is a problem of confirming or denying an individual's claimed identity, so it answers the question: "Is he the one who he claims to be?".

4.1 Machine learning

For each person recorded in the database, we have four hand geometry samples where every sample has 48 points with given information of their location on two dimensional plain. When authentication of a person is performed, the new hand geometry sample is compared to those stored in the database. In the proposed system K-Nearest-Neighbor method is applied. This method is used in the way that for every of the 48 point on the hand, four nearest neighbors are identified (figure 5).
When performing verification the samples from the database are categorized as belonging to claimed identity or not. That means we have only two classes when we do verification. If for some point are among four nearest neighbor more of those belonging to claimed identity, than those which belong to other class (in which are all points belonging to different identity), verification for that person will be successful. Otherwise verification will be unsuccessful.

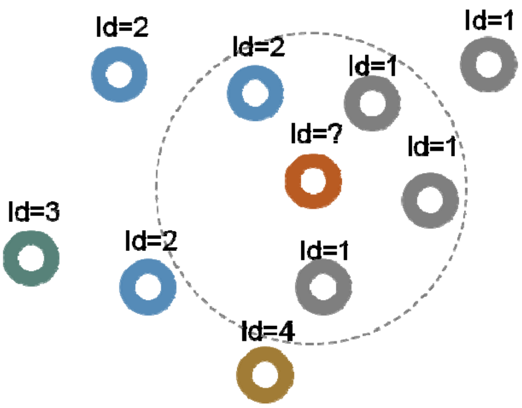


Figure 5*: Identifying 4 nearest neighbors.*

It is noticeable that even when we do verification, we do not perform one-to-one comparison (i.e. we don't compare "live" taken sample with only one sample from database), but make one-to-all comparison like when we perform identification. If we do not make categorization of point into two classes (one for point belonging to claimed identity and for points of all other identities), but keep every identity as a separate class, then with the same amount of work as for verification, we perform identification. For identification we count how many points of every class is among four nearest neighbors and classify the point to the class which has the biggest number of representatives.
Proposed hand geometry template has 48 characteristics, what means that the same K-Nearest-Neighbor method is performed 48 times (once for each position of the hand). Identification is made by number of points classified to each class, so the classification of the whole hand geometry sample is to the class to which the most of its point belong to. Classification of the whole sample can also be made on the "global level". That means that first, for each class is recorded how many points belonging to it, were among four nearest neighbors for any position of the hand. The classification of the sample is then made to the class with biggest number of these. Also, it can be recorded how many times was point of a class first, second third and fourth nearest neighbor, and put a weight making nearer neighbor more important. Method used in testing was "global level" classification without weighting.

4.2 Distance functions

Besides machine learning, distance functions were used for verification. The result of verification using these standard methods can be useful for comparing.
Absolute distance:

$$AD = \sum_{k=0}^{n} |y_i - f_i| \qquad \#$$

Weighted absolute distance:

$$WAD = \sum_{k=0}^{n} \frac{|y_i - f_i|}{\sigma_i} \qquad \#$$

Euclidean distance:

$$ED = \sqrt{\sum_{k=0}^{n} (y_i - f_i)^2} \qquad \#$$

Weighted Euclidean distance:

$$WED = \sqrt{\sum_{k=0}^{n} \frac{(y_i - f_i)^2}{\sigma_i^2}} \qquad \#$$

D1 distance [5]:

$$D1 = \sum_{k=0}^{n} \frac{|y_i - f_i|}{y_i + f_i}$$

5 EXPERIMENTAL RESULTS

The testing was done on a small sample of 10 persons with five images of hand for each, which is good enough only as initial orientation for future work. Results of verification are shown in table 1. Results are shown in CER (central error rate), which show the point where FAR (false acceptance rate) and FRR (false rejection rate) meet.

Method	CER (%)
K-Nearest-Neighbor	2,13
Absolute Distance (AD)	3,33
Weighted absolute distance (WAD)	5,88
Euclidean distance (ED)	5,88
Weighted Euclidean distance (WED)	5,88
D1 distance (D1)	5,88

Table 1: *Results of testing.*

The K-Nearest-Neighbor method outperforms the distance functions. Shown results for K-Nearest-Neighbor method are for identification, which is generalization of verification for this method. Results shown for distance functions are the results of verification.
The performance of K-Nearest-Neighbor method comes with very big cost of time. The algorithm used for verification does not have better complexity than the one used for identification, what is significant disadvantage of this method.

6 DISCUSSION

In this work K-Nearest-Neighbor method is used for verification, i.e. identification. Good performances as well as important disadvantages were identified. This makes good basis for finding a way of increasing the performance of algorithms from aspect of complexity. New and broader testing is required to get more reliable results and to identify the point where hand geometry stops being suitable for identification, as for this biometry is already well determent. As a future work analysis of testing results on particular individual samples could be useful in finding positive effect of merging verification methods. The feature extraction part of the work requires a simulation model to test the performances of this part of a system.

7 ACKNOWLEDGEMENTS

This work was partially performed during the author's stay at J. Stefan Institute in Ljubljana and Faculty of Computer and Information Science, University of Ljubljana, supported by the Slovenian Research Agency and the ERASMUS student exchange programme.

References

[1] Buratov, Yaroslav and others. Hand recognition using geometric classifiers.

[2] Jain, Anil K, Flynn, Patrick i Ross, A Arun. Handbook of Biometrics. Springer. New York. 2008.

[3] L.Wayman, James. A Definition of "Biometrics". San Jose State University. pp. 21-24. San Jose. 2000.

[4] NSTC. Biometrics.gov. "Foundation Documents". [Intrernet]. www.biometrics.gov/Documents/biofoundationdocs.pdf. 7. 8 2006.

[5] Jain, Anil K, Ross, Arun i Pankanti, Sharath. A Prototype Hand Geometry-based Verification System. *2nd Int'l Conference on Audio- and Video-based Biometric Person Authentication (AVBPA)*. pp. 166-171. Washington. 1999.

[6] Covavisaruch, Nongluk, and others. Personal Verification and Identification Using Hand Geometry., ECTI *Transaction on Computer and Information Technology*, pp. 134-140. 2005.

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Sodelovanje, programska oprema in storitve v informacijski družbi

Collaboration, Software and Services in Information Society

Uredil / Edited by

Marjan Heričko

http://is.ijs.si

13. oktober 2010 / October 13th, 2010
Ljubljana, Slovenia

PREFACE

The transition into an information and knowledge society is being accompanied by a redefinition of business processes and a renewal of information solutions and services. There is significant growth in the size and complexity of these systems. Therefore, the main challenge is how to successfully manage and control the complexity of software and services in order to enable efficient collaboration between individuals, organizations and/or systems. The CSS 2010 proceedings bring together papers that address different aspects of software and service engineering, including collaborative communication in multi-organizational and virtual environments, methods and techniques that could be beneficial for enhancing communication and interoperability between participant organizations and actors, knowledge management environments, value co-creation in electronic marketplaces and service ecosystems, e-learning and e-university projects and solutions, advanced techniques for efficient requirements elicitation, model-driven user interface development, service metrics and software metric gathering tools as well as semantic web technologies, mobile platforms, cloud computing and Future Internet challenges.

Special thanks go to the members of the Program Committee and additional reviewers as well as to the authors who have contributed to the success of this sub-conference. We hope that these proceedings will be beneficial for your reference and that the information in this volume will be useful for further advancements in research and industry.

prof. dr. Marjan Heričko
CSS 2010 – Collaboration, Software and Services in Information Society Conference Chairman

PROGRAMSKI ODBOR / PROGRAM COMITTEE

Dr. Marjan Heričko
University of Maribor, Faculty of Electrical Engineering and Computer Science

Dr. Ivan Rozman
University of Maribor, Faculty of Electrical Engineering and Computer Science

Dr. Kimmo Salmenjoki
Seinäjoki University of Applied Sciences

Dr. Lorna Uden
Staffordshire University, Faculty of Computing, Engineering and Technology

Dr. Gabriele Gianini
University of Milano, Faculty of Mathematical, Physical and Natural Sciences

Dr. Hannu Jaakkola
Tampere University of Technology Information Technology (Pori)

Dr. Mirjana Ivanović
Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad

Dr. Zoltán Porkoláb
Eötvös Loránd University, Faculty of Informatics

Dr. Gregor Polančič
University of Maribor, Faculty of Electrical Engineering and Computer Science

Dr. Aleš Živkovič
University of Maribor, Faculty of Electrical Engineering and Computer Science

Dodatni recenzenti / Additional Reviewers

Dr. Vili Podgorelec, Dr. Luka Pavlič
University of Maribor, Faculty of Electrical Engineering and Computer Science

V sodelovanju z / In collaboration with

nessislovenia
Technology Platform for Software and Services

INVESTIGATING TECHNOLOGICAL SOLUTIONS FOR ENHANCHING COLLABORATIVE COMMUNICATION IN A MULTIORGANIZATIONAL ENVIRONMENT

Jari Soini, Jari Leppäniemi, Hannu Jaakkola
Department of Software Engineering
Tampere University of Technology - Pori
P.O.Box 300, FIN-28101 Pori, Finland
Tel: +358 2 6272890; fax: +358 2 6272727
e-mail: jari.o.soini@tut.fi

ABSTRACT

In an emergency situation, different organizations become involved and have to work cooperatively in the disaster response. This type of situation requires close collaboration between different authorities and also the optimized, integrated use of management systems and resources. The high interoperability of information systems and also knowledge of the function of each of the other parties could be seen as a prerequisite for seamless communication and an enabler for successful joint operation. However, in practice, sharing information and communication in large-scale emergencies remains a major challenge. These recognized difficulties relate to deficient communication and coordination among organizations. Consequently, a two-year research project (SSMC/DDKM) was initiated at the beginning of 2008. The research focused on conceivable methods and techniques that could be beneficial for enhancing communication and interoperability between participant organizations in emergency situations. This paper briefly presents the selected research themes and the key results.

1 INTRODUCTION

In disaster situations, the need has been seen for methods and technologies which can integrate communication in order to build a collaborative understanding of what is happening and who is doing what [1, 2]. It is common that in disasters there are multiple participating responders that have to collaborate and coordinate their response actions and manage the situation together. Comprehensive knowledge concerning the situation and know-how of the right way to react to disaster situations is critically important when minimizing the consequences of a disaster.

However, sharing information and communication in large-scale emergencies remains a major challenge [3, 4]. The recognized difficulties relate mainly to deficient communication and coordination among organizations [3, 4, 5]. Previous studies have confirmed that the reasons why things go wrong in emergency management in large-scale emergencies are generally related precisely to breakdowns in information, communication, and coordination [3, 6]. Moreover, the incompatibility of the legacy systems of the authorities, as well as the low level of interoperability between them, do not support the beneficial use of existing data, which is scattered in different information systems. Even when technically separate systems are networked with each other, their ability to benefit from the data in other systems is deficient. The main technological challenges seem to relate specifically to the integration of systems, data models, and applications [7]. To address these challenges in the emergency context, the Seamless Services and Mobile Connectivity in Distributed Disaster Knowledge Management (SSMC/DDKM) [8] project was initiated to find solutions for these problems. The context of the research project was the advanced usage of situational knowledge in connection with disasters and major accidents. The research focused on conceivable methods and techniques that could be beneficial for enhancing communication and interoperability between participant organizations in emergency situations. This paper briefly summarizes the key research results - investigated and applied methods and technologies - of the research work that was carried out.

2 SSMC/DDKM PROJECT FOR STUDYING COMMUNICATION ISSUES

2.1 Goals of the Project

The research problem was approached by investigating current collaboration in accident and disaster situations between the Finnish authorities that are responsible for producing, disseminating and utilizing information in these situations. In this project, three main research topics were selected for investigation to solve the research problem. The specific research topics of the SSMC project were:

- recommendations of tools and methods to be used in process and knowledge modeling in a disaster management context (based on evaluation and practical tests). The aim was to study how well process modeling and a graphical description of the information flows between the different participants could be applied to the emergency response situation.

- specification of the principles of flexible Service Oriented Architecture (SOA) to support the loose networking of subsystems - an open reference architecture to enhance seamless co-operation and compatibility between systems and services. The aim was to study opportunities for defining a free and open source based reference architecture for a flexible service oriented system designed to support the loose connection of the information systems of different authorities.
- prototypes of new community type services for mobile phones to be applicable in a disaster context. The aim was to study, and also develop, a mobile-based service solution/system, enabling communication – the ability to produce and disseminate information in real time - in major accidents and disasters.

The uniting factor for all these research themes is the utilization of open source principles and existing open standards. For example, the mobile service system developed has been implemented using open source software based libraries and components, and is based on open and free software standards. Moreover, the process descriptions studied can be used as a basis for finding suitable services in SOA-based service provision (e.g. the greatest benefit of the investigated solution BPMN is its connectivity to an SOA environment).

The key results achieved and the final outcomes of the research are presented in Section 3.

2.2 Background of the Project and Participating Organizations

The project was initiated at the beginning of 2008 and coordinated by the Tampere University of Technology (TUT) [9], and supported with regard to disaster expert knowledge by the Emergency Services College (ESC) of Finland [10]. The starting point behind the SSMC/DDKM project was to take part in international emergency management research in the context of major accidents, disasters and natural catastrophes i.e. the project was a part of joint international project activity, in which all members of the consortium have their own projects and goals, with the common aim to co-operate and benefit from each other's research results.

During this project, *international level* research co-operation partners were NICT (National Research Institute on Information and Communication Technology) and Keio University (SFC) in Japan, the VSB Technical University in Ostrava (Czech Republic), the Christian Albrechts University in Kiel (Germany) and the University of Maribor /FERI (Slovenia). Project partners at *national level* were divided into three groups: Implementing organizations (the Tampere University of Technology (TUT) and the Emergency Services College (ESC)), Authority organizations (the Ministry of the Interior, the Prime Minister's Office, the Ministry of Justice, the Ministry for Foreign Affairs, the Emergency Response Centre Agency and the Police College of Finland), and Corporate partners (Nokia Corporation, DNA Ltd, Logica Ltd, EADS S.N. and Birdstep Technology Ltd).

The project budget was 548,000 EUR for 2008-2009; one third of the funding came from the participating organizations and two thirds from the Finnish Funding Agency for Technology and Innovation (Tekes) [11].

3 KEY RESEARCH RESULTS

Below the key results achieved on each main research topic are presented as defined at the beginning of the project (see sub-section 2.1).

3.1 Process and Knowledge Modeling in the Emergency Management Context

Formal or semi-formal models can be seen as a means towards universal communication. Models are a standardized way to specify a phenomenon (i.e. situation) in a (semi-) formal and universal manner. One of the main research topics of the project was in fact to study how process modeling can be utilized in the emergency management context. The goal was to study and develop the technology required (i.e. determine standardized description solutions) for modeling accident and disaster knowledge and related processes to support improved information and communication in disasters and accidents between response participants.

It is clear that success in interoperability and communication demands a form of activity that is carefully planned in advance [12]. A typical emergency response involves a complex flow of response tasks, resources, responder personnel, and information. The study focused on how collaboration could be implemented utilizing process modeling and a graphical description of the information flows between the various tasks or roles related to the emergency response. Business process models (BPM) [13] can be used with the help of visuals to describe complex links and tasks on many different levels and with the appropriate precision [14, 15]. In this study Business Process Modeling Notation (BPMN) [16] was chosen as an experimental tool for studying the practicality of modeling as an aid in emergency management because it is the recommended [17] graphical notation for process description in Finnish agencies.

During the project, several disaster scenarios in Finland were modeled beforehand by the process modeling method, using BPMN. Figure 1 below shows a notation-based description that was made in close co-operation with Finnish emergency authorities during the project. An example process diagram, describing the usage of a response plan (RP), was made based on interviews with emergency experts [18]. During the study literature sources (e.g. guides, operating procedures, manuals) were also used in order to become familiar with Finnish Emergency Response Center (ERC) processes in accident and disaster situations [19, 20].

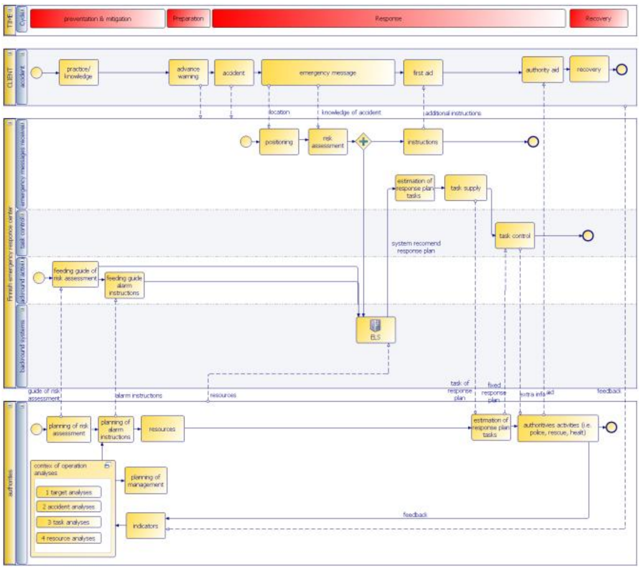


Figure 1: *An example process diagram of several tasks related to ERC response plans modeled using BPMN notation.*

The graphical description and links between the tasks and roles inside the organization and also between the participant organizations proved a useful tool for clarifying and enhancing the collaboration of authorities in the management of disaster situations. Based on the case study experiences, process modeling was found to provide the opportunity for analyzing the management of disasters systematically, studying the role of each actor and the courses of action taken in comparison with their counterparts. Moreover, process modeling was found to offer the chance to study the responsibilities, obligations, and decision-making power of various roles in the different situations described and also to consider the aims and capability of each role during the disaster. It seemed that the authorities considered the fact that describing a disaster enables each actor to assess their own role in relation to other actors as one of the more significant factors in process modeling in a multi-organizational situation. However, the organizational aspects must be taken into account when developing parallel and interconnected processes between different parties. The main challenges observed in the utilization of process modeling in this environment seemed mainly to be related to the current practices and behavioral norms existing in the various authority organizations. Also, the lack of common vocabulary and shared semantics between the response organizations has proven to be a challenge in emergency contexts [3].
To sum up, at least in this case process modeling with BPMN was found suitable for proactive preparation work, post-procedure evaluation, and as an aid for operational activities. No evidence appeared that would rule out the suitability of process modeling utilization in emergency management planning. There have also been other studies [21, 22] that have shown that emergency plans are very similar to business processes, supporting our findings that these plans can be modeled, and also that it is useful to model them, with the notations used for business modeling. Research results related to this issue are described in more detail in paper [23].

3.2 Domain-specific Service Oriented Reference Architecture for Emergency Management

Effective management of disaster situations requires information from many different sources and the decisions made should be based on the combination of several types of knowledge. The information needed is scattered in closed and open databases, registers and data sources, so that one must strive to create situational awareness through techniques of data search and data combination. This position was the starting point for designing the reference architecture for a flexible service oriented system to support the connection of the information systems of different authorities. One significant source and focal point in this work was the standardization work on SOA components being done in W3C [24].
Figure 2 below consists of the main components of the designed domain-specific service oriented reference architecture for emergency management. The main architectural layers, packages, and the possible dependencies between them are presented using UML notation. All packages can aggregate sub-packages and many services are usually composed of other services.

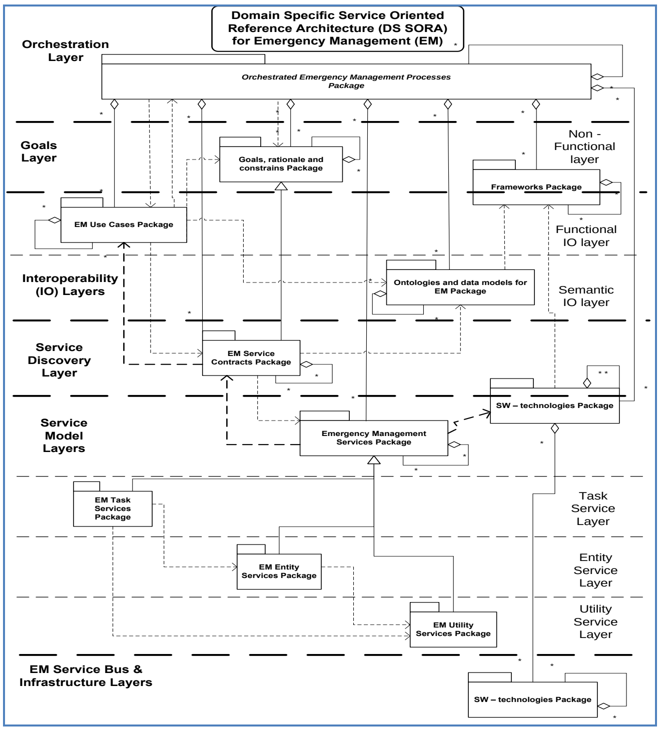


Figure 2: *Layers, packages, and dependencies of a domain-specific reference architecture for co-operation by independent stakeholders within an emergency management ecosystem.*

The packages and layers of the reference architecture that are presented in Figure 2 above are described in brief below. A more detailed explanation of this reference model has been given in [25].

Orchestrated emergency management processes. The modeled parts of the emergency management processes are abstracted for orchestration (and automation), and the supervision and monitoring of the processes are implemented by a distributed enterprise service bus (ESB).

Goals, rationale, and constraints. The goals, reasons for interoperability and constraints are located in the Goals layer. The packages in this layer comprise the functional and non-functional requirements that are used to articulate the goals and rationales within the constraints. The requirements are for both the architecture and the business process to be supported by the architecture. The packages are realized via service contracts (and ultimately services) but the Frameworks packages (described below) control the enunciation of the requirements.

Frameworks and design patterns. These packages are located in the two layers targeted for actual requirements and guidance information: 1) in the Goals layer - the Non-Functional layer - and 2) in the upper Interoperability (IO) layer, i.e. the Functional IO layer. These packages are composed of the selected interoperability frameworks, design patterns, and architecture principles and are dependent on the “Ontologies and data models” and “SW technologies” packages. Some examples of possible sources of these packages are: ATHENA framework for interoperability [26], IEC TC65/290/DC work [27], Orchestra architecture [28], OASIS service oriented reference architecture foundation and reference model [29], and the tactical situation object model which describes the concepts for situational awareness [30].

Use cases. The packages of use cases are located in the functional interoperability layer. They are an essential part of the articulation of the requirements and they are also an essential part of the logic for interoperable processes when they are orchestrated for control and monitoring.

Ontologies and data models. These packages are located in the semantic interoperability layer and are essential for achieving interoperability between stakeholders. The selected and/or developed ontologies and data models are based on the principles derived from the Frameworks package. Some examples of candidates for the ontology and data model for emergencies are Kruchten et al. [31] and the “tactical situation object” [30].

Service contracts. Service contracts are placed in the “Service Discovery Layer.” The services are expressed precisely in these packages and in such a way as to satisfy the feasible constraints and concerns of the stakeholders. The discoverability of the services, understandable functionality and also the functional and non-functional limits of the services provided are based on the service contracts.

Services. Services are positioned based on their types in corresponding layers. Maybe the simplest but still useful layering is based on ERL’s [32] division of different service types into task, entity, and utility services.

Software technologies. The chosen standards and SW technologies [e.g. 24, 30, 33] are dependent upon the services to be realized. In other words, the implementation possibilities of the derivate works of Frameworks are based on the availability and maturity of the SW technologies.

The concrete architectures will be based on the domain-specific (an emergency in this case) concepts and knowledge required, such as ontologies and data models, for a shared mission (goals) and co-operative work to strive toward that goal (described by use cases and abstracted business processes), the frameworks and design patterns for interoperability and non-functional requirements in such a way that the desired functionality (use cases, orchestrated processes) could be implemented (using SW technologies) via different types of services (task/entity/utility services), which are described and defined by service contracts according to the goals, rationales and constraints of the ecosystem. The uppermost layer is the “Orchestration Layer,” which comprises abstracted process packages that accumulate all the necessary knowledge of a domain-specific application of the reference architecture.

3.3 Prototype of a New Community-Type Service – Emergency Alert System

One aim of the project was to investigate and develop mobile solutions, which could help generate and distribute information and also communicate in disaster situations. The goal set was to develop an emergency communication system which would utilize IP-based networks – in practice the Internet. Using IP-based services in emergency communication is not a new idea, however, there are not many applications available aimed at the end users. The use of free and open software components was selected as the starting point for planning the system. The aim was to utilize existing technologies and standards in the system design.

The final outcome of the project work was an IP-based emergency alert system (client-server prototype application) for sending and receiving emergency messages. The mobile service system developed has been implemented using open source software based libraries and components, and is based on open and free software standards. This development work was made in close co-operation with a partner who is the major mobile phone technology developer in Finland (and also worldwide). The pilot version of the new system was released at the end of the project in December 2009.

The picture below (Figure 3) shows a generalized overview of the communication system that was developed. In essence, the purpose of the system is to relay messages. The system is used to relay emergency alert information, but the same basic model could be used for any kind of information.

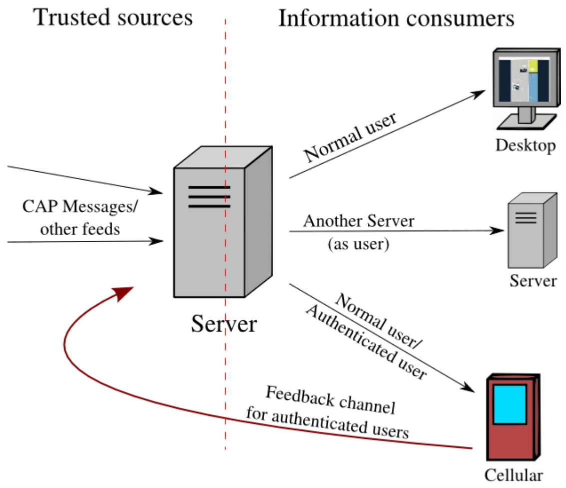


Figure 3: *Overview of the emergency alert system.*

Briefly, the communication system includes three main functions: collecting data, transmitting it to the server, and sending it to the client. The server enables the downloading and receiving of emergency alerts and emergency messages from many information sources and the server's main task is to collect, provide and relay different kinds of information feeds to client devices. The client application consists of two main system entities: a daemon application (software logic) and a graphical user interface, which are based on a Qt framework and Linux operating system. Figure 4 below shows an example of the use case view (maps search view).

Figure 4: *Google Maps map position opened by client application.*

The image above shows a Nokia N810 Internet Tablet running a browser launched by the client. If location information is available in the received feed, the client can show the location on the map. In this case, the client is using the system's default browser and Google Maps to display the map.

A more detailed description of the system itself, its operational principles, features and the technologies used are presented in [34, 35]. The development work of the system – generalization of the application in this case – is currently continuing in the ongoing follow-up project. The aim is to study how to extend the use of the implemented system and its applicability for different branches of industry.

4 CONCLUSION

This paper presents the results of a research project, in which we looked at conceivable and suitable methods and techniques that could be beneficial for enhancing communication and interoperability between participant organizations in multi-organizational environments. The SSMC/DDKM project focused on investigating emergency situations in which different organizations become involved and work cooperatively in disaster response. The paper presents briefly the results of the selected three different research approaches - system architecture (SOA), process modeling, and mobile technology - for solving the challenges observed in communication and interoperability in this kind of environment. All the presented solutions are based on open source approach principles and open standards that confirm the suitability and also the usability of these standards.

The research work is continuing with the "Positioning of Shared Services" project, by piloting the results presented in this study and in addition studying mechanisms for data retrieval and analysis to enable users to find the information that is relevant for them. The aim is to study and develop methods including semantic search methods to improve the formation of situational awareness in a disaster situation.

This follow-up project will be carried out mainly with the same project consortium as in the completed SSMC/DDKM project.

References

[1] B. van de Walle, M. Turoff. Decision support for emergency situations. Information Systems and e-Business Management, Vol. 6, No. 3, pp. 295–316, 2008.

[2] U. Ruppel, A. Wagenknecht. Improving emergency management by formal dynamic process-modeling. In Proceedings of the 24th Conference on Information Technology, pp. 559-564, 2007.

[3] B. S. Manoj, B. A. Hubenko. Communication challenges in emergency response. Communications of the ACM, Vol. 50, No. 3, pp. 51-53, 2007.

[4] L. Palen, S. R. Hiltz, S. B. Liu. Online forums supporting grassroots participation in emergency preparedness and response. Communications of the ACM, Vol. 50, Issue 3, 2007.

[5] A. Meissner, T. Luckenbach, T. Risse, T. Kirste, H. Kirchner. Design Challenges for an Integrated Disaster Management Communication and Information System. In papers presented at The First IEEE Workshop on Disaster Recovery Networks (DIREN 2002), 7 p., New York, 2002.

[6] E. Dykstra. Towards an International System Model in Emergency Management: information communication and coordination in emergency management- public and private sector approaches in different countries and systems. Communication. Public Entity Risk Institute (PERI), September 22-26, 2003.

[7] S. Mehrotra, C. Butts, D. Kalashnikov, N. Venkatasubramanian. CAMAS: A Citizen Awareness System for Crisis Mitigation. In Proceedings of the 2004 ACM SIGMOD international conference on Management of data. pp. 955-956. Paris. France. 2004.

[8] H. Jaakkola, J. Soini, J. Leppäniemi. Seamless Solutions and Mobile Connectivity – Towards location independent and flexible system architectures. Tampere University Press. In Publication No. 12. pp. 1-17. Finland. 2008.

[9] Tampere University of Technology (TUT), Retrieved 14/8/2010 World Wide Web, http://www.tut.fi/public

[10] Emergency Services College (ESC), Retrieved 14/8/2010 World Wide Web, http://www. pelastusopisto.fi/pelastus/home.nsf/pages/index_eng

[11] Finnish Funding Agency for Technology and Innovation (Tekes), Retrieved 14/8/2010 World Wide Web, http://www.tekes.fi/eng/

[12] P. C. van Fenema, B. Pentland, K. Kumar. Paradigm shifts in coordination theory. Academy of Management Annual Meeting. New Orleans. USA. 33 p., 2004.

[13] M. Treat. Business Process Management (BPM) - What is BPM Anyway?. Retrieved 11/9/2009 World Wide Web, http://www.bpminstitute.org/articles/article/ article/what-is-bpm-anyway.html

[14] C. Plesums. Introduction to Workflow. Workflow Handbook 2002. (Eds.) by Layna Fischer. Published by Future Strategies Inc. Lighthouse Point. FL. USA. 2002.

[15] E. Sivaraman, M. Kamath. On the Use of Petri Nets for Business Modeling. In proceeding of the 11th Annual Industrial Engineering Research Conference. 7 p., Orlando. USA. 2002.

[16] S. White. Introduction to BPMN. Retrieved 16/4/2010 World Wide Web, http://www.bpmn.org/ Documents/Introduction to BPMN.pdf

[17] JHS 152 - Process modeling. Retrieved 11/5/09 World Wide Web, http://www.jhs-suositukset.fi/suomi/jhs152

[18] Emergency Response Centers' authorities. Pori and Kuopio. Finland. Interviews in autumn 2008.

[19] M. Castrén, S. Ekman, M. Martikainen, T. Sahi, J. Söder. Guide of catastrophe. Jyväskylä. Finland. 2006.

[20] T. Raivio, Y. Gilbert, H. Lonka, M. Hjelt. Estimation of Emergency Response Center. Gaia Consulting Oy. Helsinki. Finland. 2007.

[21] C. Shell, I. Braun. Using a Workflow Management System to Manage Emergency Plans. In proceedings of the International Conference on Information Systems For Crisis Response and Management (ISCRAM'09). May 10-13. Gothenburg. Sweden. 2009.

[22] M. Zalewski, P. Sztandera, M. Ludzia, M. Modeling and Analyzing Disaster Recovery Plans as Business. Springer Berlin. Vol. 5219. Heidelberg. 2008.

[23] P. Linna, J. Soini, J. Leppäniemi, H. Jaakkola. Harmonizing emergency management knowledge representation. In proceedings of International Conference on Management of Engineering and Technology (PICMET'09). pp. 1047-1051. Portland. Oregon. 2009.

[24] The World Wide Web Consortium (W3C), Retrieved 14/8/2010 World Wide Web, http://www.w3.org/

[25] J. Leppäniemi. Domain Specific Service Oriented Reference Architecture (Case: Distributed Disasters and Emergency Management). Presented in IADIS International Conference Collaborative Technologies 2010. July 26-28. Freiburg. Germany. 2010.

[26] ATHENA Interoperability Framework. Retrieved 14/02/10 World Wide Web, http://www.modelbased.net/mdi/index.html

[27] ISO IEC, Retrieved 03/02/10 World Wide Web, http://www.iso.org/iso/home.htm

[28] ORCHESTRA, Retrieved 03/02/10 World Wide Web, http://www.eu-orchestra.org/

[29] OASIS, Reference Architecture for Service Oriented Architecture. Retrieved 09/02/10 World Wide Web, http://docs.oasis-open.org/soa-ra/v1.0/soa-ra-cd-02.pdf

[30] OASIS, Reference Model for Service Oriented Architecture 1.0. Retrieved 09/02/10 World Wide Web, http://docs.oasis-open.org/soa-rm/v1.0/soa-rm.html

[31] P. Kruchten, C. Woo, M. Sotoodeh, K. Monu. A human centered conceptual model of disasters affecting critical infrastructures. In Proceedings of the 4th International Conference on Information Systems For Crisis Response and Management (ISCRAM'07). March 23th. Vancouver. 2007.

[32] T. Erl. Service-Oriented Architecture: Concepts, Technology, and Design. Prentice Hall PTR. 2005.

[33] The Internet Engineering Task Force (IETF), Retrieved 14/02/10 World Wide Web, http://www.ietf.org/

[34] P. Sillberg, P. Rantanen, M. Saari, J. Leppäniemi, J. Soini, H. Jaakkola. Towards an IP Based Alert Message Delivery System. In proceedings of the International Conference on Information Systems For Crisis Response and Management (ISCRAM'09). May 10-13. Gothenburg. Sweden. 2009.

[35] P. Rantanen, P. Sillberg, H. Jaakkola, T. Nakanishi. An Asynchronous Message-based Knowledge Communication on Ubiquitous Environment. Presented in International Workshop on Ubiquitous Data Management (UDM2010). April 4th. Tsukuba. Japan. 2010.

“PROJECT ROOM” - A TAILOR-MADE COLLABORATION ENVIRONMENT

Damjan Obal
Faculty for Electrical Engineering and Computer Science
University of Maribor
Smetanova ulica 17, 2000 Maribor
e-mail: damjan.obal@uni-mb.si

ABSTRACT

The following paper presents proceedings from our research on and implementation of a tailor-made environment enabling online virtual collaboration. Actual implementation is presented through our case study (the *Challenge: Future* student competition). The environment described is called “Project Room.”

In our paper we present the development process and implementation of a virtual collaboration environment offered to students to work, collaborate and communicate on distance. Instead of using one of the existing platforms, the decision to design a new collaboration platform was made by the team developing and later facilitating the mentioned student competition platform. A simplified solution tailor-made to fit the needs of the peer community. The environment is modular, new features can be added or removed so that the collaborative work process remains as fluid as possible. The environment implemented in the *Challenge: Future* online competition platform tends to be user friendly, have a fast learning curve and appeal for the majority of highly diverse global community. In our case-study we present the so called “Project Room” which represents the working area for student groups competing. Because of its modular and adaptive nature, the environment is also used for other user groups working with the online platform, such as the judging panel, advisory board and others.

1 INTRODUCTION

Online collaboration is nowadays accepted as a standard working praxis. Not only for global corporates but also for scholars, students and others. There are several existing commercial and open/source platforms that enable online collaboration. Solutions for different target groups and meeting a variety of needs, ranging from project management (ex. Basecamp, Flowr, Huddle) to real-time collaboration environments with shared whiteboards and integrated communication tools (ex. DimDim, WebEx) and course management systems for education purposes (ex. Moodle). There are also numerous solutions like those offered by Google or Zoho, that enable group work, sharing of knowledge and collaboration through communication. Yet, the perpetual challenge of making the perfect environment for virtual collaboration still remains. Perfect for some could mean complicated, to sophisticated for others. Therefore the need to develop niche solutions for specific user groups or to enable modular structure where groups can enable or disable some features.

When developing the *Challenge: Future* competition platform we wanted to offer the students competing in teams a working environment where they could work, share and communicate as a group. While planning we continually kept in mind that this community is extremely diverse and that the environment should be easy and intuitive to use. The plan was to offer the essential modules and than design new ones according to users feed-back.

2 TAP INTO YOUR PEERS - RESEARCH PHASE

The *Challenge: Future* platform was being built entirely from scratch, which gave us the possibility to come up with unique, tailor-made solutions. We called the virtual semi-closed environment offered to students the “Project Room.” Every student team[1] should have a place to work, share, collaborate and communicate on distance. The question what features to implement arose and therefore a research and analysis was done. The competition platform itself does not tend to mimic social media or community centered platforms, but wants to act as a connection between students and global challenges. The goal being a fluid experience with just the right amount of interactions and features. Research which involved stakeholder interviews, SME (subject matter expert) interviews, user testing and observation, was done with a clear goal: identify the main features/functions. Only than the designing process could begin. Previous internal research has exposed potential problems when trying to combine too many features into one service. Problems for both the development team as well as the users.

[1] Students could compete in the competition only as teams. Therefore each students forms or joins a team according to her own preferences. The teams have 3-5 members, all students bellow the age of 30.

Results of the research have shown a clear focus on four features: a) privacy, b) virtual collaboration, c) file sharing and d) activity monitoring. Those were the main focus points that were in common in most of the research data regarding the online collaboration on the platform.

3 PROJECT ROOM – DESIGN PHASE

The "Project Room" needed to be integrated into the platform as a whole, yet remain a closed working environment available to each student team. Before the development begun, we prepared a document with the vision statement, goals and functionalities of the platform. The four features mentioned in the previous part remained the focus points in the designing process. Meaning we left out most of the functionalities and focused solely on the important [2] features to keep the working environment simple to use and easier to develop (fast implementation curve).

3.1. Privacy

Privacy is an important issue, not only for corporates and organizations involved in, but also for students competing in the *Challenge: Future* competition. Especially after the internet community is recently being exposed to warnings about illegal personal data collection through internet and other privacy issues. Issue of privacy was therefore taken with care when designing the "Project Room" and the platform itself.

We constructed different roles that represent different users with several levels of control. Still, the "Project Room" remains closed at all times and not even administrators have the insight here. Besides administrators we defined other roles to help maintain levels of privacy. Main roles are: unregistered user, registered as student, mentor, school representative, challenger, judge, advisory board member, youth advisory board member, or registered as editor. Roles are flexible and have the ability of stacking. In praxis that means one user can have multiple roles (ex. compete as student and contribute also as content editor).

With the use of different roles the platform also maintains high level of privacy. The research showed that users prefer a semi-closed environment where registration is needed to participate, over an open platform. That means unregistered users are not even able to see basic information displayed in user profiles.

3.2. Virtual Collaboration

As the teams competing are global working on distance through a virtual platform was a priority. The question was which functions to include as the goal of the managerial team is not to present a sophisticated, features-packed collaboration environment. Research proceeding indicated that the users expect simple collaboration while working on their solutions. Therefore we implemented wiki pages into the "Project Room" because they allow open collaboration on documents. Each team member can track changes made, edit the content and later, when all team members agree, the team makes the final solution (exported as a PDF from the wiki) visible to judges.

3.3. File sharing

File sharing is closely connected to virtual collaboration principles and is usually integrated in collaboration platforms. Sharing files among team members inside a single "Project Room" was also integrated in the *Challenge: Future* platform. Team members can upload all main types of documents (text, image, video) which are available to other team members. Documents are uploaded to a secure server and after successful upload they appear as icons (icon reflecting content type) in the "Project Room." Uploading is fluid through a user/friendly user interface that explains the upload restrictions (file size and type) and guides the user. Files can also be added as supplements to the solution the team is presenting to the community and judges.

3.4. Activity monitoring

Another feature the peer community expressed to see integrated in the working environment was monitoring the activity of team members and their own. Especially the internet savvy are used to following other peoples feeds and updates through social media such as Twitter or Facebook.

[2] According to data collected with internal research done in years 2009 and 2010. Interviewed and surveyed were students involved in the competition as registered users and contestants.

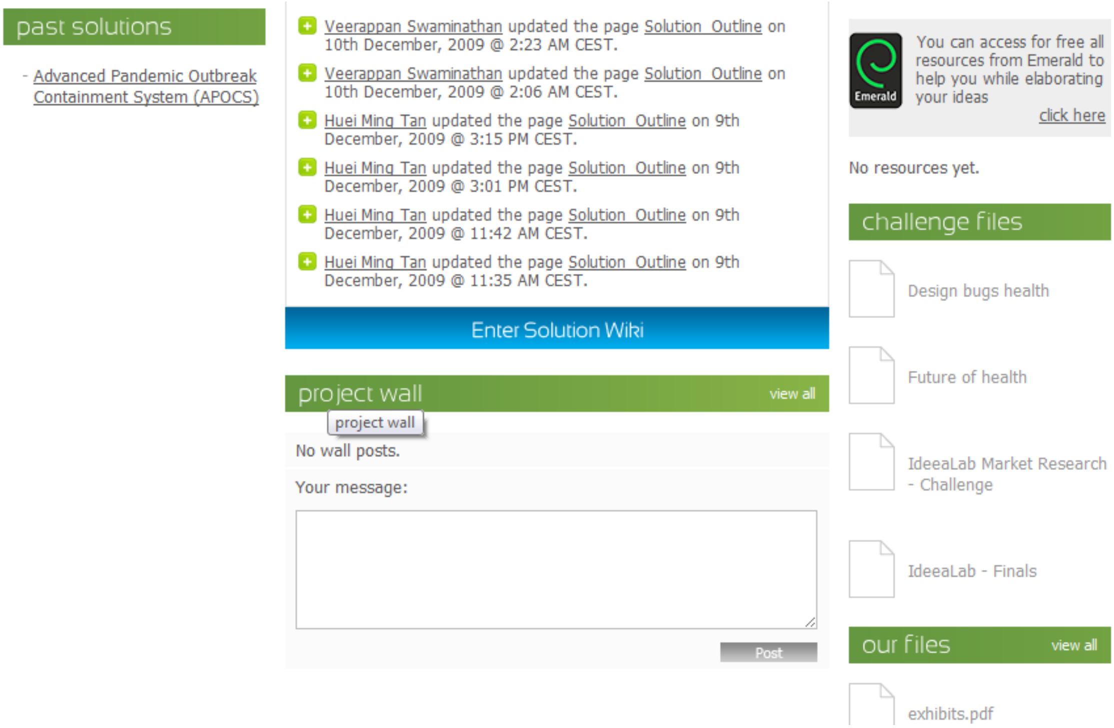


Figure 1: *Student team "Project Room" main page - view as a team creator*

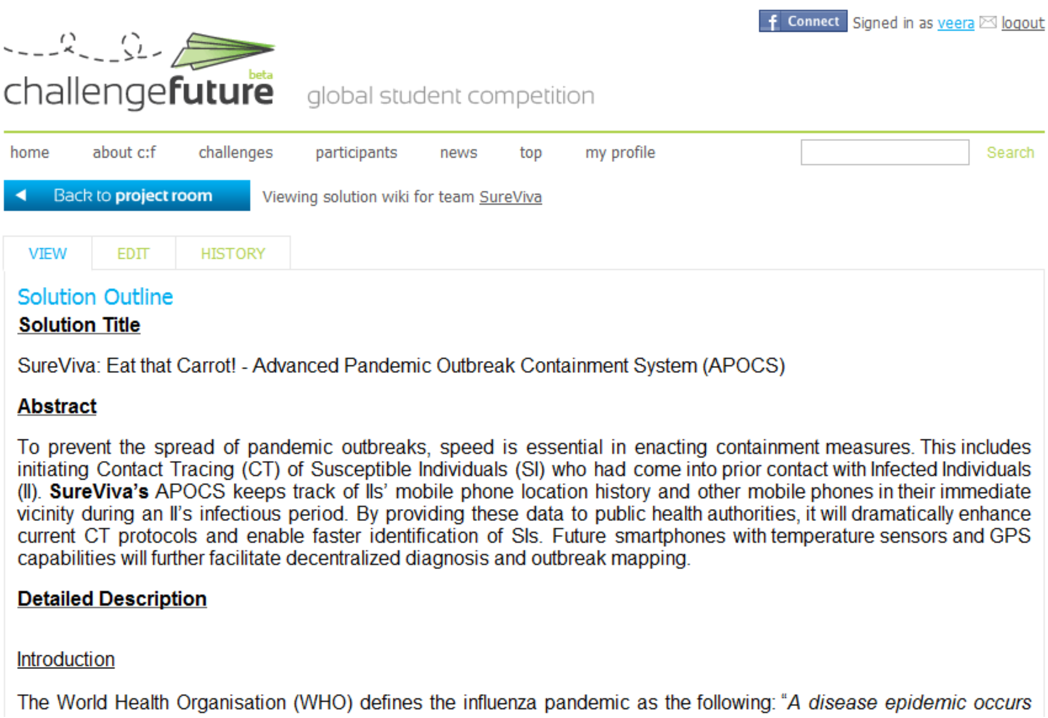


Figure 2: *The wiki page inside the "Project Room" - view as a team creator*

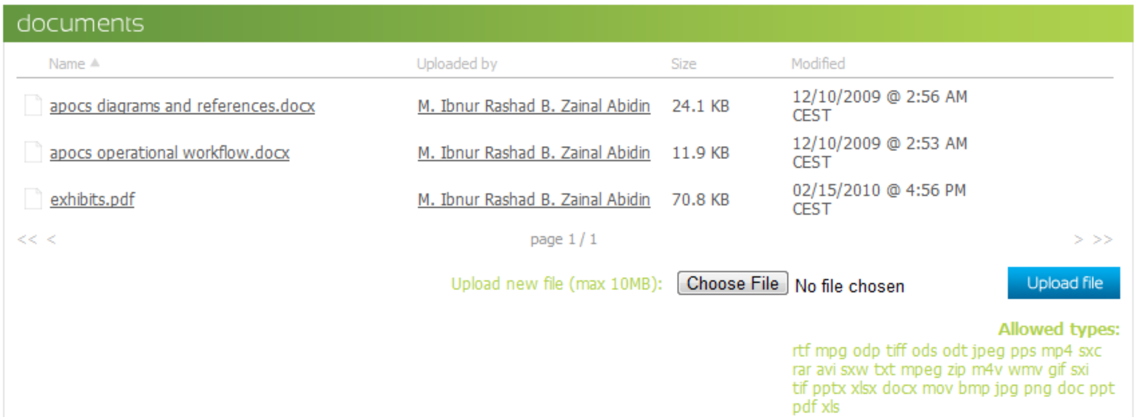


Figure 3: *File sharing and management inside the "Project Room"*

4 PROJECT ROOM – VARIATIONS

As mentioned in the introduction, the "Project Room" is designed so that it can be used by various user groups. In the *Challenge: Future* competition there were several variations of the environment used for a) student teams, b) advisory board, c) youth advisory board and d) *Challenge: Future* ambassadors. Different user groups mean different demand regarding functions of the working environment. Not only the functions, but it was also needed to redesign the "Project Room" and reposition some functions and content blocks. The information flow was namely different for various user groups.

All user groups used their "Project Room" as a communication channel, competing student team even more intense than other groups. The "Project Room" designed for the advisory board members was simplified and stripped of some functions and content blocks. The focus point was towards communication more than collaboration. Yet, the ambassadors and the youth advisory board expressed wishes to enable the collaboration and co-working functions. They were the most productive (regarding online collaboration) groups and used the functions offered to full extend. User groups other than the student teams, could form multiple wiki pages and export them as PDF documents.

5 CONCLUSION AND FURTHER DEVELOPMENT

The research done after the first season of the competition and the feedback gathered throughout the competition showed that the students as well as other user groups used the collaborative environment presented to them. Even if it meant using it only as a meeting spot where they planned further interactions. Most of the students knew each other before and were in a way forced to co-work online through the *Challenge: Future* platform, yet also for them the experience was rewarding and pleasant. They shared all the information and data in one central place, where they also received further resources and instructions. The organizers weren't sure if the students would like and get used to the "Project Room" and so the positive feedback was a much needed encouragement for further development.

Plans for a revamped collaboration environment are already in motion. At the time being, the organizers are still collecting and analyzing all the feedback data which will be later used to redesign the "Project Room" and to implement it into the revamped platform. The environment needs a more intuitive user experience with a better information flow, because some of the students complained that the interaction pattern was not as easy as expected. There was also a demand for a more sophisticated document sharing approach. A solution could be implementing an existing solution like the Google Docs. But the later would mean moving away from the basic idea of the environment. Namely the planning and design team decided to develop a simple lightweight environment which could also work only as a hub. Therefore, the wiki philosophy should remain but should contain minor changes. Like allowing other team members to edit the wiki page.

The *Challenge: Future* platform as well as the "Project Room" environment tend to remain highly adaptable and scalable which leaves the room for constant evolution. The whole designing process is user-oriented and should remain like that.

References

[1] E. von Hippel. Welfare implications of user innovation. *Journal of Technology Transfer.* 2005.

[2] E. von Hippel. Democratizing Innovation. The MIT Press. Cambridge. 2005

[3] D. Tapscott, D. Williams. Wikinomics: how mass collaboration changes everything. Portfolio Publishing, Penguin Group. New York, NY. 2006.

[4] Li, C. & Bernhoff, J. Groundswell - Winning in a world transformed by social technologies.Harvard Business Press, Boston, Massachussets. 2008

[5] Piller, Frank T. & Walcher, D. Toolkits for Idea Competitions: A Novel Method to Integrate Users in New Product Development, 2005, MIT. Retrieved July 11, 2010 from http://userinnovation.mit.edu/papers/

E-FAKULTETA: INFORMACIJSKE REŠITVE

Bojan Bedrač, Marjan Krašna
Filozofska fakulteta
Univerza v Mariboru
Koroška cesta 160, 2000 Maribor, Slovenija
Tel: +386 2 22 93 645; fax: +386 2 22 93 625
e-mail: bojan.bedrac@uni-mb.si

POVZETEK

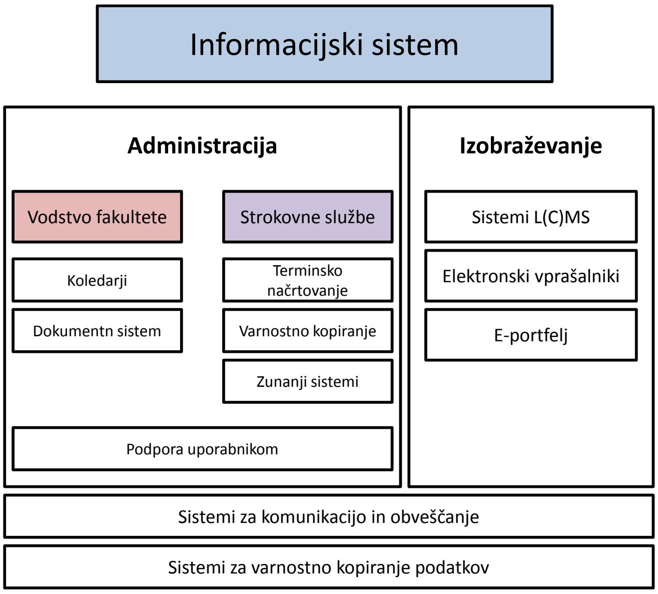
Fakultete morajo v obdobju informacijske družbe držati korak s časom na področju informatizacije in približevanja lastnih storitev različnim ciljnim skupinam. Za optimizacijo in posledično racionalizacijo poslovnih procesov smo si na Filozofski fakulteti pomagali z različnimi informacijskimi rešitvami. V tem članku so predstavljene informacijske potrebe sodobne fakultete oz. e-fakultete in nekatere rešitve, ki povezujejo različne organizacijske enote znotraj fakultete. Predstavljene so informacijske rešitve, s katerimi se zagotavlja izvajanje poslovnih procesov znotraj sodobne fakultete. Predstavljene pa so tudi informacijske rešitve, s katerimi se izvaja integracija fakultete z ožjo in širšo javnostjo. Uvedba informacijskih rešitev je pustila pečat na različnih področjih in izboljšuje poslovanje fakultete z več vidikov.

1 UVOD

Informatizacija fakultet se je seveda začela že dolgo nazaj in razne informacijske rešitve že močno prepletajo vsakdanje delovne procese znotraj fakultet. Najbolj vsakdanji informacijski rešitvi danes sta, prav gotovo, elektronska pošta in spletne strani. Ti dve sta verjetno prvi skupni imenovalec informacijskih rešitev vseh fakultet. Elektronska pošta se uporablja za usmerjeno komunikacijo med sodelavci in zunanjimi partnerji, pa tudi s študenti ter ostalimi. Spletne strani pa se uporabljajo za bolj splošno, čeprav še vedno usmerjeno, komunikacijo s širšo javnostjo. V tem članku bosta zaradi vsakdanje uporabe ti dve storitvi v svoji osnovni funkciji večinoma izpuščeni, saj je namen članka predstaviti ostale storitve, ki so morda manj razširjene v izobraževalni sferi in prinašajo v poslovne procese fakultete izboljšave tudi na ostalih področjih, kot so ekologija, geografska ločenost, redundantnost in varovanje podatkov. Storitve so predstavljene na primeru Filozofske fakultete Univerze v Mariboru, kjer se povezujejo v informacijski sistem. Storitve so razvrščene glede na področje, v katerem se uporabljajo:

- administrativni proces,
- izobraževalni proces
- in integracija širše javnosti.

Vsako izmed naštetih področij ima specifične potrebe po določenih storitvah in njihovi integraciji v delovni proces. Uporabljene informacijske rešitve, ki sestavljajo informacijski sistem Filozofske fakultete, in njihova delitev so prikazane na shemi informacijskega sistema Filozofske fakultete (Slika **1**).



Slika 1: *Informacijski sistem Filozofske fakultete*

2 MERILA IZBORA INFORMACIJSKIH REŠITEV

Integracija različnih informacijskih rešitev v poslovne procese predstavlja kompleksno nalogo, ki zahteva podrobno načrtovanje in sprejemanje kompromisov. Sprejemanje kompromisov med dodano vrednostjo, ki jo rešitev prinaša ter izbiro tehnologije, katera jo omogoča. Pri vpeljevanju informacijskih rešitev na Filozofski fakulteti sledimo nekaj osnovnim merilom, katerim mora v čim večji meri zadostiti vsaka uporabljena rešitev.

2.1 Zanesljivost

Vse informacijske rešitve, ki se uporabljajo morajo biti stabilne in preizkušene različice izbranih tehnologij. Razvojne različice in tehnologije s kratko življenjsko dobo, se praviloma ne uporabljajo, razen ko gre za nove tehnologije priznanih proizvajalcev. Vsaka tehnologija mora biti pred širšo integracijo, temeljito preizkušena v manjših skupinah. V primeru izrednih dogodkov lahko izvedemo vse tehnične posege za normalizacijo stanja storitev.

2.2 Povezljivost in razširljivost

Povezljivost je tista ključna lastnost vsake informacijske rešitve, ki poenostavlja uporabo številnih različnih informacijskih rešitev, brez večjega vpliva na uporabniško izkušnjo. Zaradi zagotavljanja povezljivosti se uporabljajo tehnologije, ki sledijo standardom, čeprav je včasih potrebno uporabiti tudi kakšno rešitev, ki najbolj ustreza potrebam in ni neposredno povezljiva z ostalimi tehnologijami. Takšni kompromisi so potrebni predvsem pri povezovanju določenih storitev z različnimi vrstami naprav. V takšnih primerih so včasih potrebne tudi razširitve oz. prireditve tehnologij. Razširitve so seveda lažje izvedljive, če gre za odprtokodne rešitve, zaradi česar je odprtost kode velik faktor pri izboru rešitve.

2.3 Zagotovljena podpora

Nekatere tehnologije se razvijajo zelo hitro in se v nekaj letih spremenijo tako korenito, da niso več združljive s prejšnjimi različicami ali pa je podpora za prejšnje različice opuščena. Zaradi zagotavljanja dolgoročnega stabilnega delovanja informacijskih rešitev, je izbor odvisen tudi od načina njihovega delovanja. Prednost imajo rešitve, ki omogočajo lokalno namestitev in tako zagotavljajo samozadostnost in neodvisnost storitve.

2.4 Nizka cena

Vsaka uporabljena informacijska rešitev mora biti finančno ustrezna, kar pomeni, da mora biti cena nakupa, implementacije in vzdrževanja storitve, upravičena in sorazmerna s prednostmi, katere prinaša. Cilj uporabe vsake informacijske rešitve je nižanje stroškov in/ali razbremenitev zaposlenih. Pri tem gre lahko za optimizacijo ali avtomatizacijo določenih delovnih nalog in nemalokrat tudi za dvig kakovosti dela. Pri izbiri informacijskih rešitev izbiramo brezplačne rešitve oz. rešitve, ki so del obstoječih pogodb in ne predstavljajo velikih dodatnih finančnih vložkov.

3 ADMINISTRATIVNI PROCES

Administrativni proces zajema organizacijske enote in delovne procese, ki so povezani z administrativnimi nalogami znotraj fakultete. Sem spadajo vse strokovne službe in vodstvo fakultete. Zaposleni v teh organizacijskih enotah običajno delujejo v skupinah, ki spominjajo na projektne skupine. Zaradi tega so tudi njihove zahteve po informacijskih storitvah v svojem delovnem procesu podobne.

3.1 Vodstvo fakultete

Vodstvo fakultete sestavljajo dekan, prodekanke, senat s pripadajočimi komisijami, poslovodni odbor in administrativno osebje.

DELJENI KOLEDARJI
Zaradi zelo zasedenih urnikov in pogostega usklajevanja sestankov, se je pojavila potreba po deljenih koledarjih in sinhronizaciji podatkov med zaposlenimi in napravami, ki jih uporabljajo za dostop do podatkov. Rešitev temelji na tehnologiji Microsoft (MS) Windows SharePoint Services 3.0 [1], ki jo je mogoče naložiti na strežnike, ki poganjajo operacijski sistem MS Windows. Dostop do deljenih koledarjev je mogoč z uporabo osebnega računalnika, ki je povezan v internet. Dostopamo lahko z uporabo poljubnega spletnega brskalnika ali odjemalca (MS Outlook 2007/2010). Sinhronizacijo z mobilnimi telefoni vršimo z uporabo brezžičnih tehnologij in sinhronizacijske programske opreme proizvajalca mobilnih telefonov.

DOKUMENTNI SISTEM LMS MOODLE
Senat Filozofske fakultete je v začetku leta 2008 začel z opustitvijo uporabe papirnih delovnih gradiv za seje senata. Vsa delovna gradiva, ki so se pred tem tiskala in razpošiljala senatorjem, so bila prestavljena v dokumentni sistem. Pri izbiri dokumentnega sistema, smo pretehtali kar nekaj možnosti, na koncu pa izbrali kot rešitev LMS Moodle [2]. Senatorji (profesorji) poznajo LMS Moodle, prav tako pa so delo z Moodlom usvojili administrativni delavci zaposleni v strokovnih službah, ki delovna gradiva pripravljajo. S prehodom na uporabo centralnega mesta, kjer se shranjujejo delovna gradiva za seje, je potrebno izpostaviti predvsem nekaj prednosti:

- Delovna gradiva so vedno dostopna na istem mestu
- Delovna gradiva so dostopna kjerkoli in kadarkoli z uporabo spletnega brskalnika
- Lažja organizacija korespondenčnih sej
- Dostop do delovnih gradiv je omejen z uporabniškimi računi
- Spremljanje dostopa do gradiv
- Sistem omogoča spremljajočo komunikacijo in obveščanje o spremembah
- Znižanje stroškov zaradi manj tiskanja oz. fotokopiranja in poštnih stroškov

3.2 Strokovne službe

Strokovne službe so tiste organizacijske enote, brez katerih fakulteta ne more normalno delovati. Nudijo vse vrste administrativnih storitev, na katerih temeljijo ostali delovni procesi znotraj fakultete. V interesu vseh zaposlenih je, da delujejo optimalno. Med strokovne službe običajno sodijo računovodstvo, kadrovska služba, referat za študentske zadeve, knjižnica in druge nepedagoške službe. Delo v teh organizacijskih enotah je večinoma skupinsko, zato spominjajo na projektne skupine. Zaradi tega smo pri iskanju optimalnih informacijskih rešitev posegli po orodjih, ki so se v preteklosti uspešno izkazala pri projektnem delu. Pred kratkim je bila na fakulteti ustanovljena tudi dislocirana enota mednarodne pisarne, ki je pokazala, da so informacijske rešitve, ki jih uporabljamo razširljive in univerzalno uporabne ne glede na geografsko ločenost. V strokovnih službah tako uporabljamo več informacijskih rešitev, ki so medsebojno neodvisne, skupaj pa tvorijo močno podporo delovnim procesom. V okviru članka predstavljamo le nekatere izmed njih.

TERMINSKO NAČRTOVANJE
Zaposleni uporabljajo deljene koledarje, ki so se izkazali za uspešne tudi v drugih organizacijskih enotah. Uporaba koledarjev je urejena hierarhično, da lahko vsak zaposleni upravlja s svojim in skupnim deljenim koledarjem, vodje

posameznih organizacijskih enot pa lahko upravljajo tudi s koledarji svojih podrejenih. Na ta način imajo jasen pregled nad obremenitvijo podrejenih, kar jim omogoča sprejemanje utemeljenih odločitev glede sprejemanja in razporejanja novih delovnih nalog. Koledarji se praviloma delijo znotraj organizacijske enote, nobenih omejitev pa ni niti za širšo organizacijo dela, ki se razprostira med več organizacijskih enot. Storitev tudi tu omogoča MS Windows SharePoint Services 3.0, ki deluje v povezavi z odjemalcem MS Outlook 2007/2010 ali poljubnim spletnim brskalnikom. Prednosti, ki jih prinaša tak način dela so:

- Poenostavljeno delegiranje delovnih nalog
- Transparentnost obremenitve zaposlenih
- Utemeljene odločitve pri terminskem načrtovanju

ZUNANJI DOKUMENTNI SISTEM GROOVE

Strokovne službe imajo opravka z velikimi količinami dokumentov. Na velikem številu dokumentov sodeluje tudi po dva ali več zaposlenih. S tem se je pojavila potreba po centralizirani shrambi dokumentov. Običajno se takšne potrebe rešujejo s shranjevanjem dokumentov v omrežne mape na datotečnih strežnikih. Takšen način zahteva varnostno kopiranje map v strežniku, po drugi strani pa v primeru izpada internetne povezave otežuje delo. Zaradi poenostavljanja teh opravil, smo uporabili drugo informacijsko rešitev, ki je geografsko neodvisna, zmanjšuje potrebo po varnostnem kopiranju in omogoča normalno delo tudi ob prekinjeni internetni povezavi. Poleg samega dokumentnega sistema pa omogoča tudi komunikacijo in visoko stopnjo razširljivosti. Rešitev se imenuje MS Groove 2007 [3] oz. novejša različica MS SharePoint Workspace 2010 [4] in je del pisarniškega paketa MS Office 2007 oz. MS Office 2010. V večji meri še vedno uporabljamo različico MS Groove 2007. Gre za zunanjo storitev, ki jo zagotavlja korporacija Microsoft, možna pa je tudi lokalna namestitev sinhronizacijskega strežnika in s tem popolna neodvisnost. Po dveh letih uporabe smo s storitvijo zadovoljni in o postavitvi lastnega sinhronizacijskega strežnika ne razmišljamo. Delovanje informacijske rešitve je zanesljivo in zelo učinkovito. Dokumenti se shranjujejo v skupen delovni prostor, do katerega dostopajo sodelujoči. Ob spremembah dokumentov so ti označeni kot spremenjeni, sodelujoči pa so o spremembah sproti obveščeni. Sinhronizacija podatkov zagotavlja, da so podatki shranjeni lokalno tudi na vseh osebnih računalnikih sodelujočih, kar zmanjšuje potrebo po varnostnem kopiranju. V primeru odpovedi računalnika katerega izmed sodelujočih, se ob naslednji prijavi podatki spet sinhronizirajo v aktualno stanje ostalih. Prednosti, ki jih tak način dela prinaša so naslednje:

- Omogoča nemoteno sodelovanje dislociranih enot
- Samodejno obveščanje o spremenjenih dokumentih
- Sinhronizacija podatkov med vse sodelujoče – zmanjšanje potrebe po varnostnem kopiranju
- Omogoča komunikacijo med sodelujočimi, ki je shranjena v delovni prostor
- Omogoča nemoteno delo med prekinjeno internetno povezavo
- Dodatno omogoča razširljivost s poljubnimi strukturami (obrazci in vnaprej pripravljenimi moduli)

SISTEMI ZA VARNOSTNO KOPIRANJE PODATKOV

Pri velikih količinah dokumentov v obdelavi je potrebno učinkovito skrbeti za varnostno kopiranje podatkov. Večina podatkov je sinhronizirana med sodelujoče ali drugače shranjena v omrežnih mapah na datotečnih strežnikih, vseeno pa ostane nekaj podatkov, ki se nahajajo na osebnih računalnikih in jih je potrebno ločeno varnostno kopirati. V ta namen uporabljamo dve informacijski rešitvi. Obe delujeta v obliki odjemalca, ki se namesti na osebni računalnik in periodično izdeluje varnostne kopije. Obe rešitvi sta tudi brezplačni in odprtokodni, ter delujeta na različnih operacijskih sistemih.

Prva rešitev je Areca Backup [5], ki se namesti v obliki odjemalca, za shranjevanje varnostnih kopij pa uporablja izmenljive podatkovne medije ali obstoječe datotečne strežnike. To so lahko datotečni strežniki, ki uporabljajo omrežne mape ali protokol FTP. Rešitev deluje ne glede na geografsko ločenost in jo s pridom uporabljamo v dislocirani organizacijski enoti. Za upravljanje z varnostnimi kopijami odjemalec vsebuje grafični vmesnik, ki delo močno poenostavi. Upravljanje s politiko varnostnega kopiranja se izvaja neposredno na odjemalcu, zaradi česar je rešitev manj primerna za večje število odjemalcev. Rešitev uporablja javansko tehnologijo, zaradi česar je uporaba mogoča na različnih operacijskih sistemih.

Druga rešitev je Amanda [6], ki ima dolgo tradicijo v razvoju sistemov za varnostno kopiranje v operacijskih sistemih *NIX, v zadnjem času pa se je razširila tudi z odjemalci za operacijski sistem MS Windows. Rešitev je izvedena kot par odjemalec-strežnik. Strežniško storitev je mogoče namestiti samo na operacijske sisteme *NIX, odjemalec pa tudi na MS Windows sisteme. Politika varnostnega kopiranja se nastavlja centralno na strežniku, slaba stran te rešitve pa je, da je težje dodajati geografsko dislocirane odjemalce. V brezplačni različici prav tako ni na voljo grafičnega vmesnika za obnovitev varnostnih kopij, je pa ta na voljo v plačljivi različici. V praksi je tudi delo brez grafičnega vmesnika dovolj enostavno za administratorja.

4 IZOBRAŽEVALNI PROCES

V izobraževalnem procesu je najbolj uporabljana informacijska rešitev LMS Moodle. Zaradi odprave nekaterih omejitev postavitve univerzitetnega LMS Moodle in možnosti zagotavljanja lastne podpore, imamo vzpostavljen lastni LMS Moodle. Samo v lanskem študijskem letu so pedagoški delavci z uporabo LMS Moodle izvajali več kot 80 predmetov. Prednosti uporabe LMS Moodle ne rabimo posebej poudarjati, spodbudno pa je, da število predmetov iz leta v leto raste, kar pomeni, da udeleženi v izobraževalni proces na daljavo izbrano rešitev sprejemajo pozitivno, posledično pa ima to blagodejen vpliv na okolje [7]. Uspešno smo integrirali tudi lastni modul za podporo spremljanja študentske prakse [8] in izdelavo študentovega e-portfelja [9]. Razlogi za uspeh so podobni, kot pri uporabi LMS Moodle v vlogi dokumentnega sistema.

5 INTEGRACIJA S ŠIRŠO JAVNOSTJO

Po nedavni raziskavi med vsemi študenti naše fakultete, so rezultati potrdili, da je največkrat uporabljen vir informacij o fakulteti spletna stran fakultete. Pred letom dni smo zaradi posebnih potreb pričeli z uporabo odprtokodne in brezplačne informacijske rešitve za upravljanje s spletnimi stranmi. Rešitev, ki smo jo izbrali po več preizkušenih rešitvah, je dotCMS [10]. Poganja jo javanska tehnologija, odlikuje pa jo odprtost, razširljivost, uporaba standardov in praktično neomejene možnosti pri oblikovanju in delu s spletnimi stranmi. Med zanimivejšimi lastnostmi, izpostavljamo le nekaj ključnih:

- Upravljanje več domen
- Možnost kreiranja lastnih podatkovnih struktur
- Možnost definiranja relacij med strukturami
- Enostavno ustvarjanje večjezičnih vsebin
- Podpora močnemu skriptnemu jeziku
- Temelji na standardnih tehnologijah
- Dobro dokumentiran sistem in programski vmesniki

S prenovo spletnih strani smo uspeli združiti vse podatke na enem mestu in tako spletne strani povezujemo z več drugimi informacijskimi sistemi. Med njimi naj omenimo nekaj ključnih informacijskih virov:

- AIPS (govorilne ure in izpitni roki)
- COBISS (bibliografije zaposlenih)
- SICRIS (raziskovalne podrobnosti zaposlenih)
- Moodle (povezave s predmeti, ki jih predavajo)
- Obvestila referata za študentske zadeve

Vsi omenjeni informacijski viri so obiskovalcem dostopni neposredno z osebnih spletnih strani zaposlenih. Statistika uporabe spletnih strani je pokazala, da največji del uporabnikov predstavljajo študenti, zato smo vsebinam namenjenim njim posvetili še posebno pozornost.

SISTEM ZA IZVEDBO ELEKTRONSKIH VPRAŠALNIKOV

V študijskem letu 2009/2010 smo pričeli z aktivnejšim spremljanjem kakovosti študijskega procesa. V okviru Akcijskega načrta, ki je nastal na podlagi smernic in predlogov Zunanje evalvacije v predhodnem študijskem letu, smo pripravili vprašalnike, ki jih študenti izpolnjujejo preko spleta. Rešitev, ki jo uporabljamo v ta namen je brezplačna in odprtokodna, imenuje pa se php Easy Survey Package [11]. Omogoča nam izvedbo javnih ali zaprtih anket in samodejno osnovno obdelavo podatkov. Pri omejevanju dostopa do izpolnjevanja vprašalnikov smo izbrali metodo, ki je dokaj varna in ne zahteva posebnega obveščanja študentov in s tem povezanih stroškov. V tem študijskem letu smo izvedli tri ankete, načrtujemo pa jih skupno šest. Izvedba je potekala brez zapletov, osnovno obdelani rezultati pa so bili na voljo takoj po zaključenih anketah.

6 ZAKLJUČEK

Skupek informacijskih rešitev, ki jih uporabljamo, je verjetno podoben tistim na ostalih fakultetah, ki uspešno vključujejo informacijske rešitve v svoje delovne procese in s tem znižujejo stroške poslovanja. Naš cilj je zagotavljati informacijske rešitve, ki dvigujejo kakovost storitev in razbremenjujejo zaposlene oz. dvigujejo kakovost dela. Menimo, da mora biti sodobna fakulteta v pravem pomenu e-fakulteta in spodbujati nadaljnjo informatizacijo. Le s takšnim pristopom bo mogoče zagotavljati racionalizacijo poslovanja.

Viri

[1] Microsoft Corporation. (2010, avgust) Windows SharePoint Services 3.0 - Overview | TechNet. [Online]. http://technet.microsoft.com/en-us/windowsserver/sharepoint/bb684453.aspx

[2] Moodle. (2010, avgust) Moodle.org: open-source community-based tools for learning. [Online]. http://moodle.org

[3] Microsoft Corporation. (2010, avgust) Groove 2007. [Online]. http://msdn.microsoft.com/en-us/office/bb308957.aspx

[4] Microsoft Corporation. (2010, avgust) Microsoft SharePoint Workspace 2010 - Microsoft Office. [Online]. http://office.microsoft.com/en-us/sharepoint-workspace/

[5] Petrucci, O.. (2010, avgust) Areca Backup - Official Website. [Online]. http://areca-backup.org

[6] Zmanda, Inc. (2010, avgust) Amanda Open Source Backup. [Online]. http://amanda.zmanda.com/

[7] Bratina, T. and Krašna, M., "Discontinuing the printed learning material in teaching statistics and research methodology," in *CSMW'09 / Computer Science and Mobility Workshop*, Klagenfurt, 2009, pp. 29-33.

[8] Krašna, M. and Kaučič, B., "Designing e-portfolio module for open source LMS," in *Computer science and technology, (Electrical and engineering series)*, WSEAS, 2007, pp. 305-310.

[9] Krašna, M. and Bratina, T., "E-portfolio of teacher training," in *MIPRO 2007 : 30th Jubilee International Convention, May 21-25, 2007, Opatija, Croatia : proceedings = Computers in education*, Rijeka, 2007, pp. 146-149.

[10] dotCMS Inc. (2010, avgust) Home - dotCMS Web Content Management System. [Online]. http://dotcms.com

[11] Butterfat, LLC. (2010, avgust) Projects/phpESP - Butterfat. [Online]. http://www.butterfat.net/wiki/Projects/phpESP

ANALYSIS OF THE SHAREPOINT APPLICATION TEMPLATES

Jan Božnik
University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova ulica 17, 2000 Maribor, Slovenia
Tel: +386 2 220 7399; fax: +386 2 220 7272
E-mail: jan.boznik@uni-mb.si

ABSTRACT

The article offers a review of workflows and application templates that are available for the SharePoint portal. Various possible uses are presented, including information on what they cover and which workflows they employ. Also in the article is the short proposal of introducing application templates to an existing company.

1 INTRODUCTION

Everything in the business world revolves around business processes that companies and organizations advertise as their core business. Business processes can be automated to the extent that they do not require human interaction, but most of them still require cooperation between man and the system. Activities that follow in sequence can be linked into a whole. This produces a workflow. Workflows can occupy a number of areas and domains, there is virtually no limit. For optimal results, however, workflows must be carefully planes and designed, and nonetheless used. With an appropriate environment these tasks would be simplified and would offer usefulness to a grater audience.

Microsoft SharePoint represents such an environment [1] and with its web orientation covers the majority of standard business functionality. Requirements and needs are different from business to business, addressing them are the application templates. With their help the environment can adapt to specific needs, and with little effort, they can be completely customized. A link with Microsoft .NET Framework and its component Windows Workflow Foundation allows deployment of custom workflows and with SharePoint Designer tool it is easy to edit and modify web pages and workflows.

2 SHAREPOINT

SharePoint, Microsoft's answer to the enterprise portal, has quickly grown up [2] since its introduction in 2001. The main selling point for a server-based portal product such as SharePoint is increased collaboration throughout departments and the enterprise. Collaboration is enabled through the use of document libraries, online calendars, announcements, alerts, and many more features. In addition, SharePoint has an extensible programming infrastructure that uses a concept called web parts [3]. *Web parts* are reusable widgets that can be placed in any SharePoint page and configured however the developer specifies.

Pretty much everything in SharePoint is a list. A *list* is what you think it would be: a collection of items. Documents are the items that live in a document library; events live in a calendar; issues live in an issue tracker list; and so on. This architecture makes for a very extensible platform that can handle new types of user-created items. One of the greatest new and sought-after features of the Office 2007 system is the addition of workflow. Organizations have long searched for a cohesive answer to the problem of workflow in Office and SharePoint, and with the help of Windows Workflow Foundation, Microsoft has provided a solution [4].

3 OUT-OF-THE-BOX WORKFLOWS

If SharePoint is the workflow runtime host, the Office applications - such as Word, Excel, and PowerPoint - are the clients. As previously mentioned, these tools are SharePoint and workflow-aware. Therefore, the functionality necessary for interacting in workflows is hardcoded into these applications. For example, Word includes the ability to start, modify, and complete workflows without ever leaving the application.

The following prebuilt workflows are provided with SharePoint out of the box:

- Approval,
- Collect Feedback,
- Collect Signatures,
- Disposition Approval,
- Translation Management and
- Three-state.

Approval

The out-of-the-box Approval workflow covers one of the most common workflow scenarios known to man: document approval. This is a natural fit for SharePoint because of its document centric nature. By default, the Approval workflow is associated with the Document content type; therefore, additional configuration to associate this workflow with a particular document library is not necessary. This workflow has some basic steps and custom forms. When an Approval workflow is started on a document, a task is created for each designated approver to either approve or reject the document. In addition, the approver can request changes to the document or reassign the task to another person.

Collect Feedback

The Collect Feedback workflow supports any process that involves obtaining comments from one or more people related to a document. Like the Approval workflow, the Collect Feedback workflow is associated with the Document content type by default. Therefore, you can manually start it on any document, and it does not need to be associated with a particular document library to be available.

Collect Signatures

The Collect Signatures workflow is specific to a feature in Microsoft Office related to digitally signing documents. To use this workflow, a user must first add one or more signature lines to a Word or Excel document from within the application. After the document has been saved to a SharePoint document library, the Collect Signatures workflow can be started from within the client application only.

Disposition Approval

The out-of-the-box Disposition Approval workflow provides functionality to allow organizations to easily manage expired content. For example, a document library can have a policy that dictates that after a certain amount of time, a document should be expired. This policy can do one of two things when the expiration occurs: delete the document or start a workflow.

Translation Management

This workflow type facilitates the process of manually translating a document from one language to any number of other languages. It is unique in that it can be associated only with a translation management library. Therefore, to use the Translation Management workflow, you first need to create a translation management library.

After a translation management library is created and a Translation Management workflow is associated with it, any new documents added to the library are duplicated for each language to be translated to as dictated by the translators list associated with the workflow. You are given an opportunity to create a new list of translators and associated languages when you're configuring the workflow's settings. This list is used when a new document is added to the document library and its language is set. Translators are assigned to source and destination languages.

Three-state

The Three-state workflow is included as a generic state-machine workflow that has three states [5]. Most likely, this workflow type will be associated with an Issue Tracker list because a simple issue tracking process would commonly have three states. For example, if an issue representing a software bug is created, it generally starts out in a state called something like New or Active. After a developer sees and fixes the documented issue, he or she generally changes the issue's status to Fixed or Resolved. The issue's originator then tests the software to make sure the bug was truly taken care of. If everything looks good, the issue transitions to a closed state.

4 RUNNING WORKFLOWS

There are different ways to start SharePoint workflows. In addition, there are different places where workflows can be initiated. Based on the settings for a workflow, an instance can be started automatically when a new item is added to a list or a new item of a certain content type is created, when an existing item is modified, or when a user chooses to explicitly start a new instance. All three of these options are mutually exclusive and can be turned on and off independently.

5 APPLICATION TEMPLATES

SharePoint Application Templates [6] are custom designed scenarios, tailored to different needs and requirements of business processes or task groups in any organization. They also represent a springboard for developing further, in-depth solutions in SharePoint environment. Application templates are closely integrated with SharePoint platform and offer additional functionality, such as lists of relevant themes and modified workflows. Here, we cannot talk about final applications, but rather a good foundation for it. The essence of the proposal is therefore to prepare the basis on which we can build interesting and successful solutions.

Application templates are divided into two categories: site templates and server administrator templates.

5.1 SERVER ADMIN TEMPLATES

These templates, compared to the site templates, are even further advanced in terms of integration into the business domain; they apply specifically tailored lists and some of them include different workflows. The following subsections briefly describe all twenty server admin templates issued by Microsoft.

a) ***Absence Request and Vacation Schedule Management*** template helps employees manage requests for out of office days. Team members post days they will be unavailable and can use the template to assign their responsibilities to others during those days.
b) ***Budgeting and Tracking Multiple Projects*** template helps manage lists of projects and their related budgets, tasks, issues and milestones. It allows users to enter project information, including estimated start and end days, project health, project owner, percent complete and budget information in both currency and days.
c) ***Bug Database*** template helps software development teams collect and track bugs in their code. The template allows bugs to be logged by both end users and members of the development team including collection of supporting data such as Reproduction Steps, Related

Bugs, Categories and related screenshots, documents or other files.

d) ***Call Center*** template helps teams manage service requests originating outside the company from inception through to cause identification, initial solution, escalation and resolution.

e) ***Change Request Management*** template helps teams manage lists of projects and their related budgets, tasks, issues, milestones and change requests. Users enter project information, including estimated start and end days, project health, project owner, percent complete and budget information in both currency and days.

f) ***Compliance Process Support Site*** template helps both teams and executive sponsors manage compliance implementation endeavors. The site allows users to specify compliance regulations that must be met as well as projects that are required to meet those regulations.

g) ***Contacts Management*** template is used by groups or teams who wish to maintain a common set of contact names, phone numbers and email addresses.

h) ***Document Library and Review*** template helps teams manage the review and feedback on documents. The template allows team members to upload documents, automatically retain old versions and check the document out for editing offline.

i) ***Event Planning*** template allows event attendees to register for sessions offered during an event. All visitors have their profile which automatically classifies them in the appropriate groups for events and sessions.

j) ***Expense Reimbursement and Approval Site*** template helps companies track and organize employee expense reports. Employees create an expense report listing trip information and expenses.

k) ***Help Desk*** template is similar to the Call Center template with the difference that end users rather than team members initiate a request by visiting the Help Desk Windows SharePoint Services site and completing a Service Request.

l) ***Inventory Tracking*** template helps organizations track inventory levels by capturing manual input of sales and new inventory purchases from suppliers.

m) ***IT Team Workspace*** template helps IT teams manage the development, deployment and support of software projects, tracking of bugs and provides functionality to help simplify the management of IT helpdesks.

n) ***Job Requisition and Interview Management*** template helps recruiters streamline the process of managing the creation of job requisitions, coordinating interviews, collecting interview feedback and tracking the results of the interviews all in one place.

o) ***Knowledge Base*** template provides a place where team members can upload and tag documents so that others can more easily find and learn from others in their organization.

p) ***Lending Library*** template helps manage the physical assets in an organization's library. The application template tracks general properties about the physical assets as well as which user has currently checked out the asset.

q) ***Physical Asset Tracking and Management*** template helps teams manage new asset proposals and track the status of existing assets. Using the built in workflow, an asset manager can approve or deny requests for new assets made by others within the organization.

r) ***Project Tracking Workspace*** template is designed to help teams list and view project issues, tasks and status from a central location.

s) ***Room and Equipment Reservations*** template helps teams organize the use of their shared resources. The site tracks room and equipment reservations, helping team members find times when specific types of equipment are available.

t) ***Sales Lead Pipeline*** template helps sales teams track the status of sales leads, opportunities and closed or lost deals.

6 INTRODUCTION OF TEMPLATES TO AN EXISTING COMPANY

For the purpose of this article let us consider an example company such as the Institute of informatics at Maribor, Slovenia.

Institute of Informatics in Maribor operates under the Faculty of Electrical Engineering and Computer Science at the University of Maribor. The Institute of Informatics has ongoing pedagogical and research activities. For their efforts to become even more competitive and attractive to businesses, it requires constant inspiration for adequate solutions within their operations. With this in mind we will present a proposal for a deployment and use of certain application templates to the Institute of Informatics.

In the first step we investigated the usefulness of the following proposal: ***Vacation and absence request, Contacts management, Event planning, Lending library, Physical asset tracking and management and Room and equipment reservations.***

Since the Institute of Informatics staff has nearly 50 employees, a need for more transparent control over vacation and absences arises. Many employees attend various conferences and business meetings, so the need is even greater. The template Vacation and absence request would provide a clearer and more comprehensive overview of who has asked for time off and why, and when is the time off supposed to be exercised. Employees could also assign their responsibilities to others while they are away. Various groups with different roles and access rights could be created, for example: executives could approve or reject a claim for absence; other users would be able to follow the status through the central calendar and thus have an overview of all missing personnel. This could lead to easier work distribution and less work arrears.

Contact information is often needed because of the cooperation with others so it would be reasonable to introduce a Contacts management template, which would store all the contact information arranged by category in one place. For organizing and attending the conferences and other various events an Event planning template would be a smart decision. Users could organize visits to the activities and events in their own calendar. They would be divided into different roles; each of these individuals would have access rights to different personalized pages (speakers, listeners, staff, supplier, etc.). The templates Lending Library and Physical asset tracking and management would take care of any needed or borrowed resources.

Users could give requests for new funds and managers would approve or reject them. All users would have an overview of the physical resources and could better coordinate their hiring; those who do not return items on time would be automatically notified of the expiration date. Because of the fact that the institute also conducts pedagogical process, the last template in the first part, the Reservation of rooms and equipment, is crucial. Designing the schedule for computer classrooms and seminar rooms usually accompanies certain overlapping or misunderstandings, so any regulation in this area would certainly be accepted with gratitude.

This would complete the first phase of the adoption of the templates, which included all those who are best suited for their business.

Assuming that the introduction of the first phase was successfully adopted by all members and users we could continue with the second template package, namely: ***Project tracking workspace, IT Team workspace, Budgeting and tracking multiple projects, Bug database, Change request and Inventory tracking.***

Main goals and guiding activities here are project management and all the steps that are associated. The situation, when carrying out several projects at once, is common in modern day companies so special care should be taken in terms of monitoring financial resources and respecting agreed deadlines. There are many methods and strategies in the management of projects [7], some are more accurate and more detailed than others and with application templates, and the SharePoint environment offers both simple and transparent but very powerful functionality of the project management. This package covered project management and related activities, leaving aside the last set of templates, which would be introduced gradually. The templates that mainly relate to supporting activities are: ***Call center, Compliance process, Document library, Expense reimbursement and approval, Help desk, Job requisition and interview management, Knowledge base and Sales lead pipeline.***

The Institute of Informatics is aware of the importance of knowledge and its proper storage and transmission plays an important role. The template Knowledge base captures knowledge and presents it in a central site via various categories, where users can search for relevant information. The Document library template would house a list of all participants and their borrowed material along with the return dates. As in every business or organization, certain costs arise; let it be travel or the costs of additional courses. The Expense reimbursement and approval template would take care of the requirements for reimbursement and approval of cost recovery. Users would add requirements to the central site. Managers could approve or reject applications and also publish relevant policies and arrangements related to disbursing resources. Without employees of course this whole thing would not work so it is necessary to hire new personnel, for which the template Job requisition and interview management is intended. After obtaining the candidates, further discussions and interviews would be carried out and results could be monitored through a dedicated bulletin board. Last but not least, to keep an overview of the potential sales opportunities, finished and failed business proposals, a template called Sales lead pipeline could be used. It would also be possible to see which operations have failed and find the causes of failure. The last set of templates would complete the process of introduction, which would also contain training courses for all users. Only with a thorough knowledge of the environment the benefits could be shown.

7 CONCLUSION

In this paper we presented a brief introduction to the world of SharePoint and its workflow based nature. Application templates, discussed in the article are just the basis for customization, optimization and utilization of the business processes that companies try to implement. Workflows have an increasing role in modern day business processes and as such are subject of special attention. With the development of the organizational portals such as Microsoft SharePoint workflows receive a powerful runtime environment for further development.

References

[1] Bruce Bukovics, Pro WF Windows Workflow in .NET 3.0, ZDA, 2007

[2] Windows IT Pro: SharePoint Everywhere, available at: http://www.windowsitpro.com/article/news2/sharepoint-everywhere.aspx, accessed August 2010.

[3] Amanda Murphy, Shane Perran, *Beginning SharePoint 2007: Building Team Solutions with MOSS 2007,* Wiley Publishing, Indianapolis, Indiana, ZDA, 2007

[4]Todd Kitta, *Professional Windows Workflow Foundation*, Wiley Publishing, Indianapolis, Indiana, ZDA, 2007

[5] Ken Scribner, *Windows Workflow Foundation, Step by step*, Microsoft Press, Washington, ZDA, 2007

[6] Microsoft, Application Templates for Windows SharePoint Services 3.0, available at: http://technet.microsoft.com/en-us/windowsserver/sharepoint/bb407286.aspx, accessed August 2010.

[7] Wikipedia, Project Management, available at: http://en.wikipedia.org/wiki/Project_management, accessed August 2010.

THE MB-UIDE AS AN EXTENSION OF A UML COMPLIANT MODELING TOOL

Jan Kryštof, Martin Cvingráf, Jiří Brázdil, Radek Leifer
Department of Informatics
Faculty of Business and Economy, Mendel University in Brno
Zemědělská 1/1665 613 00 Brno, Czech Republic
Tel: +420 545 132 241; fax: +420 545 132 245
E-mail: jan.krystof@mendelu.cz

ABSTRACT

This paper presents a solution to the construction of a Model-Based User Interface Development Environment which is built as an extension of a UML compliant adaptive modeling tool. Such environment construction is cheap and effective since standard CRUD operations are present and developers can entirely focus on the MB-UIDE desired functionality. This paper presents the architecture of such a system and other highlights relevant to the concept of adaptive modeling tools.

1 INTRODUCTION

Nowadays, there are many modeling approaches dealing with user interface modeling and related development. They have several characteristics such as notation, various types of models and of course software environments which support the entire development process. Since all approaches are different, they all use some specific environment for modeling and development. This type of environment is generally referenced as a Model-Based User Interface Development Environment (MB-UIDE). There are two fundamental requirements on the MB-UIDE defined by Schlungbaum [8]:

(1) MB-UIDEs must include a high-level, abstract, and explicitly represented (declarative) model related to the interactive system to be developed (either a task model or a domain model or both).

(2) MB-UIDEs must exploit a clear and computer-supported relation from (1) to the desired and running user interface. That means that there is some kind of automatic transformation like knowledge-based generation or simple compilation to implement the running user interface.

These requirements point to the complexity of such environments and therefore they are hard to construct. However, there is a general attempt in software production to build inexpensive environments providing appropriate features for the given scope of interest. This paper deals with the construction of the MB-UIDE according to the proposal of Kryštof and Procházka [6], which is based on utilization of an adaptive modeling tool (see fig. 1). This paper gives an overview of current MB-UIDEs in its section two. The section three describes the MB-UIDE architecture which is based on adaptive modeling tools and describes the final implementation. The section four brings information about an experimental implementation issue and related research.

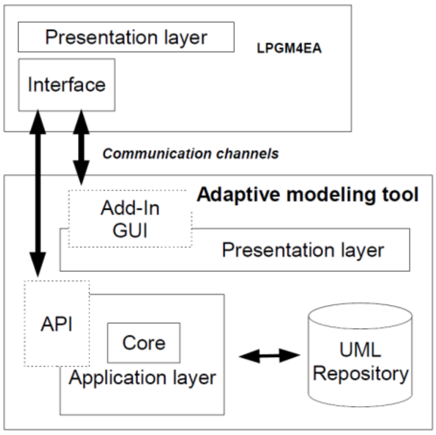


Figure 1: *The extention concept of an adaptive modeling tool*

2 CURRENT MB-UIDEs

Current modeling tools can be characterized with the respect to their origins. Some tools have been implemented as an extension of existing software while the rest has been implemented as a completely new product.

The group of environments which do not reuse any existing product can be represented by tools such as HUMANOID [9], Teallach [4], AUI [7] or by the solution presented by Blankenhorn and Jeckle [3]. Except for the last representative, those three environments support modeling approaches based on non-standardized notations. This can be an acceptable reasoning why authors developed completely new products. However, the environment created by Blankenhorn and Jeckle supports standardized UML modeling and could have been implemented as an extension of some existing software.

The other group is made up of MB-UIDEs such as XMobile [10], WebRatio [1] or ArgoUWE [2]. These tools are implemented with the use of open-source software such as Eclipse and ArgoUML. The Eclipse is an integrated development environment which can be used for different purposes. However, the Eclipse can be extended by Eclipse Modeling Framework and Graphical Modeling Framework and can be turned into a common UML compliant modeling tool. The Eclipse platform is utilized, for instance, by the WebRatio and XMobile MB-UIDEs while ArgoUWE is constructed as a reimplementation of an open-source modeling tool ArgoUML.

The last group is represented by MB-UIDEs which do not provide a common graphical environment and the development depends on text-based tools. This is the case of the SUPPLE or UIML approach. The graphical MB-UIDE for the former approach is called LiquidUI which is currently under development. However, this sort of environments is considered more as experimental rather than usable.

An overview of the chosen MB-UIDEs is presented in table 1 where modeling approaches and related software are mentioned. The column "functionality of MB-UIDE" captures, whether the given MB-UIDE satisfies two requirements given by Schlungbaum. If those requirements are not satisfied, the environment is capable only of modeling. The column "Core" contains the name of software, which is the MB-UIDE based on.

Table 1: *Comparison of current MB-UIDEs*

Approach	**MB-UIDE**	**Functionality of MB-UIDE**	**Core**
WebMl	WebRatio	Yes	Eclipse
UWE	ArgoUWE	No	ArgoUML
TEALLACH	TEALLACH	No	own
AUI	AUI	Yes	own
HUMANOID	HUMANOID	Yes	own
TERESA	TERESA	Yes	own
XMOBILE	XMOBILE	Yes	Eclipse
SUPPLE	---	Yes	---
UIML	LiquidUI	Yes	own
UMLi	ARGOi	No	ArgoUML
Blankenhorn	GuiLayout	No	own

The main advantage of utilization of third party components is the great speedup of the implementation phase when the entire MB-UIDE is being constructed. Since all fundamental components are available, there is not necessary to implement, for instance, CRUD operations, which are needed for the complete model data management or own user interface layer that is necessary for usable work with the environment. However, Kryštof and Procházka pointed out, that this way of constructing MB-UIDE can get more effective by focusing on current general-purpose modeling tools satisfying following requirements:

- The tool must provide an API (Application Programming Interface) that enables access to the model repository and consequent data processing.
- The API must provide a mechanism for establishing a communication channel in order to observe which action is being performed on the model data (element creation, model deletion, etc.).

In other words, the proposed concept employs modeling tools providing a plug-in mechanism which is a common way for adding a new functionality into software applications. The plug-in mechanism usually enables also adding some simple user interface into the original environment in order to make the new functionality accessible through a set of controls. However, the rich user interface is necessary to support common features of MB-UIDEs and therefore it is not possible to take use of the user interface offered by the plug-in. The solution suggested by Kryštof and Procházka proposes to implement the MB-UIDE as a child process of the modeling tool, particularly as the child process of the plug-in and therefore the MB-UIDE can run in own windowing system along with the original modeling tool. The next section describes the architectural design of the MB-UIDE as well as some aspects of current modeling tools which impact the resulting architecture.

3 THE ANALYSIS AND ARCHITECTURAL DESIGN OF THE MB-UIDE

Since our research is focused on development of modeling approach which is compatible with the UML, we investigated several UML compliant modeling tools in order to implement the MB-UIDE according to the proposal of Kryštof and Procházka. Afterwards, we created an architectonical design for the MB-UIDE.

3.1 The analysis of UML compliant adaptive modeling tools

The following modeling tools were investigated during the analyses: Enterprise Architect (Sparx), Poseidon UML (Gentleware), Rational Rose (IBM), Star UML (group of authors) and Visual Paradigm (Visual Paradigm). With the respect to their features, we can declare that all of them fulfill requirements on the adaptive modeling tools, i.e. they provide an appropriate API and a mechanism which enables notifying clients when an observed event occurs.

The result of the investigations is depicted in table 2 where two important aspects are included. Firstly, we included methods how API is used in order to process model data. The API can be delivered either as a native SDK and can be accessed directly or via the COM layer. However, the access to the API via the COM layer requires a COM bridge implementation (e.g. commercial JIntegra or open source Jacob for Java implementations) and is not straightforward compared to the native API and implies an extra work. The second aspect which we focused on was the ability to access UML repository in off-line mode. The off-line access enables to manipulate model data without a running instance of the modeling tool and therefore the model data can be processed as a batch job. This can be considered an important advantage: since transformations of large model data are time consuming, some automation is appropriate. The tools which do not support off-line model data processing (Poseidon UML, Start UML and Visual Paradigm) allow access to the model repository during the run-time of the modeling tool. Therefore, the only way how to obtain the access to the model data leads to a utilization of the plug-in mechanism.

Table 2: *Relevant features of MB-UIDEs*

Modeling Tool	Access to API	Off-line processing
Enterpise Architect	COM, Java	Yes
Poseidon UML	Java	No
Rational Rose	COM	Yes
Star UML	COM	No
Visual Paradigm	Java	No

3.2 The architectonical design of the MB-UIDE

The architectural design of the MB-UIDE is captured in figure 2. The MB-UIDE consists of two modules MModelerConnection and MMBUIDE which provide the access to the model repository and perform particular MB-UIDE compliant functionality respectively.

The module **MModelerConnection** relays on an instance of the modeling tool which is connected with an open project containing model data. The MModelerConnection does not perform any application logic related to modeling or transformations. The main responsibility of the module is to listen to the state of the model repository. This particularly means checking whether any model element was created, deleted or modified. If any observed event occurs, corresponding information is propagated to the module MMBUIDE where the information is interpreted and particular reaction is invoked. The module consists of two components:

- The Plugin component is a regular plug-in and the first meeting point between the original modeling tool and the MB-UIDE. The Plugin depends on a running instance of the modeling tool.
- The RepositoryProvider component provides data model repository of a particular project to external clients. This component does not depend on any running instance of the modeling tools and therefore can be used for off-line jobs. The usual off-line client is a batch process.

The module **MMBUIDE** is responsible for interpretation of the state of the model data repository. This task includes, for instance, decorating the newly created model elements by appropriate tagged values or checking whether the updated model satisfies rules of a given meta-model. The entire module is split into three components:

- The CComChannel represents a bidirectional communication channel between the MMBUIDE and MModelerConnection. This enables to send messages from the MB-UIDE to the native user interface of the modeling tool as well as receive messages from the MModelerConnection which indicate a particular change of model data repository. The CComChannel is actually an entry point for the sub process of the plug-in and manages the run of the MMBUIDE component
- The CUILayer provides a user interface layer to the MB-UIDE. The entire functionality of the MB-UIDE is accessible from a windowing system which is created and managed by this component.

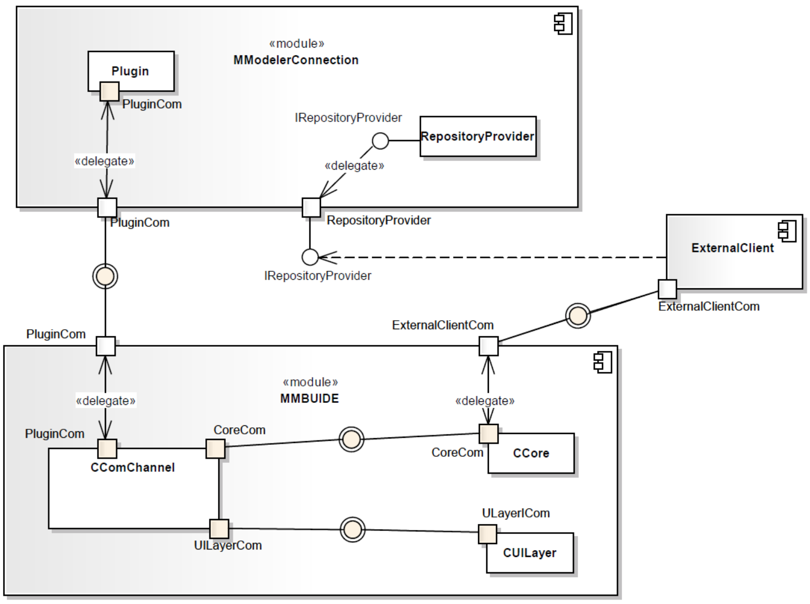


Figure 2: *The component model of the MB-UIDE architecture*

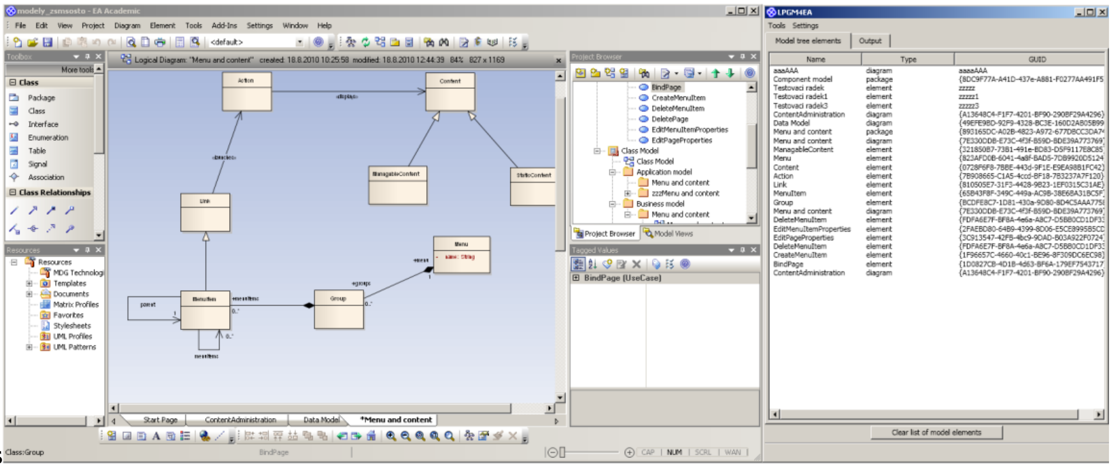


Figure 3: A s*creenshot of Sparx Enterprise Architect modeling tool (on the left) and the LPGM4EA (on the right)*

- The CCore represents the real core of the MB-UIDE. This component contains the whole functionality, especially methods for model transformations and checking the model-consistency. This component also includes common resources necessary for model based development such as templates for model-text transformations, transformation mappings, naming rules or meta-model related information (e.g. UML profiles, constraints etc.).

4 RESULTS AND RELATED RESEARCH

The architectonical design included in this paper was used for implementation of the MB-UIDE which is an extension of the Enterprise Architect modeling tool. The Enterprise Architect supports offline processing of UML data as well as provides the native API for Java and .NET. The implemented MB-UIDE is called LPGM4EA and its goal is a software support of the LPGM methodology [5]. This methodology was proposed for the designing and development of a graphical user interface. The experimental implementation is depicted on the figure 3.

6 CONCLUSIONS

This paper has presented the architectonical design of the MB-UIDEs which can be implemented as an extension of an adaptive modeling tool as it is proposed in [6]. This design offers an inexpensive solution which can be implemented in short time period because it gains a common modeling functionality from the underlying modeling tool. In our research we developed a MB-UIDE called LPGM4EA which is an extension the Enterprise Architect and proves that the concept [6] can be implemented and utilized as an MB-UIDE.

Acknowledgment

The paper is written as a part of the solution for IGA 09/2010.

References

[1] Acerbis, R.; Bongio, A.; Butti, S.; Ceri, S.; Ciapessoni, F.; Conserva, C.; Fraternali, P. & Toffetti, G. (2004), WebRatio, an Innovative Technology for Web Application Development, in 'ICWE', pp. 613-614.

[2] Alexander Knapp, Nora Koch, F. M. & Zhang, G. (2003), 'ArgoUWE: A CASE Tool for Web Applications', [online] http://www.pst.informatik.unimuenchen.de/~kochn.

[3] Blankenhorn, K. & Jeckle, M. (2004), A UML Profile for GUI Layout., in Mathias Weske & Peter Liggesmeyer, ed., 'Net.ObjectDays', Springer, , pp. 110-121.

[4] Griffiths, T.; McKirdy, J.; Paton, N. W.; Kennedy, J. B.; Cooper, R.; Barclay, P. J.; Goble, C. A.; Gray, P. D.; Smyth, M.; West, A. & Dinn, A. (1998), An Open-Model-Based Interface Development System: The Teallach Approach, in 'DSV-IS (2)', pp. 34-50.

[5] Kryštof, J. (2010), 'An LPGM method: Platform Independent Modeling and Development of Graphical User Interface', Informatica, in press.

[6] Kryštof, J. & Procházka, D. (2009), Rozšíření UML modelovacích nástroju pro potreby vývoje grafických uživatelských rozhraní, in 'Objekty 2009'.

[7] Lu, X. & Wan, J. (2007), Model Driven Development of Complex User Interface, in 'MDDAUI'.

[8] Schlungbaum, E. (1996), 'Model-based User Interface Software Tools - Current state of declarative models', Technical report, Graphics, Visualization and Usability Centre, Georgia Institute of Technology, GVU Tech Report.

[9] Szekely, P. A.; Luo, P. & Neches, R. (1993), Beyond interface builders: model-based interface tools, in 'INTERCHI', pp. 383-390.

[10] Viana, W. & Andrade, R. M. C. (2008), 'XMobile: A MB-UID environment for semi-automatic generation of adaptive applications for mobile devices', J. Syst. Softw. 81(3), 382--394

STRUCTURED APPROACH FOR GATHERING USER STORIES

Marina Trkman, Viljan Mahnič
Information systems laboratory, Software engineering laboratory
Faculty of computer and information science
Tržaška 25, 1000 Ljubljana, Slovenia
e-mail: marina.trkman@gmail.com

ABSTRACT

Our research focuses on the efficient gathering of user stories at the beginning of a software development project. This is of great importance from the point of planning the project. Based on authors' experience gained on the projects of business process modelling and on agile development, we introduce an idea how to use BPMN models to get the list of user stories. We provide a recipe how to convert activities into user stories. In such way the list of user stories is defined at the beginning of agile development project and only minor fluctuations during the project are expected.

1 INTRODUCTION

When asking the question: which of the types of software development is more successful, the traditional or agile, the answer is agile. But using that type itself does not bring success. The project must have a proper dose of agility and discipline [1]. Since the planning of the software development project depends on initial list of user stories, we looked deep into the crucial phase of gathering them.

Recent theory considers a story-writing workshop as the best option for gathering user stories. It has a weakness: it can become disorienting when pursuing each path of user stories to its end. Consequently, the list of user stories is incomplete all the way through the project. To overcome that weakness we propose the usage of BPMN models of business processes which provide an understanding of business flow and help with identifying the user stories.

The paper starts with an explanation of user stories, what does it take to make good ones and which techniques of gathering them are offered. Then we continue by defining the problem regarding to planning of the project. Next, we show how BPMN models of business processes can be used to overcome that problem.

2 USER STORIES

It seems easy to think that if everything is written down in specifications and agreed to then there can be no disagreements, developers will know exactly what to build, testers will know exactly how to test it, and, most importantly, customers will get exactly what they wanted. Cohn [2] is stating that this is wrong: customers will get the developers' interpretation of what was written down, which may not be exactly what they wanted.

Writing things down has advantages: written words help to overcome the limitations of short-term memory, distractions, and interruptions. On the other hand many sources of confusion exist from either the imprecision of written words or from words with multiple meanings. These problems fade away if we shift the focus from writing requirements down to talking about them. Naturally, some of the problems of our language exist with verbal as well as with written communications, but when developers and customers (the ordering company who will own the software and the employees who use the software) talk there is an opportunity for a short feedback loop. That is why agile development with user stories is nowadays popular. They prioritize communication with the customer throughout the development process. This is a drastic change compared to traditional developing.

A user story is a short prose specification of system interactions [3]. It is a description of a functionality that is valuable to either a customer [2]. It contains three aspects [2]:

- a description of the story as a reminder of content,
- conversations about the story to flesh out the details and
- tests which convey and document those details.

A goal of the user stories in not to document every last detail about a desired feature but to write down a few short sentences that will remind developers and customers to hold future conversations.

Advantages of the user stories [2]:

- they are comprehensible by all stakeholders;
- they are in the right size for planning releases, programming and testing;
- they are enabling iterative development among developers and future users;
- they are encouraging deferring detail so nothing like specification is needed to start the project;
- they are supportive to opportunistic development over top down approach;
- they are encouraging a participatory design over empirical one;
- they are building up the tacit knowledge because of the face-to-face communication.

2.1 How to make a list of user stories

Ideally the customer writes the stories. On many projects the developers help out, either by writing during an initial workshop or by suggesting new stories to the customer during developing. The responsibility for writing stories resides with the customer and cannot be passed to the developers. Additionally, because the customer is responsible for prioritizing the stories that will go into each iteration, it is vital that the customer understand each story.

In general, there are different techniques for creating a set of user stories such as: user interviews, questionnaires, observation of users interacting with trial releases and story-writing workshops [2]. Interviews are hold with employees (the users of the upcoming information system) which play different roles within the business. The best way for getting the essence of user's needs and an initial list of stories is to use open ended and context-free questions. From those we continue to the specific open-ended questions. To get more information about previously defined user stories we use questionnaires. With those it is easier to prioritize the stories.

A story-writing workshop is a meeting that includes stakeholders who can contribute to the list of user stories. Cohn [2] suggests that this is the most effective way to quickly trawl for stories. It combines elements of brainstorming and prototyping. The preliminary prototype is built up iteratively during the workshop as the participants brainstorm the things a user may want to do at various points while using the application. The idea is not to identify actual screens and fields as in traditional prototyping but to identify the conceptual workflows. When starting a low-fidelity prototype the first thing to decide is which user role to start with. For each user role we draw a different prototype. We start building it with an empty box and a question what actions can the user do on the main screen of the application. The meeting participants now use the brainstorm technique: they are throwing ideas of user actions which are noted not debated; focus is on the quantity not on the quality. For each user's action a line to a new box is created and a new story is written. Walking through the workflows will help everyone to think of as many stories as possible. Following questions are used to identify missing stories: What will the user most likely want to do next? What mistakes could the user make here? What could confuse the user at this point? What additional information could the user need? Weakness can emerge here: it can become disorienting when pursuing each path of the prototype to its end. The author therefore suggests the following approach: for the first component we write down its salient details and then move forward to a component connected to it and do the same. Then move to a component connected to that one rather than going back. When the stories are obtained the prototype should be put away because it is not a long-term artifact of a development process.

2.2 Problems with gathering user stories

Lee at al. [4] explored gathering of agile requirements specifications. Their research is based on a weakness of agile practices. Without a proper mechanism to capture and organize requirements it is impossible to attain full potential impact in the delivery of customer satisfying solutions. User stories gathered at the beginning are important for planning the resources and estimating the time line of project development. In order to have a good plan of project development, we need to identify the general list of user stories as early and definitely as possible. Cohn [2] detected the possibility to add/delete or change user stories during the project as positive. But we are stating differently. If user stories are not (at least) in majority defined at the beginning it is not possible to predict the duration and the scope of the project.

Most requirements elicitation for agile development takes place in different forms. One of them are the models of business processes. We present the area of business process management through the eyes of Kovačič [5] and propose a solution for structural gathering of user stories through BPMN models of business processes.

3 BUSINESS PROCESS MANAGEMENT (BPM)

Business process management presents a whole of logically connected operational and control procedures and activities which result in planned business products or services [5]. Based on the input they produce a certain output which presents a business effect for the internal or external customer. There are three kinds of business processes [5]: support, core and key business processes. They are divided according to the importance for the business. Core processes bring a certain added value for the external customer, while support processes operate as a support for the core and key processes. Key processes have a feature of support processes but have a higher importance for the existence of the business.

Modeling of business processes covers designing, making and using of a model's components. Models are images of real world. Modeling presents the reality from a specific point of view for the specific purpose. There are three basic reasons for modeling business processes [5]: documenting (requirement for example International Standardization Organization 9001), analyzing and providing information systems support. Modeling is the second step of the business process reengineering [5]. Understanding is the first step and the crucial one for finding out how the job is done and how activities complement to create added value for the customer. Within modeling we make a snap shot of the currently practiced business processes ("as-is" processes). Findings from the previous phase are put down in "as-is" models which are drawn with one of the next techniques: Data Flow Diagram (DFD), Unified Modeling Language (UML), Business Process Modeling Notation (BPMN) etc. The third step is simplifying in order to achieve higher efficiency and better performance of processes. The expected results are shorter time of waiting, lower inventory costs, less administration, etc. The fourth and last step is optimization. It results in standardization of processes. We present the desired state with "to-be" business process models. Therefore, for each "as-is" process which we want to optimize we draw a "to-be" model. That way we show where we are now and where we want to be with the advanced use of information technology (IT). The main goal of IT is to give support to the business with automatization.

With models of business processes the developers gain a good understanding of the business. Through interviews with users and formal documentation (instructions, manuals, laws…) models show a step by step recipe how to create added value for the business. Models of business processes are a good start for any IS development [5].

3.1 Creating user stories from BPMN activities – the new idea presented in this paper

BPMN is a graphical notation intended specifically for business processes while others techniques make priority to different aspects of modeling [6]. An activity is one of the main elements of the business analysis and presents a logical conclusion of tasks executed by the employee in a certain department. For creating the names of user stories out of activities we need some of the attributes a BPMN activity provides:

- a lane (or a pool) to expresses the role of an employee who executes the activity;
- the name of the activity to expresses the function which needs to be developed;
- the name of the process (to which the activity belongs to) to expresses the added value of the function for the business.

After we have obtained the information for activities which we want to support with IT, we use Connextra's template [2] to generate good names for user stories: "I as a (role) want (function) so that (business value)". Example is provided in Table 1.

Table 1: *An example how to create a user story out of BPMN attributes*

BPMN attributes of an activity			**The name of an user story**
A lane or pool	The name of an activity	The name of a process	
Student officer	Checking student's condition s of entry into the incoming school year	Student enrollment	I as a Student officer want to check students' conditions of entry into the incoming school year so that I will enroll only those students who have the right to be enrolled.

4 CONCLUSION

As the customer is responsible for defining stories, it is important for the success of the project that the customer knows what he or she wants and is ordering. The comparison of "as-is" model and "to-be" model determines the new IS needs. Moreover, with BPMN models of the business it is much easier for developers to ask relevant questions to the users.

Workgroups are in previous theory [2] considered as a preferable way of gathering user stories for agile development. They combine brainstorming and prototyping. BPMN models of business processes can be a better solution because of its structured approach of understanding the business. The advantage of BPMN models is that they are the result of many interviews with employees on different working positions. Our proposed approach enables an easier, more structured identification of user stories and thus a better estimation of the scope of the project.

References

[1] B.W. Boehm, R. Turner. Balancing Agility and Discipline, A Guide for the Perplexed. Pearson Education, Inc. 2004.
[2] M. Cohn. User stories applied: for agile software development. RR Donnelley. 2008.
[3] B. Deckter, E. Ras, J. Rech, P. Jaubert, M. Rieth. Wiki-Based Stakeholder Participation in Requirements Engineering. IEEE Software. 2007. 24 (2). pp. 28-35.
[4] C. Lee, L. Guadagno, X. Jia. An Agile Approach to Capturing Requirements and Traceability. Proceedings of the 2nd International Workshop on Traceability in Emerging Forms of Software Engineering. pp. 47-54. Montreal 2003.
[5] A. Kovačič, V. Bosilj Vukšič. Management poslovnih procesov: Prenova in informatizacija poslovanja. GV Založba. 2005.
[6] OMG. Business process model and notation (BPMN), v2.0. 2010. http://www.omg.org/spec/BPMN/2.0. pp. 550.
[7] B. Fitzgerald, G. Hartnett, K. Conboy. Customising agile methods to software practices at Intel Shannon. European Journal of Information Systems. 2006. 15 (2). pp. 200-213.

COMPARISON OF SOFTWARE METRICS TOOLS FOR .NET

Jernej Novak
University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova ulica 17, 2000 Maribor, Slovenia
Tel: +386 22207187; fax: +386 22207272
e-mail: jernej.novak1@uni-mb.si1

Gordana Rakić
University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
e-mail: goca@dmi.uns.ac.rs

ABSTRACT

In this paper we will compare tools for calculating software metrics. We will examine tools that are available free on the internet and capable of calculating software metrics for C# language. The purpose of this paper is to compare the calculations of software metrics and determine if the tools are providing the same results. We would like to discover if choosing a different tool can provide different results and consequently bring different decisions.

1 INTRODUCTION

To make software that is good and reliable we need a suitable quality model. One of the quality models is the ISO 9126, which among other things, requires software metrics [1][2]. To make a reliable quality model we need tools that can provide us with software metrics calculations. To make a solid and reliable quality model we need tools that can provide us with accurate data.

Software metrics are a mathematical definition for mapping the entities of a software system to numeric metric values. We also assume that software metric tools implement software metrics according to their definitions.

There are quite a few available tools that can provide us with software metric values for C# language or .NET assemblies. In this paper, we will not deal with whether this information is accurate or not but whether different tools give us different results.

The goal of this paper is to answer if metric tools for C# language give us different results. After a review of existing literature, we found a similar paper but this paper was testing tools for the Java language and was using a different comparison technique [3].

In the next section, we will describe what tools we used, what tools we chose for metrics calculations and we will offer a description of the analysis. The third section presents the results and an analysis of them. We conclude with a summary of the results and some questions for further work.

2 DESCRIPTION OF THE STUDY

In this chapter we will describe how we performed the experiment.

All experiments were performed on a standard PC which always met the minimum requirements expected for all software metrics tools.

The data collected from software metric tools was stored for further processing. All results were manually migrated to an Excel file for easier further processing.

First we chose tools for calculating software metrics, afterwards we selected software metrics for testing. Then we chose software on which we will performing metric calculations. At the end of this chapter, we will describe in detail how the analysis of results was carried out.

2.1 Software Metrics Tool Selection

First, we performed a free search on the internet to find available tools for the computation of software metrics. The first criteria was that the tool can calculate any software metrics for C# language or .NET assemblies. We found ten tools, of which five met all the required criteria. The criteria was that the tool is available for free, or that at least a trial version is available for a specific period of time, and that the tool has a graphical user interface from which we could manage the tool. In the following paragraphs we will briefly describe the five chosen tools for our experiment.

CodeMetrics (Add-in for Reflector) analyses and computes several code quality metrics on .NET assemblies. This tool and its source code are freely available on the [4].

NDepend is a Visual Studio tool to manage complex .NET code and achieve high Code Quality. With NDepend, software quality can be measured using Code Metrics, visualized using Graphs and Treemaps, and enforced using standard and custom Rules. Software metrics calculation is also done on .NET assemblies. The NDepend is shareware but is freely available for research projects.[5]

SourceMonitor is a freeware program that lets you look inside your software source code to find out how much code you have and to identify the relative complexity of your modules. This tool runs on C# source code and it is freely available to download over the internet.[6]

Visual Studio Code Metrics is a set of software measures that provide developers with a better insight into the code they are developing. This tool comes free with Visual Studio in a premium and ultimate edition.[7]

Borland Together (add-in for Visual Studio) enables you to analyze, design and implement flexible and maintainable software architectures that can be easily modified as requirements change. It runs on C# source code. It is available for shareware usage for 15 days.[8]

2.2 Software Metrics Selection

First, we made a list of all the available software metrics that the tools provided. Tools provide different set metrics. But after analyzing them, we recognized metrics with different names but with the same definition and calculating algorithms. We decided to choose seven software metrics, of which we could say that they represent the same software metrics definitions. After considering all the parameters we decided to include the following seven metrics in our research:

Maintainability index includes composed metrics. It is calculated with a certain formula from lines-of-code measures, McCabe measures and Halstead complexity measures[9]**.**

Average Cyclomatic Complexity is used to indicate the complexity of the program. It measures the number of linearly independent paths through the program source code [10]**.**

Average Depth of Inheritance Tree (DIT) is the maximum length of a path from class to root class in the inheritance structure of a system [11].

Average Coupling Between Objects (CBO) is the number of classes that a class is coupled to [11].

Lines of Code (LOC) is used to calculate the size of a software program by counting the number of lines in the text of the program or intermediate language [12].

Average Lines of Code represents the average LOC metrics per module.

Average Comment Ratio is calculated from the average lines of code divided by the average comments count per module**.**

For an unambiguous definition for software metrics please refer to the original sources defined in the references.

Tool	SoftwareMetrics						
Program name	**MI**	**avg CC**	**avg DIT**	**avg CBO**	**avg LOC**	**LOC**	**avg COM.**
Visual Studio	✓	✓	✓	✓	✓	✓	✗
Borland	✓	✓	✓	✓	✓	✓	✓
SourceMonitor	✗	✓	✓	✗	✓	✓	✓
ReflectorAddin	✗	✓	✓	✓	✗	✗	✗
NDepend	✗	✓	✓	✓	✓	✓	✓

Table 1: *Tools and software metrics.*

In Table 1, we can see the used tools and metrics for our experiment. The fields selected with a check show that the tool supports selected metrics and the spaces marked with an X show that the tool does not support selected metrics. All tools have a solid coverage of software metrics except the Reflector addin.

2.3 Choosing Software

To choose testing tools, we limited ourselves to testing systems written in C# language. We performed a free search on the SourceForge.net page and randomly chose five different projects. We chose only Windows projects for easier compilation and calculation of software metrics. We were also trying to choose different sizes and projects of varying origins. In the next paragraph we will present the five programs that we chose for our experiment.

CompressFolders is a windows application that allows us to compress folders and files.

FileManager is a Windows application that enables us to manage files, upload and download files from a server and to see the history of uploaded and downloaded files.

Gammit! is a Windows application that allows us to set the gamma level for an LCD monitor. It also corrects the lighting of an LCD monitor.

Palantir is a Windows application that allows us to review all remote computer connections.

ZedGraphWin is a Windows controller for drawing graphs for Windows applications.

2.4 Analysis description

To collect data, we installed all the chosen tools. Between tools there were no dependencies or side effects that we should consider.

The tested software was saved in a dedicated space so that all tools were performing an analysis on the same source code. To avoid unwanted changes or mistakes on the source code, we stored the analyzed source code prior to testing in a read-only drive.

When we installed all the tools and when all the tested software was ready, we ran an analysis with each tool for each tested software. Then we exported all data to MS Excel 2007 and there conducted some further analysis, as described later on in this paper.

Metric **k**

	Project 1	Project 2	...	Project i	Average
Tool **1**	$M_kV_{1,1}$	$M_kV_{1,2}$	...	$M_kV_{1,i}$	$\overline{M_kV_1}$
Tool **2**	$M_kV_{2,1}$	$M_kV_{2,2}$	...	$M_kV_{2,i}$	$\overline{M_kV_2}$
⋮	⋮	⋮	⋱	⋮	⋮
Tool **j**	$M_kV_{j,1}$	$M_kV_{j,2}$	...	$M_kV_{j,i}$	$\overline{M_kV_j}$

Table 2: *Calculation of average values by tool for multiple projects.*

All stored data was saved to a dedicated excel file. As shown in Table 2, we calculated the average metric values for every metric. So we had seven tables with data, each for every software metric. Most data collection and data organization was manual. We calculated the average values for each tool from 1 to **j**. For each tool and for each software metric we got average values from $\overline{M1V1}$ to $\overline{MkVj}$, where k is the number of software metrics and j is the number of tools. From these values we then made the vectors $\vec{T}_1$ until $\vec{T}_j$ and with these values we made T-Tests.

$\vec{T}_1 = (\overline{M_1V_1}, \overline{M_2V_1}, \dots, \overline{M_kV_1})$
$\vec{T}_2 = (\overline{M_1V_2}, \overline{M_2V_2}, \dots, \overline{M_kV_2})$
⋮
$\vec{T}_j = (\overline{M_1V_j}, \overline{M_2V_j}, \dots, \overline{M_kV_j})$

Table 3: *T vectors.*

Table 3 shows the average values for the software metrics for each tool. Because not all tools supported all applicable software metrics, we had to alternate some of the vectors $\vec{T}$ that we could compare with the pair T-Test.

Tools	**Tool 1**	**Tool 2**	...	**Tool j**
Tool 1		T-Test($\vec{T}_1$, $\vec{T}_2$)	...	T-Test($\vec{T}_1$, $\vec{T}_j$)
Tool 2			...	T-Test($\vec{T}_2$, $\vec{T}_j$)
⋮				⋮
Tool j-1				T-Test($\vec{T}_{j-1}$, $\vec{T}_j$)
Tool j				

Table 4: *Tool comparison table.*

In Table 4, we can see a comparison of each tool with each other with a T-test. Each vector $\vec{T}$ from one tool we compared to another tool. After conducting every comparison, we got metrics as seen in Table 4. For significantly different results, we will take those that are in an 80% of interval of trust. We decided on the value 80% after reviewing the existing literature review [13]. This means that two tools will be different if their values of T-Test are under 0.05.

3 RESULTS AND ANALYSIS

A quick glance at the results confirmed that results can differ significantly between tools. We also calculated standard deviations for each metric. In Table 6, we can see values differing especially for LOC software metrics. The most consistent results were seen with DIT software metrics. A very significant deviation in LOC software metrics can be assigned to many different poorly defined LOC software metrics.

	MI	avg CC	avg DIT	CBO	avg LOC	LOC	avg com
Average standard deviation	7.5	6.9	0.8	7.6	49	3104	16

Table 5: *Average standard deviations for specific metrics.*

3.1 T-Test results

We can look at the obtained results of the T-Test in Table 6 and if we accept an 80% trust interval, we can see that most of the tools give us significantly different results.

TTEST	Visual Studio	Borland	Source Monitor	Reflector Addin	Ndepend
Visual Studio		0,036	0,037	0,021	0,032
Borland			0,038	0,039	0,037
Source Monitor				0,056	0,038
ReflectorAddin					0,08
Ndepend					

Table 6: *Values of the T-Test for each tool.*

The T-Test fails only on results that give us the two pairs of tools. These are Reflector.CodeMetrics and SourceMonitor pair and Reflector.CodeMetrics and NDepend pair. And even this test failure can be assigned to the small vector $\vec{T}$ $\vec{T}$. Reflector.CodeMetrics had the small vector $\vec{T}$ $\vec{T}$ because it only supported three of the seven tested software metrics. All other tool results were confirmed, so that we can claim that tool results differ significantly.

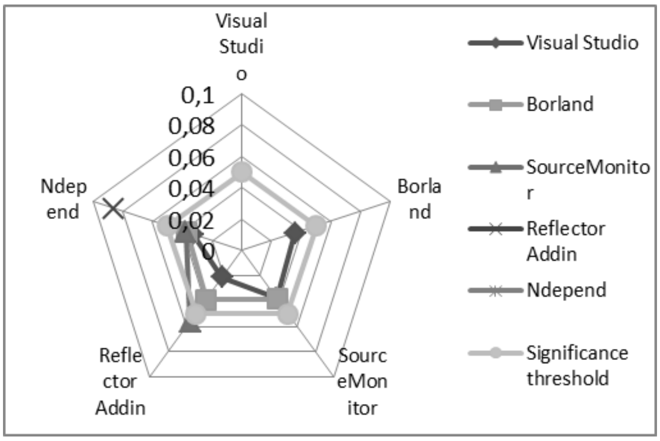


Figure 1: *Polar graph with significance threshold.*

In Figure 1 we can see that the only values of the Reflector and NDepend did not pass the T-Test so we cannot claim that these two tools gave us different results.
We showed that C# metrics tools differ from each other. So we only confirmed what was shown in the article "Comparing software metrics tools" [3] with the difference of using different tools (C# instead of Java) and by applying a different method.

4 CONCLUSION AND FUTURE WORK

Developers and engineers of software must rely on scientific data and this data should be accurate. Especially when these data are the foundation for making important decisions. If we make decisions based on inaccurate data, we can change the development lifecycle and with this we can affect the quality of a software product.
In this experiment we showed that the metrics tool gave us different results for the same inputs. Results can vary so much that engineers can adopt different decisions just because they chose another tool. So choosing the right tool is important. We also must know that we cannot compare metrics data given from different tools. We hope that in the future tools will get more consistent and will give us comparable results.
In this paper we did not answer which tool is better and give us more accurate results. In the future maybe we should mathematically review the metrics and write test cases from which we could then determine if the metrics tool calculated the correct metrics value.
Also, if we keep in mind the conclusions in this paper and the results from [14] we have to consider problems in large projects developed in several languages. In these projects, several tools have to be used to measure all the components of the project. In that case, we have to be prepared to gain inconsistent and incomparable metric results. These conclusions confirm the need for new "universal" software metric tools that would provide comparable results for the whole project independent of the languages used for the development of the project.

References

[1] ISO ISO/IEC 9126-1, *Software engineering - Product Quality - Part 1: Quality model*., 2001.

[2] ISO. ISO/IEC 9126-3, *Software engineering - Product Quality - Part 3: Internal metrics*., 2003.

[3] Jonas Lundberg and Welf Löwe Rüdiger Lincke, "Comparing Software Metrics Tools," *International Symposium on Software Testing and Analysis*, 2008.

[4] CodeMetrics add in Reflector. [Online]. " http://reflectoraddins.codeplex.com/wikipage?title=CodeMetrics&referringTitle=Home

[5] (2010) NDepend. [Online]. http://www.ndepend.com/

[6] SourceMonitor. [Online]. http://www.campwoodsw.com/

[7] Code Metrics Values. [Online]. http://msdn.microsoft.com/en-us/library/bb385914.aspx

[8] Borland Together 2005. [Online]. http://www.borland.com/us/products/together/index.html

[9] M. Vestergren Andersson, "Object-Oriented Design Quality Metrics," *Uppsala University*.

[10] McCabe, "A Complexity Measure," *IEEE Transactions on Software Engineering*, 1976.

[11] C. F. Chidamber S. R. and Kemerer, "A Metrics Suite for Object Oriented Design," *IEEE Trans. Softw. Eng.*, 1994.

[12] Metrics definitions. [Online]. http://msquaredtechnologies.com/m2rsm/docs/rsm_metrics_narration.htm

[13] George P. McCabe David S. Moore, *Introduction to the practice of statistics*.: W.H. Freeman and Company New York, 2003.

[14] Rakić G., Budimac Z., Bothe K., *Towards a 'Universal' Software Metrics Tool - Motivation, Process and a Prototype*, Proceedings of the 5th International Conference on Software and Data Technologies (ICSOFT), 2010

EVALUATION OF SERVICE-ORIENTED SYSTEMS USING SOFTWARE METRICS

Tomaž Korelič
University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova 17, 2000 Maribor, Slovenia
Tel: +386 2 2207352
e-mail: tomaz.korelic@uni-mb.si

Service-oriented systems are currently a widely adopted and modern approach for the development of enterprise systems. Although it has become the predominant architectural style of choice for many organizations, few measurements are being taken and few metrics are used for evaluating those systems and services. Previous research has shown that there are few SOA metrics designed to evaluate complexity, effort estimates and the health status of SOA solutions.
This paper addresses the issue of design time and run time evaluation of SOA systems in the areas of complexity, criticality, reliability, performance and service versioning. We extracted the most important metrics from previous research, which we believed to be crucial for the evaluation of an SOA system. After narrowing down this metric suite, we proposed a composite metric for the easier and more effective evaluation of an SOA system.

1 INTRODUCTION

The service-oriented architectural (SOA) style allows us to build distributed applications, which consist of services – independent software entities, whose functionalities are accessible through a network. Although SOA is widely perceived today, few metrics exist for measuring system complexity, reliability, resource and performance[5]. The evaluation of the system and its services is extremely important. It helps to identify poor design decisions, poor resource performance and shows the complexity of the system. The complexity is crucial in reuse and the further development of services in SOA systems.
System complexity, reliability and performance can suffer in later development phases if such problems are not identified earlier. In this paper, we will extract the most important metrics proposed by previous studies that have evaluated an SOA system in development.
Services can either be atomic or compound (business processes). Business processes will be examined as collaborations of multiple services. Certain metrics can be applied to both types, but also to the system as a whole.
Sections 2 - 6 present certain metrics ranging from the product, resource, flexibility and complexity metrics we have chosen. In each topic we stress the most important metrics that were included in our research. In section 7 we describe our research more specifically and in section 8 we conclude this paper and describe future work.

2 PRODUCT METRICS

A service-oriented system is comprised of nodes and each node contains multiple services with multiple operations. Clients communicate with services via messages synchronously or asynchronously. The services can be atomic or composite (business processes). Based on these properties, this section introduces product metrics. The following Table 1 describes the metrics for complexity, reliability and performance of a certain SOA system[3].

Table 1: Important Product metrics

Metric	*Type*	*Description*
NCY	Complexity	Network cohesion in the system
NSIC	Complexity	Number of services involved in the compound service
SIY	Complexity	Services independence in the system
AIS	Reliability	Absolute importance of the service.
ADS	Reliability	Absolute dependence of the service
ACS	Reliability	Absolute criticality of the service
RC	Reliability	Reliability of the Compound service-

The first three metrics measure the system's complexity and the last four its reliability.
NYC counts the direct (unidirectional) ties between the nodes in the system, where NSIC counts the services involved in the compound service (both directly and indirectly). Similarly, SIY is the count of pairs of services that depend on each other.
Reliability is measured by counting the clients that depend on the service, i.e. which invoke its operation (AIS), the count of other services this service depends on (ADS) and RC is inversely proportional to the weakest link of a chain, i.e. to the reliability of the most critical service that takes part in the compound. Additionally, ACS measures the absolute criticality of a service, with the product of AIS and ADS[3].

3 RESOURCE METRICS

The following section presents the metrics associated with resources in a SOA infrastructure (network, system as a whole, nodes, services, operations). They can be broken down into performance, service versioning and reliability of these resources. Table 2 presents them as the authors in [4].

Table 2: Resource metrics

Type	*Metric*	*Description*
SV	CVS	Count of simultaneously deployed versions of the service.
SV	ALTVS	Average lifetime of versions of the service.
SV	ACSVY	Average count of services' versions in the system.
SV	ALTSVY	Average life time of services' versions in the system.
SV	MCFS	Metadata change frequency of the service.
P	MRY	Overall message rate in the system
P	NTY	Overall network traffic in the system during a unit of time.
R	SLACS	SLA compliance of the service, measured as the fraction of time during which all SLA fulfillment indicators of the service lie in green and/or yellow areas.
R	SLAVDS	SLA violation danger of the service, measured as: *(fraction of time in yellow area) / (fraction of time in green and yellow areas)*
R	FRO	Fault rate of the operation per one unit of time calculated as: *count of faults/count of received messages*
R	FRY	Overall fault rate in the system.

SV: Service Versioning, P: Performance, R: Reliability

A distinguishing feature of SOA systems is the possibility of having different versions of the same service be active simultaneously. If this issue is not addressed correctly, it can have negative consequences of poor versioning organization and service maintenance. Two types of resources are appropriate in this context – the installed service versions and the services' metadata. A metadata repository (or registry) is an essential component of service-oriented infrastructures, serving for loose coupling and dynamic binding, and enabling agility in that way[4].

The metrics considering service versioning measure the count of simultaneously deployed versions of the service (CVS), the average lifetime of versions of the service (ALTVS), the average count (ACSVY) and lifetime (ALTSVY) of services' versions in the system and the metadata change frequency of the service (MCFS)[4].

The performance metrics help us identify performance issues in the network infrastructure. These two included metrics give a "bird's-eye view" of the system's performance. They measure the overall message rate in the system and the overall network traffic in the system during a unit of time.

The other important features of SOA systems are their ability to assure Service Level Agreements (SLAs)[9]. Service –level agreements are parts of service contracts and uniformly describe non-functional requirements of these services. Certain metrics address these requirements and they describe faults that happen in the system[4].

4 FLEXIBILITY AND COMPLEXITY METRICS

Flexibility, agility and complexity metrics, can diagnose the "health" of an SOA solution based on the information of services, its operations, architectural properties, flexibility, agility and complexity.

The proposed metrics in [5] are divided into two categories: the design-time and run-time metric. They result from the experience of multiple SOA developers. We will not get into too much detail for all the metrics presented in [5]. They can be calculated individually, but by combining their results into SCI, SVCI and FAI (table 3) we get a much more simplified look at the flexibility, agility and complexity of the system.

Design-time metrics can be calculated during the identification, specification and realization phases. On the other hand, run-time metrics are derived during implementation and installation [5].

Table 3: Aggregate SOA indexes

SOA index	*Description*
SOA complexity Index (SCI)	Measures a part of an SOA solution complexity including security, management and SOA Governance, all of which offer significant benefits but also increase the complexity of the overall SOA.
Services Complexity Index (SVCI)	Measures complexity, but also looks at the individual complexities of each of the composed services.
Flexibility and Agility Index (FAI)	Tracks the flexibility and agility of the SOA solution.

In order to calculate the three SOA indexes we have to use multiple smaller metrics. The design-time metrics include measurements like the weighted number of exposed interfaces, or operations per service, as defined in the WSDL documents, the fraction of services which are stateless as opposed to stateful; the fraction of transaction-aware services in relation to the overall number of transaction-aware and non transaction-aware services within the SOA solution; the fraction of services that are realized through indirect exposure in respect to the total number of services that are realized using both indirect exposure and direct exposure; and the fraction of tasks as part of a business flow that are manual (human tasks). We will not get into too much detail for these metrics, which were proposed by [5].

These design-time metrics measure the flexibility, agility and complexity of a solution in respect to design time decisions. Metrics are first computed for individual services and then compiled into a more global metric applied to the FAI, SVCI and SCI indices[5].
Another research paper [10] proposes a very similar set of design time metrics measuring coupling, cohesion, design size, complexity, service granularity, parameter granularity and consumability. These metrics measure similar properties for an SOA system as the metrics proposed in [5].
The run-time metrics can help us identify overly complex architectures and provide insights into the flexibility factor of an SOA solution. These are also computed for individual services and then applied to FAI, SVCI and SCI. SOA run-time metrics include, for example, the total number of services that comprise a SOA solution, the fraction of web services which are composite, the number of services that are selected dynamically over the total number of services that are selected dynamically or statically and the total number of versions over the total number of services within the SOA solution[5].
In our research, we only applied the complex FAI, SVCI, and SCI metric for measuring complexity, flexibility and agility of an SOA solution.
Another noteworthy piece of research was conducted in [7], but was not specifically included in our paper. Among other metrics, a set of metrics was presented to measure the sensitivity of the SOA solution to the removal of one or more services.

5 MEASURING COMPLEXITY

The following topic addresses the overall complexity of an SOA system. Of a total of 12 metric presented in [1] we extracted the following 6, which we thought were the most relevant ones. The complexity of an SOA system affects its reusability. The other 6 metrics were simpler and were also needed for the calculation of the six chosen ones. In total, two metrics for the assessment of a system's complexity and four metrics for measuring the addressing of complexity are discussed in table 4 [1]. Another research focuses on the reusability and proposes an overall metric for measuring it[6]. We used one metric out of the research (RE). Table 4 summarizes a total of 7 metrics, which we included in this category.

Table 4: Complexity metrics

Metric	*Full name*
SCF	Service coupling factor
SSC	System's service coupling
EOA	Extent of aggregation
SCZ	System's centralization
DOA	Density of aggregation
ACZ	Aggregator centralization
RE	Reusability of SOA services

SCF is calculated as the sum over all the single COS-values (Coupling of service) of a system's service consumers is set in relation with the maximum couplings that could occur in a system. COS is defined as the count of services a given service calls operations on. SCF is a metric designed for component-oriented systems and that the author in [1] applied to service-oriented systems. Consequently, this metric cannot incorporate service oriented principles.
SSC measures the degree of coupling in a given system with regards to its modifiability.
The EOA metric relates the count of channels between non-aggregative consumers and aggregators with the overall count of channels from non-aggregative consumers to arbitrary service providers. Aggregators are services that act as a service provider and service customer. Important to note is that EOA relies on a system of heavily using consumers. If an application is solely triggered by a service consumer, the result might not be representative of the complete system.
SCZ describes to what extent a system is centralized. Whenever an application uses aggregators in order to adapt to external services, aggregators might be used extensively while the control is centralized in one component. In such cases, the SCZ value can be misleading.
DOA indicates to what extent the aggregation in a system combines more basic services to more complex services. Especially in combination with the SCZ, the DOA value can indicate whether centralization of a system goes along with good aggregation.
ACZ indicates the degree of centralization in a system by considering the use of mediating services. Of course, this metric has also to be carefully applied. This is because it also solely puts an interpretation of an externally visible structure over the actual internal structure of a component that determines the visible part [1].
RE measures the services reusability and is calculated using different aspects, such as business commonality, adaptability, standard conformance, modularity and discoverability [6].

6 COUPLING METRICS FOR SOA

The coupling is presented by the relationship between services, which shows their dependency. The more relationships that are present in a service, the coupling between this service and others will become tighter. The authors in the [2,8] proposed two suites of metrics to evaluate a service's quality to its ability of coupling. We use coupling metrics to measure the maintainability, reliability, testability, and reusability of services. These metrics are operated in run-time, which provide more exact results[2].

Table 5: Coupling metrics

Metric	*Description*
CBS	Coupling between services
IMS	Instability metric for service
DC2S	Degree of coupling between 2 services
DCSS	Degree of coupling within a given set of services
Coupling	Coupling between services and clients

CBS is built directly from the CBO (Coupling Between Objects) metric. It is calculated based on the number of relationships between a service and other services in a system. IMS shows the interaction between a service and the others in a system through sending and receiving messages. The CBS metric is only applied on all of the services in the system. The DC2S metric identifies the relationship between two services to detect their dependency on each other. DCSS reflects the coupling between all services in a system[2]. The metric *Coupling* refers to coupling between services and clients (service customers)[8].

7 SOA METRICS COMPOSITION

Our goal was to extract the most important metrics out of many metric suites, as proposed by the authors in the references. We eliminated the duplicated metrics and, in our estimation, the least important ones. With the remaining 28 metrics, we tried to create a composition of metrics that comprised the remaining ones. The composition was prepared as a tree-like structure, where each main branch contains metrics from each category (complexity, flexibility, reliability). The higher the metric is, the more important we believe it is.

Due to the similarity of reliability metrics in chapter 2 and 6, we used only the ones in chapter 6 for the reliability branch. The metric included in the upper tree branches are the complex SCI, SVCI and FAI indexes as stated in [5]. The upper reliability branch contains metrics that measure the maintainability, reliability, testability, and reusability of services as described in chapters 2 and 6. The child nodes in the complexity branch contain simpler metrics for measuring the complexity of services. Additionally, the child node to FAI contains the relating service versioning and finally the child node to the reliability node contains service-level agreement metrics. A special case in this structure is the performance metric, which we inserted in an independent assistant node. Figure 1 presents the composition of the extracted SOA metrics.

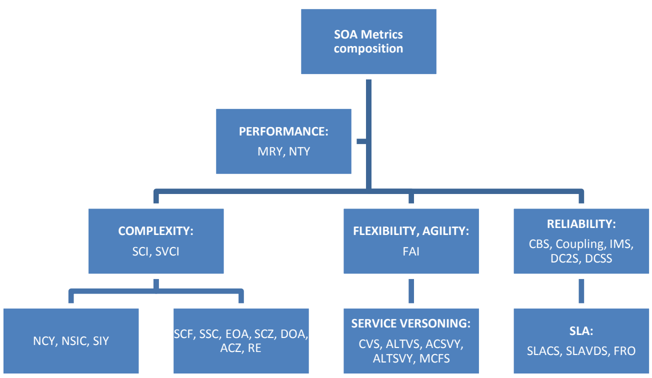


Figure 1: *SOA metrics composition*

8 CONCLUSION AND FUTURE WORK

The metrics presented in this paper are an extraction of initially proposed metrics by other authors. These metrics can help to identify poor design, resource quality, and motivate a re-engineering of a design prior to realizing the application. Due to the high amount of proposed metrics and the duplication of particular ones, we extracted the most important ones and conducted an SOA metric composition in a tree-like structure. One of the problems with these metrics is the difficulty of empirical validation and calibration across a number of different projects. Therefore, we need to weigh each metric to provide more accurate measurements. Additional work is required to empirically validate the composition metric across a broad spectrum of SOA applications. With our research, we tried to narrow down the available SOA metrics to a smaller, more useful set that equally measure SOA flexibility, agility, complexity and performance.

References

[1] H. Hofmeister, G. Wirtz, *Supporting Service-Oriented Design with Metrics*, 12th International IEEE Distributed Object Compution Conference, 2008.

[2] P. T. Quynh, H. Q. Thang, *Dynamic coupling Metrics dor Service-Oriented Software*, International Journal of Computer Science and Engineering 3:1 2009.

[3] D. Rud, A. Schietendorf, R. R. Dumke, *Product Metrics for Service-Oriented Infrastructures*, IWSM MetriKon 2006.

[4] D. Rud, A. Schietendorf, R. R. Dumke, *Resource Metrics for Service-Oriented Infrastructure*, www.cs.uni-magdeburg.de/~rud/papers/Rud-13.pdf, last visit may 2010.

[5] M. Hirzalla, J. Cleland-Huang, A. Arsanjani, *A Metrics Suite for Evaulation Flexibility and Complexity in Service Oriented Architectures*, ICSOC 2008, LNCS 5472, pp. 41-52,2009.

[6] S. W. Choi, S. D. Kim, *A Quality Model for Evaluating Reusability of Services in SOA*, 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008.

[7] M. Monsalve, *Managing the QoS of E-Government: Metrics for Large Scale SOA*, XXVI International Conference of the Chilean Computer Science Society, 2007.

[8] W. Xiao-jun, *Metrics for Evaluating Coupling and Service Granularity in Service Oriented Architecture*, ICIECS 2009.

[9] R. Anantharangachar, P. Anantharamaiah, K. Thalapathy, *An SOA Compass for Next Generation Networks*, IEEE International Conference on Services Computing (SCC'06), 2006.

[10] B. Shim, S. Choue, S. Kim, S. Park, *A Design Quality Model for Service-Oriented Architecture*, 15th Asia-Pacific Software Engineering Conference, 2008.

E-MARKETPLACES AND CUSTOMERS' PERCEIVED VALUE

Dijana Močnik
University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova 17, 2000 Maribor, Slovenia
Tel: +386 2 220 7000; fax: +386 2 220 7272
e-mail: dijana.mocnik@uni-mb.si

ABSTRACT

The focus of this paper is to present the complex nature of customers' perceived value in electronic marketplaces (EMP). Customer value is an antecedent of customer satisfaction and loyalty. A customer's perception or assessment of value is based on results produced and received, in relation to the price and other related costs incurred by the customer in acquiring the product or service. Such an understanding of customer value will contribute to the successful utilization of EMPs. For network participation, customers' perceived positive net value is imperative; however, this value relates to the trade-off between relevant past, present, and future customer benefits and sacrifices in a complex way, which is the reason that to date there are no established measures for customer perceived value in electronic environments.

1 INTRODUCTION

Technological development—particularly the rapid development of information technology—is one of the main forces changing the current approaches to various business activities. New services exploiting electronic marketplaces (EMPs) are constantly being created. The Internet's provision of low-cost, efficient interconnectivity has had a dramatic influence on the way in which online activities (services) are conducted [1]. The opportunity for organizations to manage their own processes using electronic means has led to the development of new inter- and intra-organizational relationships and, consequently, the firm network phenomenon. However, creating commercially viable business services in EMPs is challenging as technological innovations alone do not guarantee commercial success, which explains why EMPs are not as popular as originally anticipated. Therefore, it is vital for EMP promoters to investigate the determinants of EMP customers' perceived value, which helps customers decide whether to use EMPs or not.

The next section reviews the background of the customers' perceived value, followed by a discussion of EMP customers' benefits and sacrifices. Next, the dynamic nature of EMP customers' perceived value is presented. Finally, the co-creation or network aspect of EMPs and their impact on customers' perceived value is discussed.

2 CUSTOMERS' PERCEIVED VALUE IN EMPs

Every transaction between a seller and buyer aims to create (or generate) value for both parties. This very logic is the same whether the transaction is made in the traditional or electronic way. However, the perception of value for customers in EMPs is much more complex than for their offline counterparts. In EMP settings, not only the product or service itself, but also the functioning of the EMP, the Internet channel, the mutual interdependence between different actors within the EMP, and time influence the value perception; this is true for both customers (whether businesses or end users) and other involved partners (e.g., infrastructure providers, network operators, application providers, device manufacturers). Each partner has his/her own perceptions of the benefits he/she may expect as well as sacrifices he/she is required to make. Distinctive characteristics impact customers' decisions to participate in the EMP rather than in the traditional market.

The product, the service, and the relationship have all been used as starting points for customers' value definition. Woodruff's [2] often-cited definition states that a customer's perceived value is that individual's perceived preference for and evaluation of those product attributes, attribute performances, and consequences that arise from use and that facilitate or block the achievement of the individual's expected goals. This definition is primarily valid for products as it is based on the evaluation of the products. Somewhat different is the description of the value of the service, which consists of service quality and customer satisfaction [see, e.g., 3]. Parasuraman et al. [4, 5] suggested that service quality should be defined by the customer with reference to how well the service is delivered during interactions between the service provider and customer. Furthermore, Grönroos [6] suggested that service quality stems from a comparison of what customers expect from the service (i.e., their expectations) to the seller's actual service performance. This means that existing studies on perceived value of services have mostly focused on the value at the level of customers' expectations and perceptions [see, e.g., 2]. The assessment of customers' perceived value of executing or using different activities through EMPs involves more than simply an outcome dimension as the perceived value of the service depends also on technical (what is delivered) and functional (how the service is delivered) quality. Therefore, focusing solely on the value of the product, the relationship, or the value of the service would not provide sufficient understanding of the perceived value for the EMP customer [e.g., 4, 6, 7].

Customers' perceived value thus represents the complex trade-off between multiple benefits and sacrifices that EMPs bring to and gain from customers in comparison to the net value of an alternative (or competition) [e.g., 8–13]. The aspect of sacrifices has been absent from existing research of customers' perceived value of goods and services in traditional markets, where such value was based primarily on benefits. The net value (or give-and-get components) [see 16] highlights not only the benefits gained, but also the sacrifices or investments that an EMP customer has to make in order to gain the benefits. Thus, the better the combination of sacrifices and benefits, the higher the customers' perceived value [e.g., 17].

Measuring customer perceived value is essential in assessing current services and for the development of further ones, because customer segments may have different motives to use services and thus perceive different value in them. Value-oriented segmentation is proven to be better than the traditional approaches based on demographics or tracked buying behavior in EMPs, which are new to customers. However, to date there are no established measures for customer perceived value in electronic environments. The majority of existing empirical research on customer perceived value is based on traditional services, or on consumption experiences of goods [18–22]. Modifications may be needed to depict the spontaneous use of self-services in specific situations and the convenience compared to other alternatives.

The following section discusses EMP customers' benefits and sacrifices.

3 CUSTOMER BENEFITS AND SACRIFICES IN EMPs

Various authors have argued that the net value of using EMPs consists of the trade-off between two basic value elements: benefits and sacrifices [e.g., 23–25]. The sources of benefits are identified as experience, commercial success, pioneering and useful service features [14]. Experience refers to getting to know the new service, testing it, learning to use it, and seeing how it works. Commercial effectiveness refers to financial gains and improved sales that customers of EMPs expect from using the service. For some customers, getting new customers and increased revenues were the only reason to start using the EMP; some expected their use of the EMP would primarily improve the public's awareness of the company and subsequently result in increased sales. Pioneering refers to the status of being among the first companies to use an EMP, thereby reflecting a willingness to be among the forerunners. Related to this is the expectation that using the EMP delivers an image of a future-oriented company that is well aware of the current development trends. This is different from gaining experience with the service in that pioneering is concerned with the image of the modern firm, whereas experiences are more practical level benefits associated with the usage of the new service. Useful service features refer to the potential benefits directly related to the EMP service as well as its usage or features, including the improved ability to reach customers and send personally tailored messages. Meanwhile, sources of sacrifice can be classified as monetary or non-monetary [14]. Monetary factors refer to direct or indirect costs of using the EMP. Direct costs are related to prices paid by the customer (investments in infrastructure and functioning of the systems), while non-monetary factors consume time, effort, and learning. Thus, the anticipated investments also include the effort and learning required to be able to use the EMP. These value sub-elements are interrelated in a complex way. Certain sacrifices made by customers may increase the benefits they perceive, whereas some benefits can increase the sacrifices, thereby reducing the customers' perceived net value.

Another important characteristic of the value creation in the EMP is learning, which has also been ignored in the traditional setting. Due to the novelty and technological character of the EMP, the individual should learn how to utilize technology effectively, which in turn creates a requirement for investment from the customer side [25]. In other words, the customer needs to commit time and effort to learn to use the EMP and be able to derive value from it [26]. Service may also be used as a self-service to a certain extent, increasing the importance of the involvement and competences of the customer [see 9]. Moreover, since technology is constantly developing, the customer needs to keep up with such development and continuously learn to operate and fully exploit the new features of the service.

4 DYNAMIC NATURE OF CUSTOMERS' PERCEIVED VALUE IN EMPs

Another important characteristic of value creation in EMPs is the value that is to be realized in the future, which has not received attention in traditional markets. The specialty of EMPs is the changing perception of value. Value changes and takes different forms during the service delivery and usage process (i.e., before actual service consumption, during it, and even afterwards). Hence, value is seen as a dynamic concept that has past, present and future dimensions [9].

Because of the (relative) novelty of EMPs (as, for example, mobile advertising), customers lack previous experiences and are therefore uncertain about the expected value [see 27, 28], which may decrease customers' willingness to participate in the EMPs. Expected value (i.e., the benefits and sacrifices customers expect to occur before they start using the service) plays an important role in value creation as it strongly influences customers' willingness to try a new service, continue to use it, and form a relationship with a new service provider [9, 14]. However, as previously discussed, in the context of EMPs the expected value emphasizes that there are not only expected benefits, but also expected sacrifices in relation to both the service and the relationship in which the service is produced [14]. On the other hand, the customers' perceived value also emphasizes the realized value, which refers to actual benefits and sacrifices the customer perceives when evaluating the service experience. Therefore, realized value refers to the trade-off between benefits and sacrifices in relation to the net value of an alternative perceived by the customer when evaluating the service experience [14].

As many EMPs are subject to continuous development, customers also have future expectations about EMP use [29]. This potential value is the trade-off between the benefits and sacrifices that customers expect from the service in the future. Thus, both elements of value—benefits and sacrifices (and their sources, as previously discussed)—have past, present, and future dimensions [14]. Expected, realized, and potential value are represented in Figure 1.

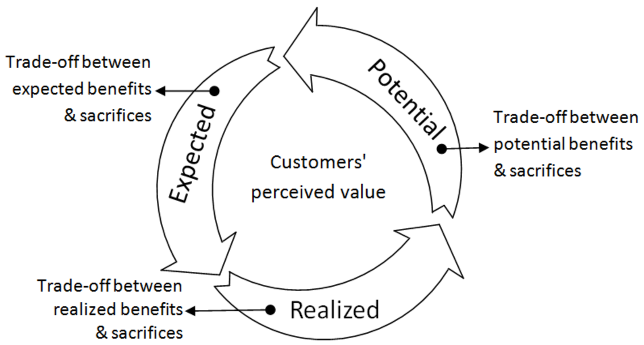


Figure 1: *Expected, realized and potential perceived value of a customer.*

5 VALUE CO-CREATION IN EMPs

The special feature of the functioning of EMPs is the need for the customer to cooperate and interact with other EMP actors (e.g., infrastructure providers, network operators, application providers, device manufacturers, and products/services receivers). All these network actors influence the final service offering and its value [see 9]. The services involving new technological innovations generally require close collaboration among many partners in order to combine several competences. Thus, a variety of factors influence customers' value perceptions and make the formation of the value both uncertain and difficult to assess in advance [29]. Furthermore, in EMPs, the service process takes place through a technical interface, which means that the role of the customer increases and part of the service can in fact be self-service [see, e.g., 9]. This makes the role and competencies of the customer essential to the success of the EMP. Therefore, the role of the service provider in training and supporting the use of the service is also an integral part of evaluation.

As a result, the success of EMPs is achieved through co-creation (or co-production) [e.g., 6, 30]. The basic idea is that the customer is no longer "a target" that receives the value determined by the service provider; instead, value can only be created with and defined by the customer [31]. However, customers make the evaluation and simultaneously influence it. Therefore, their perceived value is always influenced by their personal qualities and the way in which the evaluation is made. Hence, it is not just the service that determines the perceived value, but the customers themselves, whose value perceptions are therefore subjective [14].

6 CONCLUSION

This paper discussed the concept of EMP customers' perceived value to clarify the complex nature and essence of the value concept and create a more profound understanding of the multifaceted character of customers' perceived value in EMPs. Unlike traditional markets, customers' perceived value in EMPs assumes more than just how well the service is delivered during interactions between the service provider and customers. Customers' perceived value also depends on technical and functional quality as well as on the complex trade-off between multiple benefits and sacrifices that EMPs bring to and gain from the customers at different points in time. According to the time aspect, there are perceived expected, realized, and potential value of EMP customers. In addition, these perceived values are subjective because they depend on the personal characteristics of the customers making the valuation while simultaneously influencing it.

References

[1] Goldsby TJ, Eckert JA (2003) Electronic transportation marketplaces: a transaction cost perspective. Industrial Marketing Management 32: 187– 198.

[2] Woodruff RB (1997) Customer value: the next source of competitive advantage. Journal of the Academy of Marketing Science 25(2): 139–153.

[3] Ulaga W & Chacour S (2001). Measuring customer-perceived value in business markets. Industrial Marketing Management 30: 525–540.

[4] Parasuraman A, Zeithaml V & Berry L (1985) A conceptual model of service quality and its implications for future research. Journal of Marketing 49: 41–50.

[5] Parasuraman A, Zeithaml V & Berry L (1988) SERVQUAL: a multiple-item scale for measuring of consumer perceptions of service quality. Journal of Retailing 64(1): 12–37.

[6] Grönroos C (1978) A Service-oriented approach to marketing of services. European Journal of Marketing 12(8): 588–601.
[7] Patterson P & Spreng R (1997) Modelling the relationship between relationship value, satisfaction, repurchase intentions in a business-to-business services context: an empirical examination. International Journal of Service Industry Management 8(5): 414–434. DOI: 10.1108/09564239710189835.
[8] de Chernatony L, Harris F & Dall'Olmo Riley F (2000) Added value: its nature, roles and sustainability. European Journal of Marketing, 34(1/2): 39–56.
[9] Komulainen H, Mainela T, Tähtinen J & Ulkuniemi P (2007) Retailers' different value perceptions of mobile advertising service. International Journal of Service Industry Management 18(4): 368–393.
[10] Möller K & Törrönen P (2003) Business suppliers' value creation potential. A capability based analysis. Industrial Marketing Management 32(2): 109–118.
[11] Ulaga W (2003) Capturing value creation in business relationships: a customer perspective. Industrial Marketing Management 32(8): 677–693.
[12] Walter A, Ritter T & Gemünden HG (2001) Value creation in buyer-seller relationships. Theoretical considerations and empirical results from a supplier's perspective. Industrial Marketing Management 30(4): 365–377.
[13] Zeithaml V (1988) Consumer perceptions of price, quality, and value: A means-end model and synthesis of evidence. Journal of Marketing 52(1): 2–22.
[14] Komulainen H (2010) Customer perceived value of emerging technology-intensive business service. Dissertation. Faculty of Economics and Business Administration, Department of Marketing, University of Oulu. Accessible at http://herkules.oulu.fi/issn14552647/.
[15] Folkes V (1994) How consumers predict service quality. In: Rust RT & Oliver LR (eds) Service Quality. Thousand Oaks CA, Sage: 108–122.
[16] Woodall T (2003) Conceptualising 'value for the customer': An attributional, structural and dispositional analysis. Academy of Marketing Science Review: 1–42. Accessible at http://findarticles.com/p/articles/mi_qa3896/is_200301/ai_n9201442/.
[17] Rokkan AI, Heide JB & Wathne KH (2003) Specific investments in marketing relationships: expropriation and bonding effects. Journal of Marketing Research 40(2): 210–224.
[18] Babin BJ, Darden WR & Griffin M (1994) Work and/or fun: measuring hedonic and utilitarian shopping value. Journal of Consumer Research 20(4): 644-656.
[19] Caruana A, Money AH & Berthon PR (2000) Service quality and satisfaction – the moderating role of value. European Journal of Marketing 34(11/12): 1338-1354.
[20] Cronin JJ, Brady MK & Hult GTM (2000) Assessing the effects of quality, value, and customer satisfaction on consumer behavioral intentions in service environments. Journal of Retailing 76(2): 193-218.
[21] Kerin RA, Jaine A, Simmons HC, Cox EL & Howard DJ (1992) Store shopping experience and consumer price-quality-value perceptions. Journal of Retailing, 68(4): 376.
[22] McDougall GHG & Levesque T (2000) Customer satisfaction with services: putting perceived value into the equation. Journal of Services Marketing 14(4/5): 392-410.
[23] Kothandaraman P & Wilson D (2001) The future of competition: value-creating networks. Industrial Marketing Management 30(4): 379–389.
[24] Menon A, Homburg C & Beutin N (2005) Understanding customer value in business-to-business relationships. Journal of Business-to-Business Marketing 12(2): 1–38.
[25] Curran JM & Meuter ML (2005) Self-service technology adoption: comparing three technologies. Journal of Services Marketing 19(2): 103–113.
[26] Trkman P & Baloh P (2003) Use of a model for information technology education. Journal of Information Technology Education 2: 379–392.
[27] Hogan JE (2001) Expected relationship value. A construct, a methodology for measurement, and a modelling technique. Industrial Marketing Management 30(4): 339–351.
[28] Hibbard JD, Hogan JE & Smith GR (2003) Assessing the strategic value of business relationships: the role of uncertainty and flexibility. Journal of Business and Industrial Marketing 18(4/5): 376–387.
[29] Möller K (2006) Role of competencies in creating customer value: A value-creation logic approach. Industrial Marketing Management 35(8): 913–924.
[30] Lusch RF & Vargo SL (2006) Service-dominant logic: reactions, reflections and refinements. Marketing Theory 6(3): 281–288.
[31] Edvardsson B, Gustafsson A & Roos I (2005) Service portraits in service research: a critical review. International Journal of Service Industry Management 16(1): 107–121.

E-LEARNING MATERIALS DEVELOPMENT PROJECTS

Marjan Krašna
University of Maribor, Faculty of Arts
Koroška cesta 160, 2000 Maribor, Slovenia
e-mail: marjan.krasna@uni-mb.si

ABSTRACT

During last few years many projects for development of e-learning materials were sponsored by the Government of Slovenia and European Social Fund. We have been part of development team that produced four e-learning materials for primary and secondary education and we have participated in the evaluation of e-learning materials.

From the perspective of project management these projects are extremely sensitive and volatile. There are problems of team assembly, time schedule, motivation, technology, and knowledge. In this article we are going to present the life cycle of the e-learning materials, team building and internal and external risk factors.

1 INTRODUCTION

This year the projects for e-learning materials are going to end. In two years' time only successful developers of e-learning were asked to upgrade their materials and verify their educational values. Faculty of Arts was to prepare the e-learning material for environmental studies for secondary schools. Faculty of Education participate in the preparation of e-learning materials for primary schools. Our team members have also years of experiences with the preparation of learning materials for the university level of education. In fact we were the first who started e-learning on the University of Maribor. We could even say that we tested educational videoconferences in previous millennia. Anyway we prepared e-book and multimedia learning materials even when Windows NT was new.

Preparing e-learning materials is not an easy task. There are differences in the learning population. Kids behave different than students and adults have different preferences than all others types of students. Those different preferences of students and their different age are the unique constraints in the development of the e-learning materials.

In the 2007 the Government of Slovenia and ESF (European Social Funds) prepare the public tender for e-learning materials production projects. Due to external reasons the projects were very short. At that time projects goals were to develop e-learning materials for environmental nature studies. It covers the learning material for the third year of primary school (8 to 9 year old) and elective subject in secondary school (14 to 17 years old). In the third year of primary school we need to pay additional attention to the text [1]. Pupils are unable to read large amount of text. Everything needs to be simple, evident, fun, exciting, nice, didactically suitable, and must encourage the research spirit in the pupils [2][3]. This approach is not suitable for the secondary school students. They regard themselves as grown up and such approach would seem to them as childish.

We have pretty good overview of what kind of ICT equipment is used in the schools. Therefore we need to prepare the e-learning materials according to the limitations of average school's computer and internet access. There are also some environments that use Linux instead of Windows and in some environment we see increasing amount of Apple computers. Despite the fact that most of schools use Microsoft Windows we discover that Internet Explorer is not used as default browser on all of them. Close to 40% of the users use Firefox web browser which is not entirely compatible with the Microsoft IE. Therefore it is not suitable to use Microsoft specific multimedia distribution system. Other options were classic HTML with JavaScript or Flash. According to the research on web for the good practice projects the Flash was our favourite [4]. It is portable and packed into one file. It works even if we have problems with SCORM packages but it has some disadvantages. The file may grow if multimedia materials are included into the flash video. Time to transmit is much longer then acceptable 8 second [5][6]. Problem with the Flash player upgrade can render Flash video unplayable and waste all our effort. The decision was not easy but we choose Flash anyway and successfully finish the project.

In 2009 we were contacted from the Ministry of Education and Sport to upgrade our previous e-learning materials due to high quality products we produced in the first place. Since we participate in the projects of two different faculties they choose different ways to conduct projects. Faculty of Education was a subcontractor to outside company which provide technical support and implementation the e-learning materials but on Faculty of Arts we were in the position of project leader and hire some externals experts for some specific tasks.

In a span of just four years we finished four different projects with different employees; different approaches; and different objectives.

2 LIFE CYCLE OF E-LEARNING MATERIALS

As any product even e-learning materials have their life cycle. First major part is the production; and second part is service. In the first part we often start with the **requirement**

specifications where we want to assess the scale of the project. The scale of the projects is in the case of e-learning material measured with numbers of screens. Technical requirements provided by the Ministry of Education and Sport further specify the average computer screen. On average screen at least two interactive elements and sound and video (measured in seconds) are required. As we know what the price for building blocks we can easily calculate the price of whole project. On the other hands we have problems with the assessment of the number of screens for specific tasks.

The second phase is the **scenario** phase. Special didactics professionals (authors) know the content that needs to be in the e-learning materials. They often start their job with text based scenario with the description of multimedia elements. If they are capable to use design tools they can even suggest how to prepare the computer screens. In most of our projects we used PowerPoint as the design tool. It is simple; everybody is sufficiently familiar with it and knows its capabilities.

The third phase is the **design and production** phase where designers prepare the e-learning materials and test the material with the authors. In the design we apply different kind of professionals. Audio and video production starts in this phase and software professionals prepare simulations and animations according to the scenario.

When the e-learning materials are prepared project enters the **verification** phase. Appointed professionals conduct a walkthrough the e-learning material and assess the didactical values and appropriateness of the content.

Dissemination is the next phase of the project. Learning materials are subjected to the teachers in schools. At the beginning they attend seminars how to use e-learning materials and then they use this e-learning materials with their students. Their reports are the final step in the verification process.

Finalization is the last stage of development of e-learning materials. All faults discovered are corrected and e-learning materials are given to the general public to use them. E-learning material comes to the **operation** phase where only user support is active for a period of three years. At the end of this period the projects can be terminated.

3 WORK TEAMS

In the projects of development of e-learning materials we need participants with different knowledge and skills [7]. First team consist of **authors** of e-learning materials. They are didactics specialists covering different scientific areas. In our case we have specialists from biology, chemistry, physics, geography and technology. They start their work preparing the scenarios.

Team of designers consist of different participants with really different skills. We need **graphic designers**, **audio and video production specialists**, **computer scientist**, and **communication specialists**. Their work is not only to fulfil wishes from the scenario scripts. They need to suggest changes to the authors which are not always technically skilled and they often have wrong perspective to the outcome of their work.

In our project when we prepare learning material for primary school we have two cartoonists, one fine art specialist that suggest position of elements on screens and balanced the colour metrics. We often include students of educational computer science which provide the support to the unskilled authors to prepare better presentations. A problem solver, who gathered all unresolved technical problems, analyse them, solve them, and report back the solutions is also needed.

4 DESIGN AND PRODUCTION

As we have two different types of students we also have two different approaches to the development of e-learning materials [8]. For primary schools we have to prepare everything from the beginning. Team members for the primary schools were also less skilled in the design process then the team members for the secondary school e-learning materials. Therefore we decide that the simplest way to produce the e-learning material for the primary school is a two steps process. First authors of e-learning material prepared PowerPoint presentation. These presentations were later used by the team of implementers to produce the Flash video.

With this in mind we decided to give free hands to the authors. They should prepare the materials up to their limitations, the descriptions what should be on the slide, and correct transitions between slides.

We were much more successful in the preparation of the e-learning materials for the secondary schools. Materials were already prepared, most of the materials were in digital form and authors knew exactly what they want to do with the materials. It was also easier since we did not have to worry too much about the text constraints.

5 DEVELOPMENT TOOLS

In our projects we use Flash and SCORM packages. There are different Flash development tools. Some of them transform PowerPoint presentations to the Flash other are for pure Flash production. We have discovered that freeware software was useless for many reasons: producers' logotypes were always displayed on the screen; they enable only basics effects and transitions; and results are not totally different than the original. We have bought two development tool Articulate Studio and iSpring Pro. In our case the more appropriate was iSpring Pro which enables us to prepare Flash video without integrated Flash player.

For the preparation of SCORM packages we wanted to find the tool that transform text to SCORM but it was impossible to acquire the product. The vendor gives us only option to pay 3000 € for educational corporate license. Therefore we used the freeware tool "Reload packager" and transform the text to the SCORM. For the tests we also used different tools: HotPotatoe, eXe and Articulate QuizMaker. Interactive Flash pages were made with the Macromedia Director.

6 DIFFERENCES BETWEEN THEORY AND PRACTICE – PROCESS OF DEVELOPMENT

In theory everything is clear but in reality project start with the contract. The pre-project activities are costly and are covered from the funds of the development team – in our case Faculty covered the expenses.

The projects are planned and suppose to start on predefined date or whole projects shifts on the calendar. Well this is just a theory. In all our projects which should last around nine months we did not receive contract on time. When the contract was signed three months after the project should start the end date was fixed on predetermined time schedule. There were no negations with the Government about the delay. Their response is always – finish the projects on time or you are not going to be played. If we reject the project we know that we would not be able to participate in any additional public tenders. The reduction of the project time for one third requires drastic measures.

7 CHANGES IN THE DEVELOPMENT PROCESSES

After the scenarios were prepared the design of screens began and all the preparation instructions for the multimedia and interactive elements started [9][10].

Since we have problems with the time constraints we wanted to speed up the processes. Only option for the increased development was in the verification (see Figure 1) where we provide students of educational computer science to help authors with the preparation of screen in PowerPoint. When the presentations were prepared they were also exactly according to the authors expectations. No additional verification of the processes was needed.

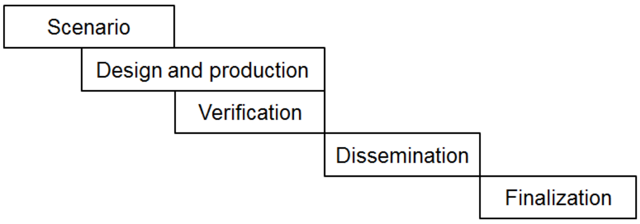


Figure 1: *Phases of the project*

8 E-LEARNING PARADIGM

From the production of first learning material till the last one we have to change the e-learning paradigm. We have discovered that retention of knowledge where everything is just a mouse click away is too small and from this perspective we failed the objective. Therefore we have changed the paradigm and we prepared different type of interactions in new learning materials that stimulates students' creativity and tries to teach them competences instead of content (see Figure 2) [11].

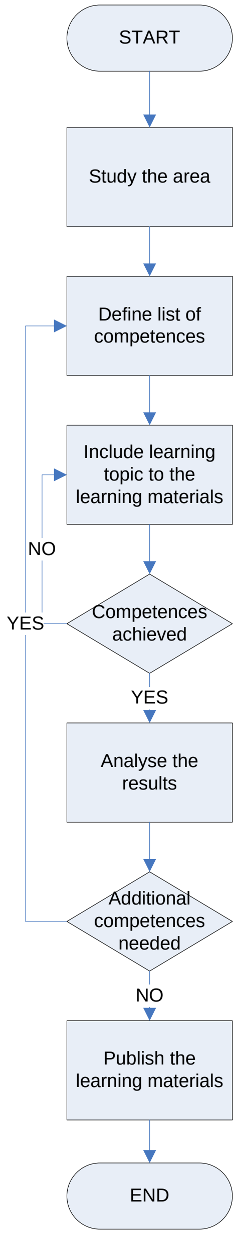


Figure 2: *New e-learning paradigm*

9 RISK ANALYSIS AND OBSTACLES

We have started the risk analysis discussion in the topic where we describe the changes between theory and practice. Here we are going to describe internal and external obstacles that influence the risk. The real risk analysis always requires the understanding of development processes also also support processes which the project depend upon. Therefore we cannot regard this topic as clean topic. Interpersonal relations are very important in this matter and we have to know basics of diplomacy to release growing tensions between members of the projects' teams.

8.1 Internal obstacles

Tight **schedule** and a great **workload** always bring to the frustration and overload. If the management is not aware of this the project will suffer and in worst case it will collapse.
Communication is another risk topic. Communication is needed but on the other hand too much communication stalls the work. Management needs to channel the communications and monitor them.
Control is mandatory but too much control is not productive. If participants start feeling very uncomfortable with their workload it is easy for them to say: "*I don't need this money I'm out.*"
Gold plating is another problem in the project. Participants overestimate their capabilities and think what can be done during the weekends. **Blaming others** is common "solution" in such environment. We need to apply communication skills to reduce the tension between collaboration parties. Time constraints were so tight that people were really frustrated. In one moment we have to drop the tight collaboration since authors blame students for their own failures. The obvious accelerators steps prove to be unsuccessful. Hiring new students would be disastrous since they would need too much time to adjust and learn their job. To protect the students we need sometimes an open hard discussion with the project leader and prove him wrong.

8.2 External obstacles

Most of the external obstacles were present because of the communication with the Ministry of Education and Sport. Every project starts with the delay because they do not have financial funds to cover national part of the project. Delays are frequent and many authors do not get paid on time.
Preparing reports is another topic that requires attention. During the duration of a project it often happens that report forms changes from one report to another. What is good for one reviewer is totally unacceptable for another reviewer. In some circumstances contracts with the participants on the projects were changed many times before they become valid and administration personal were completely frustrated.
Some external obstacles to the project are also internal obstacles in the faculty. The top management of the faculty insist that something is made opposite to the requirement of the Ministry. They even try to enforce their view despite the fact they were proven wrong.

10 CONCLUSION

Development of e-learning materials follows the rules established in software project management. Processes just look simple and are often underestimated from the technical perspective where "really big problems are". But the product of e-learning development is often used by more persons than other software products and they still needs to be of high quality, didactically suitable and attractive. Project participants have different knowledge and digital skills and much attention need to be paid to the communication. Authors on the other hands have a free will and traditional methods of forcing them to work are unsuccessful. Successes of such projects are often very tangible enterprise and require great management skills. Management have only incentives for participants to work. Despite these facts we managed to successfully finish four projects with "many or even more" problems.

References

[1] Moreno R, Mayer RE, Verbal redundancy in multimedia learning: When reading helps listening, Journal of educational psychology, Volume: 94 Issue: 1 Pages: 156-163 Published: Mar 2002

[2] Wang N, Johnson WL, Mayer RE, Rizzo P, Shaw E, Collins H, The politeness effect: Pedagogical agents and learning outcomes, International journal of human-computer studies, Volume: 66 Issue: 2 Pages: 98-112 Published: Feb 2008

[3] Mayer RE, Fennell S, Farmer L, Campbell J, A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style, Journal of educational psychology, Volume: 96 Issue: 2 Pages: 389-395 Published: Jun 2004

[4] Jedrusik P, Preisack M, Dammann F, Simplified production of multimedia based radiological learning objects using the Flash Format, ROFO-Fortschritte auf dem gebiet der rontgenstrahlen und der bildgebenden verfahren, Volume: 177 Issue: 7 Pages: 1009-1015 Published: Jul 2005

[5] Nina, B., Anna, B., & Allan, K. Integrating User-Perceived Quality into Web Server Design. 9th International World Wide Web Conference. 2000

[6] Galletta DE, Henry RM, McCoy S, Polak P, When the wait isn't so bad: The interacting effects of website delay, familiarity, and breadth, Information systems research, 2006

[7] Krašna M, Bratina T, Kaučič B; Implementation of Distance Learning Materials, Mipro 2009, Opatija, Croatia, 2009

[8] Krašna M, Bratina T, Kaučič B, Differences in e-learning material production according to the learners' age and comprehension level, 20th Central European Conference on Information and Inteligent Systems, September 23-25, 2009, Varaždin, Croatia, 2009

[9] Hardin PC, Reis J, Interactive multimedia software design: Concepts, process, and evaluation, Health education & behavior, Volume: 24 Issue: 1 Pages: 35-53 Published: Feb 1997

[10] Krašna M, Bratina T, Active elements in e-learning materials, Sixth International Conference on Informatics, Educational Technology and New Media in Education, Sombor, Serbia, 2009

[11] Krašna M, Bratina T, New Paradigm in Preparing E-Learning Materials, MIPRO 2010, Opatija, Croatia, 2010

CONSUMING LINKED DATA ON MOBILE DEVICES

Jože Pfeifer
Hermes SoftLab d.o.o., a ComTrade company
Zagrebška cesta 104, SI-2000 Maribor, Slovenia
E-mail: joze.pfeifer@comtrade.com

Vili Podgorelec
University of Maribor, FERI
Smetanova ul. 17, SI-2000 Maribor, Slovenia
E-mail: vili.podgorelec@uni-mb.si

ABSTRACT

In this paper we introduce the concept of linked data, what it means for the web and how it can be used on the devices that we carry with us all the time – mobile phones. Because of physical and hardware limitations of mobile phones (e.g. smaller screen, slower speed), some adjustments have to be made, but some advantages are still prevalent, like having up to date location information and a constant connection with the web. The proper combination of mobile phones and semantic information dynamically obtained from various open resources on the web provide us with a powerful tool for the next generation of services.

1 INTRODUCTION

Mobile devices have already been used for semantic web services a couple of years ago [10]. In our research we focused on newer generation smartphones, such as the Apple iPhone, and a very recent trend on the semantic web – the linked data concept. All recent smartphones with touch input and GPS location information represent a unique opportunity for using linked data information to provide the most relevant and up to date information to the user based on his whereabouts in an easy to use and interactive way.

2 LINKED DATA

The basic concept of Semantic Web is a Web of Data – of any data one might conceive of. To make Linked Data a reality, it is important to have a huge amount of interrelated data on the Web available in a standard format, reachable and manageable by Semantic Web tools. This collection of interrelated datasets on the Web can also be referred to as Linked Data, Web of Data or the Semantic Web [3]. The terms therefore refer to a set of best practices for publishing and connecting structured data on the Web.

Key technologies that support Linked Data are:
- URI (Uniform Resource Locator – a generic means to identify entities or concepts in the world).
- HTTP (Hypertext Transfer Protocol – a simple yet universal mechanism for retrieving resources or descriptions of resources).
- RDF (Resource Description Framework – a generic graph-based data model with which we structure and link data that describes things in the world). It is based on XML.

In 2006 Tim Berners-Lee outlined four principles of Linked Data in his “Design Issues: Linked Data” note [1]:
1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs so that they can discover more things.

The problem back then was that a surprising amount of data wasn’t linked yet, because of problems with one or more of the steps. He provided solutions to those problems, details of implementation and factors affecting choices about how to publish your data.

3 AVAILABLE DATASETS

The web has come a long way since the initial principles were introduced, as the diagram in figure 1 clearly shows. Only a part of the linked data diagram [5] is visible, but the total number of different datasets is quite impressive (over 90). As the diagram hasn’t been updated for more than a year, there should be even more linked data candidates that joined the community project since then.

What the diagram does show are some of the datasets that are published in the Linking Open Data community project [4] and are interlinked with at least one other dataset in the cloud. The FOAF profiles (Friend Of A Friend) circle does not represent datasets published by a single provider, but is an aggregate of many individual datasets. It is shown as a single circle for historical reasons.

The size of each circle corresponds to the number of triples in each dataset. The numbers are usually provided by the dataset publishers and are sometimes rough estimates.

To summarize what a triple is, it is a statement about a Subject, Predicate and Object (i.e. “The sky is blue”, where “The sky” is the subject, “is” represents the predicate and “blue” represents the object).

The direction of the arrows indicate the dataset that contains the links, e.g. an arrow from A to B means that dataset A contains RDF triples that use identifiers from B. Bidirectional arrows usually indicate that the links are mirrored in both datasets. The thickness of the arrow corresponds to the number of links.

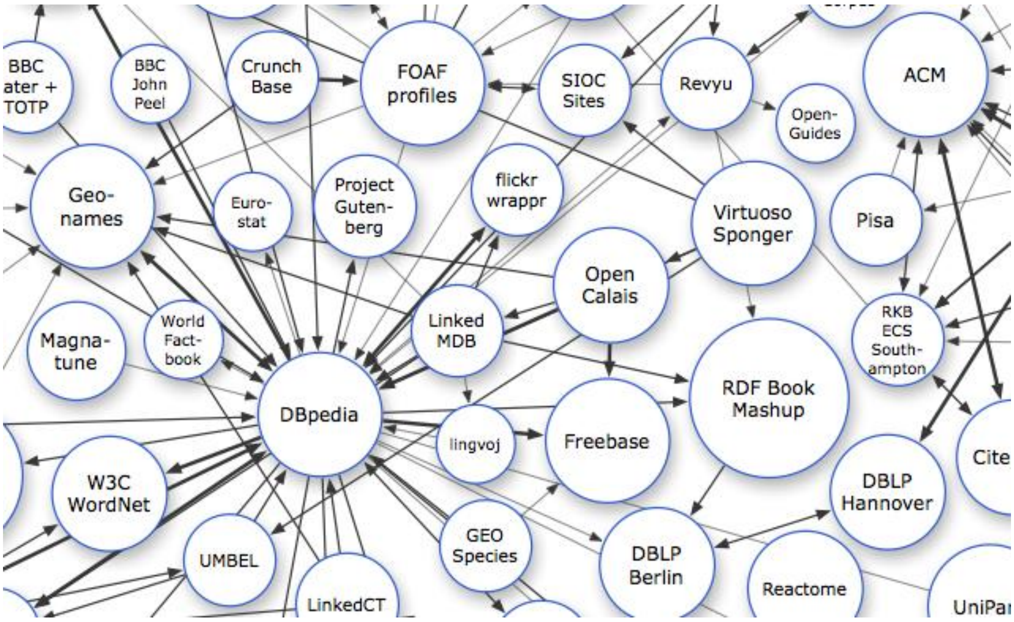


Figure 1: *A part of the Linked Data diagram [5]*

4 USED DATASETS

The datasets that were used in the developed prototype were from a variety of services, including DBpedia, GeoNames, Freebase and Flickr. Because of all the connections between the various datasets, one could easily include other datasets as needed.

4.1 Dbpedia

DBpedia is the Semantic Web version of Wikipedia and enables you to ask more sophisticated queries against data from Wikipedia using the SPARQL language (SPARQL Protocol and RDF Query Language) [6]. In figure 1 it can be seen as the most centric point in the whole linked data concept as it was developed early on in 2007 and linked by most datasets, which came after it.

Due to the usage of the SPARQL language for executing queries on triples, sophisticated queries can be written, whereby relationships between subjects are very important.

An example of a simple query would be to get all Cities in Germany. You could surely find those cities using a simple search engine like Google, but what if you wanted something more sophisticated, like all the persons in the database, that were born in Hamburg before current chancellor of Germany, Angela Merkel (born on the 17.07.1954), sorted by birth? No keyword search engine could return relevant results, but with DBpedia and SPARQL you can get those same results without a problem[1]. The required query can be seen in figure 2.

The results can be returned as HTML for viewing in your browser, or as a spreadsheet, XML, JSON, Javascript, NTriples or RDF/XML for offline viewing.

[1] http://dbpedia.org/sparql/

```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?person ?name ?birth WHERE {
  ?person dbo:birthPlace
        <http://dbpedia.org/resource/Hamburg>.
  ?person foaf:name ?name.
  ?person dbo:birthDate ?birth
  FILTER (?birth <"1954-07-17"^^xsd:date).
}
ORDER BY ?birth
```

Figure 2: *SPARQL query example for DBpedia*

In our case we used DBpedia to get more information about the location returned from the GeoNames service[2].

4.2 Geonames

GeoNames is a geographical database that covers all countries and contains over 8 million place names that are available for usage free of charge [7]. It provides 36 web services[3], which enable the user to find the nearest place name, get the location time zone, weather conditions, etc.

Most results can be returned in XML or JSON format, but some web services offer RDF, CSV, TXT and even RSS and KML.

We used the "findNearbyPlaceName" web service to get the nearest place name. The results, which can be seen in figure 3, include standard information, such as toponym name, country name and population.

Amongst the results there is also a geonameId attribute, which helps us to get to the RDF document for the place[4]. That document contains even more information about it,

[2] http://dbpedia.org/resource/Maribor
[3] http://www.geonames.org/export/ws-overview.html
[4] http://www.geonames.org/3195506/about.rdf

such as the name in various languages, Wikipedia articles in various languages and also an owl:sameAs element, which contains the URI to the DBpedia article about it.


```
{ "geonames":[ {
    "countryName":"Slovenia",
    "adminCode1":"J2",
    "fclName":"city, village,...",
    "countryCode":"SI",
    "lng":15.6466667,
    "fcodeName":"populated place",
    "distance":"0.5837",
    "toponymName":"Maribor",
    "fcl":"P",
    "name":"Maribor",
    "fcode":"PPL",
    "geonameId":3195506,
    "lat":46.5547222,
    "adminName1":"Maribor",
    "population":89056
  } ]
}
```


Figure 3: *JSON returned from GeoNames*

With the help of the owl:sameAs element everyone can connect their information with someone else's. What this basically means is that everyone then knows that those different documents describe the same subject.

4.3 Freebase

Freebase also offers additional information about various subjects. If we continue with our previous example, we saw that we can get the URI about Maribor on Freebase[5] from DBpedia. This query has to be written in MQL (Metaweb Query Language), which is specific to Freebase.

With MQL we can make a query about the first 20 places and their URIs within Maribor, as seen in figure 4.

```
{ "id":  "/en/maribor",
  "mid": null,
  "/location/location/contains": [{
    "id":    null,
    "name":  null,
    "limit": 20
  }]
}
```

Figure 4: *MQL query for Freebase*

We specify all known data by providing the values and also the data we wish to obtain by writing "null" as the value. Freebase then runs the query and replaces those null values with found data.

The returned data from Freebase is in JSON format, so that it can easily be parsed and displayed on a mobile device.

[5] http://www.freebase.com/view/en/maribor

Freebase was recently purchased by Google, who explained that they plan to use the acquired knowledge to further expand their search possibilities (e.g. queries like "actors over 40 who have won at least one Oscar" may soon become a possibility [8]).

4.4 Flickr

Flickr enables users to search for public or their own private photos based on their current location, either by location name or by GPS coordinates. The whole array of API methods they provide can be found online[6].

The Flickr wrappr in figure 1 returns all photos taken in a given location. We took the second approach and used the "flickr.photos.search" method to search for photos, taken near our current location. This method does have a drawback, as a lot of photos don't have GPS information yet and therefore aren't returned by the service. But you have a wider array of filters with which you can search for those photos.

The returned data is in JSON format, similar to figure 3. With it we can reconstruct the URL of the image location and read its metadata, such as the title and location, where it was taken.

5 LINKED DATA ON MOBILE DEVICES

We wrote a prototype application to communicate with all the described datasets and display the information in an interactive way. All queries were done using REST (Representational State Transfer) over HTTP with JSON as output. The application workflow can be seen in figure 5.

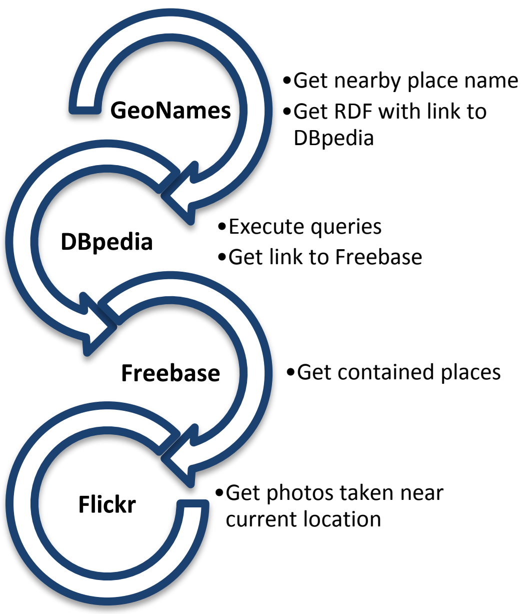


Figure 5: *Prototype application workflow*

First, the user's current location was acquired via GPS and sent to the GeoNames service, so the current place name

[6] http://www.flickr.com/services/api/

could be determined. Based on the returned results the GeoNames RDF document could be obtained, which contained the link to the DBpedia resource about the location. As DBpedia also provides an URI to the resource on Freebase, we could then easily query Freebase for all the places that are part of the current location. Based on the GPS coordinates (or resource name) from DBpedia we could then request the public photos from Flickr. All this acquired data could then be browsed and further interacted with. A screenshot of the developed prototype can be seen in figure 6.

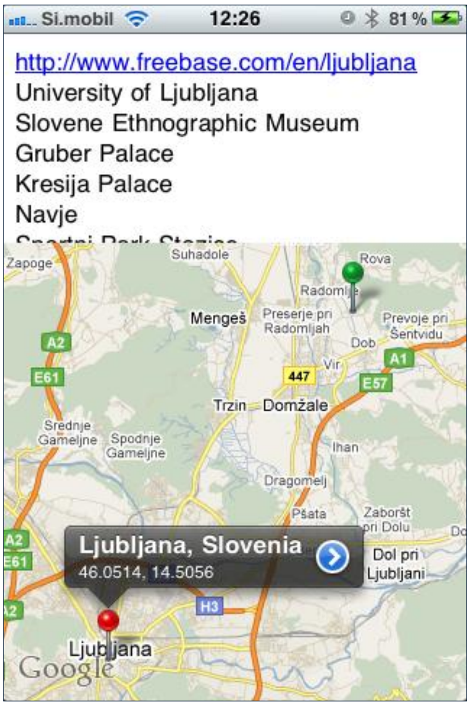


Figure 6: *Prototype iPhone application results*

The red pin represents the found location, returned from GeoNames. If we tap the blue arrow, we are taken to the information obtained from DBpedia, such as the abstract about the displayed location. On top of the application we can also see the URI to the resource on Freebase, with all contained resources being displayed below it. The green pin on the top right of the map represents a photo that was found on Flickr and can also be tapped and viewed in a similar fashion as the red pin.

6 DISCUSSION & CONCLUSIONS

In this paper we introduced the concepts behind Linked Data that are necessary to consume various datasets on modern mobile devices. A similar approach with multiple datasets has already been done with DBpedia Mobile[7] [9] which unfortunately hasn't been updated in over two years, but employs the same idea as this paper.

The development of applications that communicate over HTTP isn't problematic with newer mobile operating systems such as iOS and Android. As REST is supported on those platforms, and the obtained RDF (XML) and JSON can easily be parsed and analysed, no intermediate server is required, as all those queries can be made directly.

During development we encountered a few drawbacks, as it could happen that a given service could suddenly become unavailable because of technical difficulties. This happened with DBpedia a couple of times, which returned an error upon requesting the required results. There was even the problem with caching, as DBpedia uses caching for faster results but this meant that it could happen that your current location was returned just fine one hour, but the next it wasn't cached anymore and therefore you got an empty result. Even GeoNames sometimes returned an error.

Those problems are hard to circumvent as there are usually no backup services with similar functionality available, but all this should improve with time as seen with Freebase, which started offering a similar functionality as GeoNames, so that you can search for nearby places by providing your current location.

This of course means that you have to be more cautious when developing your application and anticipate such events, but the uptime of datasets will also most certainly improve with time, so that this will be less of an issue.

Despite the mentioned problems we think that in the near future mobile devices will play an even more important part in our daily lives as we consume all that connected knowledge on the go, where we want to, when we want to and how we want to.

References

[1] W3C, "Linked Data – Design Issues", http://www.w3.org/DesignIssues/LinkedData
[2] Linked Data, "Connect Distributed Data across the Web", http://linkeddata.org/
[3] W3C, "Linked Data", http://www.w3.org/standards/semanticweb/data
[4] W3C, "Linking Open Data", http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[5] R. Cyganiak, "About the Linking Open Data dataset cloud", http://richard.cyganiak.de/2007/10/lod/
[6] DBpedia, "DBpedia", http://www.dbpedia.org/
[7] GeoNames, "GeoNames", http://www.geonames.org/
[8] Blogspot, "Deeper understanding with Metaweb", http://googleblog.blogspot.com/2010/07/deeper-understanding-with-metaweb.html
[9] Ch. Becker, Ch. Bizer, "DBpedia Mobile: A Location-Enabled Linked Data Browser", 1st Workshop about Linked Data on the Web (LDOW2008), Beijing, China, April 2008.
[10] M. Sheshagiri, N.M. Sadeh, F. Gandon, "Using Semantic Web Services for Context-Aware Mobile Applications", Proceedings of the Workshop on Context Awareness, MobiSys 2004, 2004.

[7] http://wiki.dbpedia.org/DBpediaMobile

SYMBIONODE DATA CARRIER IN DELAY AND DISRUPTION TOLERANT NETWORKING (DTN)

Sašo Vrbinc, Boštjan Grašič, Marija Zlata Božnar, Primož Mlakar
MEIS storitve za okolje d.o.o.
Mali Vrh pri Šmarju 78
SI-1293 Šmarje - Sap
Tel: +386 1 3663226; fax: +386 1 3663227
e-mail: saso.vrbinc@meis.si

ABSTRACT

The work of N4C project involves a development of a future Internet in the areas without telecommunication infrastructure. Delay and disruption tolerant networking (DTN) was designed to replace the "legacy" Internet in the areas, where data mules are used as data carriers instead of a classic wiring. Data mules are typically computers with wireless communication device, but in this paper we describe different and very cost effective approach of data transfer over a simple storage device, such as USB key or memory card, in a DTN.

1 INTRODUCTION

DTN is a network solution for the area where no continuous network connection exists mostly due to the extreme environmental difficulties in which wired infrastructure is ineffective, or GSM and satellite network would be to expensive [1],[2],[3].

The N4C research group continues with the development of DTN2 bundle protocol, which is a standard implementation of a DTN architecture components primarily developed by the Universities [4],[5]. The indispensable part of the architecture comprises a routing protocol for best possible data transfer. PRoPHET is a Probabilistic Routing Protocol using a history of encounters and transitivity [6]. In research presented in this paper a Prophet implementation developed in Luleå University of Technology has been used that is not based on standard bundle protocol [7].

Every DTN network consists of a set of static nodes and dynamic nodes (data mules) [8],[9],[10]. A typical data mule has connectivity with other nodes over Wireless LAN (WLAN) or other type of communication, such as Bluetooth. In contrast to a typical data mule, SymbioNode (described in this paper) is a storage device, and the data exchange is possible only when it's physically connected to another node. Therefore, some manual interaction is required for successful data transfer. All other tasks are autonomous and completely automatic with help of the Prophet functionality. To enable such functionality, many problems arise, especially with automatic device detection and prophet running environment.

1.1 Basic concept

When a storage device is attached to the node, the system needs to assure appropriate running environment for new prophet instance. Every prophet instance requires one network interface for communication with other nodes. Since the storage device has no networking capabilities, a new or free existing network interface on the node is required.

Figure 1 describes the concept of a solution, where two prophet instances see each other in the network and possibly exchange data. From Prophet #1 perspective, another node appeared in the network and communication can be established just like with the "real" node.

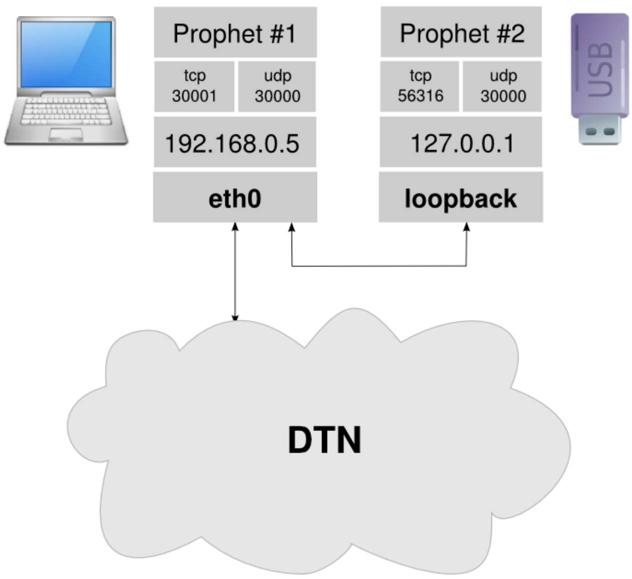


Figure 1: *SymbioNode concept*

SymbioNode represents Prophet #2 instance, which occupies the loopback device. It can communicate with Prophet #1, or with any other DTN node in the network.

Many different solutions have been tried to create a virtual network interface, including Virtual LAN(VLAN), but the only workable solution was loopback interface that successfully established communication with other nodes in the network. The given solution supports only one attached SymbioNode at a time since only one loopback device exists on the system. To avoid such limitation, the virtual IP addresses can be used, but its use in our test environment the bundle transfer occasionally failed for unknown reason.

1.2 Automatization requirements

The operating system must provide a system for auto detection of newly attached or detached storage devices. Most OS, provide such system, which are in principal more or less similar, however, in contrast to Windows Autorun, Linux has more convenient and configurable Hotplug system [11].

Each node should include a special "Hotplug" script that ensures appropriate environment for running SymbioNode Prophet #2 instance, that later makes an exchange of data between the local node, or any other in the network.

The following summary of requirements must be provided by the OS and the script for complete automatic data transfer:

- auto detect an USB key when added or removed
- mount the device to appropriate location
- provide network virtual interface or loopback device
- start prophet, and watchdog for maintenance processes

1.3 Pros and Cons

The SymbioNode technology has some great advantages over standard DTN node. Generally, SymbioNode gives the following advantages and disadvantages.

Advantages:

- very low cost system
- especially suitable for "production" of several mule nodes for test bed
- complete DTN functionality
- quick installation & configuration
- very suitable for any kind of "long range" transportation, for example, using postal services for storage device delivery between the nodes or deliver storage device between tipi and village router
- current script implementation has only been tested with Prophet, however it should be feasible with DTN2 as well
- support for different OS on the same storage device

Disadvantage:

- Physical contact is needed through USB interface which can be a difficulty if routers are placed outdoor (due to precipitation, condensation, …)

2 AUTOMATIC USB KEY DATA TRANSFER

The target platform was chosen based on the requirements, and OS functionality. Linux provides a convenient and very configurable system for running services on device plug/unplug. Moreover, it also offers a quick and powerful scripting language Bash, which can be used for implementation of all necessary requirements summarized in the previous section. Therefore, support for automatic data transfer can be enabled with appropriate Linux system configuration and custom Prophet Script.

2.1 Prophet script configuration

The script was primarily written for "Hotplug" system, which is running on openWRT [12]. Moreover, it's fully compatible with "udev" [13] and no customization is required. Configuration of the prophet startup script is done over variables at the beginning of the script. Full description of the prophet script variables is in Table 1.

SCRIPT_LOCATION SCRIPT_PATH	Location directory of the startup script. Full path of the startup script.
DELAY	The script automatically searches for mounted device and since the mount is not instantaneous, the variable defines waiting period for the device become available by the kernel.
VIP_ENABLED	If you want to use Virtual IP then set this flag to 1
LOG_PATH	This log holds information about running prophet invoked by the SymbioNode. Each entry has the following information: <prophet_proc_id> <network_if> <device_name_path> <watchdog_proc_id>
DEVICE_DIR	System device directory. Standard location is /dev/
SEARCH_DEV	Regular expression for searching the attached device.
PROPHET_DIR PROPHET_BIN PROPHET_CONF	When device is mounted at location defined in fstab, the prophet binary is executed with command: <mount_point>$PROPHET_DIR$PROPHET_BIN. IP1 defined in $PROPHET_CONF is required if $VIP_ENABLED flag is set to 1.
PHYSICAL_IF_NAME	If $VIP_ENABLED this variable defines an interface used for Virtual IP configuration.
VLAN_IF_PREFIX	VLAN interface name can vary with a platform. Most often used syntax is ${PHYSICAL_IF_NAME}.<id> or just vlan<id> The <id> is found automatically.
WATCHDOG_DELAY	Watchdog live period in seconds. When time expires prophet is terminated and storage device unmounted for safe removal.
SUSPEND_CMD	This variable is part of the power management and it's used as part of the locking system, described later in this section.

Table 1: *Description of Prophet script(prophet.sh) variables*

The script is also relying on mount configuration of the storage device. Therefore, an entry in /etc/fstab is required to ensure consistency with the absolute paths in prophet.ini(configuration file of prophet), such as DFTPPATH or STORAGEPATH, otherwise Prophet will not function properly. Both paths are mandatory in prophet configuration, since they define the location of

incoming/outgoing traffic and storage location where received bundles are finally stored.
Standard fstab entry for USB key looks like this:

```
/dev/sdb1       /MEIS2  ext2    rw,sync 0       0
```

File: */etc/fstab*

Sometimes it's safer to address partition by ID and consequently avoid complications with standard notation. Most Linux distributions provide partition addressing by ID and it's recommended to use it. The usual location that contains links to all storage devices and its belonging partitions is in /dev/disk/by-id/. Every device has its own unique ID and unlikely any conflicts will ever appear. The syntax in fstab is the same as in previous example except the device partition location and name is different.

```
/dev/disk/by-id/usb-Verbatim_STORE_N_GO_07980809E465303B-0:0-part1
/MEIS2  ext2    rw,sync 0       0
```

File: */etc/fstab*

Mounting the device to appropriate location is crucial for SymbioNode to work correctly. When standard fstab notation is used for addressing device name, an incorrect mount point can be assigned by the kernel if mount point already exist in the kernel. This will lead to inconsistency with prophet configuration, and result is a broken prophet.

2.2 System Autorun configuration

On different Linux distributions and platforms, the Autorun systems differ slightly. Most frequent systems are udev [13] and Hotplug2 [14] that offers customization of actions, invoked by kernel events from various physical devices. Events are device specific, and in the case of USB key two types of events(attached/detached) can be generated. In general, every removable device produce the same type of events, but all event information are specific to a different device.

2.3 Autorun Prophet Script using udev

Udev functionality gives a convenient way for selective running of user space scripts when device is attached or detached. Basically, udev is a rule based manager and a special rule to handle attached/detached USB device is needed. The default location of rules is in "*/etc/udev/rules.d/*", and usually already exist some rules provided by the distribution. All rule files are in form:

```
<number[0-99]>-<name>.rules
```

If more than one exists, they are handled according to sequence number[0-99]. Custom rules have usually higher numbers. Therefore, prophet rule file should look like "*88-prophet.rules*". Rules syntax will not be described here since many good references exists on the web.
SymbioNode can be configured for various types of storage devices(USB key, USB hdd, CF, SD cards, …), indeed if host machine supports it. We used USB keys for testing purposes and udev configuration should be as follows:

```
KERNEL=="sd[b-z][1-9]", ACTION=="add|remove",
RUN+="/lib/udev/prophet.sh %k"
```

File: */etc/udev/rules.d/20-symbionode.rules*

KERNEL – Name of a device defined by the Kernel. Usually one storage device is already present as sda or hda, and the rule should be appropriately adjusted. USB storage devices are represented in the system in form "sd[a-z][1-9]". The rule above applies for all added storage devices on SCSI interface except for the existing 'sda' device which is usually the system disk, and its partitions.
ACTION – With this condition it's possible to catch only certain events from kernel, in our case attach or detach of a certain device.
RUN – Path to the application or script that executes operations for the attached device. The operator **%k** is evaluated by the kernel, and stands as a name of a device, which is passed to the script as an argument.

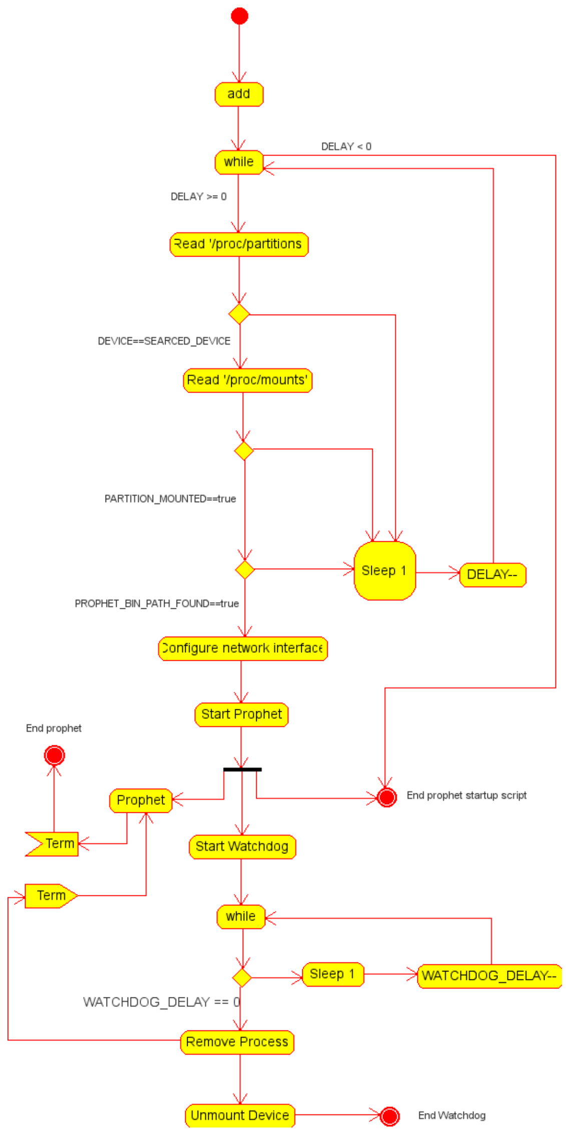


Figure 2: *Prophet script activity diagram*

2.4 Autorun Prophet script with Hotplug2

The “Hotplug2” system is predecessor of udev, and therefore slightly more rigid. It can still be found on some Linux distributions, mostly intended for embedded devices.

Primarily, “Hotplug2” doesn't have a strong rule based system, and therefore target device needs to be found with custom program, usually a bash script. System only separates Kernel events based on device subsystem type. Each subtype has a belonging directory in /etc/hotplug.d/, such as usb, that contains scripts for handling newly added/removed device. Nevertheless, script name paradigm and calling routines are same as in “udev”.

In contrast to udev, the Prophet script has one additional task to perform when Hotplug2 is used. It is necessary to find attached storage device in order to run prophet successfully. Figure 2 shows activity diagram that depicts a whole procedure of the prophet script, when a new storage device is added. If storage device was found, the script configures network interface(if virtual IP is selected), executes prophet binary, and starts the Watchdog. The watchdog ensures that Prophet process terminates, and unmounts the storage device after certain time period expires. Afterwards, the device is ready for safe removal. If device was removed before expiration time, watchdog attempts to remove the prophet process, and cleans the mount points, usually with a 'lazy' unmount. Sometimes device partitions remains in the kernel even after 'lazy' unmount. This will prevent to mount the device to appropriate location on next insertion. To avoid such flaw, it's recommended to change mount points in fstab configuration to appropriate device unique ID, indeed if it's supported by the system.

3 CONCLUSION

We developed script application and appropriate system configuration which makes automatic data transfer possible using Prophet with a storage device as data carrier. We used such system in our testbed μ-GaRaMo (Portable Gamma Radiation Monitor) [10],[15],[16] and the results are very promising. The system of data transfer showed a great reliability without any loss or corrupted data. Even when the system was not used correctly, it remained stable without damaging the filesystem of the SymbioNode nor the native system.

The system features are suitable for DTN region, where a large amount of data can be transfered over great distances with very low expenses. In the near future the system will also be upgraded with support of DTN2 reference implementation.

4 ACKNOWLEDGEMENTS

The work was done under the contract:
SEVENTH FRAMEWORK PROGRAMME GRANT AGREEMENT No 223994, “Networking for Communications Challenged Communities: Architecture, Test Beds and Innovative Alliances”, Collaborative Project

References

[1] V. Cerf, R. Kahn. A Protocol for Packet Network Intercommunication, IEEE Trans. Communications, vol. 22, no. 5, pp. 637–648, May 1974

[2] S. Farrell, V. Cahill. Delay- and Disruption-Tolerant Networking, Artech House, ISBN: 1-59693-063-2, 2006

[3] K. Fall, S. Farrell. DTN: An Architectural Retrospective, IEEE J. Selected Areas in Communications, vol. 26, no. 5, pp. 828–836, 2008

[4] K. Scott, S. Burleigh. Bundle Protocol Specification, IETF RFC 5050, www.rfc-editor.org/rfc/rfc5050.txt, Nov. 2007

[5] Delay-Tolerant Networking Research Group. DTN2 code, http://www.dtnrg.org/wiki/Code, Jul. 2010

[6] A. Lindgren, A. Doria, E. Davies, S. Grasic. Probabilistic Routing Protocol for Intermittently Connected Networks, http://tools.ietf.org/html/draft-irtf-dtnrg-prophet-06, Jul. 2010

[7] Luleå University of Technology. PRoPHET Routing Protocol home page, http://prophet.grasic.net/, Jul. 2010

[8] E. Davies, A. Doria. D2.2: Functional Specification for DTN Infrastructure Software, Folly Consulting and Luleå University of Technology, N4C project, http://www.n4c.eu/Download/n4c-wp2-023-dtn-infrastructure-fs-12.pdf, Jul. 2010

[9] A. Doria, M. Udén, D. Pandey. Providing connectivity to the Saami nomadic community", Proceedings of the 2nd International Conference on Open Collaborative, Design for Sustainable Innovation (dyd 02), Bangalore, India , December 2002

[10] E. Davies, D2.1 N4C System Architecture, Folly Consulting, N4C project http://www.n4c.eu/Download/n4c-wp2-004-sys-arch-04.pdf, Apr. 2009

[11] Linux Hotplug Project. Linux Hotplugging, http://linux-hotplug.sourceforge.net/, Jul. 2010

[12] OpenWrt Wireless Freedom. Homepage, http://openwrt.org/, Jul. 2010

[13] G. Kroah-Hartman. Kernel Korner-udev & mdash Persistent Device Naming in User Space, Linux Journal, No. 122, Jun. 2004

[14] Hotplug2 project, http://isteve.bofh.cz/~isteve/hotplug2/, Jul. 2010

[15] M. Udén, S. Grasic, J. Näslund, M. Z. Božnar, B. Grašič, S. Vrbinc. D8.3 Test bed creation Methodological report, Luleå University of Technology, N4C project, http://www.n4c.eu/Download/n4c-ltu-40-D8_3-1_0.pdf, Feb. 2010

[16] M. Z. Božnar, A. Doria, M. Udén. D8.1 PART A: Preparation and Planning of Tests PART B: Technical Kick-Off Plan PART C: Performance / research indicators for project review, MEIS d.o.o., Luleå Institute of Technology, N4C project, http://www.n4c.eu/Download/D811_1_Test_plan_w_tech_meet_n4c-wp8-1-2_6.pdf, Sep. 2008

OPEN SOURCE SOFTWARE FOR CLOUD COMPUTING

Bojan Novak
University of Maribor, Faculty of Electrical Engineering and Computer Science
Institute for Informatics
Smetanova ul. 17, 2000 Maribor Slovenia
e-mail: novakb@uni-mb.si

ABSTRACT

Open source software alternative to commercial software for Cloud computing is presented. In this paper is described a Linux-based software architecture that implements scalable, efficiency-enhancing private and hybrid clouds providing Infrastructure as a Service (IaaS). Users can provision their own collections of resources (hardware, storage, and network) via Eucalyptus' self-service interface on an as-needed basis. A Eucalyptus cloud is deployed across an enterprise's "on-premise" data center. The cloud is accessed by users over enterprise intranet. Eucalyptus consists of Cloud Controller (CLC), Walrus Storage Controller (WS3), Elastic Block Storage Controller (EBS), Cluster Controller (CC) and Node Controller (NC). User defined images are deployed on the cloud.

1 INTRODUCTION

Open-source software expands at many businesses during the extended economic downturn, and one of the areas where it is starting to offer companies a lot of flexibility and cost savings is in cloud computing. Cloud deployments can reduce costs, free businesses from vendor lock-ins, and offer flexible ways to combine public and private applications. Among most popular implementations, using cloud technology are: EUCALYPTUS, OpenNebula, Red Hat's cloud, Traffic server, Cloudera, Puppet, Enomaly, Reservoir in Zoho [1]. In the paper the implementation of the EUCALYPTUS cloud platform is presented. Company has got its name, from the combination of the first letters from the phrase: "Elastic Utility Computing Architecture for Linking Your Programs to Useful Systems" [2].

Eucalyptus Systems is an open-source software infrastructure for implementing cloud computing on clusters of computers linked together in a way that allows the group to work as one computer. Companies use software from Eucalyptus to build private, public, or hybrid clouds. You can use Eucalyptus software along with IT resources (servers, networks, storage) to form your own data center into a private cloud. Many companies building private clouds are concerned about the lack of standards in this emerging area. They don't want to make a choice now that will lock them in to a specific vendor and make it hard to switch in the future. Eucalyptus provides an integrated set of *application programming interfaces (APIs)* that are compatible with Amazon Web Services, including Amazon's Elastic Compute Cloud (EC2), Amazon Simple Storage Service (S3), and Amazon Elastic Block Store (EBS). Eucalyptus Systems supports developing a large ecosystem of third-party software solutions to help increase demand for its private cloud platform. So far, Eucalyptus has partnered with some key cloud companies including Upscale, Canonical, CohesiveFT, Rightscale, rPath and Ylastic, providing solutions that extend the usability and accessibility of Eucalyptus.

2 CLOUD PLATFORM

The architecture of Eucalyptus (Figure 1), has been designed as modular set of 5 elements that can be easily scaled:

- Cloud Controller (CLC)
- Walrus Storage Controller (WS3)
- Elastic Block Storage Controller (EBS)
- Cluster Controller (CC)
- Node Controller (NC)

Each element is acting as an independent web service that exposes Web Service Description Language (WSDL) document defining the API to interact with it.

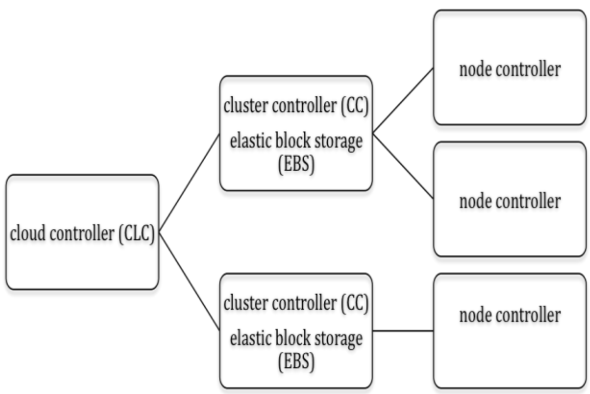


Figure 1: *The cloud platform structure.*

2.1 Cloud Controller – CLC

The Cloud Controller (CLC) is the most visible element of the Eucalyptus architecture, as it is providing the interface with which users of the cloud interact. This interface is comprised of a standard SOAP based API matching the Amazon EC2 API, a simpler "Query Interface" which euca2ools uses and a traditional web interface for direct user interaction. The CLC communicate with the Cluster Controllers (CC) and makes the top level choices for allocating new instances. This elements holds all information linking users to running instances, the collection of available machines to be run, and information about the load of the entire system.

2.2 Walrus storage controller -WS3

The Walrus Storage Controller (WS3) implements a REST (Representational State Transfer) and a SOAP (Simple Object Access Protocol) API which are compatible with Amazon Simple Storage Protocol (S3). It is used for:

• Storing the machine images (MI) that can be instantiated on our cloud,
• Accessing and storing data.

WS3 should be considered as a file level storage system. While it does not provide the ability to lock a file or portion of a file, users are guaranteed that a consistent copy of the file will be saved if there are concurrent writes to the same file.

2.3 Elastic block storage controller - EBS

The Elastic Block Storage Controller (EBS) runs on the same machine(s) as the Cluster Controller and is configured automatically when the Cluster Controller is installed. It allows to create persistent block devices that can be mounted on running machines in order to gain access to virtual hard drive. Storage volumes behave like raw, unformatted block devices, with user supplied device names and a block device interface. You can create a file system on top of EBS volumes. EBS also provides the ability to create snapshots of volumes, which are stored on WS3. At the network level, the block device is accessed using ATA over Ethernet (AoE).

2.4 Cluster Controller - CC

The Cluster Controller (CC) is between the Node Controller and the Cloud Controller. As such, it needs to have access to both the Node Controller and Cloud Controller networks. It will receive requests to allocate MI (machine images) from the Cloud Controller and in turn decides which Node Controller will run the Minst (machine instance). This decision is based upon status reports which the Cluster Controller receives from each of the Node Controllers. It can also answer requests from the Cloud Controller asking for its left overcapacity to run specific instance types, hence allowing the Cloud Controller to decide on which cluster to run new instances. The Cluster Controller is also in charge of managing any virtual networks that the Minst run in and routing traffic to and from them. The Cluster Controller also runs the EBS Controllers. The group formed of one Cluster Controller and EBS Controller and a variable number of Node Controller constitutes the equivalent of Amazon's "availability zones".

2.5 Node controller - NC

The Node Controllers' (NC) software runs on the physical servers on which the MI will be instantiated. The NC software role is to interact with the OS and hypervisor running on the node, as instructed by the Cluster Controller. In our case we use XEN type of virtualization the same as AMAZON [4]. The Node Controller collets data about the environment on which it runs in term of available resources (disk space, type and number of cores, RAM). On the NC virtual machines. are running. They could be started independently of the NC, CC, and CLC. The Node Controller perform any requested tasks from the Cluster Controller (start and stop instances) or replies to availability queries. When requested to start a MI, it will:

1. Verify the authenticity of the user request;
2. Download the image from WS3 (images are cached. Starting multiple instances of the same machine image only downloads that image once),
3. Create the requested virtual network interface,
4. Start the instance of the machine image running as a virtual machine (VM).

Stopping a virtual machine corresponds to performing the opposite operations in the order 1, 4, 3.

From the cloud structure on the figure (Figure 2), a simple implementation, running on the single server is obvious. It was build as teaching aid and for demonstrational purposes. By adding hardware resources (more CCs and NCs) it could be easily expanded to the form suitable for real life applications.

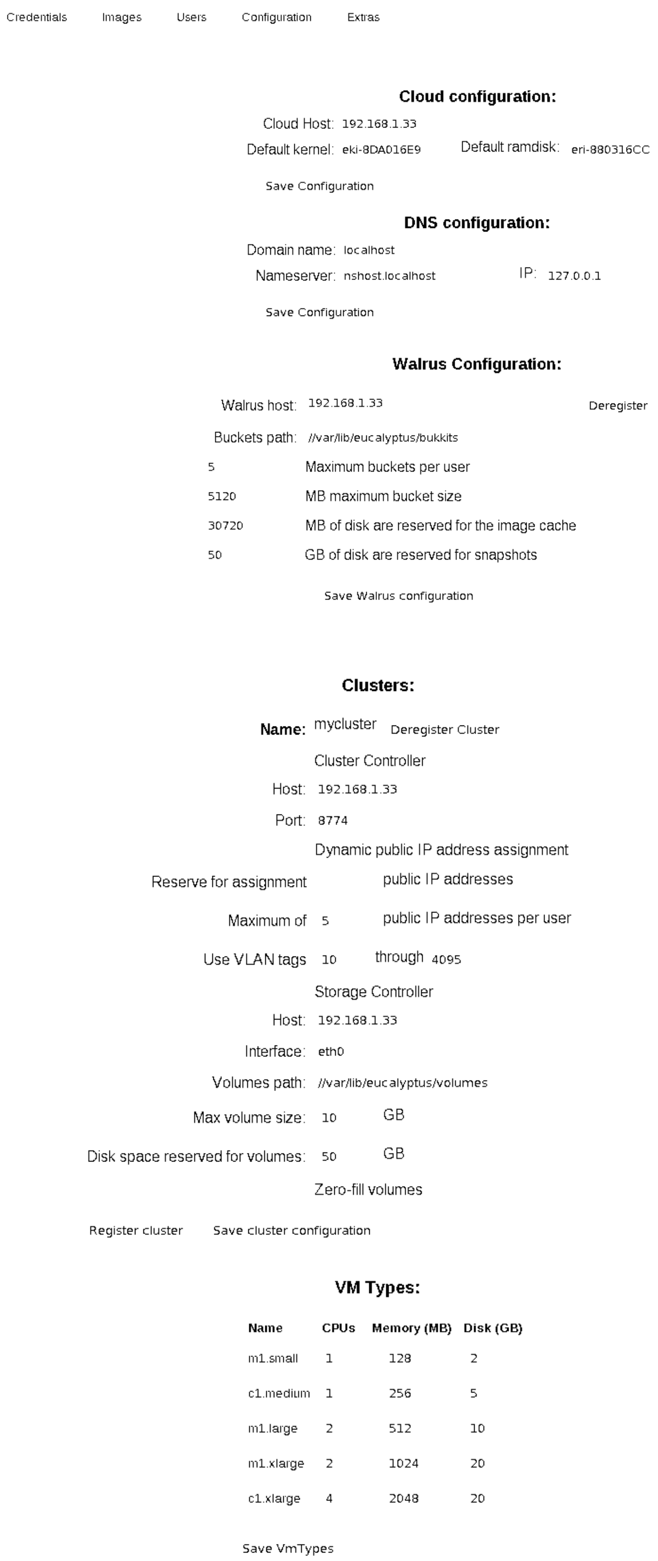


Figure 2: *Cloud configuration*

3 MACHINE IMAGES DEPLOYMENT

Machine images (MI) are created from prebuilt images for the particular operating system [3]. In our case it is the euca-centos-x86_64.tgz. This archive consists of centos-x86_64.img, and two files initrd and vmlinuz for XEN and KVM virtual eviroments. Thus a complete oprating system is provided, that can be deployed on the cloud.

Cloud administrator provides every user his own resources: number of processors, RAM, disk space on the node controler and unique key for every machine image and its unique identification number.

User deploys machine image with the command:

euca-run-instances $EMI -k mykey -t c1.medium.

For the machine image, that runs on the NC an abbreviation Minstis used.

DHCP server automaticaly provides IP address for the particular Minst. User can access idirectly his Minst using ssh command and provided key (mykey):

ssh -i ~/.euca/mykey root@$IPADDR

From now on it depends only on the user for what purpose is he going to use his Minst. He must install additional software for his applications.

4 CLOUD IMPLEMENTATION FREQUENCY

Technically cloud computing can be implemented on various free linux distributions and freeBSD. Very important factor is the type of a virtualization that is implemented on the node controller. The Ubuntu edition support KVM only. But KVM, from my own experience, performances scale very badly with more than 2 CPUs per Minst. But speed decrease of an application that runs in XEN virtual machine, is almost linear with the rising number of CPUs. The most reliable implementation of XEN virtualization is provided in the CentOS linux distribution. In the Table 1. Is presented the share of different linux distributions for cloud platforms [2]:

CentOS	25%
Debian	17%
openSUSE	4%
RHEL	8%
SLES	2%
Ubuntu (UEC)	44%

Table 1: *Cloud implementation share among distributions*

5 CONCLUSION

Cloud deployments can reduce costs, free businesses from vendor lock-ins, and offer flexible ways to combine public and private applications. Among most popular implementations, using cloud technology are: EUCALYPTUS, OpenNebula, Red Hat’s cloud, Traffic server, Cloudera, Puppet, Enomaly, Reservoir in Zoho [1]. In the paper the implementation of the EUCALYPTUS cloud platform is presented. Companies use software from Eucalyptus to build private, public, or hybrid clouds. Eucalyptus provides an integrated set of *application programming interfaces (APIs)* that are compatible with Amazon Web Services, including Amazon’s Elastic Compute Cloud (EC2), Amazon Simple Storage Service (S3), and Amazon Elastic Block Store (EBS). So far, Eucalyptus has partnered with some key cloud companies including Upscale, Canonical, CohesiveFT, Rightscale, rPath and Ylastic, providing solutions that extend the usability and accessibility of Eucalyptus.

Machine images (MI) are created from prebuilt images for the particular operating system [3]. In our case it is the euca-centos-x86_64.tgz. This archive consists of centos-x86_64.img, and two files initrd and vmlinuz for XEN and KVM virtual eviroments. Thus a complete oprating system is provided, that can be deployed on the cloud.

Cloud administrator provides every user resources: number of processors, RAM, disk space on the node controler and unique key for every machine image and its unique identification number. Once image is deployed it depends only on the user for what purpose is he going to use it.

References

[1] http://gigaom.com/2009/11/06/10-top-open-source-resources-for-cloud-computing/

[2] http://www.eucalyptus.com/

[3] https://help.ubuntu.com/community/UEC

[4] http://aws.amazon.com/ec2/

44

COLLABORATION in the VIRTUAL CENTER for ENTREPRENUERSHIP

Tatjana Welzer
University of Maribor
Faculty of Electrical Engineering and Computer Science
Maribor, Slovenia
welzer@uni-mb.si

Metka Zorič Venuti
University of Maribor
Centre for E-learning and Lifelong Learning
Maribor, Slovenia
zoric-venuti@uni-mb.si

Anthony. E. Ward
University of York
Centre for Excellence in Teaching and Learning in Enterprise
Heslington, U.K.
aew6@york.ac.uk

ABSTRACT

The Virtual Centre for Entrepreneurship (VCE) is a virtual education centre for the development of entrepreneurial skills and competencies. The centre is preparing intensive vocational courses (e-learning modules), mostly for employed learners with the workload ranges 1 or 2 ECTS. The VCE provides learners with the opportunity to develop their competencies in small pieces (steps) as and when they need them or as a part of an overall learning strategy. Providers of e-learning modules are well-known higher education institutions all over Europe, which ensures a multilingual approach and the need for cultural awareness from all participants (developers, producers, teachers and learners). From the collaboration point of view we have to be aware that participants are coming from different institutions, from different language groups and cultural environments and at last but not least they have different roles in the system.

In the paper we will present the VCE more detailed as well as different roles of participants and first experiences of participants in those different roles with stress on languages and cultural awareness. Importance of both for successful education in the frame of VCE will be also shortly presented.

1 INTRODUCTION

Successful collaboration is nowadays mostly more dependent on the global communication systems, which are enabled by information and communication technology as on partners in the process. It seems that partners do not need to know each other personally, but it is enough that they have a good technical support in computerized collaborative tools and software. This also means that collaboration (cooperation) is irrespective of the geographical location and can be spread all around the world. The collaboration that we present in this paper will be concentrated in the first phase mostly to the Europe, while partners, 60 of them, which collaborate in the ELLEIEC (Enhancing Lifelong Learning for the Electrical and Information Engineering Community) project, are coming from 30 European countries.

The main objective of the ELLEIEC is to establish a virtual education centre for the development of entrepreneurial skills and competences. The centre is called Virtual Centre for Entrepreneurship (VCE) [1]. The centre is preparing intensive vocational courses (e-learning modules), mostly for employed learners. [1]. Providers of e-learning modules are partners in the project and coming from higher education institutions all over Europe. This also means that they are coming from different language and cultural environment, what can have also an important influence on collaboration of all involved parties.

Further, we will present a structure of the VCE and VCE modules and stress the importance of diversity of languages as well as cultural awareness. At last but not least we will present experiences of partners collaborating in the VCE activities according to the roles that they have in the system.

3 VCE – DESCRIPTION

The VCE provide training in the field of Electrical and Information Engineering. The facilities of the centre are open to students at all levels of their education and to individuals of all ages who wish to engage. [1]. The main activities supported by the VCE are the following [4]:

- E-learning system that individuals can register and select a module that would like to work on it. Whereby learners can also develop their language skills in a foreign language as part of their learning.
- Teaching resources (VCE e-modules) available for partners for freely use for their own teaching purposes;
- A reference repository of relevant research publications in the field of enterprise in Electrical and Information Engineering and in Pedagogy and Assessment;

Each VCE e-module has clear defined aims presented through professional content, communication, collaboration and differentiation (Figure 1) [12]. Additional we have to point out possible obstacles existed because of different languages, multicultural participation and using of an e-learning environment (Figure 1) [12]. Especially important is multilingual and multicultural environment. Learners and teachers/tutors have to be aware of them.

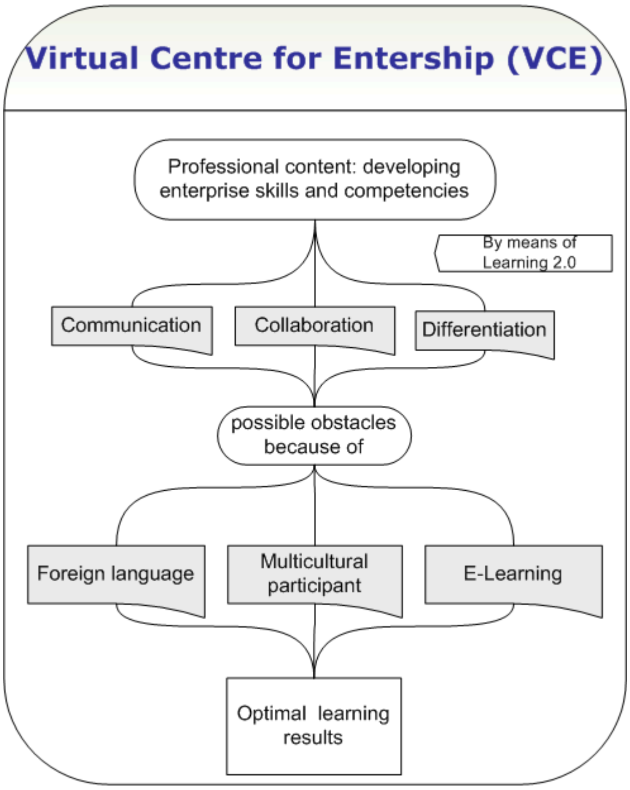


Figure 1: *VCE e-module concepts [12]*

Because of multilingual and multicultural participants, we expect usage of different languages, what is giving the VCE possibility for collaboration all around the world in its second phase. The VCE welcome page is already at this phase of development available in 5 different languages: English as the basic language, French, Spanish, Turkish and Slovene. Figure 2 is presenting just three of them – English, Spanish and Slovene.

Figure 2: *VCE welcome page in 3 languages*

3 FOREGIN LANGUAGES and CULTURAL AWARENESS

Language has an extremely important role in teaching and learning processes. Misunderstandings can happen and cause serious mistakes if words are imprecise [12]. In case of VCE, most participants will take an e-module in a most probably foreign language. Exception will be English native speakers while our VCE e-modules are at the moment only in English language. But independent to this fact all students could be conflicted with the language problem, while tutor can speak different language as student and already before mentioned misunderstandings can happened, some ideas can get lost, learner's role in collaborative learning activities can be missed. Due to language barriers, student also need to have more time for learning and understanding learning material as well as answering or preparing some other materials demands in the course. We have to deal also with a lower level of communication and participation in different forms of discussions. On the other side all participants besides the described language obstacles have a great opportunity to improve the language skills and experts vocabulary.

According to R.D.Lewis is language poor communication tool unless each word or phrase is seen in its original cultural context [7]. This means that we have to take care also about the cultural awareness. In literature, we are confronted with many definitions of culture. Hofstede defines culture as a collective phenomenon, because it is shared with people who live or lived within the same social environment. For him culture consists of unwritten rules of social game. It is the collective programming of the mind that distinguishes the member of one group or category of people from others [5], [6]. Even more than definition of culture is important to clarify cultural awareness. No or poor cultural awareness means poor understanding of the intercultural dialogue, which can lead to blunders and damaging consequences, especially in business, management and advertising [8], where cultural awareness seems to be of key importance for success. However, engineering and many other areas are also not immune against it [9]. For that reason, knowledge of cultural awareness should not be missed in any study program. If cultural awareness is the key of business, it should not be missing in other areas, including engineering and medicine.

Cultural awareness is the foundation of communication, and it involves the ability of observing our cultural values, beliefs and perceptions from the outside [2]. It becomes important in communication with people from other cultures, and we have to understand that people from different cultural environments can see, interpret and evaluate things in different ways [12], [11]. Becoming aware of culture is a difficult task since culture is not conscious of us – it is like water to fish – we live and breathe through it [10]. To avoid these cultural obstacles and follow the mentioned definitions, teachers as experts with the knowledge of culture diversity have to introduce into VCE e-modules examples and discussions that are covering different cultural points of view offering different way of communication, collaboration and expression supporting cultural diversity.

4 PARTICIPANTS in the VCE

In the previous chapters we learned that participants in the VCE have different roles or that they can appear as students, tutors, teachers,...Some of participants can appear as the same person also in different roles, may be in extreme cases also in all possible. So far we have discovered the following roles [4]:

- Learner – participant engaged with the VCE, with the intention of undertaking learning.
- Mentor/tutor/teacher – participant approved to mentor/tutor/teach students in one or more VCE e-modules.
- Mentor moderator – participant approved to view the credentials of a prospective mentor and approve the individual to be a mentor.
- Developer/producer – participant who is developing/producing a VCE e-module.
- Resource moderator/editor – participant approved to view submitted learning VCE e-modules and approves them as usable.
- Visitor – participant who wishes to access the VCE and has to registered also only to visit the VCE and not take the part in learning.
- Administrator.

We would like to stress experiences that we have collected from first learners and tutors/teachers as well as developer/producer – the whole system is still in the prototype phase and just few VCE e-modules have been developed until now. Other roles except administrator are not fully functional at the moment and administration seems not to be involved in language and culture obstacles, while it will be done on one place in one language and culture. Of course the problem could appear between administrators and project leaders, tutors/teachers and developers/producers but not from the point of collaboration in the VCE, but collaboration in the project.

The comments that we collected from learners are the following: they prefer this approach, they prefer the amount of learning material, they have some problems with e-learning tool, they expect more examples (culture oriented), more detailed explanations, tutor's availability and life contact, language itself seems not to be problem at least for most of learners already registered in VCE.

Tutors/teachers/mentors as well as developers/producers are faced with other problems: teaching face to face is easier; development of VCE e-modules is time consuming. Tutors/teachers/mentors have to be trained how to mentor and also how to teach, even more intensive training is needed for developer/producer while preparing an e-module (lecture) is much more pretending as in the case of face to face teaching and at last but not least quality assessment of courses is needed. No real language or culture problems appear.

5 CONCLUSION

In the paper, we have tried to emphasis the importance of collaboration in systems which are active in multicultural and multilingual environments and which can easily assume international character, just as VEC does. In such e-environments, we have to be aware of culture and appearance of foreign languages in the dialogue between teachers and learners, teachers and developers as well as among teachers, developers and learners itself.

To sum up, cultural awareness in general means being open to ideas of changing cultural attitudes or being sensitive to the difference between how we would like to be perceived by others and how we are actually perceived by others. Cultural awareness recognizes that we are shaped by our cultural background, which influences how we interpret the world around us, perceive ourselves and how we relate to other people [8], [9].

Also problems with understanding different languages can appear. The collection of available comments did not point out that either learners, either teachers/tutors or developer would have any significant problems with foreign languages. Of course opinions have been collected on the relation learner-teacher and teacher – developer but not between teachers, developers and students. At last but not least more information have to be collected also for cases that one participant can take different roles – will he/she change his/her opinion on some problems according to the role? This could probably change results, which are because of a small number of available answers not really relevant, but are good background for further development of collaboration in the VCE.

References

[1] H. Yahoui. ELLEIEC – Enhancing Lifelong Learning for the Electrical and Information Engineering Community, Project Report, 2010

[2] S.C Schneider, J-L. Barsoux, *Managing Across Cultures*, Prentice Hall, Harlow, 2003.

[3] T. Welzer, M. Družovec, P. Cafnik, M. Zorič Venuti, H. Jaakkola. Awareness of Culture in e-learning. In: *ITHET 2010*, IEEE, 2010, pp. 312-315.

[4] C. Perra, H. Yahoui, T. Ward. An Enterpreneurship Center in the lifelong learning spirit. In: *ITHET 2010*, IEEE, 2010, pp. 215-218.

[5] G. Hofstede. *Culture's Consequences, Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Sage Publications, Thousand Oaks, 2001.

[6] G. Hofstede, G.J. Hofstede. *Cultures and Organizations: Software of the Mind: Intercultural Cooperation and its Importance for Survival*. McGraw-Hill, New York, 2004.

[7] R.D. Lewis. *When Cultures Collide, Managing Successfully Across Cultures*. Nicholas Brealey Publishing, London, 2007.

[8] B. Farsides. Cultural Awareness and Common Unferstanding: The Key to Informeds Consent? http://tbethic.org/conf18.htm, last visit June 17, 2010.

[9] T. Welzer, M. Bonačić, M. Zorič Venuti. Cultural awareness in information society. In: Proceedings of *20th EAEEIA annual conference Innovation in education for electrical and information engineering*, 2009, IEEE, 2009, 4 f.

[10] T. Welzer. Cultural and security issues in knowledge management. In: Proceedings of *Znalosti 2009, (Edícia zborníkov Informatiky a informačných technológií)*, 2009, pp. 25-29.

[11] S.Quappe, G. Cantatore. What is Cultural Awareness anyway? How do I build it?. Cultuosity.com.

[12] M. Zorič Venuti, T. Welzer, A.E.Ward. Key Issues Concerning the Simultaneous Development of Technical and Linguistic Components in a single E-Module In: *DeSE 2010*, IEEE, 2010.

[13] T. Welzer, M. Zorič Venuti, A.E.Ward, H.Yahoui. Implementation of an E-learning Module in Virtual Centre for Entrepreneurship: The development of cultural awareness in students In: *DeSE 2010*, IEEE, 2010.

ACKNOWLEDGMENT

Thank to the European Commission for the grant to the ELLEIEC project in the Lifelong Learning Programme. Project Number: 142814-LLP-1-2008-FR-ERASMUS-ENW

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Kognitivne znanosti

Cognitive Sciences

Uredili / Edited by

Urban Kordeš, Matjaž Gams, Olga Markič

http://is.ijs.si

13. oktober 2010 / October 13th, 2010
Ljubljana, Slovenia

PREDGOVOR

Letos že sedemnajstič nadaljujemo tradicijo vsakoletnih srečanj kognitivnih znanstvenikov v okviru Slovenskega društva za kognitivne znanosti in že dvanajstič pod okriljem multikonference »Informacijska družba«.

Interdisciplinarno raziskovalno področje kognitivnih znanosti vključuje discipline s celotnega znanstvenega spektra: nevrologijo, psihologijo, računalništvo, lingvistiko, filozofijo in družbene vede (predvsem antropologijo in sociologijo). Področje raziskovanja se je v zadnjih letih razširilo od raziskovanja miselnih in višjih duševnih procesov tudi na čustva in na drugi strani na družbeno "kognitivno" delovanje. V delo kognitivnih znanosti se vključujejo tudi znanstveniki s področja fizike, kemije, biologije, ekonomije itd.

Temeljno vprašanje kognitivnih znanosti kot "dežnika" vseh disciplin, ki se v svojem raziskovanju dotikajo duševnih procesov, je, kako integrirati raznolike pristope, saj se vsaka od disciplin problemov loteva s svojega zornega kota in uporablja svoj strokovni jezik in svoje metode.

Uvedba ustreznega načina povezovanja spoznanj znanstvenih disciplin s spektra kognitivnih znanosti bi že sama po sebi rešila nekaj perečih problemov, na čelu s tako imenovanim "težkim problemom" (povezavo med duševnim in telesnim). Ravno tovrstno povezovanje je osnovni cilj kognitivne konference, ki jo prirejamo v okviru multikonference Informacijska družba.

Z vzpodbujanjem sodelovanja bi radi vzpodbudili vznik interdisciplinarne raziskovalne paradigme, ki bo združevala in nadgrajevala raziskovalne pristope, ki so se uveljavili v dosedanjem razvoju kognitivnih znanostih.

Zaradi tega vsako leto - in letos ni bila nobena izjema - poskušamo zbrati čim več raziskovalcev s širokega področja raziskovanja kognicije. Letos nas še posebej veseli, da konferenca (in z njo pričujoči zbornik) spet postaja bolj mednarodna z udeležbo kolegov z Dunaja in Zagreba.

Posebna tema letošnje konference je IGRA. Ta zanimiv kognitvni fenomen omogoča skoraj neskončno pogledov, vidikov opazovanja, definicij itd. (Med znanstveniki, ki se ukvarjajo s tem fenomenom, večkrat slišimo, da je igra tako lahko prepoznavna in preprosta , pa vendar skoraj nedostopna znanstvenemu razumevanju.) Poleg tega se igra navezuje na celo paleto izjemno zanimivih sorodnih tem s področja kognitivne znanosti kot so zabava, radost, estetska percepcija, umetniški izraz, pa tudi čustva in njihove modulacije. Prispevki, predstavljeni v tem zborniku seveda ne zajemajo vseh možnosti, ki jih letošnja igriva tema ponuja, kljub temu pa dobro odražajo interdisciplinarno bogastvo možnosti raziskovanja človekove duševnosti. Prepričani smo, da bo bralec prišel na okus in da bodo teksti vzpodbudili radovednost.

Prispevki so razdeljeni v dve skupini. V prvi so tisti, ki so (pa četudi zelo na široko) povezani z igro, estetsko percepcijo itd. Izbor letošnje teme nas je pripeljal do odločitve, da temu delu dodamo še razmišljanje dveh umetnic, ki morda ne sledi vsem znanstvenim pravilom, dodaja pa pomembno transdisciplinarno začimbo.

Seveda so bili tudi letos dobrodošli prispevki, ki se ukvarjajo z drugimi kognitivnimi fenomeni. Še posebej takšni, ki se katerega od zanimivih vprašanj lotevajo interdisciplinarno. Slednji so zbrani v drugem delu zbornika.

Urban Kordeš
Matjaž Gams
Olga Markič

PLAYING IN THE COGNITIVE SCIENCES – AN APPROACH TO MAKING SENSE OF THE WORLD

Elisabeth Zimmermann
Department of Philosophy
University of Vienna
Universitätsstr. 7, A-1010 Wien, Austria
elisabeth.zimmermann@univie.ac.at

ABSTRACT

The article aims to depict, why play is an interesting phenomenon to be studied by cognitive science. Being a phenomenon that is hard to define, it involves many other cognitive abilities, like reasoning, decision-making, memory, consciousness, etc. The paper especially focuses on the relation of play and epistemic processes, based on an enactive approach to cognition.

1 INTRODUCTION

Why should a cognitive scientist devote his/her time to research the phenomena of play and games? Is it worth to deal with a phenomenon that is so wide and so hard to grasp and is often discredited as a scientific subject especially for that reason as well as for its subjective, experiential dimension?

Probably many people are lured to study play because they can relate to play experiences in their lives, which were enjoyable and therefore stayed in their memory. Maybe it is the levity of childhood that we sometimes long for and therefore want to remember and experience as grown-up persons. Or it simply is the fun we have in playing games with friends and colleagues, with our children or even with our pets, that encourages us to have a closer look on the phenomenon of play?

The interesting aspect of play for cognitive science is the fact that it presupposes and involves a lot "cognitive" phenomena, like reasoning, planning, decision-making, consciousness, emotions, etc. Researchers from many of the core disciplines of cognitive science have studied aspects of *play*, *playing* or *games* in one or the other way in the last decades and years, e.g. in computer science/artificial intelligence the design of computer games has become a major issue (for an introduction to this topic see Salen & Zimmerman, 2004), developmental psychologists have studied play development in children for a long time (e.g., Vygotsky, 1978 (1930); Piaget, 1975 (1959); Smith, 2010), in ethology play has more recently become a topic, especially because of the cognitive abilities that it involves (e.g. being able to understand intentions of conspecifics, using signals, cooperating), which are often difficult to study in animals. (Fagen, 1981; Burghardt, 2005; Allen & Bekoff, 1997)

There also have been attempts to study play on a more theoretical and conceptual level, trying to point out characteristics of play, relate different approaches and establish a kind of framework, in which different forms of play can be located. The well-known play researcher Brian Sutton-Smith has provided a wonderful overview in his book *The Ambiguity of Play* (Sutton-Smith, 1997), where he draws the attention to the fact, that there are many ways to talk about play, many rhetorics one can use, depending on the focus (play as progress, imaginary play, self play, fate play, power play, identity play, frivolity play). While the first three rhetorics are more modern ones and currently used to describe and explain play behaviour, the later four ones have been more common in former times. It is useful to be aware of the kind of rhetoric authors rely on, when trying to study play in an interdisciplinary manner.

The fact that play often appears to be purposeless behaviour has instigated researchers to search for a functional explanation for play behaviour. There have been various theories on the function of play. The idea, that play is a kind of training for later life, dates far back in time. Already Plato wrote in *The Republic* and *The Laws* about the importance of play in education, especially for learning skills that are later needed in the work life (Plato, *The Laws*; cited in Smith, 2010). The idea was further developed and became famous with Karl Groos' *instinct practice theory* (Groos, 1898), where he described play as a necessary behaviour to perfect instinctive mechanisms, thereby increasing the reproductive fitness of an organism and its chances for survival.

While *instinct practice theory* advocates an ultimate function of play behaviour, the *surplus energy theory* by Herbert Spencer proposes a more proximate function of play. Spencer suggested in his book *The Principles of Psychology* (1872, cited in Burghardt, 2005) that evolutionary "higher animals" (mammals) developed more efficient ways in dealing with everyday necessities (finding food resources, mating, etc.) and due to that have an energy surplus, they need to get rid of. Play behaviour is one way to discharge this surplus in artificial exercise.

More recently an even more immediate function of play has been proposed by Jaak Panksepp. In his opinion „[I]t is worth considering that the main adaptive function of play may be the generation of positive emotional states. In such states animals may be more willing and more likely to behave in flexible and creative ways. " (Panksepp, 1993). Alongside creating positive emotional experiences play may also modulate the arousal level of an organism. This may be an important aspect for learning processes. I will later come back to this point, when describing the differences of play and explorative behaviour.

In this article I want to focus mainly on one aspect of play, that seems to be especially interesting from the view of a cognitive scientist having an embodied, situated, enactive approach to cognition, namely, its relation to movement and activity. Bodily-mediated cognition has been emphasised in opposition to classical cognitive science ideas of cognition as internal computation of symbols. "… embodiment means that mind is inherent in the active, worldful body, that the body is not a puppet controlled by the brain but a whole animate system with many autonomous layers of self-coordination and self-organization and various degrees of openness to the world that create its sense-making activity." (Di Paolo et al., 2010)

Understanding cognition as an active process, in which an organism constructs knowledge about the world (Maturana & Varela, 1980), one cannot overlook the importance of playful behaviour in the process of learning about the world one lives in, about oneself and about the relation between oneself and the surroundings, the "Umwelt".

Before going into more detail here, I want to introduce some definitions of play.

2 DEFINING PLAY

In 1938 Johan Huizinga, a Dutch historian, published his highly influential book *Homo Ludens*, in which he studied play as basic element of our culture. He defines play as "a free activity standing quite consciously outside 'ordinary' life as being 'not serious' but at the same time absorbing the player intensely and utterly. It is an activity connected with no material interest, and no profit can be gained from it. It proceeds within its own proper boundaries of time and space according to fixed rules and in an orderly manner. It promotes the formation of social groupings which tend to surround themselves with secrecy and to stress their difference from the common world by disguise or other means." (Huizinga, 1938; English translation: Homo Ludens: A Study of the Play Element in Culture. Boston: Beacon Press 1955). Huizinga's definition captures some of the interesting ambiguities of play. On the one hand play is *outside ordinary life* and *not serious*, but on the other hand the player is totally *absorbed* in the activity, so he takes the play seriously. Play further is a *free activity*, but also bounded by *temporary and spatial constraints* and following *rules*. Those characteristics can also be found in later definitions of play. Especially the aspect of activity is part of many definitions. For example, the ethologists Marc Bekoff and John Byers give the following definition: „Play is all motor activity performed postnatally that appears to be purposeless, in which motor patterns from other contexts may often be used in modified forms and altered temporal sequencing." Bekoff & Byers, 1981, cited in Allen & Bekoff, 1997). Being behavioural biologists, the authors have to rely on observable data; therefore their definition is very behaviour/activity-centered. The interesting point is the *modification of sequences*; this aspect can be related to the freedom mentioned in other definitions like e.g. Huizinga's.

Another definition by the psychologist Rod Martin emphasises further characteristics of play. He defines play as „…an enjoyable, spontaneous activity that is carried out for its own sake with no obvious immediate biological purpose. […] a state of mind rather than a characteristic of certain types of activities. Thus, one can engage in almost any activity in a playful way, as long as one has a nonserious, activitiy-oriented (rather than goal-oriented) mental set." (Martin, 2007). Again the elements of activity, non-seriousness and absence of a goal are included in this definition, but furthermore it mentions the experience of *fun* and *enjoyment*, a characteristic that is hard to grasp, because of its subjective nature. The other important point is that Martin considers play rather to be a *state of mind* than a special form of behaviour, and therefore any ordinary activity can be performed in a playful way, an aspect that also Salen and Zimmerman (2004) consider in their classification of play.

3 DIFFERENCES IN LUDIC ACTIVITIES: "GAME" VS. "PLAY"

Before going into the details of the categorisation of Salen and Zimmerman (2004), I want to stress the fact, that the English language uses two different words for ludic activities: *game* and *play*, whereas, e.g., in German and Slovene there is only one word (*Spiel* or *Igra*). The words *game* and *play* also have different meanings. Games can be considered to be a subcategory of play, as they are a more formalised, rule-based form of play. On the other hand the experience of play(ing) is one aspect of a game. A game can be defined as " … a system in which players engage in an artificial conflict, defined by rules, that result in quantifiable outcome." (Salen & Zimmerman, 2004).

Based on this distinction Salen and Zimmerman (2004) establish three categories of play, (a) *game play*, (b) *ludic activities*, and (c) *being playful*. They describe (a) game play as the formalised interaction that takes places when players play a game, following the rules. (b) Ludic activities are play activities that also include non-game behaviours, as e.g. kittens playing with a yarn. (c) Being playful refers to activities that are not typically play, but are done with a playful state of mind. One example they mention is being playful with words, when e.g. creating nicknames. In these actions "the spirit of play infuses otherwise ordinary actions" (Salen & Zimmerman, 2004). This category relates to

Martin's (2007) definition, that every activity can be performed in a playful way.
The authors then give a very spare and abstract definition of play and show how it accounts for all three categories: "Play is movement within a more rigid structure" (Salen & Zimmerman, 2004). The word *play* is often used in mechanical devices in this way to describe the freedom of a part within the structures of a system. The constraints of the system are an essential property that forms the space in which play can emerge. Thus, the relationship between the structures and constraints of a system and the form of play that emerges within them, the form of play they enable, is of key interest. Notice that also this definition stresses the activity/movement aspect of play.
In Salen and Zimmerman's categorisation the structures of (a) game play are easy to recognise; usually there is some kind of physical setting (e.g. dice, board, computer-joystick, etc.) and the rules of the game, building a rigid formal structure. For (b) ludic activities the structures are less formal, nevertheless existent. Playing with a ball the player is constrained by the physical surroundings, the material of the ball, gravity and also his own physical appearance and bodily skills. "To *play* with the ball is to play with all these structures, testing their limits and boundaries, finding ways of moving around and inside them." (Salen & Zimmerman, 2004). Even for (c) being playful the definition holds. Playing with words one also orients oneself along the structures of grammar.
Seen in this way, play can be understood as an epistemological activity oriented at experiencing structures and constraints, boundaries and limits of the physical, social and/or virtual environment (Umwelt), revealing new perspectives and finding new ways of interaction. The environment provides a structure and opens up a kind of *enabling space* (Peschl & Wiltschnig, 2008*)*, which allows for and enhances certain ways of interaction, thereby enabling the construction of new knowledge.
Salen and Zimmerman (2004) stress that play does not only exist because of more rigid structures but also in opposition to them. Their example is that slang is only slang because it departs from the grammatical norm, or bouncing a ball against a wall is play in contrast to a more utilitarian use of the architectural space. But what about walls that are explicitly made for bouncing balls against them (e.g. on a playground)? There is no more utilitarian way of interaction than bouncing a ball against them. I furthermore question, if this opposition is something that is experienced by players, especially by children. While playing everything around is the structure that is constraining, but also supporting and enabling certain ways of playful interaction that are to be discovered. It is probably only after an adult tells the child "this carpet-beating bar is not made for exercising on it", that he/she thinks of a different use than playing on it. In the very moment the structure only serves the purpose of the play-world.

4 **PLAYING AND SENSE-MAKING**

Interacting with the world, "organisms cast a web of significance on their world" (Di Paolo et al., 2010). In their way of regulating the structural coupling, organisms establish a certain perspective on the world, they make sense of the world. In a way their actions are the source for generating meaning. Di Paolo et al. (2010) describe this kind of meaning generation as being a formative activity, rather than an extraction of information from the world around. This enactive view on meaning generation in an organism is related to the autonomy of organisms. Autonomous systems are not totally determined by predefined rules, but are able to change their limitations due to their internal dynamics. "Cognitive systems are also autonomous in an interactive sense in terms of their engagement with their environment as agents and not simply as systems coupled to other systems (…) …they not only respond to external perturbations in the traditional sense of producing the appropriate action for a given situation, but do in fact *actively regulate* the conditions of their exchange with their environment, and in doing so, they enact a world or cognitive domain." (Di Paolo et al, 2010)
In play this regulation is altered in a way that "normal", standard interaction with the world and its objects (which can be observed in other people and is often learned from them via imitation) can be changed. Thereby also internal (inherent to or initiated by the agent) perturbations are evoked, which can lead to a different enaction of the world, a different reality, maybe "only" a transient play reality, but nevertheless a valid one for that moment. For example, think of children playing with a brush and using it as a microphone. In playful behaviour sensorimotor loops are newly created, constructed and regulated. Established relationships to objects, persons and surroundings can be altered and seen in different way.
Salen and Zimmerman (2004) talk about the transformative character of play, i.e. that the free movement of play can change the rigid structure. The structure, as already said, can be the rules of a game, but also the environment, one's own body, one's own skills, the relationship to other players, etc. These changes are an essential part of developing and learning.

5 EXPLORING AND PLAYING – TWO DIFFERENT WAYS OF LEARNING ABOUT THE WORLD?

When talking about play and its relation to development and learning, one comes across the question: when does play start?; on the one hand during ontogenetic development of an organism and on the other in the concrete situation of being in an environment and starting to interact with it.
A common distinction, which many psychologists draw, is the distinction between play and exploration. Exploration is mainly associated with learning about perceptual properties, whereas play is associated with active manipulation of e.g. objects. Smith (2010) admits that the distinction is difficult

in very young children during their sensorimotor development, but can be better studied in preschool years. Her refers to an experiment done by Corinne Hutt (1966, cited in Smith, 2010), where she gave children a novel toy, a box they could sit on, which had a lever that could ring a bell or a buzzer. Children (aged 3 to 5 years) were rather serious, when introduced to the toy, and started to feel/touch it and move the lever. They explored what the object could do. After some time their behaviour would change. They would sit more relaxed on the box, repeatedly using the lever to make noises. This was considered to be more playful activity. (Smith, 2010)

The idea is that during exploration children learn simple concepts of their environment (mostly via perceptual experience), whereas in play they combine different action schemas into more complex ones. This distinction seems a bit artificial to me, especially when one gives up the linear sense-think-act model of cognition and replaces it by a circular model, where perception and action presuppose each other in turn.

Roberta Collard (1979) sums up the results of this study in the following way: “Corinne Hutt (1966) has suggested that the difference between exploration and play is that exploratory behaviour occurs when novel stimuli raise one’s arousal level and the organism seeks to reduce stimulation, whereas play occurs when the arousal level is low and the organism seeks to increase its level of stimulation.” (Collard 1979). As already mentioned above, a specific emotional state and arousal level seem to be important pre-conditions for play to occur, as they set the player in a kind of *flow* state (Csikszentmihalyi, 1991). But what exactly evokes this change from exploration to play behaviour remains a question for further research.

6 CONCLUSION

I have tried to give a short overview on how play can be understood as an activity that is very much involved in epistemic processes, in making sense of the world we life in, creating meaning and also changing it. Play as a transformative activity enables us to at least try out how we can change through playful interaction. Through it we might be able to see things from a different perspective. Or as Di Paolo et al. put it: “Play is precisely *not* a problem requiring a solution. In fact, play is the breaking of this pattern; or rather its re-deployment into an active construction of meaningful action where no such sense-making is directly demanded from the environment or from definite internal needs. “ (Di Paolo, 2010) But exactly this kind of sense-making could alter our physical and mental capabilities in an interesting way.

References

1. Allen, C. & Bekoff, M. (1997). *Species of Mind.* MIT Press.
2. Burghardt, G.M. (2005). *The Genesis of Animal Play. Testing the Limits.* MIT Press.
3. Collard, R.R. (1979). *Exploration and Play.* In Sutton-Smith, B. (ed.) *Play and Learning*. Gardner Press.
4. Csikszentmihalyi, M. (1991). *Flow. The Psychology of Optimal Experience.* Harper Collins Publishers.
5. Di Paolo, E. A., Rohde, M. and De Jaegher, H. (2010). *Horizons for the Enactive Mind: Values, Social Interaction, and Play.* In J. Stewart, O. Gapenne and E. A. Di Paolo (eds), *Enaction: Towards a New Paradigm for Cognitive Science,* Cambridge, MA: MIT Press.
6. Fagen, R. (1981). *Animal Play Behavior.* Oxford University Press.
7. Foerster, H., von (1993). *Wissen und Gewissen.* Suhrkamp Verlag.
8. Groos, K. (1898). *The play of animals.* New York: Appleton.
9. Huizinga, J. (1956 [1938]). *Homo Ludens. Vom Ursprung der Kultur im Spiel.* Rowohlt Taschenbuch Verlag.
10. Hutt, C. (1979). *Exploration and Play (#2).* In Sutton-Smith, B. (ed.) *Play and Learning*. Gardner Press.
11. Martin, R.A. (2007). *The Psychology of Humor*. Elsevier Academic Press.
12. Maturana H. R. & Varela F. J. (eds.) (1980). Autopoiesis and cognition: The realization of the living. Reidel, Dordrecht.
13. Panksepp, J. (1993). *Rough and tumble play. A fundamental brain process.* In MacDonald, K. (ed.) *Parent-child play*. State University of New York Press.
14. Peschl M. F. & Wiltschnig S. (2008). Emergente Innovation und Enabling Spaces. Ermöglichungsräume für Prozesse der Knowledge Creation. In: Lucke U. et al. (eds.) Proceedings der Tagungen Mensch & Computer 2008, DeLFI 2008 und Cognitive Design. Logos, Berlin: 446–451.
15. Piaget, J. (1975). Nachahmung, Spiel und Traum. Klett-Cotta Verlag.
16. Salen, K. & Zimmerman, E. (2004). *Rules of Play. Game Design Fundamentals.* MIT Press.
17. Smith, P.K. (2010). *Children and Play.* Wiley-Blackwell Publishing.
18. Sutton-Smith, B. (1997). *The Ambiguity of Play.* Harvard University Press.
19. Vygotsky, L.S. 1978 (1930). *Mind in Society.* Harvard University Press.

The pleasure of enacting meaning: Aesthetic perception and pretend play

Claudia Muth
Middle European Interdisciplinary Master Programme in Cognitive Science
Institute of Philosophy
University of Vienna, Universitaetsstrasse 7 A-1010 Wien, Austria
Tel: + 49 8158331
e-mail: die-claude@gmx.de

ABSTRACT

This paper describes an attempt to create a model of aesthetic perception with regard to theories of active perception as well as to links between aesthetic perception and play. Enactivism allows an interesting perspective on pretend play and aesthetics, since both can be understood as special forms of meaning re-enaction. Assumingly, both also share the feature of self-rewarding processing. It is claimed that two seemingly opposing features lead to pleasure and preference: the fluency of processing one the one hand and the perceptual struggle itself on the other hand. This distinction is reflected in reports of aesthetic experience and is referred to as the concepts of immersion and engagement.

1 INTRODUCTION

The artwork "jeu" by Kristof Georgen shows a football made of concrete lying on a cloth with a round cut out. When confronted with that object, a contradiction is created by the material because we cannot fulfill the action that it demands: kicking the ball into the goal. It shall be used here to introduce this paper on various levels.

2 AESTHETICS AS THE PLEASURE OF (RE-) ENACTING MEANING

The irritation mentioned above reflects quite well what accounts of "active", "interactive", or "action-relevant" perception claim: Perception is strongly bound to action, as it is an action (eye- and body-movements) and as to perceive means as well to associate or anticipate action [1], [2], [3]. Perception can even be understood as an active search for opportunities to act. In this way, objects "afford" actions [3]; we even define things according to the associated actions [4]. For enactivism, perception is an active sense-making - an enaction of meaning. Aesthetics could be, quite as pretend play (pretending that something is something else), a re-enaction of meaning [5]. This seems to be a promising direction for a model of aesthetic perception – thinking for example of readymades, industrially produced objects, that turn into artworks because of their new exhibition-context. The spectator in this sense would be able to re-create the objects' meaning.

In art production, the enactive approach holds as well: "In art, the playful attitude becomes interest in the transformation of material to serve the purpose of developing experience." (Dewey 2005, p. 284) [6]. In that sense, Noë calls artists "experience-engineers" [7].

The concept of "re-enaction" could be understood as a way to connect here-and-now-cognition with "higher" cognition [5]. However, this does not explain why we can "break free" entirely of the here-and-now, for example when we think or imagine.

3 AESTHETICS AS SELF-REWARDING PROCESSING

A collection of reports on aesthetic experience revealed one contradiction that is of high interest here: People reported experiences of intellectual insights but, at the same time, of losing themselves entirely in an artwork. This reminds of the concepts of immersion and engagement, which, according to Douglas and Hargadon (2000), are connected in flow experiences. Immersion, in this context, means being absorbed by a familiar schema, whereas engagement is understood as the recognising of contradictory schemata without being absorbed. According to Douglas and Hargadon, pleasure is triggered by two factors: challenge (for example surprises in a plot) and the possibility of relating something to already known schemata. Barriers in adapting the perceived to a known schema would remind the perceiver of his or her „role as (...) sensemaker“ (Douglas & Hargadon 2000, p. 155) [8]. The described irritation when looking at "jeu" could be taken as an example of this feature. If it is claimed that they entail contradictions, are ambiguous or associating conflicting schemata – engaging with artworks often seems to be a challenge for the perceiver. Zschocke (2006) argues that in these cases, irritation induces a reflexive engagement [9].

Is this engagement – similarly to play – motivated by the process itself? This puts the process of aesthetic perception as being rewarding in the centre of investigation and leads to an understanding of aesthetics neither as a subjective nor as an objective but as an interactionist phenomenon. The process itself is „aesthetic“. Further it includes the possibility to "catch ourselves in the act of exploring the world" [7].

3.1 REWARDING FLUENCY

The difference between immersion and engagement could be seen paralleled in two interactionist theories of aesthetics: One assumes that the fluency of processing is responsible for gaining pleasure: The easier or more fluent the processing, the more pleasure we gain. The authors argue that high fluency speaks for a progress, error free processing, accurate knowledge-structures and familiarity. The latter could show, that if I already experienced a similar situation that didn't harm me, this time it wouldn't harm me either and would therefore elicit positive feelings. The mere-exposure-effect as well as priming-effects can be taken as arguments in favour of this explanation [10]. But the predictions of the model contradict findings linking novelty and reward [11].

Wittmann and Bunzeck (2007) assume that this link "might well serve to reinforce exploratory behavior, enabling animals to find new food sources and encode their location, thereby enhancing survival“ (Wittmann & Bunzeck 2007, p. 200). Berlyne's curve illustrates that preference is low if there is too little but also if there is too much novelty [12]. Carbon & Leder (2005) were able to demonstrate that the preference for innovative objects changes with their elaboration [13].

Artworks can be ambiguous but very popular at the same time, which contradicts the theory of "the easier the better" and could point to the fact, that the perceiver is actually seeking for challenge when approaching an object of art. Furthermore, Schmidt (1992) claims that some artworks provide multistability on a continuing basis, enabling the perceiver to keep him- or herself open for meaning [14].

3.2 REWARDING STRUGGLE

One understanding of aesthetic perception that takes more the engagement-side into account, points to so-called "eureka-moments", experiences after a kind of struggle with stimuli like when we recognize an object within seemingly unstructured dots. "Rewarding struggle" would therefore mean that the creation and manipulation of sense itself is rewarding [15]. Assuming that it is the processing itself that is pleasing, explains why we like challenging stimuli, however, it does not answer the question why we can enjoy an artwork again and again.

Whereas the fluency approach provides empirical evidence, this assumption exists only on a hypothetical basis.

3.3 MERGING BOTH APPROACHES

An attempt to unify both approaches – besides Douglas and Hargadon's proposition – is to assume that elaborating challenging artworks means reducing their ambiguity. In a special way, this fits in with the notion of perception as active meaning enaction. At the same time it is plausible to include the importance of fluency as the processing of an ambiguous object would get easier with growing elaboration: Once one got (partial) insights, the artwork would be more familiar. But this only holds if we assume aesthetic processing to be marked by progress.

4 PROBLEM STATEMENT

The general conflicts underlying the described debates are: Is it familiarity or challenge, immersion or engagement [8], is it the fluency of processing [10] or the process itself [15] that influences pleasure and preference in aesthetic perception?

5 A MODEL OF AESTHETIC PERCEPTION

The model depicted in image 1 assumes that it is the process of aesthetic perception itself that is special and pleasing, without excluding that features of the object or attitudes of the perceiver are important preconditions or influences.

Aesthetic perception is in its core "sense-making". It is an ongoing process of meaning-enaction and -manipulation, not always fully dissolving into recognition of a pattern.

The ambiguity of an artwork elicits a tension that can:

1. motivate further engagement as re-enaction of meaning
2. lead to immersion into the perceived (both together can culminate in flow-experiences)
3. result in an awareness of the own perceptual activity
4. be accompanied by (partial) insights
5. turn into a state where instability of meaning is consciously retained.

Pattern-recognition is understood as the basic neuronal mechanism of perception. Ramachandran and Hirstein (1999) hypothesise that the binding of features – achieved through neuronal synchronization – is connected to the reward system, which is to the knowledge of the author not experimentally tested yet. Literature gives no clear answer if pleasure and preference are gained through the fluency of processing or the processing itself. The completion of a pattern could stand in forthe experience of an insight even if modern artworks hardly provide the possibility of understanding in conceptual terms. Challenge seems to be a key word when talking about flow, respectively about the interplay between immersion and engagement.

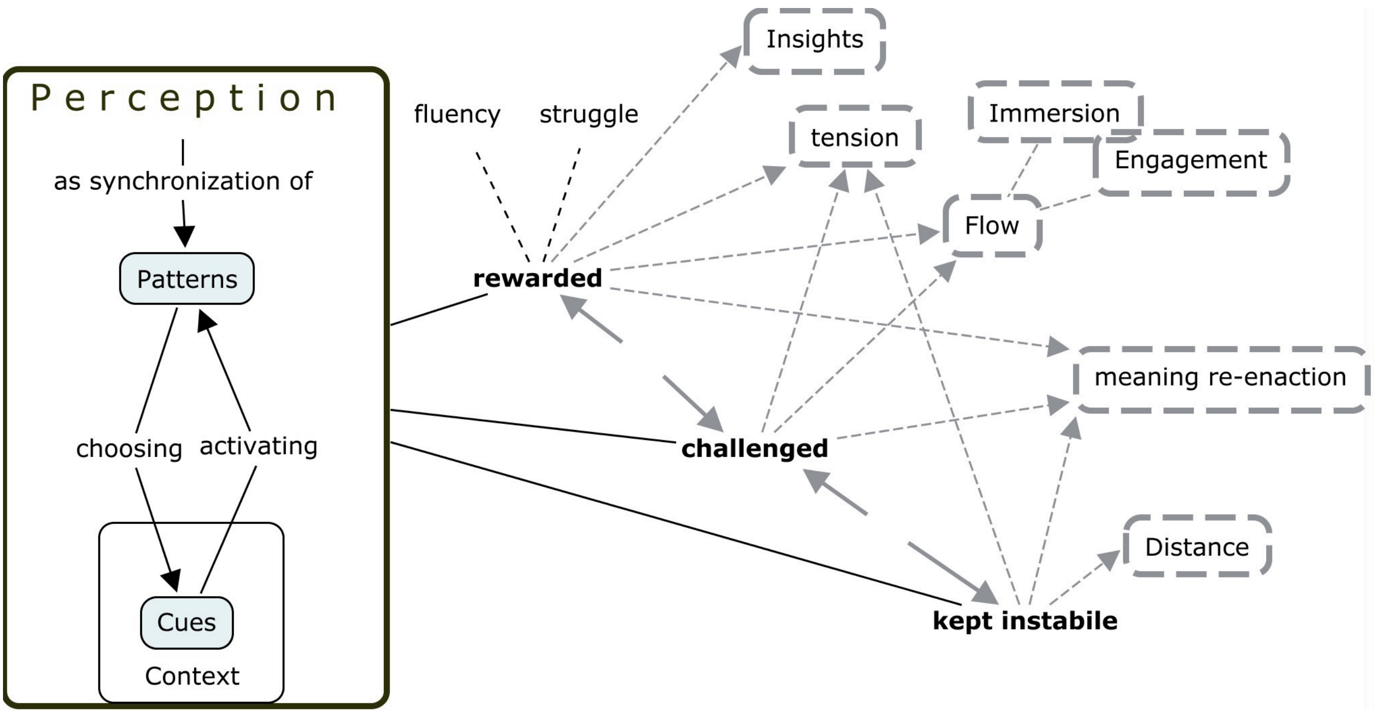


Image 1: A Model of aesthetic perception.

Artworks, attitudes or contexts might pose a challenge that could even be reacted to by retaining instability of meaning. This is implied in the notion of "distance". It gets obvious how this is related to play, especially pretend play where meaning re-enaction creates a kind of multistability, the paradox that something is used as something else [16]. In this way, the concepts in grey lines can be understood as a view on play as well.

6 NOTES

Grkinic (2010) made an important point describing a case of vandalism against an artwork. The photography was understood as insulting religious convictions. Certainly, the attacker's mastering of the artwork included insights but can it be called aesthetic just because we assume processing to be self-rewarding [17]? This highlights the importance of subjectivist aspects, like "disinterestedness" or "distance", and makes clear that not every aesthetic experience is reflected in this model; also not every single condition is necessarily met in one example.

Further, not only artworks can elicit the described effects. The big question is how we get into a "playful" state of meaning re-enaction. Other questions refer to the quality of insight and understanding. What role does simulation play, which are the possibilities of non-verbal understanding, do we find a similar experience in general problem-solving, are there different forms of understanding? Are there specific moments of aesthetic experience or is it a lasting experience? Is insight possible without fully dissolving the problem and how is insight related to flow?

Furthermore, a clarification of the relation between sense-making and the brain's reward-system would be necessary.

Many questions about how to relate the model to reports of aesthetic experience have to stay unanswered. Besides the ones already described, people reported for example mystical experiences, somatosensory perceptions, feelings of fright, presence and notions of beauty. Those categories could lead further research into different directions and should remind of the broad variety of experiences subsumed under aesthetic perception that is not nearly embraced by this work.

References

[1] Churchland, P.S., Ramachandran, V.S., and Sejnowski, T.J. (1994). A Critique of Pure Vision. In: C. Koch, J.L. Davis. *Large-Scale Neuronal Theories of the Brain*. MIT Press, Cambridge, Massachusetts, London.

[2] Findlay, J.,M. ; Gilchrist, I.D. (2003). *Active Vision*. Oxford University press, New York.

[3] Clark, A. (1998*). Being There: Putting Brain, Body, and World Together Again* (Reprint.). Mit Press.

[4] Holzkamp, K. (1973). *Sinnliche Erkenntnis. Historischer Ursprung und gesellschaftliche Funktion der Wahrnehmung*. Athenäum, Frankfurt.

[5] Di Paolo, E., Rohde, M., & De Jaegher H. (2008). *Horizons for the enactive mind: Values, social interaction, and play.*Forthcoming, MIT Press.

[6] Dewey, J. (2005). *Art as Experience*. Perigee.

[7] Noë, A. (n.d.). *Interdisciplines: Art as enaction*. Retrieved from: http://www.interdisciplines.org/artcog/papers/8

[8] Douglas, Y., & Hargadon, A. (2000). The pleasure principle: immersion, engagement, flow. In *Proceedings of the eleventh ACM on Hypertext and hypermedia* (p. 153-160). ACM, San Antonio, Texas.

[9] Zschocke, N. (2006). *Der irritierte Blick*. München: Wilhelm Fink Verlag.

[10] Reber, R., Schwarz, N., & Winkielman, P. (2004). *Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?* Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1295305

[11] Wittmann, B. C., Bunzeck, N., Dolan, R. J., & Düzel, E. (2007). Anticipation of novelty recruits reward system and hippocampus while promoting recollection. *NeuroImage*, 38(1), 194-202.

[12] Berlyne, D. E. (1974). *Studies in the new experimental aesthetics.* Hemisphere Publishing, New York, Washington.

[13] Carbon, C. C., & Leder, H. (2005). The Repeated Evaluation Technique (RET). A method to capture dynamic effects of innovativeness and attractiveness. *Applied Cognitive Psychology*, 19(5), 587–602.

[14] Schmidt, S.J. (1992). *Der Kopf, die Welt, die Kunst.* Böhlau, Köln.

[15] Ramachandran, V., & Hirstein, W. (1999). The Science of Art - A Neurological Theory of Aesthetic Experience. *Journal of Consciousness Studies*, 6(6-7).

[16] Sutton-Smith, B. (2001). *The Ambiguity of Play* (2nd ed.). Harvard University Press, Cambridge, Massachusetts, London.

[17] Grkinic, B. (2010). *Katharsis and aesthetic emotion.* Retrieved from http://othes.univie.ac.at/8178/

ANALYSING TEAM GAMES AS SOCIOPHYSICAL SYSTEMS

Josip Stepanić, Barbara Gavranović
Faculty of Mech.Eng. & Naval Arch. – University of Zagreb
I. Lučića 1, HR-10000 Zagreb, Croatia
Tel: +385 1 6168592; fax: +385 1 6168 466
e-mails: josip.j.stepanic@fsb.hr, barbara.gavranovic@fsb.hr

ABSTRACT

Sociophysics is an interdisciplinary field in which societal characteristics and relations are being formalized and their relations quantified.

In this paper it is argued that team games could be viewed as optimal departure point in further development of sociophysics notions. In particular, team games resemble repetitive experiments from sciences, which is in the basis of its methodology. Furthermore, team games could be either used in existing forms, or in a myriad of possible forms obtained after modification of rules of existing games in variable amounts.

The paper analyses analogues of fundamental physics' quantities and possibilities of their reliable determination in team games.

1 INTRODUCTION

Societal processes can be described and analysed using different methodologies, such as sociophysics [1]. Sociophysics is an interdisciplinary discipline which quantifies elements and relations within a given, usually social system. In accordance with its name, formalism of theoretical physics is applied onto models of social systems or its parts. Contributions to origins of sociophysics are found throughout the history of recorded science. However, the particularly intense activity in the sociophysics is currently observed in literature.

But, application of theories obtained for physical systems, onto social systems, is by no means straightforward, and in fact starts with a fundamental problem: in methodology of physics assumes experimental verification of stated theoretical results, in such a way that whoever wants in principle can establish the appropriate experimental setting. Moreover, the experiments in question can in principle be repeated indefinite number of times, always with the same initial conditions, within the same precision. It is assumed that in society, one cannot prepare some social system or its part in a given initial state, and one cannot repeat that setting even once, not mentioning indefinite number of repetitions.

During time, presumably for experiments within social disciplines, researchers tended to establish repeatable, thus verifiable, settings and initial conditions of a system tested. There are several mutually exclusive underlying factors. In particular, the observers and experimental set-up should not interfere with observations. That requires careful planning of a single occurrence of the experimental setup, and severely influence the possibility of repetition of experiments with same participants.

One type of activity in which repeated behaviour is regularly observed is a game.

In this paper we analyse the possibility of using games as a source of realistic data about aspects of behaviour, to be utilized as input for diverse sociophysics' models.

In the second section we describe in more details basic sociophysics notions, emphasise social mass and use it further for development of the relation between the games and sociophysics. In the third section games are described and their characteristics important for sociophysical modelling discussed in more detail. Fourth section concludes the paper and describes several perspectives of future development of this topic.

2 BASIC NOTIONS FROM SOCIOPHYSICS

Sociophysics is a rather broad field already. Let us concentrate on the macroscopic, meta-theoretic approach to description of societal dynamics [2]. The basis of this approach is a similarity between dynamics of some social system and of some physical model. The physical models are usually conceptually simple, yet of rich dynamics. In that way the optimal combination of a small number of parameters used and a large number of described characteristics, is obtained. However, there is no general receipt of connecting the social and physical situations, hence the physical model exploited should be chosen in accordance with additional requirements [3, 4]. Furthermore, important and measurable quantities in physical models and social systems are not aligned.

A social system is a complex system, characterised with a large information content and a finite amount of resources. On the realistic time scales, its evolution includes the response to the external influences. Implicit characteristics of a social system are its relation with the organisation, structure, and selectivity [5]. A formalised description of social systems initially relied on the notion of the equilibrium, the inappropriateness of which partially motivated development of non-equilibrium-based models. Here we use the concepts valid in both types of models. The functions introduced there we consider as guidelines for describing a system evolution.

Previously, one of the authors formulated basics for setting the more efficient approach to social system description [4]. It is motivated with the fact of the similar bysis of the description of thermodynamic systems. It is possible using several variables, although they contain a large number of entities. This means that there is a great reduction in the number of variables needed for description of the system, compared with the number of variables needed for description of the same number of independent entities. The basis of the approach for societal phenomena description is that the quantities used should [1]:

- refer to the whole system,
- have recognised meaning,
- be independently measurable,
- be little in number.

The set of quantities used in description of social systems' state consisted of social entropy and social free energy. Details of these quantities are available elsewhere Along with these, other quantities were interpreted from the point of view of sociophysics.

It is not accidental that in sociophysics, in a significant part, approaches exploiting thermodynamics were being developed. The thermodynamics itself was being developed intensively in the epoch in which systems were manipulated in a given sequences, while their microscopic content and micro-dynamics was uncertain, or unknown.

Examples of sociophysical interpretations of other societal quantities, like social capital and social trust, are available elsewhere [6].

Fostering the analogy between thermodynamics and social phenomena further, one would like to argue about the possibility to introduce, in the context of societal phenomena, analogues of other, in particular microscopic physical quantities. In that sense, defining the social analogue of mass is one of the encountered problems.

Mass, a measure of inertia, relates response function (acceleration) of a body subjected to external influence (the force). In such a way, two bodies differing only in masses will have different responses subjected to external forces.

In defining social analogue of mass – shortly: social mass, problems encountered are how to obtain repeatability necessary for measurements, and how to extract situations in which a social system unit (e.g. individual human, organisation) is found with a probability different from zero in a same state during a significant amount of time. Formulating of such a system would bring forward the possibility to define a social mass, which would provide significant impact to further development of sociophysics, and to deepening of our understanding of societal phenomena at the quantitative level.

One such a type of societal and other activities are organizations and economic processes. Their sociophysical interpretation is given elsewhere [7, 8]. Another suitable type of societal activities are games.

3 GAMES

Games are structured activities, characterized by interaction, challenge and rules. They can be individual or team, of different duration, performed in diverse spaces with a broad range of additional equipment included. Their rules serve to channel participants' interaction in such ways so that challenge is achieved, or at least tried to. Underlying to all that, is the fact that in a large number of games, repeated situations constantly occurs. Moreover, games are considered as activities in which participants can be rather easily be brought into states of considerable duration in which they do not interact significantly with observers and experimentalists, i.e. in which they participate in games as if they were not observed in an experiment.

In that way, one can think of games with sufficient involvement of participants as of a regularly encountered, well-known, broadly adopted, realization of a particular type of societal phenomena. Rules are important elements of games. Each rule represents a particular restriction of all available behaviours of participants. That contributes to lowering the entropy of the participant, or participants. But, rules are subject to change, and we witness constant (naturally, not often in a particular game) change of rules in games available. Same games played in different locations, for different age groups and different levels of achievements can differ also in attributed set of rules.

The moment that rules can be modified if that is agreed upon by participants, furthermore contribute significantly to the suitability of analyses of game playing for defining and determining values of sociophysical quantities.

While analyses of team games provide us with an efficient step forward in development of sociophysics, that approach is far from being universal. Its scope is determined with the broadness of results reachable from that, large set, but still a small subset of available societal phenomena.

4 DISCUSSION AND CONCLUSIONS

Sociophysics is rather propulsive scientific discipline which repeats in crucial steps development of e.g. thermodynamics and which is considered to be further developed onto a rather advanced level.

An interesting, and potentially useful approach in development of sociophysics and our understanding of societal phenomena, is sociophysical analyses of team games. The fact that in these kind of games several or many participants are involved brings about the possibility to extract typical behaviours.

References

[1] J. Stepanic, G. Sabol, M.S. Žebec. Describing social systems using social free energy and social entropy, Kybernetes **34**(5-6), 857-868, 2005.

[2] J. Stepanic, H. Stefancic, M.S. Zebec, K. Peračković. Approach to a Quantitative Description of Social

Systems Based on Thermodynamic Formalism, Entropy **2**, 98-105, 2000, http://www.mdpi.org/journals/entropy,
[3] J. Kasac, H. Stefancic, H and J. Stepanic. Comparison of social and physical free energies on a toy model. Physical Review E **70**, 016117, 2004,
[4] J. Stepanic. Social Equivalent of Free Energy, INDECS – Interdisciplinary Description of Complex Systems **2**(1), 53-60, 2004, http://indecs.eu,
[5] N. Luhmann. Social systems (in Croatian), Globus, Zagreb, 1981,
[6] J. Stepanić, P. Rodik. Operationalising social free energy, CS – Autopoiesis, Information Society, 2004.
[7] J. Stepanic, I. Bertovic, J. Kasac. Mediated Character of Economic Interactions, Entropy **5**, 61-75, 2003, available from http://www.mdpi.org/journals/entropy,
[8] R. Fabac, J. Stepanić. Modeling Organization Design - Applying a Formalism From Theoretical Physics Model. Journal of Information and Organizational Sciences **32**(1), 25-32, 2008.

THE EXPERIENCE OF FLOW, PLAY AND LEARNING

Toma Strle
University of Ljubljana, Faculty of Arts
Aškerčeva 2, 1000 Ljubljana, Slovenija
Tel: +386 31 338 268
email: toma.strle@guest.arnes.si

ABSTRACT

I argue that conditions necessary for the state of flow to occur can be transferred and achieved in learning. If these conditions are met the state of flow can emerge while learning, making learning more enjoyable and phenomenologically more play-like. But learning is not all flow and play. It seems it also requires phases of reflection in which learning in flow is not likely (or impossible) to occur according to Dietrich's hypofrontality hypothesis. But if flow is achieved when learning, learning becomes more enjoyable, play-like and fun activity.

1 INTRODUCTION

In the present paper I will consider what is characteristic of flow, which conditions are necessary for the state of flow to occur and whether learning can actually occur in the state of flow. I will mainly focus on the possibility of transferring the conditions of flow to learning and the possibility that autotelic (something that is an end in itself) character of flow also emerges in the process of learning.

The experience of flow Csikszentmihayli (1997) is describing represents a state where the activity we are performing is intrinsically enjoyable, fun and similar to play. The conditions of flow (very clear steps of the way, immediate feedback to one's action, balance between challenge and skill) and other elements of flow (merging of action and awareness, exclusion of distractions from consciousness, no worry of failure, disappearance of self-consciousness, distortion of time and, the autotelicity of the activity) as we shall see, seem to offer a usable framework which we can try to use in learning environments. If successful a possibility to experience learning as play and hence more enjoyable and meaningful emerges.

But learning is not all flow. Reflection is also an intrinsic and important part of learning and Dietrich argues that flow cannot occur during reflection. He claims we have two distinct information processing systems in the brain: the explicit system tied to conscious awareness which is associated with higher cognitive functions and the implicit system which is associated with 'effortless' information processing characteristic of the flow state (Dietrich, 2004). If we are to experience flow, a temporary suppression of the analytical and meta-conscious capacities of the explicit system must take place.

In the end I will briefly mention the topic of spontaneous idea generation which is brought about by the state of flow and which according to Csikszentmihayli gives us the feeling of discovering something new producing the feeling of enjoyment. (Csikszentmihayli, 1997).

2 THE EXPERIENCE OF FLOW

In his book, *Creativity: Flow and the Psychology of Discovery and Invention*, Csikszentmihayli asks what it is that separates very creative (and usually successful) people from others. Interviewing many different people from rock climbers to scientists and businessmen, he discovered one element they all have in common: They all love what they do and are not doing it for fame or money but because it is fun and enjoyable (Csikszentmihayli, 1997). He found out that it is "not what these people do that counts but *how* they do it." (Csikszentmihayli, 1997, 107)

2.1 Elements and conditions of flow experience

When Csikszentmihayli interviewed people that loved what they were doing he found that "what kept them motivated was the quality of the experience they felt when they were involved with the activity." (Csikszentmihayli, 1997, 110) He also found that people were talking about activities that were not dull, easy or passive but rather difficult, very involving and had an element of novelty and discovery in them. He termed such enjoyable, optimal experiences as flow (called also peak experiences by Maslow or 'the zone').

He reports of nine main elements of these experiences (Csikszentmihayli, 1997, 111-113). Elements 1, 2, 3 and sometimes 4 can also be understood as conditions for flow:

1. *There are very clear steps of the way.* In flow we always know what needs to be done from moment to moment of our activity.
2. *There is immediate feedback to one's action.* We know exactly how well we are doing. In science for example this is harder to achieve than in climbing because one needs to

internalize rules and criteria of a certain field in order to give feedback to him/herself.

3. There is a balance between challenge and skill. The task or activity should not be too dull or too easy but is must also not be too challenging, exceeding our skills so that we would feel frustrated. Csikszentmihayli nicely describes this balance as being on the fine line between boredom and anxiety.

4. Action and awareness are merged. Our minds are not wandering and we are completely focused on what we do. This is made possible by the clarity of goals and immediate feedback.[1]

5. Distractions are excluded from consciousness. We are aware only of what is relevant because of our intense focus on the present activity. Being aware only of what is relevant relieves us from the usual depression and anxiety we experience sometimes in everyday life.

6. There is no worry of failure. Some describe it as the feeling of total control. It occurs simply because we are not concerned with failure – our involvement with the activity is too great.

*7. Self-*consciousness *disappears.* Csikszentmihayli claims it is possible we have an experience of stepping out of the boundaries of the ego, and become a part of a larger entity when experiencing flow.

8. The sense of time becomes distorted. In flow time can shrink so that an hour feels like a minutes or time can stretch so that a couple of milliseconds seem like seconds. This depends on what we are doing.

9. The activity becomes autotelic. In flow we are doing the activity because of the activity itself (enjoyment of doing something) and not because of some external goals such as getting rich, getting a better grade or being successful.

Elements 4 to 9 are also very prominent in experiencing play as we will see in the next chapter so two main questions arise: Can we learn in flow and secondly, which conditions of flow can be met in learning so that learning would be more like flow and consequently more like play?

3 AUTOTELIC CHARACTER OF PLAY

M.A. Rod defines play as "…an enjoyable, spontaneous activity that is carried out for its own sake with no obvious immediate biological purpose. … a state of mind rather than a characteristic of certain types of activities. Thus, one can engage in almost any activity in a playful way, as long as one has a non-serious, activity-oriented (rather than goal-oriented) mental set." (Rod, M.A., 2006) If we take the characterization of play from M.A. Rod, we see it greatly resembles the autotelic character of flow.

Psychologists Rubin, Fein and Vandenberg in 1983 tried to summarize the state of the research in the field of child[2] play. As Sutton-Smith tells us in his book they "chose to describe it in a way that corresponds less with typical research concerns in their field, about play as a kind of development or progress, and more in terms of player's disposition …" (Rubin, Fein, Vandenberg, 1983; from Sutton-Smith, 2001, 188) When trying to find out why the players like to play, they found certain characteristics of play that makes it enjoyable and some of these characteristics resonate very well with characteristics of the state of flow, namely the autotelic character or intrinsic enjoyment, intense concentration on the activity and the lack of external goals: "the hallmark of play is that it is intrinsically motivated (that it is fun)", "Play is characterized by attention to means rather than ends", "Play is guided by organism-dominated questions rather than context-dominated questions.", "Play behaviors are not instrumental.", "Freedom from externally imposed rules is necessary." and "players are actively engaged in their activity." (Rubin, Fein, Vandenberg, 1983; from Sutton-Smith, 2001, 188-189)

I see the state of flow – if we are able to achieve it in the process of learning – as a way of being able to experience learning as a more fun and play-like activity.

4 CONDITIONS OF FLOW TRANSFERRED TO LEARNING

If learning includes clear goals, immediate feedback and our skills match the challenge, we are on a good way learning becomes more autotelic and intrinsically enjoyable, more play-like.

4.1 There must be clear steps in the learning process

In flow we always know what needs to be done from one moment to another of our activity. It is especially important to set clear immediate goals of what we want to learn next and of how we are going to do it.

4.2 Immediate feedback in learning

Csikszentmihayli reports that "Students get flow from group work, from individual tasks, and from quizzes much more often than they do from listening to the teacher or from watching audiovisuals." (Csikszentmihayli, 2002, 14)

[1] This element is also dependent on our ability to focus (Dietrich, 2004) and the reason why it can be considered as a condition for flow (see Ch. 4.4).

[2] I should note here that these characteristics are probably also valid for adult's play but they occur during different activities.

Learning passively (just listening to lectures) makes the activity less likely to produce feedback than doing it with others or having a specific task in which we are engaged.

4.3 Matching skill and challenge in learning

One important condition for experiencing the state of flow is that our skills match the challenge. One could argue it is problematic (or impossible) to achieve that when we are learning a completely new thing. At first it seems that acquiring new skills or knowledge would prevent us to experience flow. But on the other hand most of the learning process we do not encountered only new things but already possess some skill or knowledge we are trying to improve. Gross Richard's psychological definition of learning nicely points to this: "Learning, therefore, normally implies a fairly permanent change in a person's behavioural performance. … So, if a change in behaviour is to be counted as learning, the change must be linked to some kind of past experience (regardless of whether there was any attempt to bring about that change)." (Gross, 2005, 171) This implies that a very important part of learning is getting better at something based on previous experience. If we presuppose a certain degree of skill or knowledge of the learner and accommodate the challenge so that it is not exceeding skill (but is still challenging) it should be possible to achieve flow in learning.[3] Even if we don't achieve flow in learning we can avoid anxiety (if the challenge is too great) and boredom (if the challenge is very low) (see figure 1).

Luc Steels (talking about computer games in learning environments) for example claims that "it is equally important that the individual is in control of the challenge level, so as to increase challenge when the skill has become higher, otherwise activities become boring and self motivation is lost." (Steels, 2004, 5) A good way to achieve this condition is to implement computer games in learning environments. Furthermore, Paras and Bizzocchi argue: "The goal of successful game design is the creation of meaningful 'play' which is achieved by creating game-play that enables discernable and integrated interaction by the player. … Through their use of immersive experiences, games provide opportunity for play which can result in flow experiences." (Paras, Bizzocchi, 2005, 3) We will briefly touch the on this topic when talking next about merging action and awareness.

[3] It is also very important we are able to choose our own goals in order to achieve the best balance between skill and challenge which is many times not the case in our learning and working environments. (Csikszentmihayli, 2002)

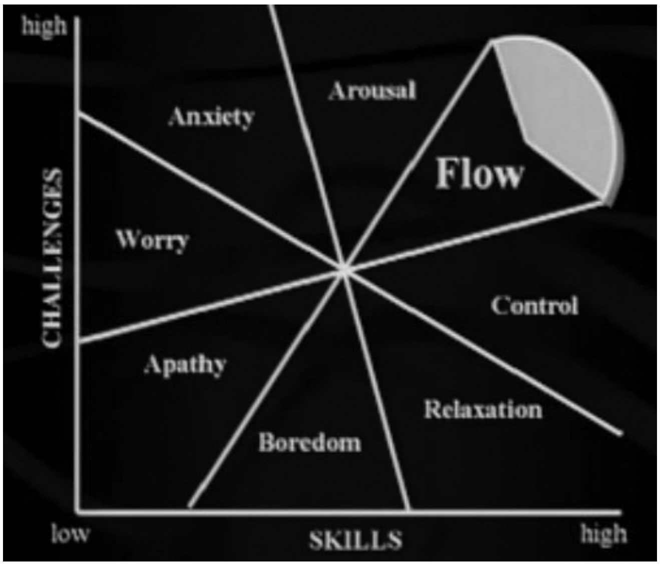


Figure 1: *Balancing challenge and skill*

4.4 Merging action and awareness

According to Dietrich action and awareness merge if enough volitional control over the executive attentional system is achieved. As a consequence the focus of attention narrows "to exclusively buffer the task at hand, eliminating other phenomenological features computed by the explicit system to enter consciousness." (Dietrich, 2004, 758) It would be interesting to investigate which mental techniques used for training attention – such as meditation – would enable better control over the executive attentional system prolonging and/or making the experience of flow to occur more often in learning.

5 ENGAGEMENT Vs. REFLECTION IN LEARNING

Paras and Bizzocchi nicely point to the fact, that learning is not 'all flow' and engagement but requires reflection which seems to be incongruent with the state of flow. "Research also shows that reflection is an important part of the learning process and while in the state of flow, players rarely reflect on the learning that is taking place." (Paras, Bizzocchi; 2005, 1) "Rather than a linear process, learning follows a cyclic pattern: experiencing, reflecting on that experience, drawing conclusions based on these reflections, forming a plan for new action based on these conclusions, then acting again, and so on." (Paras, Bizzocchi, 2005, 6) They actually propose an "endogenous implementation of reflection" in the learning game so that the flow experience would not be broken at all – not even while reflecting. I would argue this is somehow unrealistic since sometimes we have to 'step out' and see how we are progressing, which errors we are making, what new goals should we set, etc. And this can probably only be done using our meta-conscious capacities (eg. reflection) which we cannot use in flow as Dietrich (2004) claims with his hypofrontality hypothesis.

5.1 The hypofrontality hypothesis

It seems that meta-conscious capacities described by Dietrich (Dietrich, 2004) require the activation of brain circuitry that is incompatible with activation of circuitry that produces the state of flow.

Dietrich argues we have two distinct information processing systems in the brain that acquire, memorize and represent knowledge (Dietrich, 2004). The explicit system tied to conscious awareness is associated with higher cognitive functions of the frontal lobes (dorsolateral prefrontal cortex) and the implicit system which is associated with 'effortless' information processing (characteristic of the flow state) of the medial temporal lobe structures (like basal ganglia). If we are to experience flow, a temporary suppression of the analytical and meta-conscious capacities of the explicit system must take place which renders us incapable of reflecting upon the learning process, setting further goals, etc. Suppressing the meta-conscious, reflective cognitive functions brings us to another characteristic of the state of flow: spontaneous idea generation.

5.2 The state of flow and spontaneous idea generation

The state of flow enables spontaneity because when in flow we forget outer goals, our problems, we lose ourselves in the activity, our actions become automatic and not limited by our own conscious rational frames and thoughts, ideas and reflections. I the state of flow, information processing becomes more 'effortless' – and thus more spontaneous – not limited by our conscious cognitive capacities (Dietrich, 2004). When we 'turn off' our 'criticizing conscious reflection', novel and original ideas are more likely to emerge. Producing novel and original ideas is also very important in learning because it enables us making interesting connection between things we learned (concepts, pictures, words) and it gives us the feeling of discovering something new which produces the feeling of enjoyment as Csikszentmihayli (1997) claims.

6 CONCLUSION

If we manage to implement the conditions of flow in the learning process (clearly defined steps and goals of learning, immediate feedback on how well we are doing and balancing between challenge and skill) we can experience flow during learning. Learning could not, as we saw, be all flow. It also requires our meta-cognitive abilities and reflection brought about by our prefrontal cortical structures. But when our implicit system of effortless information processing is most active we can experience learning as a more enjoyable, play-like and fun activity.

Ideas and suggestions about the possibility of learning being more like play given in the present paper are mostly hypothetical. It would be of great interest and very necessary to dwell in the topics in more detail in order to be to answer some further important question. Are the conditions to achieve flow enough to achieve flow in learning? To which degree can flow actually occur in learning? How big of a role does the ability to focus play in achieving flow in learning? How important is reflection in the process of learning? If flow and autotelic character of learning can be achieved what implications does this have on our learning and working environments – after all we all wish to experience our activities more as play and fun.

References

[1] Csikszentmihayli, M. (2002). Do Students Care About Learning? A Conversation with Mihaly Csikszentmihayli (Interview with Marge Scherer). *Educational Leadership,* Vol. 60, No. 1, 12-17.

[2] Csikszentmihayli, M. (1997). *Creativity: Flow and the psychology of Discovery and Invention*. Harper Perennial, New York, Ch. 4, 5, 77-126.

[3] Dietrich, A. (2004). Neurocognitive mechanisms underlying the experience of flow. *Consciousness and Cognition 13*, 746–761.

[4] Gross, R. (2005). *Psychology, The Science of Mind and Behaviour.* Hodder Arnold, London.

[5] Paras, B., Bizzocchi, J. (2005). *Game, Motivation, and Effective Learning: An Integrated Model for Educational Game Design.* Proceedings of DiGRA 2005 Conference: Changing Views – Worlds in Play.

[6] Rod, A. M. (2006). *The Psychology of Humor: An Integrative Approach.* Elsevier Academic Press, Burlington, MA

[7] Steels, L. (2004). The Architecture of Flow, from Tokoro, L. (ed.) & Steels., L. (ed.). *A. Learning Zone of One's Own: Sharing Representations and Flow in Collaborative Learning Environments.* IOS Press, Amsterdam, Ch. 7, 135-153.

[8] Sutton-Smith, B. (2001). *The Ambiguity of Play*. Harvard University Press.

KAKO IGRAJO MOŽGANI - SONIFIKACIJA BIOLOŠKIH SIGNALOV

Viktor Avbelj
Odsek za komunikacijske sisteme
Institut Jožef Stefan
Jamova 39, 1000 Ljubljana, Slovenija
Tel: +386 1 4773824; fax: +386 1 4773111
e-pošta: viktor.avbelj@ijs.si

ABSTRAKT

Človek učinkovito analizira zvok, saj ga uporablja za sporazumevanje, ko posluša glasbo in drugje. Nekateri signali, ki so prisotni v živih organizmih, nimajo narave zvoka, a jih lahko v zvok spremenimo. Takšni so signali, ki jih oddajajo možgani in jih merimo s pomočjo EEGja in tudi signali, ki jih oddaja srce in jih vidimo v EKGju. Spremenjeni v zvok so ti signali dostopni človeškemu sluhu in sposobnostim, ki jih imamo. V prispevku predstavimo nekaj primerov, kjer nam poslušanje odpira drugačen vpogled v svet bioloških signalov.

1 UVOD

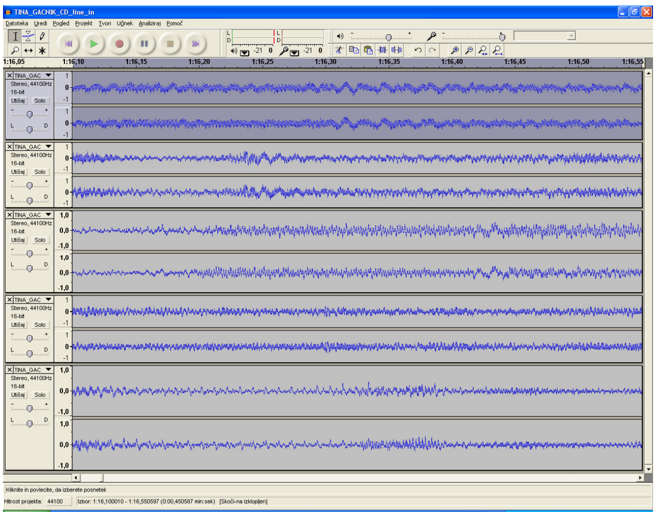
Kadar signal, ki po svoji naravi ni zvok, spremenimo v zvok, govorimo o sonifikaciji. Sama beseda elektroencefalografija (EEG) pomeni, da elektične signale možganov narišemo in s tem prikazujemo časovni potek izmerjenih električnih potencialov. Podobno velja za elekrokardiografijo (EKG), kjer je v ospredju srce. Grafe običajno najprej vizualno pregledamo in potem določamo časovne intervale, višine valov, frekvence, spektre, povezanost med več sočasno posnetimi signali in drugo. EEG signale je težko "razumeti", saj gre za večkanalne posnetke z veliko šuma. Šum tukaj imenujemo vse, kar nas trenutno ne zanima, čeprav je ta šum lahko le rezultat množice opravil, s katerimi se možgani vzporedno ukvarjajo. Človek je s svojima ušesoma in poslušanjem sposoben analizirati prav takšne težavne signale. Le frekvence, ki nastopajo v EEGju in EKGju, niso najbolj primerne za neposredno poslušanje. Ko posnete EEG ali EKG signale ali le posamezne značilnosti teh signalov sonificiramo, se nam odprejo nove možnosti, ki nudijo drugačen vpogled, kot je vizualno pregledovanje grafov.

V nadaljevanju predstavimo nekaj primerov. S prvim prikažemo sposobnost človeških možganov, ki iz zvočnega posnetka brez težav izluščijo signal, ki je globoko skrit v šumu. Drugi primer obravnava sonifikacijo utripov srca noseče matere in otroka v njej. Tretji primer prikaže sonifikacijo EEG.

2 PRIMERI

2.1 Signal globoko v šumu

Za poskus, s katerim ponazorimo veliko sposobnost možganov za sledenje zvokom, skritih v šumu, smo uporabili glasbeni CD, na katerem je več pesmi, ki jih ob spremljavi glasbil poje ista pevka. Nastane vprašanje, ali lahko človek sledi eni pesmi, če 5 pesmi predvajamo skupaj (predvajanje na predstavitvi)? Tehnično lahko rečemo, da je signal globoko skrit v šumu (slika 1). Prikazano sposobnost človeka lahko uporabimo pri analizi EEG, če le zmoremo signale uspešno sonificirati tako, da lastnosti EEG prenesemo v slišno področje človeka.



Slika 1: *Skladba, ki ji želimo slediti (zgornje temnejše področje) v šumu, ki ga sestavljajo 4 sočasno predvajane skladbe (spodnje svetlejše področje). Prikazan je majhen del posnetka (0,45 s), ki obsega le en zlog v prvi skladbi.*

2.2 Utrip src matere in otroka v njej

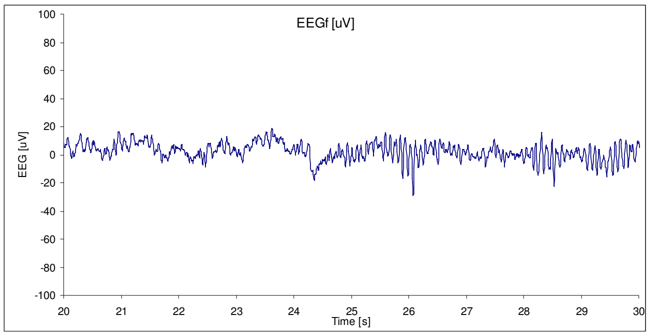
Drobna naloga, s katero se ukvarjajo možgani, je uravnavanje bitja srca. Srce sicer že vsebuje ritmovnik, ki daje osnovno frekvenco utripanja. Poleg drugih počasnejših mehanizmov, ki vplivajo na utrip, imajo možgani neposreden dostop do srca preko direktnih nevronskih povezav. Najhitreje lahko vpliva deseti možganski živec - imenovan klatež (vagus), ki omogoča uravnavanje hitrosti bitja srca od utripa do utripa, kar je posebej opazno pri mladih ljudeh. S tem, ko smo se omejili na ritem srca, smo se sicer oddaljili od možganov in EEG, vendar nam je ostala sicer drobna a pomembna stvar, s katero se ukvarja avtonomni živčni sistem - pa ne samo ta, saj se strinjate ali ste tudi že občutili, kar govorijo besede "ko ga je zagledala, ji je srce poskočilo". Nedavno so ugotovili, da sta utripa src matere in otroka povezana [1, 2]. To so ugotovili s klasično analizo signalov ob spreminjanju dihanja, ki ga je izvajala mati. Je mogoče tako povezavo src slišati na sonificiranem posnetku?

EKG posnetek noseče matere, kjer sta bila skupaj posneta elektrokardiograma matere in otroka, je bil sonificiran na sledeči način [3]. V izbrani lestvici sta igrala dva inštrumenta (čelo=mati, flavta=otrok). Višino tona je določala dolžina intervala med utripoma srca tako, da se je višina povečala, kadar je bil interval krajši in nasprotno. Če je interval ostal enak, se višina tona ni spremenila, trajanje tona pa se je podaljšalo do naslednje spremembe intervala. Vsak ton inštrumenta se začne v času utripa srca matere ali otroka. Tako je iz EKG posnetka nastala skladba, dostopna človeškemu dojemanju glasbe (slika 2). Lahko ob poslušanju te glasbe zaznamo preplet utripov obeh src (predvajanje na predstavitvi)? Komponisti te skladbe so možgani matere in otroka in še vse drugo, kar je vplivalo na utripa src.

Slika 2: *Skladba, ki je nastala ob sonifikaciji utripov srca matere in otroka (10 taktov).*

2.3 Sonifikacija EEG

Ker so frekvence EEG, ki jih posnamemo na glavi osebe v področju, ki je večinoma izven obsega frekvenc, ki jih zaznava sluh, je potrebno surov EEG spremeniti [4]. Že predvajanje posnetka z večjo hitrostjo (npr. 50 krat), prestavi frekvence v slišno področje. Uporabimo lahko frekvenčno modulacijo, pri kateri spreminjamo frekvenco oscilatorja z amplitudo posnetega signala. Tako spremenjeni signal lahko predvajamo v realnem času (slika 3). S spektralno analizo zaporednih kratkih delov posnetka lahko iščemo posamezne kratke pojave oscilacij pri določenih frekvencah (valovi alfa, beta, gama, delta in theta) in jih preslikamo v tone glasbenih inštrumenov. Obstaja še veliko drugih možnosti, v katerih lahko uporabimo tudi sposobnost prostorske zaznave zvoka pri ljudeh.



Slika 3: *Prikazani EEG signal je razlika potencialov med elektrodama Pz in Cz (kratek del posnetka). Sonifikacijo smo izvedli s hitrejšim predvajanjem.*

3 ZAKLJUČEK

Stetoskop uporabljajo v medicini že vse od leta 1816, ko ga je prvič uporabil zdravnik Laennec v Parizu. Še danes je to inštrument, brez katerega si zdravnika ne znamo predstavljati. Ob velikem razvoju drugačnih postopkov preiskav, je zvok obtičal v ozadju. Pri sposobnosti človeka, da lahko učinkovito analizira zvok, je to zapostavljanje neupravičeno. Kako preslikati biološke signale v zvok, da bodo dostopni človeškim sposobnostim pa ostaja vprašanje.

Reference

[1] Van Leeuwen P, Geue D, Thiel M, Cysarz D, Lange S, Romano MC, Wessel N, Kurths J, Grönemeyer DH. Influence of paced maternal breathing on fetal-maternal heart rate coordination. Proc Natl Acad Sci U S A. 2009 Aug 18;106(33):13661-6.

[2] Ivanov PCh, Ma QD, Bartsch RP. Maternal-fetal heartbeat phase synchronization. Proc Natl Acad Sci U S A. 2009 Aug 18;106(33):13641-2.

[3] Falkner M. Fetal Heart Music Project of the Paracelsus Clinic. http://www.michaelfalkner.de/herzmusik/heartmusicproject.html.

[4] Hermann T, Meinicke P, Bekel H, Ritter H, Mueller HM, Weiss S. Sonifications for EEG data analysis. Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, July 2–5, 2002.

IGRE IN IGRANJE

Olga Markič
Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana, Slovenija
e-mail: olga.markic@guest.arnes.si

POVZETEK

V prispevku je predstavljena opredelitev pojmov igre in igranja ter njuna medsebojna relacija. Čeprav je Wittgenstein skušal pokazati, da klasična definicija pojma igre ni mogoča, je Suits predlagal abstraktnejšo opredelitev igre (game) in pokazal, v kakšnem odnosu je do igranja (play).

1 UVOD

Običajno se ob omembi besed »igra« in »igranje« ne sprašujemo po njihovem pomenu, ampak si zamislimo kaj konkretnega: na primer najljubšo računalniško igro, igro med dvema ognjema, igro s punčkami, šah, košarko, lutkovno igro, škropljenje z vodo, ravbarje in žandarje, igranje kitare ali pač karkoli nam je že blizu. Kadar igramo igro ali se igramo, se ponavadi ne sprašujemo, kaj točno počnemo, ampak smo potopljeni v samo dejavnost. Včasih pa se, če je v nas vsaj malo filozofa, skušamo povzpeti eno stopničko više in ugotoviti, ali je vsem tem igram kaj skupnega. Ali lahko najdemo lastnosti, ki so skupne vsem igram in jih razločujejo od ne-iger. Kaj je tisto, kar določa igranje? Ali takrat, kadar se igramo, vedno tudi igramo igro? Vprašanja se na prvi pogled morda zdijo lahka, a ko začnemo iskati odgovore hitro spoznamo, da zadeva ni tako preprosta.

V pričujočem prispevku bom predstavila predlog Bernarda Suitsa (2005), kako opredeliti igro in igranje ter njuno relacijo. Verjetno bi tudi njegovo pojmovno analizo lahko uvrstili v eno od retorik. Suitsova analiza je zbudila veliko pozornosti predvsem v razpravah v filozofiji športa, Hurka (2005) pa je kot pisec uvodne študije v drugo izdajo izpostavil predvsem etiško komponento. Čeprav se morda zdi pojmovna analiza preveč oddaljena od empiričnih raziskovanj v kognitivni znanosti, se mi zdi potreben prvi korak do celovitejšega razumevanja področja, ki se ga lotevajo raziskovalci različnih disciplin.[1]

[1] Prispevek predstavlja zgoščeno predstavitev analize pojmov igra in igranje predstavljenih v Markič (2009).

2 IGRE

Prav pojem »igre« je v filozofiji postal paradigmatski primer težav, ki jih imamo, če hočemo poiskati definicijo, pri kateri navedemo nujne in zadostne pogoje za uvrstitev pod določeni pojem. Wittgenstein je v *Filozofskih raziskavah* na primeru »igre« dokazoval, da je iskanje klasične, Aristotelovske definicije brezupno početje. Takole pravi v znamenitem 66. paragrafu Filozofskih raziskav (Wittgenstein 1953): »Obravnavaj npr. enkrat procese, ki jih imenujemo »igre«. Menim igre na igralni deski, igre s kartami, igre z žogo, borilne igre itd. Kaj je vsem tem skupno? – Ne reci »Morajo imeti nekaj skupnega, sicer jih ne bi imenovali 'igre'« – temveč glej, ali je vsem njim kaj skupno. – Kajti če jih pogledaš, sicer ne boš videl ničesar, kar bi bilo skupno vsem, videl pa boš podobnosti, sorodnosti, in sicer cel niz. Kot rečeno: ne misli, temveč glej! – Poglej npr. igre na igralni deski z njihovimi raznolikimi sorodnostmi. Sedaj preidi na igre s kartami: tukaj najdeš marsikatere analogije z ono prvo skupino, mnogo skupnih potez pa izgine, pojavijo se druge. Če sedaj preidemo na igre z žogo, se ohrani mnogo skupnega, veliko pa se izgubi. – So vse zabavne? Primerjaj šah s križci in krožci. Ali vsepovsod obstaja zmagovanje in izgubljanje, ali konkurenca med igralci? Pomisli na pasjanso. V igrah z žogo obstaja zmagovanje in izgubljanje; ko pa kakšen otrok vrže žogo v zid in jo spet ujame, se ta poteza izgubi. Glej, katero vlogo igrata spretnost in sreča. In kako različni sta spretnost pri šahu in spretnost pri tenisu. Pomisli sedaj na igre v krogu: Tukaj je element zabave, koliko drugih značajskih potez pa izgine! In tako lahko gremo preko mnogih, mnogih drugih skupin iger in vidimo, kako se podobnosti pojavljajo in izginjajo. Izid tega razmišljanja pa se sedaj glasi: Vidimo neko zapleteno mrežo podobnosti, ki se prepletajo in križajo. Podobnosti v velikem in malem.«[2]

Po analogiji s podobnostmi med člani družine, kot so postava, poteze obraza, barva oči, hoja, temperament itd., Wittgenstein te skupne lastnosti imenuje »družinske podobnosti«. Če so družinske podobnosti največ, kar lahko odkrijemo pri igrah, potem je nadaljnje iskanje opredelitve pojma igre brezupno početje. A vsi ne delijo tega mnenja.

[2] Z dovoljenjem povzeto po še neobjavljenem prevodu B. Cerkovnika.

Kanadski filozof Bernard Suits je v knjigi *The Grasshopper: Games, Life and Utopia*, ki je izšla že leta 1978[3], pokazal, da se je Wittgenstein prehitro vdal in ni sledil svojim lastnim nasvetom, da mora vsak najprej pogledati in se prepričati, da skupnih lastnosti res ni mogoče najti. Suits se je potrudil pogledati globlje pod površinske podobnosti in poiskati tisto, kar je skupno igram na bolj abstraktni ravni. Svoje razumevanje odnosa med igro (angl. game), igranjem (angl. play) in športom je izčrpno predstavil tudi v seriji člankov (1977, 1988, 2004).

Suits svojo zgodbo postavi v Utopijo, kjer so na voljo vse instrumentalne dobrine. V njej se postavi v bran kobilico iz Ezopove basni. Kobilica se vse poletje igra, rezultat pa je, da ostane brez hrane za zimo. A kot pove svojemu učencu Skeptiku, se je odločila za igranje, čeprav ve, da bo to imelo zanjo pogubne posledice, Tako odločitev je sprejela zato, ker ve, da se tako posveča dejavnosti, ki ima najvišjo intrinzično vrednost. Kobilica v dialogih s svojimi učenci, predvsem Skeptikom, zagovarja igranje iger kot najvišje dobro in ob tem predlaga definicijo igre, ki jo zagovarja pred številnimi protiprimeri.

Suits je predlagal naslednjo definicijo za igro (game): »Igrati igro je poskus doseči določeno stanje stvari [predigralni[4] cilj] in pri tem uporabljati zgolj sredstva, ki jih dovoljujejo pravila [igralna[5] sredstva], kjer pravila prepovedujejo uporabo bolj učinkovitih v prid manj učinkovitih sredstev [konstitutivna pravila], in kjer so pravila sprejeta prav zato, da omogočijo tako aktivnost [igralna naravnanost].«, ali še bolj preprosto »igrati igro je prostovoljni poskus premagati nepotrebne ovire«. (Suits 2005, 54 - 55)

Njegova analiza pojma igre ima štiri glavne elemente, ki so nujni in skupaj zadostni pogoji, ki jih zahtevamo, da neko dejavnost smatramo za igro. To so (1) *predigralni cilj*, (2) *igralna sredstva*, (3) *konstitutivna pravila* in (4) *igralna naravnanost*. Pri prvem elementu, predigralnem cilju, gre za to, da igralec vedno stremi k cilju, ki ga lahko opredelimo neodvisno od same igre. Na primer, pri golfu je ta cilj spraviti žogico v luknjo v zemlji. Pri tem se sme uporabljati samo sredstva, dovoljena za to igro (igralna sredstva), v našem primeru palico za golf. V igri se mora držati konstitutivnih pravil igre. Pri golfu tako ne smeš prijeti žogice, jo odnesti do luknje, in jo z roke spustiti vanjo, temveč jo moraš udariti s palico od tam, kjer je ležala, itd. Ob upoštevanju teh pravil pa uspeh v igri tipično zahteva kar se da učinkovito doseganje predigralnega cilja. V našem primeru pomeni to spraviti žogico v luknjo v čim manj primerih. Vendar ti trije elementi še niso dovolj za igro. Za to, da nekaj šteje kot igra, je po Suitsovem mnenju pomemben še četrti element, to je prostovoljno sprejemanje konstitutivnih pravil. Igralec sprejme pravila prav zato, ker le ta omogočijo igro in s tem izrazi svojo igralno naravnanost. Tako igralec golfa sprejme omejitev, da žogice ne sme prenašati z roko ali kako drugače izboljševati njene lege, ker pač želi igrati golf. Spoštovanje pravil je tako prostovoljna zaveza igralca.

Suits poudarja, da lahko med igre uvrstimo tudi dejavnosti, ki jih sicer z besedami ne imenujemo igre, ki pa vseeno sodijo mednje. Na primer, tek na 400 metrov na atletskem tekmovanju in plezanje na goro. V obeh primerih lahko najdemo vse štiri elemente. Pri teku je cilj preteči ciljno črto, to pa mora atlet storiti z lastnimi močmi in po za to določeni progi. Odreči se mora izbiri bližnjic in prostovoljno sprejeti pravila. Pri alpinistih je predigralni cilj, stati na vrhu gore, potrebno doseči v okviru konstitutivnih pravil alpinizma – izstop iz helikopterja ali vožnja z žičnico prav gotovo ne šteje kot vzpon v skladu s pravili o uporabi dovoljenih sredstev. Po drugi strani pa Suitsov predlog nekatere dejavnosti, ki jim v vsakdanjem govoru sicer rečemo igra, ne uvršča med igre. Na primer »Ringa, ringa raja, muca pa nagaja«, ki bi jo mnogi imenovali igra, po mnenju Suitsa sodi bolj med koreografirane pesmice oziroma gledališke predstave.

Suits (2005) se je v svoji analizi omejil na igro kot »game« in tudi v angleškem prevodu *Filozofskih raziskovanj* je v 66. paragrafu uporabljena beseda »game«. Vendar na ta način, strogo gledano, z definicijo Kobilice ni odgovoril Wittgensteinu v celoti. Tako slovenska beseda »igra« kot nemška »das Spiel« namreč pokrivata oba angleška izraza »game« in »play«.

3 IGRANJE

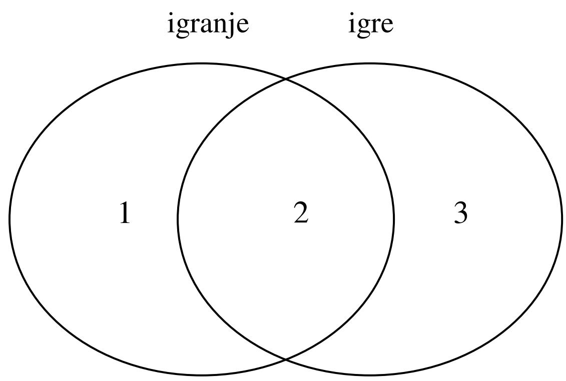
Tudi če sprejmemo predlagano analizo igre (game), se nam pomen igre kot igranja (play) še vseeno izmika. Ali bi lahko rekli, da so vse igre tudi igranje, da so prve podmnožica drugih? Suits (1977, 120) meni, da ne. Pokaže, da med igro in igranjem ni logične relacije, čeprav se morda zdi to na prvi pogled presenetljivo. Res je, da »igrati igro« pogosto lahko razumemo kot igranje, vendar iz tega ne moremo sklepati, da ker je x primerek »igrati igro«, je x torej tudi primerek igranja. Prav tako ne velja niti obratno. Ko se otroci pri mizi igrajo in z žlico brodijo po pire krompirju, ne igrajo igre. Prav tako ne igramo igre, ko se igramo z lasmi in jih zavijamo s prsti. Po drugi strani pa, kot poudarja Suits, za profesionalne športnike, ki nastopajo v igri za plačilo, ne moremo brez kakršnih koli omejitev reči, da se igrajo. Zanje to pomeni delo. Igrati igro tako ni vedno igranje. Morda nas zavajajo besede, zato Suits predlaga, da se pri iskanju opredelitve igranja osredotočimo na tiste uporabe besede igrati, ki jih ne moremo nadomestiti z drugimi. Po drugi strani pa bi v primeru sklopa »igrati igro« besedo »igrati« brez težav lahko nadomestili s »sodelovati« v igri.

[3] Tu se nanašam na drugo, dopolnjeno izdajo (2005).
[4] Prelusory - izpeljanka iz latinske besed ludus, ki pomeni igro.
[5] Lusory

Suits se je pri iskanju splošne definicije oprl na Schopenhauerjevo idejo, da gre pri igranju za presežek energije, ki je sicer potrebna za sledenje instrumentalnim ciljem. Želel je najti dovolj splošno definicijo, ki bo veljala tako za ljudi kot za živali. Pojem energija je zamenjal s pojmom viri in predlagal naslednjo definicijo: »*X* se igra če in samo če je *x* izvedel začasno premestitev virov, ki so prvenstveno dodeljeni instrumentalnim smotrom, k avtoteličnim[6] dejavnostim» (Suits 1977, 124) Če uporabimo predhodni primer: hrana je v prvi vrsti namenjena prehranjevanju, to je njen glavni smoter. A lahko jo uporabimo tudi za druge cilje, na primer za gradnjo cest in hribčkov, kot to otroci včasih počnejo s pire krompirjem. Suits kot posebej zanimiv vir izpostavi čas, saj kadarkoli želimo nekaj storiti, potrebujemo za izvedbo dovolj časa. Prva prioriteta glede časa so bile vedno instrumentalne dejavnosti, povezane z vzdrževanjem in izboljševanjem pogojev eksistence. Toda kadar imamo več kot dovolj časa za doseganje teh ciljev, potem čas lahko začasno trošimo za intrinzično cenjene cilje. Na primer, po službi in gospodinjskih opravkih gremo na košarko, beremo roman ali pa si kaj zaigramo na kitaro.. V obeh primerih, tako pri otroku, ki je ustvarjal s pire krompirjem, kot z odraslim, ki je igral košarko, bral roman ali igral na inštrument, je šlo za igranje. Šlo je za dejavnost z intrinzičnim smotrom in za premestitev virov, ki so bili prvotno dodeljeni instrumentalnim smotrom (pire krompir hranjenju, čas pa izboljševanju pogojev preživetja).

Po Suitsovem mnenju je smiselno razlikovati med dvema vrstama igranja. Prvo je poimenoval »primitivno igranje«, ki je samo igranje, drugo pa »sofisticirano igranje«, ki je še nekaj več. (Suits 1988, 1) Pri primitivnem igranju veščina ni bistvena sestavina. Kot primer navede otroško igranje z vodo pri kopanju. (Suits 1988, 2). Pri tem početju ne gre za delo, v njem ni nič instrumentalnega, a hkrati to početje tudi ni povsem brez smotra in naključno, saj otrok izvaja gibe tako, da uspe razpljuskati vodo. Ob pljuskih vode uživa, zato še naprej nadaljuje z gibi, ki privedejo do novih pljuskov. Bolj ko se ukvarja s to dejavnostjo, boljše veščine si bo pridobil. Suits poudarja, da se veščine razvijejo ob tem, ko otrok ponavlja gibe, ki ga vodijo do pljuskanja in niso same po sebi tisto, kar otrok išče. Pri primitivnem igranju tako pomembno vlogo igra uvajanje novih izkušenj, ki večinoma izhajajo iz radovednosti in raziskovanja novega, manj pa je pomembne vadba in uživanje ob obvladovanju veščin. Ponavljanje pa seveda lahko pripelje do razvijanja novih veščin, ki kasneje lahko postanejo vrednota same po sebi. Kadar se to zgodi, potem se premikamo od primitivnega k sofisticiranemu igranju, torej do iger in morda še k čemu. (Suits 1988, 2). Med sofisticirano igranje bi tako lahko vključili različne športne dejavnosti, npr. plavanje, igranje z žogo (npr. amaterska košarka), pa tudi gimnastiko, ples, igranje glasbenega inštrumenta itd. Tem oblikam igranja Suits ne pravi igra (game) ampak zanje uporablja izraz »performance«.

Suits je pokazal, da odnos med igrami in igranjem lahko ponazorimo s pomočjo Vennovih diagramov.



Pri tem je opozoril, da je treba biti pozoren na razliko med Eulerjevimi in Vennovimi (Suits, 2004). Gornji diagram moramo brati kot Vennov diagram, v katerem posamezni deli diagrama, označeni s številko, predstavljajo eksistenčne trditve o posameznih primerkih. (Suits 1988, 7-8)

(1) Obstaja *x*, tako da je *x* igranje, a ni igra. (primitivno igranje ali amaterske športne predstave, npr. gimnastika)

(2) Obstaja *x*, tako da je *x* igranje in igra, (sofisticirana igra ali amaterske športne igre, npr. amaterska košarka)

(3) Obstaja *x*, tako da je *x* igra, a ni igranje (profesionalne igre, npr. bridge, ali profesionalne športne igre, npr. profesionalna košarka)

4 ZAKLJUČEK

Analiza pojmov igra in igranje je pokazala, da gre za različna pojma. Čeprav rečemo »igrati igro«, to ne pomeni, da gre nujno za igranje, lahko pa se igramo, a ne »igramo igre«. Igra je ciljno naravnana dejavnost, ki ima predpisana sredstva za dosego cilja in svoja konstitutivna pravila, ki jih igralci prostovoljno sprejemajo. Tudi igranje iger je lahko igranje, a v zadnjem času kognitivni znanstveniki ugotavljajo, da je za razvoj ne samo motoričnih in kognitivnih sposobnosti, temveč predvsem za razvijanje socialnih veščin izjemnega pomena igranje, ki je čim manj omejevano s pravili (Wenner, 2009). S tem so mišljene predvsem različne oblike igranja prevzemanja vlog kot tudi fizična merjenja moči z igrivo naravnanostjo. Podobno kot pri otrocih take oblike igranja opazimo tudi pri živalih.

[6] Izraz izhaja iz grške besede *autotelēs* in pomeni imeti smoter v sebi, ne izven sebe

Nevroznanstvniki tako ugotavljajo, da je potreba po igranju evolucijsko zelo stara. Jaak Panskepp (2008) je razvil model štirih temeljnih čustvenih sistemov pri sesalcih: STRAH, ISKANJE, PANIKA IN BES, temu pa je dodal še tri specifično namenske sisteme, ki so nastopili v različnih stopnjah razvoja sesalcev: POŽELENJE, SKRB IN IGRANJE.

Kobilica pa naj se, Skeptiku navkljub, še naprej igra.

Literatura

Hurka, Thomas (2005). *Introduction*, v Suits, Bernard: *The Grasshoper; Games, Life and Utopia*, Ontario; Broadview, str. 7-20.

Markič, Olga (2009). "Igre, vrednote in šport", v M. Velikonja, P. Stankovič in G. Starc (ur.), *Kalejdoskop športa, Maribor: Aristej, str. 133-145.*

Panksepp, J. (2008). »The Affective Brain and Core Consciousness«. V Lewis, M., Haviland-Jones, J.M. in Feldman Barrett, L. *Handbook of Emotions*. The Guilford Press, New York, London, str. 47-67.

Suits, Bernard (1977). *Words on Play*, Journal of the Philosophy of Sport, IV, str. 117–131.

Suits, Bernard (1988). *Tricky Triad: Games, Play, and Sport*, Journal of the Philosophy of Sport. XV, str.1-9,.

Suits, Bernard (2004). *Venn and the Art of Category Maintenance*, Journal of the Philosophy of Sport, XXXI, str. 1-14.

Suits, Bernard (2005). *The Grasshoper; Games, Life and Utopia*, Ontario; Broadview.

Sutton–Smith, Brian (1997). *The Ambiguity of Play*, Cambridge, London: Harvard University Press.

Uršič, Marko, Markič, Olga (2003). *Osnove logike,* Ljubljana: Filozofska fakulteta

Wenner, Melinda (2009). »The Serious Need for Play«. Scientific American Mind; Feb/Mar2009, Vol. 20 Issue 1, str. 22-29.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*, G.E.M. Anscombe in R. Rhees (ur.), G.E.M. Anscombe (prev.), Oxford: Blackwell.

ČUSTVENI PROCESI IN IGRA KOT NAČIN MODIFIKACIJE ČUSTVENE IZKUŠNJE

Maja Smrdu,[1] *Janez Mlakar*[1]
[1]Psihiatrična klinika Ljubljana, Studenec 48, Ljubljana

Povzetek

Čustvovanje je eden izmed pomembnejših vidikov človeškega življenja, ki določa njegovo kvaliteto in morda tudi smiselnost. Zato ne preseneča, da so se z njim ukvarjali že Antični filozofi, in da še danes potekajo številne raziskave ter nastajajo vedno nove teorije, ki ga poskušajo osvetliti in razložiti. Tako se s čustvenimi procesi ukvarjajo različne vede, zlasti psihologija, nevrologija, filozofija, evolucijska biologija in tudi ekonomija. Vendar pa ob mnogih analizah, predstavljenih z različnih vidikov, manjka sinteza, ki bi pripomogla k ustreznemu razumevanju tega kompleksnega dela človeškega življenja.

V prispevku sprva predstaviva nekatere ključne teorije čustvovanja ter jih v nadaljevanju poskušava združiti in jim dodati novo dimenzijo k razumevanju tega pojava. Ob tem poudarjava, da je za razumevanje čustvenih procesov ključno zavedanje njihove iterativnosti in sovplivanja vseh posameznikovih, do nekega trenutka doživetih čustev. Opozoriva tudi na doslej vse preveč spregledana prospektivna čustva, katerih pomen je bistven za možnost modifikacije čustvovanja - tako preko psihoterapije kot preko drugih terapevtskih tehnik, med katere uvrščamo tudi igro.

1. Pregled izbranih teorij čustvovanja

Človekovo čustvovanje se ne začne in konča z nekim dogodkom, ki v posamezniku vzbudi čustvene procese, ampak se začne z začetkom posameznikovega življenja in se nadaljuje in oblikuje vse do njegovega zaključka. Zlasti ko na čustvovanje pogledamo skozi kognitivne procese, nam postane jasno, da zanje velja kompleksnost, interaktivnost in iterativnost. Številni avtorji so že poskušali razložiti in opredeliti procese, ki določajo čustvovanje, a še nikomur ni uspelo odkriti in podati zares prave razlage tega vsakodnevnega dela človeškega delovanja.

William James (1884) je predlagal razlago, da gre pri čustvih za pomembno obliko čutenja, ki ga povzročajo spremembe v fiziološkem stanju telesa glede na njegovo avtonomno in motorično aktivnost. To pojmovanje je razširil Carl G. Lange (1885), ki je čustvo opisal kot zaznavo telesnih odzivov na nek dogodek in zavedanje teh odzivov. Kot nasprotni pol te James-Langove teorije pa je Walter Cannon (1929) postavil teorijo, da je za nastanek čustva ključno psihično doživetje, ki ni odvisno od fizioloških reakcij. Do tega sklepa je prišel z ugotovitvijo, da so zaznane reakcije v drobovju enake pri različnih čustvih, denimo pri strahu in jezi, zato na podlagi enakih fizioloških reakcij ne moremo razlikovati različnih čustev. Pomemben prispevek je v razumevanje čustvovanje prispevala Magna Arnold (1960) z uvedbo pojma ocenitve [appraisal] položaja. Šele ta povzroči, da nekaj postane privlačno ali odbojno in dobi čustveno konotacijo. Čustvo je občutena težnja k nečemu, kar je intuitivno ocenjeno kot dobro (koristno) ali proč od nečesa, kar je ocenjeno kot slabo (škodljivo). Poznejše teorije ocenitve (Lazarus 1991, Fridja 1986, Scherer 2001) so v raznih smereh nadgradile in podrobneje razgradile proces ocenitve.

V okviru kognitivnih terapevtskih pristopov se zelo poudarjajo bazična spoznanja in stališča, ki se skrivajo za človekovim čustvovanjem in ravnanjem. Kognitivne teorije tako poudarjajo spoznanje o stanju stvari, ki je formulirano kot trditev [proposition], čustva pa so opredeljena z implicitnimi trditvami, ki jih vsebujejo (Mlakar 2009).

Novejše teorije čustev zelo poudarjajo spoznavne vidike čustvovanja (ocenitev, stališča, prepričanja) in hkrati vključujejo fiziološke vidike. Nekatere raziskave kažejo, da imajo nekatera čustva razpoznavno različne profile telesnih odzivov (LeDoux 1996, Panksepp 1998). Zelo vplivna je tudi teorija somatskih markerjev (Damasio 1994), ki govori, da vsaka situacija, ki ima za posameznika čustven pomen, in ki ga ta kognitivno oceni, sproži telesne odzive (somatske markerje). Ti vodijo do telesnih občutkov, ti pa po avtomatizmu vplivajo na odgovor posameznika. Tako je telesno stanje pomemben dejavnik izbire akcije.

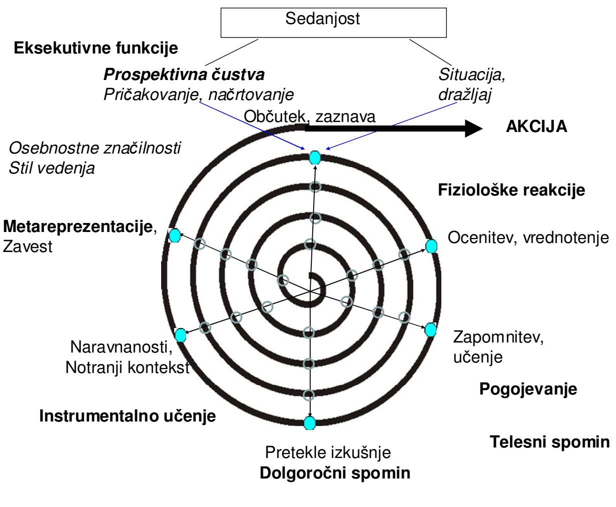
Vrsta avtorjev je že zgodaj povezovala čustva tudi s procesi instrumentalnega učenja. Med njimi zlasti Gray (1975, 1981) poudarja, da je pri tem ključno učenje vedenjskih odgovorov, ki izzovejo nagrado ali kazen. V primeru čustvovanja je nagrada vse, kar človeka spodbuja k delovanju, in kazen vse, kar ga odvrača, čemur se poskuša izogniti.

2. Sinteza psihološke analize čustvenih procesov

Odziv na vsako trenutno situacijo, v kateri se znajdemo, poteka preko dveh poti: ene, usmerjene z voljo, in druge, usmerjene z neposrednim dražljajem. To povzroči občutek oziroma zaznavo, ki sproži takojšno fiziološko reakcijo in proces prepoznavanja ter vrednotenja [appraisal], ki poteka paralelno s semantično potjo. Ob tem sodelujejo talamična jedra, hipokampus in amigdala. Nevtralno ocenjeni dražljaji privedejo do inhibicije behavioralnega sistema, medtem ko pozitivno ali negativno ovrednoteni dražljaji sprožijo nadaljnjo aktivacijo.

Predelava informacije temelji na izkušnjah, na podlagi katerih smo se naučili povezovati specifične dražljaje, zato tudi vsak nov dražljaj povratno vpliva na nadaljnje učenje oziroma pogojevanje (anterio medialni cingularni korteks in hipokampus). S pomočjo procesa dolgoročnega spomina se ohranjajo pretekle izkušnje (temporalne strukture, frontalni reženj), pri čemer je treba poudariti, da je čustven spomin specifičen, saj ni deklerativen, semantičen in epizodičen. Ti čustveni spomini vzbudijo notranji kontekst - kar je jasno razvidno pri travmatiziranosti -, ki usmerja asociacije in akcije. Na tej točki se začnejo oblikovati tudi vzorci reagiranja in obnašanja (prefrontalni korteks) ter tipični stili vedenja. Vidnejšo vlogo prevzemajo voljno inicirane aktivnosti, zavestno postavljeni cilji, metareprezentacije in prospektivna čustva.

Prospektivna čustva so vse premalo zajeta v teorijah čustvovanja in avtorji večinoma omenjajo zgolj po eno izmed čustev iz te skupine, najpogosteje zanimanje. Iz neposrednega opazovanja pa vidimo, da ljudje v vsakodnevnem življenju opisujejo tudi upanje, pričakovanje, predvidevanje idr. Ravno ti, s prefrontalnimi korteksom generirani procesi, pa so ključni za zavestno funkcioniranje človeka in njegovo sposobnost dojemanja sebe ter posledično modifikacije čustvenega odziva oziroma spremembe čustvene izkušnje. S psihoterapevtskimi metodami vstopamo ravno na tej točki čustvenih procesov. Igra kot terapevtska tehnika posega v različne faze omenjenega procesa, saj zaradi možnosti prekategorizacije preteklih izkušenj v varnem okolju ustvarja nove čustvene reakcije in spomin.



Slika 1: Spirala psihološke analize čustvenih procesov

Zaključek

S prispevkom želiva združiti različne teorije čustvovanja, ki posegajo v različne vidike čustvenih procesov, in pokazati, da je globlje razumevanje teh procesov mogoče samo, če jih dojemamo celostno. Pri tem se zdi, da se procesi spoznavanja katerega koli pojava v človeškem okolju najprej začnejo s podrobnejšo analizo in se nato usmerijo k vse bolj celostni sintezi, ki pa je na drugem nivoju razumevanja kot je bilo razumevanje pojava pred vsemi analitičnimi postopki in teorijami. Zajema lahko nešteto zavojev, podobno kot prikazana spirala psihološke analize čustvenih procesov, a se noben zavoj zares ne ponovi, temveč vsak predstavlja novo dimenzijo.

Literatura

Arnold, M. (1960). *Emotion and Personality*. New York: Colombia.

Cannon, W.B. (1929). *Bodily changes in Pain, Hunger, Fear and Rage*, 2nd edn. Appleton, New York.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. Grosset/ Putnam, New York.

Fridja, N.H. (1986). *The Emotions*. Cambridge University Press, Cambridge.

Gray, J.A. (1975). *Elements of a Two-Process Theory of Learning*. Academic Press, London.

Gray, J-A. (1981). Anxiety as a paradigm case of emotion. *British Medical Bulletin* 37: 193-197.

Lange, C. (1885). The Emotions, *v* E. Dunlap (ur.), *The Emotions*, 1922edn. Williams and Wilkins, Baltimore.

Lazarus, R.S. (1991). *Emotion and Adaptation*. Oxford University Press, New York.

James, W. (1884). What is an emotion?, *Mind* 9: 188-205.

LeDoux, J.E. (1996). *The Emotional Brain*. Simon and Schuster, New York.

Mlakar, J. (2009). Čustvovanje-osnovni pojmi, *v* P.Pregelj & R. Kobentar (urs.), Zdravstvena nega in zdravljenje motenj v duševnem zdravju. Rokus Klett, Ljubljana, str. 191-196.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oford University Press, New York.

Rolls, E.T. (2005). *Emotion explained*. Oxford University Press, New York.

Scherer, K.S. (2001). The nature and study of appraisal. A review of the issues, v K.S. Scherer, A. Schorr & T. Johnstone (urs.), *Appraisal Processes in Emotion*. Oxord University Press, Oxford, str. 369-391.

IGRA NA GLAGOLSKI NAČIN

Tomaž Strgar
Oddelek za filozofijo
Filozofska fakulteta Univerze v Ljubljani
Aškerčeva 2, 1000 Ljubljana, Slovenija
e-mail: strgart@hotmail.com

POVZETEK

Prispevek v obliki dialoga teoretično in praktično izpeljuje znanstven pristop k proučevanju občega pojma igre. V izhodišče jemlje klasično Huizingovo delo *Homo ludens*, pri čemer pa težišče obravnave s statičnega objekta 'igre' premešča na 'igranje' kot dejavnost. Ker standardna znanstvena metoda v svoji osnovi predpostavlja nek od raziskovalca neodvisen in stabilen predmet raziskave, se besedilo loteva sintaktično-semantičnih struktur, ki jih z *igro* in *igranjem* tvorijo različni jeziki. Ob tem še posebej izpostavlja povratno rabo glagola 'igrati' v slovanskih jezikih, kjer se v okvirih razmerja med osebkom in njemu zunanjim predmetom *igranja* glagolsko dejanje usmerja nazaj na prvega. S tem subjekt postane objekt znanosti.

***

PRVI: Takole piše Huizinga: »Ko govorimo o semantičnih povezavah besede 'igra' najprej opazimo tole: v nemščini sicer lahko rečemo tudi 'ein Spiel treiben', v holandščini 'een spelletje doen', a najustreznejši je vendarle glagol *igrati*. *Igramo* neko *igro*. Z drugimi besedami: ob označitvi glagola moramo ponoviti pojem, izražen v samostalniku.«[1]

DRUGI: Ok. In?

P: To dejstvo, ki sicer velja tudi v slovenščini, se mu zdi nekaj zelo posebnega…

D: Se ti to zdi čudno?

P: Čudno. Ne razumem, zakaj mu je to čudno. Mogoče je slab prevod… Hrvaško izdajo 'mam. V slovenščino je preveden samo del knjige. Citat je iz poglavja *Dojemanje pojma »igra« in njegovo izražanje v jeziku*, ki ga ni v izboru. Angleškega prevoda nimam, holandščina mi je pa španska vas.

D: Angležom je 'španska vas' *double Dutch ;-)* … Ga ni na *Google Books*? Saj to je ja stara zadeva…

P: Leta 1944 je bila prva angleška izdaja. Menda je besedilo celo sam prevedel… To je cela štorija; poglej 'Homo ludens' na Wikipediji… Skratka: »ni predogleda«.

D: Ja, s prevodi je križ… »Semantične povezave« so v različnih jezikih pač različne. Tudi meni se to ne zdi ravno nenavadno. Po moje je osnova »igre« glagolska. Vedež?

VEDEŽ: Vseslovanska beseda *igra* se navezuje na rekonstruirano indoevropsko glagolsko osnovo **Haig-* v pomenu 'sunkovito se premikati, tresti se'.[2] Pomensko sorodna so tudi izhodišča sodobnih besed za igro v romanskih in germanskih jezikih, ki se »razprostirajo nad skupinami pojmov, ki pomenijo gibanje ali delovanje.«[3] Nemški *das Spiel* je denimo substantiv glagola *spielen*, ki je v svoji najstarejši poznani obliki pomenil 'živahno gibanje, ples'. Lahko rečemo, da je osnova jezikovnih izrazov za pojem igre v evropskih jezikih praviloma glagolska.

P: Saj zato! Mogoče Huizingov jezikovni občutek… No, takole gre v hrvaškem prevodu: »Po svoj prilici to znači da je taj čin tako osobit i samosvojan da se uopće ne uklapa u obično djelovanje. Igrati nije činiti u uobičajenom smislu.« Se pravi: »taj čin«, *igra*, je »nekaj tako zelo posebnega, da se sploh ne 'uklapa' v običajno delovanje.« Če prav razumem je neobičajno tu to, da ob označitvi glagola ponovimo »pojem, izražen v samostalniku.«

D: Hm…

P: Moti me tudi ta »čin«… Igra zame ni 'čin'.

D: Kaj pa zate jé 'čin'?

P: V slovenščini predvsem vojaški rang in pivski vzklik, včasih tudi 'dejanje'; v hrvaščini poleg ranga in 'dejanja' še 'dejstvo', 'obred' ali 'opravilo'. *Igra* ni nič od tega; kvečjemu bi lahko rekel, da gre za 'dejavnost'…

[1] Huizinga, *Homo ludens*, str. 38.

[2] Snoj, *Slovenski etimološki slovar*, gsl. *igra*.

[3] Huizinga, *ibid*.

D: No ja… Lahko bi recímo rekli, da je *igra* v osnovi glagolsko dejanje… Najbrž te tu dovršnost moti…

P: Dobro, saj je tudi dejstvo; v smislu: 'obstaja nekaj, čemur rečemo igra'… Ja, dovršnost me moti! Po moje je bistvo igre v gibanju! Ko je igre konec, igre ni več! Bistvo igre je v tem, da se igrá! Tako kot je bistvo vozila, da se vozi, svetlobe da sveti, póka da póči… Hm… … Kakorkoli: meni se zdi logično, da v primeru glagolske osnove samostalnik zgolj fiksira glagolsko dejanje.

D: Ja, najbrž gre za 'izgubljeno s prevodom'… Vprašanje, če sta v originalu »čin« in »činiti« tudi iz iste osnove… Tovrstnih jezikovnih nelogičnosti je cel kup v prevodni literaturi. Primer: ko Hegel nekje pravi, da *Person* nima nobene *Besonderheit*, se to pri nas sliši kot da *oseba* nima nobene *posebnosti*, kar pa neposvečenim milo rečeno nekoliko čudno zveni… Kakorkoli: menim, da bi bilo zadevo moč navezati na Grke.

P: Ma nèmoj…

D: Poglej. Temelj zahodne kulture je antična Grčija. Temelj zahodnega koncepta znanosti je starogrška filozofija, znana po svoji »gonji za bitjo«: išče se stabilna osnova, nekaj trdnega, na čemer bo moč graditi nadaljnjo vednost. Mogoče tudi zato Huizingi samostalnik *das Spiel* deluje primarneje kot glagol *spielen*; takšna naravnanost je morda v sodobnem jeziku prekrila zgodovinsko dejstvo, da je samostalnik *igra* le substantivacija izvorno glagolskega pojma. In osnova vsake znanosti je nek samostalnik; to velja vse tja do Velikega Póka, imena bogá moderne znanstvene kozmologije.

P: A?

D: Vsaka znanost v svojem temelju predpostavlja neko substanco. Tudi ko proučuje gibanje, proučuje gibanje *nečesa*. Fizika v osnovi proučuje gibanja osnovnih delcev. Ampak ti osnovni delci so vedno le nekakšna delovna hipoteza, kar je pokazal tudi razvoj subatomske fizike.

P: In kaj 'ma tu véze fizikalna teorija o nastanku vesolja?

D: Prikaže meje tovrstnega pristopa. Fizika se z drobljenjem časa vse bolj približuje temu *dogodku*, ampak do konca ne bo prišla nikoli. *Pók* ni stanje, četudi je samostalnik. Ni substanca, temveč substantiv. Ali po filozofsko: *pók* nima lastne biti.

P: Ima pa svoje bistvo in svoj pojem… Dobro, tu bi se zdaj lahko šla dialektično razgibavat' skozi Heglove tri *Logike*, ampak: o *igri* se pogovarjamo, če se še spomniš.

D: Spomnim. Zdi se mi, da beseda *pók* tvori podobne »semantične povezave« kot beseda *igra*. Pri teh pa sva začela.

P: Potem sva spet na začetku… *Kako filozofsko !-(* … Hecam se… Misliš: prav kakor *igri* najbolj prilíči glagol *igrati*, tako je *póku* najustrezneje *póčiti*. Ampak podobnosti se tukaj tudi končajo. Če se že reče, da »*igramo* neko *igro*«, pa nikoli ne rečemo, da je »*póčil* nek *pók*«… Rečemo le: »*póčilo je*«….

D: Poglejva Jakopinovo *Slovnico ruskega knjižnega jezika*, poglavje *Glagolski način*. Citiram od začetka poglavja: »Glagolski način združuje v sebi več sestavin, ki imajo to skupno potezo, da so tako ali drugače povezane z glagolskim pomenom; obenem pa gla-«

P: *Rusko slovnico??* Pretiravaš. Zakaj za vraga bi šla pa zdaj Jakopinovo *Rusko slovnico* gledat?!

D: Zakaj pa ne?? … Dobro… Izbor je do neke mere arbitraren. Jakopinovo zato, ker mi je pač pri roki; poleg tega je pisana jasno in jedrnato. *Rusko slovnico* pa zato, ker je ruščina pač najbolj razširjen izmed našemu podobnih jezikov. To pomeni: »semantične povezave«, ki jih tvori ruščina, zvenijo logično precejšnji masi ljudi.

P: 'Nas pa Rusov je tristo miljonov' praviš… Lahko bi denimo vzel angleško… No ja: próbat' ni greh, pravijo… Ajde!

D: Ok. Nadaljujem tam, kjer si me prej prekinil: »obenem pa glagolski način kaže na odnos med povedkom, osebkom in predmetom.«[4]

P: V našem primeru torej na odnos med *igranjem*, *igrajočim* in *igro*… Ali po filozofsko: obravnava razmerje med subjektom in objektom, ki se izraža v glagolskem dejanju. To je: *nekdo igrá igro*…

D: Huizinga in, kot rečeno, tudi znanost nasploh najprej določi svoj predmet kot nekaj obstoječega. V primeru *igre* je to ideja v platonskem smislu: znanstvenemu raziskovanju je dostopna le prek svojih instanc v fizičnem, torej v različnih oblikah *igre*.

P: Drži… Po moje druge poti niti ni; vsaj takšne ne, ki bi lahko obveljala za znanost.

D: Najbrž res. Vsaj če ne gremo pretirano spreminjat vsebine pojma 'znanost'. Kako Huizinga določi predmet svoje raziskave?

P: V prvem poglavju: *Bistvo in pomen igre kot kulturnega pojava*.

D: Kot kulturni zgodovinar torej. To je še ena obča značilnost znanstvenega pristopa. Vsakršna znanost je vedno neka določena znanost. Znanost *o nečem*.

[4] Jakopin, *Slovnica ruskega knjižnega jezika*, str. 243.

P: Kar pa je v bistvu posledica prvega koraka: določitve predmeta.

D: Imaš prav. Ker hočeva bit' znanstvena, morava najprej določiti predmet raziskave. Vzemiva tu za predmet kar besedo sámo. Seveda ne samó eno – to bi bila potem bolj henološka razprava. Bistveno je, da primerjamo njene izraze v različnih jezikih.

P: *Enološka ;-)* Hecam se… Sicer pa to naredi tudi Huizinga! Spórni citat o 'čudnih' »semantičnih povezavah besede igra« je iz razdelka *Oznake za igro v germanskih jezikih*… Mogoče pa te res delujejo čudno le tam… Skratka: v predhodnih razdelkih na kratko pregleda jezikovne izraze pojma *igre* v grščini, sanskrtu, kitajščini, pri Indijancih, v japonščini ter znotraj semitskih in romanskih jezikov. Zanimivo: v istem poglavju posebej izpostavi še pojem *igre* v razmerju do 'treh univerzalij': borbe, glasbe in erotike… Išče predvsem skupne poteze: »Izhajamo lahko zgolj iz skupnega pojma igre, ki ga ob manjših odstopanjih označujejo besede, ki se pomensko prekrivajo v večini evropskih jezikov.«[5]

D: Pa obravnava tudi »semantične povezave«, ki jih z *igro* ti jeziki tvorijo?

P: Hm… Ja, ampak nikjer tako neposredno kot v omenjanem citatu. Tu se tudi najbolj čudi…

D: Slovanskih jezikov ne obravnava?

P: Ne… Ampak v bistvu se »ob manjših odstopanjih« tudi tu izraz »pomensko prekriva« s tistim »skupnim pojmom igre«, ki je predmet Huizingove raziskave…

D: … ki ga je zamejil že poprej. Evo: z *Rusko slovnico* bova prišla do »semantičnih povezav«, ki jih z lahko z *igro* tvori najbolj množično govorjèn slovanski jezik.

P: Imaš srečo… Ampak: to bo potem obravnava nekakšne 'ruske igre'…

D: *Ruska ruleta ;-)* Ne; strukture teh »povezav« bova sproti primerjala s slovenskim in še kakim jezikom…

P: Obravnavava torej pojem igre, kot ta funkcionira v ruščini. Primerjalno.

D: Recímo. In ker sva že poprej ugotovila, da je osnova *igre* glagolska, bova v izhodišče vzela glagol *igrati*.

P: Z glagolom v izhodišču… Ajde, pa naj bo *Ruska slovnica*.

D: Ok. »Glagolski način je torej pomensko-sintaktična jezikovna kategorija. Izraža se v *prehodnosti* in *neprehodnosti*, *nepovratnosti* in *povratnosti* ter *tvornosti* in *trpnosti* zvez ruskega glagola.«

P: Je *igrati* prehodni ali neprehodni glagol?

D: Hm… Glagolom, ki »v stavku lahko imajo ali celo morajo imeti ob sebi predmet v tožilniku pravimo *prehodni* (tranzitivni glagoli), ker se dejanje prenaša ali dogaja na predmetu, tako da nastane razmerje: *osebek* ali *povzročitelj dejanja – povedek* ali *dejanje – neposredni (direktni) predmet*. Брат читáет газéту.« Brat bere časopis.

P: Sestra igrá nogomet. *Igrati* je potemtakem prehodni glagol.

D: Se strinjam. Neprehodni namreč »ne morejo imeti predmeta v tožilniku brez predloga. Takih je veliko osnovnih glagolov, predvsem pa spadajo sem vsi povratni: Сестрá спит. Мáльчик ýчится.« Sestra spí. Deček se učí.

P: Ampak: 'Otroci *se igrajo* slepe miši…' Predmet je tu v tožilniku, glagol pa v refleksivni obliki… Ker so vsi povratni glagoli neprehodni, se ruski otroci ne smejo igrati slepih miši *:-o*

D: Kaže, da res ne; morajo *jih* igrati. Lahko pa *se* igrajo: 'мáльчик игрáтся'. Medtem ko lahko angleški otroci zgolj *igrajo* – 'children are playing', ima ruski jezikovni izraz na voljo povratno obliko glagola. V ruščini ima kar »slaba polovica glagolov tudi svoj povratni par. Povratni člen na -ся imajo skoraj vsi prehodni glagoli.«

P: Ja, angleščina povratnih glagolov sploh ne pozna… Kolikor vem tudi drugi zahodni jeziki *igranja* običajno ne uporabljajo v povratni glagolski zvezi, razen morda španščine…

D: …

P: Pravzaprav so tudi v slovenščini redke igre, ki dovoljujejo povratno rabo glagola v kombinaciji z *igro* v tožilniku…. Hm… Zdi se, da takšna raba pride v poštev le za najenostavnejše otroške igre… Morda za igre posnemanja…

D: Morda celó zgolj za otroške igre posnemanja… Če se gredo 'slepe miši' odrasli, temu ne rečemo več *igra* temveč *slepomišenje :-o* No, je pa povratna raba glagola *igrati* odraslim včasih le dovoljena: namreč v primeru *erotičnih* iger… Poglejva zdaj, kaj povratnost izraža glede na »odnos med povedkom, osebkom in predmetom«.

P: To pomeni: kakšno je ob *igranju* razmerje med subjektom in objektom; med tistim, ki *se igra* na eni ter *igro* na drugi strani…

D: Je tako. V primeru, ko so objekt *igranja* 'slepe miši', subjekt pa 'otroci', raba povratne oblika glagola res nekoliko čudi: glede na to, da povratnost usmerja

[5] Huizinga, *Homo ludens*, str. 31.

glagolsko dejanje nazaj na subjekt, se tu 'slepe miši' zdijo kar nekako odveč. A vendarle: tako angleška neprehodna 'children are playing', kot tudi slovenska povratna 'otroci se igrajo' raba glagola predpostavlja nekakšno *igro*, čeprav je ta zamolčana. Gre za to, da v teh primerih pravila *igre* niso posebej ostra. Ob prehodni rabi glagola je poudarek na neki določeni *igri*, angleška neprehodna raba poudarja sámo *igranje*, povratna oblika glagola pa to dejavnost usmerja nazaj na osebek oziroma subjekt *igre*. Ali še drugače: v prvem primeru je najpomembnejša tista *igra*, v drugem prevzame to vlogo *igranje* sámo, v zadnjem pa v ospredje stopijo tisti, ki *se igrajo*: »važno je sodelovati, ne zmagati«.

P: …

D: Caillois govori o 'izroditvi igre': »Načelo igre se izrodi. Tu moramo biti pozorni na to, da se ne izrodi zaradi navzočnosti goljufov ali profesionalcev, ampak izključno zaradi okuževanja z realnostjo.«[6]

P: Zveni smiselno. Ampak tu kot objekt nastopa *igra*, ne *igranje*…

D: Vsekakor. Z glagolskega stališča lahko rečemo drugače: 'načelo igranja se izrodi, ko glagol preide v prehodno obliko'. Ko objekt postane pomembnejši od subjekta. V slovanskih jezikih se ta prehod še določneje izraža skozi opuščanje povratne rabe.

P: Hm… To je pravzaprav osnova za raziskavo tistega dela vsebine pojma *igranja*, ki ga Huizinga obide: »igre«, ki so lahko 'predmet igranja v tožilniku' »je lažje opisati,« saj ob poskusu definiranja »primitivnejših iger skoraj takoj trčimo ob neizpeljivo lastnost igranja, ki pa je po našem mnenju nedostopna analizi.« V hrvaškem prevodu se na tem mestu »suočujemo s neodredivim svojstvom onih koji se igraju«…[7]

D: Angleščina je razvila dve besedi za igro: *play* in *game*. Druga je ob tem povsem objektne narave. Kot glagol funkcionira le za označevanje iger na srečo. *Gambling* je tu v sorodu. Igralec na potek igre nima vpliva, hkrati pa lahko njen izid še kako vpliva na njegovo realnost…

P: Mhm… Tisto »okuževanje z realnostjo« je običajno odraščanje: odrasli *se* ne smejo igrati, oni morajo igrati. To je stvar kulture. Če vzamemo dobesedno: odrasli ne smejo igrati sebe. Igrati morajo nekaj drugega. Nekoga drugega. 'Persona', izraz za osebo v sodobnih germanskih in romanskih jezikih je izvorno pomenil gledališko masko, medtem ko vseslovanska beseda 'oseba' izraža naravnanost nase… Samó: vsak čas se bo začelo naslednje predavanje… Greva?

D: Greva.

[6] *Teorije igre pri Johanu Huizingi, Rogerju Cailloisu in Eugenu Finku*, str. 181.
[7] Huizinga, *Homo ludens*, str. 14.

Literatura

[1] Huizinga, J. *Homo ludens*, Zagreb: Naprijed, 1992.

[2] Jakopin, F., *Slovnica ruskega knjižnega jezika*, Ljubljana: DZS, 1968.

[3] Snoj, M., *Slovenski etimološki slovar*, Ljubljana: Modrijan, 2003.

[4] ---, *Teorije igre pri Johanu Huizingi, Rogerju Cailloisu in Eugenu Finku*, Ljubljana: Študentska založba, 2003.

IGRA BESED

Ada Holcar
Zavod Republike Slovenije za šolstvo
Poljanska 28, 1000 Ljubljana, Slovenija
Tel: +386 1 30 05 162; fax: +386 1 30 05 199
e-mail: ada.holcar@zrss.si

POVZETEK

Pričujoči prispevek vsebuje razmišljanja o bogastvu pomenov, doživljanj, izrazov, delovanj ..., ki se skrivajo in prepletajo v kolažu pripovedi pod pojmom »igra«. Kljub temu, da igro opredeljujejo številne definicije in pravila, se z njo srečujemo v najrazličnejših kontekstih in jo razumemo vsak na svoj način. Vprašanje, kje se igra začne in kje konča, nas vodi do ugotovitev, da so meje odvisne predvsem od naše perspektive razumevanja le-te.

Ključne besede: igra, glasba, umetnost, qualia.

1 UVOD

Pojem igra je večplasten. Ne skriva le mnogo različnih pomenov (združuje na primer angleška play in game, pa še mnogo mnogo več) - zdi se, kot da ga vsak razume po svoje. In vsako razumevanje je »pravo«, povsod se prilega. Nekateri v igri vidijo glasbo, drugi arhitekturo in tretji življenjske procese …

Za razmislek o glasbi lahko uporabimo Nietzschejev (1995) tekst Rojstvo tragedije iz duha glasbe. »Sokrat, ukvarjaj se z glasbo!«, zapiše Friedrich Nietzsche in nadaljuje: »Vse do zadnjih dni se je (Sokrat) pomirjal z mnenjem, da je njegovo filozofiranje vrhunska umetnost muz in ni mogel prav verjeti, da ga kakšno božanstvo spominja na ono drugo »navadno, popularno glasbo«. Nazadnje pa se mu v ječi dogodi nekaj presenetljivega: sanjal je namreč, kako ga je bog-Apolon posvaril, naj se ukvarja z glasbo, in v takem razpoloženju je spesnil proemij Apolonu v čast in spravil v stihe nekaj Ezopovih basni. K temu ga je gnalo nekaj demonično svarilnemu glasu podobnega, njegovo apolinično spoznanje, da kot barbarski kralj ne razume plemenite božje podobe in da je v nevarnosti, da se pregreši proti božanstvu – če ga ne bo razumel. Besede iz Sokratovih sanjskih podob so edino znamenje pomišljanja čez meje logične narave: Mogoče – tako se je moral spraševati – pa meni nerazumljivo vendarle še ni takoj tudi nerazumno? Mogoče obstaja kraljestvo modrosti, iz katerega je logik izgnan? Mogoče je umetnosti celo nujno potrebno vzporedništvo in dopolnilo znanosti.«

Oglejmo si torej, kako na deželo umetnosti gleda njen prebivalec, ki je na obisku v deželi logike, znanosti in kognitivnih razmislekov, z njegovimi besedami.

2 UMETNOST IN IGRA

»… tisti, ki so na drugem koncu sveta kupili drage vstopnice, so pričakovali briljantno in veličastno igro simfonikov …«

»… igra je bila popolna, igranje je prevzelo mimoidoče …«

»… arhitekt je z veliko ustvarjalnega duha in igrivosti ustvaril fasado, v kateri je bilo mogoče prepoznati nenavadno igro tonskih višin, skritih v …«

Dvignila se je zavesa in zaslišali so se zvoki ubranega simfoničnega orkestra, izpod taktirk Stokowskega in Serebriera, ki je izvajal eno najtežjih simfoničnih del Charlesa Ivesa, *Simfonijo št. 4*. V simfoniji je bilo mogoče prepoznati linearno gradnjo glasbenega stavka, zavestno spajanje raznorodnih elementov, ki tvorijo zapletene politonalne in poliritmične strukture. Izvedba je bila dragocena, sporočilna in ganljiva. Dirigenta sta vse svoje sile usmerila v uresničenje skladateljevih zamisli in intencij. Navdušen aplavz so si glasbeniki pošteno zaslužili z močno izrazno igro, ki je bila uravnotežena, tonsko zlita, sposobna najsubtilnejših odtenkov in mogočnih stopnjevanj.

Na trgu Praça da Figueira v Lizboni so glasbeniki postavljali svoje ozvočenje. Mimoidoči so se ozirali in spraševali, kakšen koncert se pripravlja. Sodeč po videzu glasbenikov ne bo nič resnega, so razmišljali, medtem ko so hiteli naprej. Pat Metheny se je začel uigravati, pridružili so se mu ostali instrumentalisti in igranje se je zlilo v celoto. Mimoidoči so obstali. Igranje, ki je pričakovano dopuščalo precej improvizacijske svobode, lahkotnosti in odklonov od ustaljenih form, je kot se za tovrsten jazzovski nastop spodobi, šele v zgodnjih jutranjih urah sklenil jam session, najpristnejši izraz tovrstne glasbe.

Zanimivost hiše Weitmanngasse 1-3 na Dunaju so fiksna senčila, ki so narejena iz nerjaveče pločevine, v katero so izrezane odprtine v grafičnem vzorcu. Skozi odprtine v senčilih v notranjost hiše prihaja svetloba, ki na stenah »riše« zanimive grafične vzorce, igro svetlobe in sence. S horizontalno postavitvijo oken v različne višine nevidnega notnega črtovja, se svobodno oblikovana fasada poigrava s neizpetimi toni štirih tonskih višin. Če bi jih uglasbili bi šlo za edinstven primer v arhitekturi in glasbi. Ob pogledu nanje si v mislih zapojem melodijo, ki lepo zazveni. »Je to slučajno?« se mi poraja vprašanje. »Ne, ne, arhitekt ne pozna notnega črtovja«, slišim za sabo, »vse je le igra«. »Potem je popolna«, zaslišim v daljavi.

Dva velika dogodka sta se zvrstila, na katerih so igrali izvrstni glasbeniki, predani svoji glasbi. Džez in klasika, improvizacija in forma. Dve različni zvrsti glasbe, dve vrsti občinstva, dvoje različnih pričakovanj in dve zelo raznoliki igri. Klasična glasba zahteva natančno odigrane note, popolno intelektualno in čustveno udeležbo posameznika. Po definiciji Milesa Davisa pa so džez vse tiste note, ki (ni)so napisane vmes. In vse to se razkriva v igri ...

V svoji pisarni se je poigraval arhitekt. »Nova zgradba, nov dizajn in nov izziv«, se je pogovarjal sam s sabo, medtem ko je risal po papirju. Tokrat bo igra se je odločil, še ne vedoč kakšna.

Simfoniki sledijo dirigentu in vsemu predpisanemu. Vsak ton, vdih in fraza so natančno določeni in morajo biti odigrani v vsej svoji preciznosti. Odstopanj ni. Le dirigent ima kanček svobode, dovolj, da mu to omogoči uvrstitev med velike »maestre« in preveč, da si s tem lahko uniči sloves. Instrumentalisti so zavezani skladatelju, obdobju v katerem je delo nastalo, pravilom, ki so takrat veljala. Igrajo na številnih koncertih z vrhunskimi solisti in dirigenti, toda njihova igra je polna strogih pravil, v njej ni prostora za zabavo in ustvarjalne intencije. Poslušalec, ki spremlja določen glasbeni vzorec, ga interpretira po svojih vrednostnih in čutnih značilnostih, ki so stvar njegovih ustaljenih miselnih vzorcev. Lahko ga prevzame moč igre, vibracij, ritma in zvoka.

Za džez je značilno poigravanje z zvoki, eksperimentiranje z glasbili, ustvarjanje imaginarnih zvočnih predstav, izvirno muziciranje, ustvarjalno in eksperimentalno druženje, improvizacija, zvočno raziskovanje in še in še. Džezovska glasba ni nikdar ostajala znotraj meja umetnih tvorb. Izvajalci si podajajo glasbene zamisli svobodno, njihova igra ni povsem natančno določena. Je pristna ter predana trenutku v katerem nastaja v interakciji z občinstvom. Lahko je zabavna, igriva …

Nekateri poslušalci koncert obiščejo, ker so slišali, da bo dirigiral veliki »maestro« in ker so vstopnice enormno visoke. Na takšnih prireditvah se moraš pojaviti in odigrati svojo vlogo, menijo tisti, za katere je obisk koncerta le igra in ne umetniški užitek. Džez koncert običajno privabi publiko, ki ji je tovrstna glasba užitek in jih resnično zanima. Glasbi se želijo predati in v sebi prebuditi igrivost, čudenje …

Dobra arhitektura govori zase, v sebi združuje umetnost in formo. Lahko nas pritegne ali odvrne, včasih strogo klasična, drugič igriva.

Ideja iz glasbe se v trenutku poslušanja lahko prenese v trenutek v življenju. Pri poslušalcu lahko spodbudi nov način komunikacije, razmišljanja.

3 IGRA OTROK

Na televiziji so predvajali posnetek koncerta *4. Simfonije* Charlesa Ivesa. To so bili trenutki namenjeni njej, čas v katerem se je predala glasbi in pozabila na vse. V sobo je vstopil njen petletni sin, ki je ob pogledu na dirigenta zmajal z glavo in vprašal: »Le kakšna igra je to? Bi se jo šla tudi midva? Poglej, vse kar potrebujeva so kitajske palčke« in odvihral je proti kuhinji. »Le kako naj mu razložim, da je koncert umetnost in da je njihovo igranje …« se je spraševala medtem ko se je vrnil, začel poigravati s palčkami in se neznansko zabaval. »Mama, zakaj si tako resna? Igraj se z mano!« Namesto igranja je začela pogovor o igri. »Igra je predvsem namenjena zabavi in ima svoja pravila …« Spomnila se je veliko zanimivih primerov, sin pa jo je poskušal razumeti. Iz službe se je vrnil oče in začel pripovedovati: »Z investitorji smo se pogajali celo popoldne, toda imel sem občutek, da je bila vse le igra. Ne vem, kaj bo s projektom.« »Si rekel igra?« se je oglasil sin. »Igra v službi? Si prepričan? Potem bom tudi jaz arhitekt.«

Med večerjo sta se s sestro začela igrati s kitajskimi palčkami. Kmalu sta z njimi izvajala zanimive ritmične vzorce. Mama se je oglasila: »Tišina, dovolj je te igre«. Takrat pa je hčerka užaljeno odvrnila: »Mama, saj to ni igra. Nisi slišala, da je šlo za dialog poln sinkop?« In njen brat je začudeno gledal. »Le kaj je potem igra?« se je še naprej spraševal. Sestra se je poigrala s svojim mobitelom in začela na glas brati, kar je pisalo v Wikipediji: »Igra je razvedrilna dejavnost, ki vključuje enega ali več igralcev in je določena:

- po cilju, ki ga igralci poskušajo doseči, ali
- po množici pravil, ki opredelijo, kako igralci igrajo.

Igre vključujejo enega igralca, pogosteje pa pomenijo tekmovanje med dvema ali več igralci«.

Je delo, ki ga opravljate vi razvedrilno? So vas najprej seznanili s cilji, ki jih je potrebno doseči in množico pravil: kdo, kje, kako in do kdaj? Je pomembno kdo prvi pride na cilj?

4 QUALIA

Je delo lahko igra? Je igra lahko delo? Kaj pa umetnost? Ali ni najvišja umetnost hkrati oboje: delo do poslanstva in igra do onemoglosti?

Ali ne velja to tudi za znanost?

Profesor Denis Noble (2006) nas v svoji knjigi The Music of Life sprašuje, kako vidimo svet? Odloči se, da nam bo povedal zgodbo protagonistov jaz in ti. Medtem ko piše poglavje, se jaz zazre v tisk in reče: »tisk je črn, stran pa je bela«.

»Kako«, vpraša ti, »kako veš, da jaz vidim enako črno in belo kot ti? Morda je tisto kar vidim jaz, ko berem tvojo knjigo, kar bi ti videl kot modro na roza ali zeleno na ultramarin ali pa katerokoli od milijon možnih kombinacij, vključno z barvami, ki jih še nikoli nisi videl! Vse kar lahko zares rečeva je, da se mora moje doživljanje barv vedno povezati s tvojim na tak način, da se bova vedno strinjala, kako poimenovati, kar bova videla. Uporabljala bova enaka imena, pa čeprav bova morda videla drugače.«

»Ne bodi neumen«, odvrne jaz, »saj naju je vendarle mati naučila, kaj je črno in belo in kaj so preostale barve!«

»Že, že«, reče ti, »toda to ni to, kar mislim jaz. Ko naju je mati učila, sem mislil, da gledava enake stvari in da jih morava videti na enak način. Potem pa sem prebral nekaj filozofije in nevroznanosti in sedaj do konca življenja ne bom razumel, kako lahko kdorkoli, vključno z mamo ve, kaj vidim jaz, ko gledam tisk. Moje doživljanje je znotraj mene, v moji glavi, v mojih možganih. Nihče drug ga ne vidi. Zato jaz morda vidim svet povsem drugače kot ti.«

»To se mi zdi že prav smešno«, poreče jaz, »ah, saj te bo minilo.«

»Ne, ne, ne bo me minilo«, odvrne ti. »Jaz sem jaz in mislim resno. In ti si ti. In ne moreš vedeti, kaj doživljam v svojem privatnem življenju.«

»*V svojem privatnem življenju?* Kje na svetu pa je to?«

»Ne igraj se iger z mano. V mojo glavo pač ne moreš videti.«

»Da ne? Pravzaprav obstaja način, ti povem kako? Lahko posnamem nevrone, skeniram možgane in vidim spremembe v pretoku krvi in lahko storim še mnogo drugih stvari … in ti bi lahko storil enako za mene, veš v najinih glavah bi našla podobne stvari.«

»Ja, ja, to mi je jasno. Ne mislim, da sem drugačen, ker sem narejen drugačen – čeprav očitno nisva fizično identična. Je le da … torej, jaz sem jaz in ti si ti. Kaj ne vidiš tega?«

»Ja, seveda vidim to, ne razumem pa, zakaj govoriš o svojem privatnem življenju in zakaj ga ne smem poznati.«

»Ah, daj no, saj veš da se ne obračam k svojim nevronom, pretoku krvi ali čemur koli drugemu kar je fizične narave. Govorim o svojih doživetjih. Jaz imam svoja doživetja in ti imaš svoja. Veš, dandanes imajo celo svoje ime. Ljudje jim pravijo *qualia*. Tudi ti jih imaš zagotovo. Samo poglej eno izmed črk na tej strani. Poglej belo na črni *quali-ji*.«

»Tvoj pogled postaja že prav dualističen. Misliš, da je v tebi nekaj kar ni fizično?«

»Ah, ne, ne, sploh ne!! Te stvari so zgrajene iz mojih nevronskih procesov, morda na način da preprosto *so* moji nevronski procesi, vsaj počutim se kot da bi jih imel. Ne mislim, da je moja duša v interakciji z možgani, pravzaprav kolikor vem, jaz *sem* moji možgani. In ti ustvarjajo tisto kar vidim, čutim, slišim.«

»Res? Jaz pa sem mislil …«

4 ZAKLJUČEK

Igra ima svoja pravila, to pa še ne pomeni, da jih ne moremo spremeniti. Pravilo je lahko celo, da pravil sploh ni. In nasploh so igre, ki potekajo zgolj po strogih pravilih, brez vsake svobodne volje in nenadnih presenečenj, dolgočasne. Ali imajo sploh pravico igra se imenovati, se lahko povprašamo.

Ali si upamo? Opustiti vse drugo in se igrati? Postati eno z igro in hkrati z vsem.

Literatura

[1] F. Nietzsche. *Rojstvo tragedije iz duha glasbe.* Karantanija. 1995, str. 92-94.

[2] D. Noble. *The Music of Life.*Oxford University Press. 2006, str.114-116.

IZ BOLEČINE V KVANTNO IGRADOST POEZIJE

Vida Mokrin-Pauer
(pesnica in pisateljica, samozaposlena v kulturi)
vida.mokrin.pauer@gmail.com

Abstract

Čudovita, čarobna je tema/svetloba: IGRA. Še posebej s pesniško umetnostjo s(m)o besede z različnimi pomeni v različnih kombinacijah oziroma komunikacijah. V *Slovarju slovenskega knjižnega jezika* (SSKJ) je pod gesli »igra«, »igrati«, »igrivost« opisana množica različnih pomenov in rab teh besed [21]. V tem prispevku bom govorila o tistih pomenih, ki se povezujejo z besednimi sklopi: "otroška igra", "igra besed" in "igra naključij", vse skupaj pa bom prežemala s kvantnim pogledom na svet in radostjo ustvarjalnosti v igri-poeziji življenja.

Uvod in razprava

Poljudnoznanstvene knjige in filmi o vesolju in kvantnosti so v temelju spremenili moj pogled na svet.
Vidim oziroma doživljam in razumem in soustvarjam kvantnost hkrati kot: (1) izven sebe (kar ni res, a se pač zlažem, da lahko "pametno" opazujem "zunanjost izven sebe") in (2) v sebi in s sabo, kajti zaupam kvantnikom in z njimi strinjajočimi se, npr. kognitivnim znanstvenikom, da je zavest opazovalca hkrati tudi zavest opazovanega [12, 17].

Čarobna kvantnost, kjer ni trdnih objektov, so samo možnosti; kjer je vse neločljiva celota, je eno in je hologram; kjer se kvantni skoki dogajajo brez poti skozi prostor in čas; v kvantnosti so vsi časi in prostori skupaj istočasno vsepovsod; dogodek je istočasno snov in brezsnovna energija; v kvantni resničnosti se dogodek odvije šele, ko opazovalec usmeri vanj/zanj pozornost; v kvantnem svetu je vse *IN IN*, ne pa *ali/ali*, ni izključevanja, hierarhiziranja; v kvantnem svetu so delci na mnogih krajih istočasno … – ko sem vse to brala/nabirala na kvantnem polju neskončnih možnosti …, mi je bilo, kot da berem/nabiram najbolj čudežno (= skrivnostno) cvetje spoznanj v vesolju oziroma v multiverzumu z mičnimi kozmičnimi strunami in M teorijo in sintropijo in vseživostjo in izvorno zavestjo in … vred [2, 3, 4, 5, 6, 7, 8, 10, 17, 22, 23, 24, 25, 26] .

In sem se ovedela, da je ta taka čudovita zavest, da je to vedenje pravzaprav že vedno vsepovsod vpisano v tisto umetnost (zdaj govorim predvsem o literarni umetnosti), ki govori o soudeleženosti v s pravili, pričakovanji in cilji nezamejeni igri vse(h)možnosti in to pomeni tudi vse(h)živosti in vse(h)smiselnosti stvarstva [11]. In take literature je izredno malo, sploh pa na Zahodu, kjer je večina miselnosti-literature še vedno ujeta v newtonovski pogled na svet. V zahodnjaški družbi je literatura s tako svobodnim pogledom na svet potisnjena v okvirček otroške literature oziroma literature za otroke, ali pa so take igrarije (bile) dovoljene le umetniškim avantgardam [1]. Oboje pa kot da nima vrednosti tako imenovane prave resničnosti, prave umetnosti. Kadar igro oziroma umetnost povezujemo z otroškim, ji je odvzeta pomembnost. (Npr. v *SSKJ*: »igra« – otroška dejavnost, navadno skupinska, za razvedrilo, zabavo.) In tudi pomen besedne igre (v *SSKJ*: "besedna igra": besedna figura, ki obstoji iz dveh, blizu stoječih enakih ali podobnih besed z različnimi pomeni) je zožen npr. na figuro, na izrazno sredstvo.
Ker razmišljam o igri, kvantni fiziki in poeziji, v kateri je izredno pomemben jezik, navajam par delčkov jezikoslovca Tomaža Sajovica študije *Jezik, umetnost in kvantna fizika* [20]:

- 3. 6.1964 je kvantni fizik Werner Heisenberg na hribu Pniks v Atenah svoj nagovor sklenil z besedami: »Če je harmonija v družbi odvisna od skupne interpretacije »enega«, enotnosti, skrite za pojavi, utegne biti jezik pesnikov pomembnejši od jezika znanstvenikov.«
- Kvantna fizika je tako rehabilitirala naravni jezik, najpomembnejše pojme v njem ter njihovo vlogo pri spoznavanju neskončnih področij resničnosti.
- Jezik vsakdanjega življenja oziroma naravni jezik, ki oblikuje podzavestno razumevanje nas samih in okolja, v katerem živimo, je jezik komplementarnosti – dinamični odprti sistem.

Torej, nekateri kvantni fiziki so se poklonili jeziku pesnikov, ki izvira iz naravnega jezika, ki ima v sebi tudi podzavestno razumevanje in komplementarnosti, odprtosti, česar besedni jezik znanstvenikov nima. (Za jezik umetnosti je izredno pomembno tudi samonanašanje.) Meni pa se zdi krasno, če bi se tudi pesniki poklonili kvantnim fizikom, pa ne zaradi njihovega besednega jezika, ampak zaradi njihovega vse bolj razvitega matematičnega jezika (ki ga pesniki, večina ljudi ne razume, zato rabimo prevode v besede, v prevodih pa se ogromno pomenov izgubi), s katerim na drug način prihajajo do razodetij o čarobnosti stvarstva.

Vse bolj in bolj uživam v branju poljudnoznanstvenih knjig o kvantnosti kot v branju (sodobnega) leposlovja za odrasle – katerega večina mi je ali hladno opisovalna ali mučno bolečinska. Kajti vanjo je vpisan newtonovski pogled na svet s svojo delitvijo na subjekt-objekt, izključevalnostjo in hierarhijo vrednostnih sistemov. Seveda boli ogromno sveta, dogajajo se grozote, ampak dogajajo se zato, ker se ne znamo osvoboditi starih pogledov in načinov (samo)ustvarjanja. Kvantni pogled na svet fantastično osvobaja, saj usmerja pozornost na spregledovano čarobnost stvarstva. Kvantni svet je igriv. Gotovo se tudi zato kvantni fiziki na veliko fotografijah smejejo kot navihani otroci. Max Planck, začetnik kvantnosti, je bil baje čisto obupan nad svojim odkritjem kvantne nelogičnosti, ampak igrivost – to je vsemožnost in vsepovezljivost v kvantnosti – je za sodobne znanstvenike, ki so jo sprejeli za dejstvo, postala neskončno zanimiva in s tem radostna. Ja, igra/igrivost v smislu vsemožnosti-vsepovezljivosti - vsesmiselnosti stvarstva.

Vedenje o kvantnosti je spremenilo tudi moj način ustvarjanja in mojo poezijo, tako tisto za otroke in odrasle (če bi se tudi odrasli odprli poeziji oziroma vsemu leposlovju za otroke in ga veliko brali, bi se svet gotovo spremenil na boljše), kot tudi tisto, ki je dojemljiva le odraslim. Obe sta postali igriviigriviigrivi, polni humorja, nepredvidljivosti … Poezija mi je postala radostni sam nad sabo začudeni »izum« stvarstva, ne pa žalostno »odkritje« o stvarstvu [16]. Nastaja pa tako, da jo prikličem. Z avtohipnozo, afirmacijami in različnimi tehnikami meditacije (krasno deluje npr. blebetavščina, ki jo včasih uporabljam tudi pri vodenju literarnih šol) se priklopim na tako imenovano polje kvantnega vakuuma oziroma na izvorno zavest [9, 13, 14, 19]. In pride navdih, pesem se kot da sama na-piše, jo kanaliziram; ponavadi se potem več dni ukvarjam z brušenjem njenega diamanta v briljant, a najprej pesem kar priteče. Opazila sem, da kadar hočem v pesmi kaj opisati ali izpovedati – kar se mi je zgodilo, preden sem šla v pisanje, ali kar opazujem okoli! sebe ali v sebi kot ločeni od trenutnega navdiha iz vsemožnosti – takoj zapadem v izključevanje, hierarhiziranje, tehtanje, v vzvišeno resnobnost ali v bolečinsko javkanje in jamranje [del pes. zbirke 15]. Ne, najboljše pesmi mi nastajajo iz zdaj-a in to kot da predvsem kar iz igre jezika samega. Kot da se jezik zavesti in podzavesti in zavest in podzavest jezika – priklopita na kvantno vsezavest stvarstva. Ogromno je besednih iger (taka poezija je najtežje prevedljiva v druge jezike) in, glej čudo, ves ta igrivi živžav nastopajočih in ogromno med seboj povezanih različnih pomenov oziroma perspektiv – se mi kar sam zlije v večinoma zelo tekoče in z lahkoto razumljive pesmi [del pes. zbirke 15, 16]. Ki so tistim, ki imajo radi igrivost, mnogopomenskost, otroškost, kvantnost – radostno osvabajajoče, medtem ko tisti, ki zahtevajo resnost in tehtnost, kritičnost, odgovornost do sveta (kot da je svet zunaj njih) ne pristajajo na mojo frivolnost. Igrivost kot izumljanje novega je v nasprotju z obupom, obsojanjem. Davidu Bohmu in mnogim je svet hologram in da so vse perspektive na svet oziroma sveta resnične oziroma enakovredne! Odvisno, katero izbereš, da bi jo gledal, in jo vidiš. In jaz pač izbiram perspektivo igrivosti, radoživosti, humorja … In se Vidim, da je moja poezija in vse, kar se mi dogaja oziroma kar dogajam – bistveno povezana z mojim odnosom do sebe. Ta odnos do »same« sebe se mi potem povnanja v »zunanji« svet, ga zrcalim v zunanji svet [18]. Če izberem, da sem igriva, ustvarjam igriv svet, če izberem, da sem zamorjena, obsojevalna, žrtev, ustvarjam zamorjen svet z žrtvami in obsodbami … Seveda se tudi jaz včasih počutim pobita, tragična na tem banalnem groznem svetu. Ampak verujem in zaradi kvantnosti tudi vem, da je enako možno tudi povsem drugače in obratno vsepovsod in vedno.

Anton Zeilinger je izjavil, da je po njegovem mnenju najpomembnejše odkritje 20. stoletja to, da so posamezni dogodki neodpravljivo naključni. Torej je svet igra naključij. Ampak meni se pa zdi, da je še bolj tako, da je vse zavest in vse živo (torej je tudi smiselno?!), da je tako, kot piše Andrej Detela [8]: »Pokaže se, da je zavest ontološka lastnost samega fizikalnega polja. Za kvantno dinamiko velja povsem ista matematika kot za procesiranje informacije v živih bitjih – vključno z višjimi duševnimi procesi (na primer mistična doživetja). Ta povezava nam morda vendarle odpira možnost znanstvenega vpogleda v fenomen izvorne zavesti.«

Anton Zeilinger [23] svojo knjigo zaključuje s stavkom: »Svet je vse, kar je, pa tudi vse, kar je lahko.«

In svet je lahko tudi neskončna radost-igra-poezija, ki se brez začetka in brez konca sveti vsa lahkotna in igriva.

Literatura:

[1] Hans Christian Andersen: *Bedak Jurček* (MK, Ljubljana, 1966)

[2] Gregg Braden: *Božanka matrika* (Cangura, Ljubljana, 2009)

[3] Barbara Ann Brennan: *Moč zdravilnih rok* (Quatro, Ljubljana, 1987)

[4] Fritjof Capra: *Tao fizike* (Opus, Beograd, 1989)

[5] Deepak Chopra: *Sinhrousoda* (Pivec, Maribor, 2009)

[6] Deepak Chopra: *Kvantno liječenje* (V.B.Z., Zagreb, 1999)
[7] Andrej Detela: *Magnetni vozli* (Littera picta, Ljubljana, 2002)
[8] Andrej Detela: *Zavest kot fenomen subkvantnega sveta* (Kognitivna konferenca, 2004)
[9] Erhard F. Freitag: *Podzavest-centrala moči* (M. Beranič, Maribor, 1992)
[10] Brian Greene: *Tkanina vesolja* (Učila international, Ljubljana, 2006)
[11] Hafis: *Zapleši z Menoj* (Ljubljana, 2005)
[12] Urban Kordeš: *Od resnice k zaupanju* (Studia humanitatos, Ljubljana 2004)
[13] David Lynch: *Kako ujeti veliko ribo – meditacija, zavest in ustvarjalnost* (Tozd, Ljubljana, 2008)
[14] Osho: *Meditacija: Umetnost ekstaze* (Fran Burgar, Ljubljana, 1994)
[15] Vida Mokrin-Pauer: *Upoštevaj kvante!* (pesniška zbirka, Apokalipsa, Ljubljana, 2007)
[16] Vida Mokrin-Pauer: *Poleti poleti!* (108 pesmi na kartah za uvidevanje, Sanje, Ljubljana, 2009)
[17] Mitja Peruš: *Vse v enem, eno v vsem* (DZS, Ljubljana, 1995)
[18] Bobbi DePorter: *Kvantno učenje* (Glotta Nova, Ljubljana, 1996)
[19] Peter Russell: *TM : transcendentalna meditacija* (Stella, Šmarješke toplice 2003)
[20] Tomaž Sajovic: *Jezik med umetnostjo in znanostjo; poglavje Jezik, umetnost in kvantna fizika* (Založba ZRC SAZU, Ljubljana, 2005
[21] Slovar slovenskega knjižnega jezika, Ljubljana, DZS, 1994
[22] Lynne McTaggart: *Polje* (Ara, Ljubljana 2004)
[23] Anton Zeilinger: *Einsteinova tančica – novi svet kvantne fizike* (Zavod RS za šolstvo, Ljubljana 2005)
[24] Danah Zohar: *The Quantun Self* (Flamingo, London, 1991)

- Le nekaj iz množice izobraževalnih filmov:

[25] SCIENCE & SPIRITUALITY - Holographic World
http://www.youtube.com/watch?v=kVt9VIvY4Ds&feature=related
[26] QUANTUM UNIVERSE
http://www.youtube.com/watch?v=0l6ZBZnOBCA&feature=related
[27] BBC science: Who lives in the eleventh dimension? - Parallel Universes 1-5
http://www.youtube.com/watch?v=nS734OF1PiI&feature=related
http://www.youtube.com/watch?v=T02H02D40EA&feature=related
http://www.youtube.com/watch?v=97LQSWVEk6s&feature=related
http://www.youtube.com/watch?v=NweKK7_nOBE&feature=related

http://www.youtube.com/watch?v=jjzj4RvJFgo&NR=1

GUT FEELINGS: THE UNCONSCIOUS STRIVE FOR COHERENCE MAXIMIZATION AS A KEY ELEMENT IN COMPLEX DECISION-MAKING

Sebastian Bachmann
Department of Philosophy
University of Vienna
Dr. Karl Lueger Ring 1
1010 Vienna, Austria
Tel: 43 650 5050784
e-mail: s.bachmann@vol.at

ABSTRACT

An increasing body of research suggests that decision-making processes are heavily influenced by gut feelings and affective moments. I examine recent theories of human thought and cognition to suggest that unconscious thought follows inference principles of coherence maximization due connectionist constraint-satisfaction processes within complex decision-making circumstances.

1 THE PHENOMENOLOGY OF DECISION-MAKING

Commonly decision-making circumstances encompass the existence of simultaneous competing alternatives in a situation that is coined by disorientation. The following phase is characterized by the person's evaluation and virtual simulation of future possibilities, prospects, risks or obstacles to find a new sense of coherence and a new orientation in the progress of life [1]. Thomae characterizes the phenomenology of the decision-making process further as to *feel* one's way towards an imaginary situation: How would it be if I would do this or that? How would I feel after going to the party instead of finishing my work? He emphasizes that the essence of making a decision is the process of coming closer to the experience of consistency and the congruency between virtual alternatives and the updated concept of the self - containing once owns motives, experiences, penchants and wishes. This process of evaluation about which possibility is the best to choose is then characterized by active-searching cognitive, and passive-receptive intuitive moments. The interaction between the former and the latter contributes to a perceptible crystallization of the "right" alternative that enables the person to identify with its decision. In case of success the decision-maker experiences congruency: "That is the right thing for me", "That's the way it should be".

2 WHY SHOULD INTUITIVE DECISION-MAKING BE PREDICTIVE?

Fuchs argues against a psychological deterministic viewpoint in which the calculation of decision outcomes in advance becomes in principal possible [2]. Spaemann argues that it is wrong to assume ideas, desires or motives as independent variables [3]. The decision-making process must be seen as a dynamic process that encompasses the modification and the reciprocal influence of these variables. They cannot exist as fixed determinants. As Tomae outlines further, they rather exist as functions of an actual experienced situation in which they are shaped and continuously modified by the situations` direction, intensity and meaning. Further, memories, feelings or thoughts are unique entities that do not exist in the exactly same way twice. Therefore it is impossible to describe the relation between an action and the mental state that caused it in a law like fashion.

In contrast, Paul Thagard proposes a general theory of coherence as the satisfaction of multiple interacting constraints and shows that this theory has numerous psychological and philosophical applications since much of human cognition could be understood in terms of coherent constraint satisfaction [4]. It has been well studied in the field of cognitive science and as a general theory to describe the world. Within, Thagard discusses how people make sense of each other and the world they live in on the basis of fitting something puzzling into a coherent pattern of mental representations, such as concepts, beliefs, goals and actions; and that this coherent states can be exactly computed.

A more holistic view on decision-making admitting the vital role of emotional moments found its way into modern approaches to account for the phenomenon that differ significantly from standard rational utility models. Philosophers, neuroscientists and psychologists have recently pointed out the inherently emotional nature of decision-making (e.g. [5], [6]). In fact, latest findings in decision neuroscience suggest that our judgments are initiated by the unconscious weighting of emotional tags associates with our memories rather than by the conscious weighting of rational pros and cons [7]. In other words we often start to *feel* something before we are conscious of having *thought* anything. The findings about the powerful influence of gut feelings on our unconsciousness and behaviour increase interest in the scientific community.

Dijksterhuis & Nordgren proposed a theory about human thought named the unconscious-thought theory (UTT) [8]. It distinguishes the thought process into an unconscious and a conscious mode of thought. Here, conscious thought is defined as a cognitive thought process that is object-relevant or task-relevant and occurs while the object or task is in one's conscious attention. In contrast unconscious thought is an object-relevant or task-relevant thought process that occurs while conscious attention is directed elsewhere.

The two modes differ in its characteristic and are therefore preferable under different circumstances. Interestingly, the theory suggests that when dealing with simple issues one should use conscious thought, whereas decisions in complex circumstances should be tackled by unconscious thought.

3 HYPOTHESIS

I offer the admittedly speculative hypotheses that unconscious decision-making processes underlie the existence of coherence mechanisms that are key for their better understanding and description. In this section I discuss evidence for this claim and the relation between coherence theory and unconscious moments in the process of decision-formation.

I support the hypothesis by presenting evidence about the role of coherence as an inseparable part in complex decision formation. Next, an evaluation of the usefulness of constraint-satisfaction models for the sake of a better understanding of decision-processes is being offered. Finally I argue that unconscious moments within the thought process involved in complex decision-making follow a function of maximal coherence among the underlying inferences; and that it can be modelled using a connectionist framework.

4 COHERENCE MECHANISMS IN COMPLEX DECISION-MAKING

Evidence for the existence of coherence mechanisms within decision-making processes under complex circumstances comes from Simon and colleagues [9] and Holyoak and Simon [10]. The latter have shown that tasks such as deciding which job offer to accept involve sets of complex inferences that need to be integrated within the decision. They examined such multiple-inference based decision-making by asking college students to render a verdict in a complex legal case. Their principle finding was that the decision-making process is accompanied by a systematic change in the evaluation of the inferences toward a pattern of coherence within the emerging decision. Simon and colleagues replicated and extended these basic findings in showing that the processing of complex tasks is accompanied by a change in inferences that increases coherence with the decision made, leading to a coherent representation of the situation. That is, the inferences that supported the chosen decision became stronger, and the inferences that supported the rejected decision decreased in their level of acceptance. They conclude that when faced with tasks of high ambiguity, conflict and complexity - conditions that might otherwise be experienced as insurmountable - the increase of coherence in support for one of the decision alternatives enables and facilitates the making of confident decisions.

5 THE USEFULNESS OF CONSTRAINT-SATISFACTION MODELS

Over the past decades connectionist theories of cognition and in particular constraint satisfaction mechanism reformulated established consistency theories.

Simon & Holyoak claim that the capability of connectionist representations capture rich and large conceptual structures whereas he relation to the person's background knowledge constitutes important progress over the restrictive dyads and triads of yesteryear [11]. Furthermore they argue that interactive constraint-satisfaction algorithms provide a more realistic and nuanced means of resolving consistency than the crude mathematical rules used by consistency theorists. Further, as suggested by Read and colleagues, connectionist-based models of thought based on constraint satisfaction offer a conceptual framework that overcomes the limitations that hobbled cognitive consistency theories, most notably, the difficulty of generalizing such theories to achieve coherence among large networks of beliefs [12]. These findings lend support to the assumed possibility that coherence-driven mechanisms of constraint satisfaction can also be applied in connectionist terms for approaching inferences in the setting of unconscious decision-making processes.

Dijksterhuis and Nordgren argue in the same direction: "Up to now, we have discovered that unconscious thought leads to polarization and that people are better able to organize information in memory with unconscious than with conscious thought. This knowledge, however, represents no more than the tip of the iceberg, and there is much more about the processes involved that remain to be discovered. For now, it is perhaps best to conceive of unconscious thought as a computational process, as slowly calculating what is best."

6 COHERENCE FORMATION IN UNCONSCIOUS MOMENTS

In the following section I emphasize evidence that coherence mechanisms are involved in unconscious moments of the decision-making process.

I speculate that the "heureka-effect", the process where the unconscious suddenly delivers the solution of a problem to the conscious, might be better understandable within the framework of coherent constraint satisfaction principles. Doing so, one could gain the insight that the effect might be connected to the moment when unconscious evaluation and the process of coherence maximization reaches a specific status of slowly integrated information that forms a coherent summary judgment.

I further suggest that people who face complex circumstances in making a decision, the unconscious part of thinking plays a significant role in setting the scene for arriving at a "good" decisions. Further, in my point of view, these underlying processes of multiple inferences between encoded mental representations of unconscious moments such as propositions, desires, feelings, goals or attitudes behave in a way characterizable in terms of connectionist constraint satisfaction principles that maximize the overall state of coherence.

Since I am following UTT theory and its perspective on human unconscious and conscious thought, it seems as there is an inconsistence dealing with it's outlined rule principle. The principle suggests that conscious thought is precise and can follow strict rules, whereas unconscious thought gives rough estimates and cannot follow rules. How does this fit with the assumption that the unconscious crystallisation process of an alternative within the decision scenario obeys rules of constraint satisfaction and coherence maximization?

Claxton made the argument in his book on consciousness, the unconscious and creativity that the distinction between rule-based and associative thinking largely maps onto the distinction between consciousness and the unconscious and is thus in line with unconscious-thought theory that comes to similar conclusions [13]. However, Sloman distinguished between following rules and merely *conforming* to them [14]. Dijksterhuis and Nordgren underline this important distinction with the example that an apple conforms to gravity by falling down rather than up, but it does not actively follow a rule in doing so.

7 CONCLUSION

To conclude, the ability to follow rules or criterions allows for precision in the context of a decision and is one of the strength of conscious deliberative thought. However even if unconscious thought cannot follow strict rules such as for example arithmetics, it still might have the ability to conform to rules. Therefore, in my point of view, it is valid to assume unconscious mechanisms as conforming to rules of coherence maximization principles.

Nonetheless the available evidence thus far does not make it fully transparent what unconscious thought really is and how it works. Understanding the exact process by which unconscious thought forms a judgment and sets the scene for a decision to occur will take additional experimenting and time.

References

[1] Thomae, H. 1960. Der Mensch in der Entscheidung. München. Unger, R.M. (1986).

[2] Fuchs, T., 2005. Was heisst ‚sich entscheiden'? Scheidewege, 35.

[3] Spaemann, R., 1996. Personen: Versuche über den Unterschied zwischen 'etwas' und 'jemand'., Klett-Cotta.

[4] Thagard, P., 2000. Coherence in Thought and Action, MIT Press.

[5] Finucane, M.L. u. a., 2000. The affect heuristic in judgments of risks and benefits. Journal of Behavioral Decision Making, 13(1), 1–17.

[6] Lerner, J.S., Small, D.A. & Loewenstein, G., 2004. Heart Strings and Purse Strings. Psychological Science, 15, 337–341.

[7] Damasio, A.R., 1994. Descartes error, Putnam New York.

[8] Dijksterhuis, A. & Nordgren, L.F., 2006. A theory of unconscious thought. Perspectives on Psychological Science, 1(2), 95.

[9] Simon, D. u. a., 2001. The Emergence of Coherence Over the Course of Decision Making. Learning, Memory, 27(5), 1250–1260.

[10] Holyoak, K.J. & Simon, D., 1999. Bidirectional reasoning in decision making by constraint satisfaction. Journal of Experimental Psychology: General, 128(1), 3–31.

[11] Simon, D. & Holyoak, K.J., 2002. Structural dynamics of cognition: From consistency theories to constraint satisfaction. Personality and social psychology review, 6(4), 283.

[12] Read, S.J., Vanman, E.J. & Miller, L.C., 1997. Connectionism, parallel constraint satisfaction processes, and gestalt principles:(Re) introducing cognitive dynamics to social psychology. Personality and Social Psychology Review, 1(1), 26.

[13] Claxton, G., 1997. Hare brain, tortoise mind: Why intelligence increases when you think less, Fourth Estate.

[14] Sloman, S.A., 1996. The empirical case for two systems of reasoning. Psychological bulletin, 119(1), 3–22.

WITTGENSTEIN'S CRITICISM OF THE COMPUTATIONAL METAPHOR OF THOUGHT

Andrej Ule
Department of Philosophy
Faculty of Arts
Ljubljana University
Aškerčeva 2, 1000 Ljubljana
e-mail: andrej.ule@guest.arnes.si

Abstract

I develop the criticism of several assumptions that are crucial for a large part of cognitive science on basis of some Wittgenstein's remark on mind and psychology. The theoreticians of cognitive science have not yet seriously considered Wittgenstein's criticism so they, quite surprisingly, frequently confuse the question 'how does it work?' with 'what does it do?' But their most 'deleterious' mistake is the confusion of the internal computational (or parallel) processes taking place in the brain (which possibly causes mental states) with socially-based, everyday criteria of recognition and classification of, and knowledge about, the content of mental states.

In his *Remarks on the Philosophy of Psychology*, Wittgenstein explicitly stated that no assumption seems more natural to him than the assumption that association i.e. thinking is not any kind of brain process, so the understanding of mental processes on the basis of brain processes is impossible (Wittgenstein, *RPP*, I., par. 903). He also adds that the prejudice about the parallelism of the psychological and the physical is a result of our primitive interpretation of grammar. Because, if we allow that psychological phenomena follow a principle of causality that is not physiologically mediated, then we assume that there exists a kind of soul *alongside* the body, a ghostly mental nature (*RPP*, I., par. 906).

This is not to say that Wittgenstein suddenly turned into an advocate of dualism, i.e. the thesis about a spiritual substance that exists apart from the body and enables spiritual phenomena. In fact, he was simply stressing, perhaps in a slightly exaggerated manner, a profound difference between human thinking and all other human processes or states, either material or non-material. In Wittgenstein's view, it is not possible to speak of thinking as a private process; it can be viewed only in the context of people's social activities. In other words, thinking takes place only within the context of rule following, within various linguistic games and within the interplay of human actions. These relationships cannot be translated back into the processes taking place within individuals, or into individual behaviours.

The point here is that this is not solely a conceptual rift between the substrates of mental states presumably entirely explicable by computation, but a rift between how 'processors' work, meaning processors that constitute the physiological-physical basis of mental states, and *what and how we are* when experiencing these states and expressing them in our social practice. The 'how we are' is essentially connected with the fact that our experiencing and functioning has *meaning for us*, that we *feel our existence* precisely through experiencing and functioning. The difference between the two is not just conceptual, but it is a difference in categories, and we have not the slightest clue as to how this rift could be overcome.

The very notion of the working of a processor involves an ambiguity pertaining to description of the causal process, i.e. the sequence of the processor's states vs. the rules of processors' working. The former can be attained by listing the physical laws that regulate the physical flow of events and initial states of the processor at particular moments; to attain the latter, we could, for example, present the operation table featuring the processor's transitions from one state to another, depending on the input and the current state of the processor. Yet such a table would be a completely idealized creation, showing how the process *should* work in ideal circumstances rather than how it actually works. Yet whichever approach we adopt in determining the rules of processor's working, we have to keep in mind the difference between the idealized, *operational working* and the actual *flow* of events i.e., the actual sequence of the processor's physical states. The operational working of the processor is the kind of working we *ascribe* to it when we *interpret* it as the implementation of a certain rule or tabular description of operation.

Wittgenstein made in the *Remarks on the Philosopphy of Psychology* the interesting remark on the Turing 'project' of calculating machine: "Turing machines are humans who calculate" (*RPP* 1, par. 1096). It seems Wittgenstein didn't understand Turing at all because Turing project may try just the opposite, namely to interpret humans as calculating machines. Stuart Shanker made a long article on this topic where he tried to indicate some good points in the implicit Wittgenstein criticism of the Turing's project in 'mechanizing thinking'. Shanker means Wittgenstein was rejecting Turing's idea of completely non-conscious purely mechanical calculation because it wouldn't be a case of rule-following, and thus it wouldn't be a calculation. According to Shanker Wittgenstein allows 'mechanical' rule-following in the sense of non-reflecting rule-governed behaviour but not rule-following without any thought (Shanker, 1987). An agent's behaviour counts as rule-following if he offers a rule as the g*rounds* for his behaviour and that include at least the logical possibility of reflection. For Wittgenstein the capability of justifying (correcting, explaining, etc.) one's past action *by reference to the* rule is a constitutive aspect of rule-following. Turing machines or pure mechanical algorithms do not have this possibility. A machine can thus operate *in accord to a rule* but not *to follow it*. I don't enter into the difficult discussion on the interpreting Turing or Church but to emphasize some other aspects of Wittgenstein potential criticism of the classical computer model of mind.

I think the first sources of Wittgenstein's criticism of the 'mechanistic' concept of computation may implicitly lie even in *Tractatus*, namely in his criticism of the equalising of operation and predication, or operation and facts (events):

"A function cannot be its own argument, whereas operation can take one of its own results as its base" (*TLP*, 5.251).

In principle, we have to distinguish between operations and predicates (propositional functions), which represent some properties or relations in the world of facts. Operation represents formal relations, i.e. relations between forms and not relations between the individual representatives of forms.

Wittgenstein's claim that operations are not functions (predications) also means that operations are not any state of things or facts, but they are our methods of representing the relations between the forms of the representatives of the state of things. In this sense, computations, too, if they are true operations represented by the Turing machine's schemes, do not occur in nature but within the symbolic systems of representation.

In his later (and unpublished) philosophical considerations Wittgenstein drew attention to a similar distinction between the causal working of the machine and operational 'working' that occurs within the representation of a machine as a symbol. These reflections are especially relevant for discussion of the 'thinking machines,' i.e. any 'strong' program of artificial intelligence. Wittgenstein says that the machine can be conceptualised idealized, meaning by abstracting possible breakdowns, malfunctions etc. In such a case it seems "that a machine has (possesses) such-and-such possibilities of movement" (*P1*, par. 193. See also *RMF,* I. par. 125). In short, a machine-as-symbol is a special *expression of the rules of working*, i.e. working according to the rule, rather than actual working. In a similar way a machine for conjunction represented by means of a corresponding table of the transition of states, would be a symbolic representation of a working by rule rather than the factual working of some 'logical machine', for example, a simple processor that implements conjunction.

Obviously, we have again arrived at the principled difference between operations within some symbolic system (model) and actual processes in the world. Regardless of how perfectly a machine works, (for example, a machine that implements some logical operation), that still does not mean that that machine actually executes an operation, because even the most perfect machine can break down; its working may be interrupted, or it may be damaged, while in principle the execution of the operation excludes such a possibility. For Wittgenstein, these possibilities are categorically, grammatically excluded.

However, Wittgenstein denied the possibility that any expression of the rule, i.e. rule following, represents the ultimate or original interpretation or explanation of rule following. Actually, in so far as 'computational interpretation' is only a symbolic expression of the mapping of the system's physical states onto the algorithmic states of the computation system, it is not possible to speak of computation. Only the *execution* of this mapping, that is, actual following of an algorithm is computation. In this sense, computation is just an actual computational interpretation rather than, for example, an 'imagined' formal interpretation that only describes the algorithm of computation. The table of mappings or the table of transitions is not yet an operation but a symbolic expression. If we combine the terminology used in *Tractatus* and *Philosophical Investigations*, we could say that actual computation only manifest itself through the use of a specific algorithm in an environment such as is a social, publicly accessible practice (*PI*, par. 202), technique (*PI,* par. 199) or custom (*ibid.*) in which we ourselves must participate logically, while to *describe* it we have to use some formal symbolism.

For Wittgenstein, the rules, i.e. rule following as well as operations and execution of operations, are an inseparable component of the human form of life, i.e. "the whole hurly-burly (Gewimmel) of human actions." For this reason, rule following cannot be abstracted and observed separately from this background - for example, in such a way as to ask whether some machine or organism in itself follows the rule.

Since thinking implies the ability to follow rules, Wittgenstein argues that it is not sensible to ask whether, for example, the machine thinks. Or, to be more precise, such a question would imply a certain resemblance with the human being.

"Could a machine think? Could it be in pain? – Well, is the human body to be called such a machine? It surely comes as close as possible to being such a machine" (*PI*, par. 359).

"But a machine surely cannot think! Is that an empirical statement? No. We only say of a human being and what is like one that it thinks. We also say it of dolls and no doubt of spirits too. Look at the word 'to think', as a tool" (*PI*, par. 360).

Wittgenstein, therefore, does not reject completely the possibility that a machine could think, but only under the assumption of a certain resemblance with humans, if only a virtual one. But whence could such a resemblance arise? Of course, from our representation that the working of a machine is part of the human form of life, that it is a logical part of the net of human agency, for example, the net of human communication, cooperation among people in common projects, games etc.

REFERENCES:

Kripke, S. (1981): *Wittgenstein on Rules and Private Languag*e. Harvard Univ. Press, Cambridge MA

Shanker, S. G. (1987): Wittgenstein versus Turing on the Nature of Church's Thesis. *Notre Dame Journal of Formal Logic*, Vol. 28, Nr. 4: 615-649.

Ule, A.(1997): *Operationen und Regeln bei Wittgenstein.* P. Lang Verl., Frankfurt/M.

Williams, M. (1999): *Witggenstein, Mind and Meaning. Towards a Social Conception of Mind.* Routledge, London, New York.

Wittgenstein, L. (1972): *Remarks on the Foundaments of Mathematics (RMF)*, MIT Press, Cambridge/M

Wittgenstein, L (1974): *Tractatus logico-philosophicus (TLP).* Routledge & Kegan Paul, London

Wittgenstein, L. (1976): *Philosophical Investigations (PI).* B. Blackwell, Oxford

Wittgenstein, L. (1980): *Remarks on the Philosophy of Psychology (RPP). Vol. 1.* B. Blackwell, Oxford

Wittgenstein, L. (1981): *Zettel.* B. Blackwell, Oxford

PODROBNO PREISKOVANJE ALI STRATEŠKA USMERITEV?

Matjaž Gams
Odsek za inteligentne sisteme, Institut Jožef Stefan
Jamova 39, 1000 Ljubljana, Slovenija
Tel: +386 1 4773900; fax: +386 61 1251038
e-mail: matjaz.gams@ijs.si, http://ai.ijs.si/mezi/matjaz.html

Povzetek: V nekaj desetletjih raziskovanja smo se ukvarjali s preiskovalnimi algoritmi, še posebej patologijo in analizami, pri katerih pogojih se globje preiskovanje ne izplača. Analizirali smo le formalne igre kot šah ali igro 8 ploščic. Te analize kažejo zanimivo povezavo s kognitivnim odločanjem v realnem življenju – da je včasih bolj smotrno slediti strateškim načelom, čustvom ali navdihu kot podrobnemu premisleku – formalni analizi. Za realne primere analiziramo iskanje oaze v puščavi, razmišljanje pralovca pri lovljenju antilope in odločitev, koliko otrok imeti.

1 UVOD

V odseku za inteligentne sisteme na Institutu »Jožef Stefan« že desetletja raziskujemo metode in tehnike umetne inteligence in inteligentnih sistemov. Med pogostimi raziskovalnimi temami je tudi preiskovanje algoritmov, še posebej v zapletenih primerih, ko med preiskovanjem ne naletimo na obetavno sled. Takrat se pogosto zgodi, da preiskovanje v večjo globino da slabše rezultate, kot če bi se odločali na osnovi bolj plitvega preiskovanja.

To protislovje so poimenovali patologija. Pri tem so bile tako prve kot naslednje raziskave vezane na algoritem minimaks, ki se uporablja kot najosnovnejši algoritem za preiskovanje dreves iger dveh igralcev. V tem prispevku ne bomo podrobno ločevali med patologijo pozicije in patologijo poteze, ampak nas bo zanimalo vprašanje bolj splošno – **ali se bolj splača podrobno premisliti, ali pa odločiti na osnovi nekih »strateških« napotkov.**

V formalnih igrah se patologija ne pojavi, kadar minimaks preišče vse možnosti do konca, tj. kadar je globina preiskovanja globlja kot vsi konci iger, pa tudi pri pomembnih spremembah, recimo pri izgubi pomembne figure v šahu ne. Pač pa se pojavi v igrah, kjer se situacija bistveno ne spremeni v korist enega ali drugega igralca po številu potez, ki je manjše od globine preiskovanja. Primer bi bila neka izenačena pozicija, ki po nekaj potezah ostaja enako izenačena.

V teh razmerah je Beal (Beal, 1980) prvi opisal patologijo na algoritmu minimaks V naslednjih desetletjih je prišlo do vrste razlag, zakaj se pojavi patologija (Nau, 1983; Bratko & Gams, 1982; Pearl, 1983; Luštrek et al., 2005, 2006; Sadikov et al., 2005), vse te razlage pa so tipično uporabljale en sam mehanizem oz. parameter in so se ukvarjale predvsem s patologijo dveh igralcev.

V Odseku za inteligentne sisteme smo v zadnjih letih razvili modele patologije (Kaluža, 2008; Piltaver, 2008), ki so pokazali odvisnost treh ključnih parametrov, tj. razvejitve b, števila vrednosti hevristične funkcije g in podobnosti s. Ti modeli so pokazali, da je pri velikem b, malem s in malem g zelo velika verjetnost patološkosti. Poleg tega so pokazali patološkost na treh realnih igrah: 8 ploščic, šahovske končnice, kalah, torej igrah enega in dveh igralcev. Letos je v objavo sprejet še naš drugi prispevek na to temo v najprestižnejši reviji s tega področja Artificial Intelligence (Nau et al., 2011). V tem prispevku smo preiskovali faktorje, ki vplivajo na uspešnost preiskovanja. Pokazali smo, da so pogoji uspešnosti v igrah dveh igralcev zelo podobnih igram enih igralcev. Pri premišljanju o formalnih igrah se je utrnila drzna misel, da je to morda podobno kognitivnemu razmišljanju pri realnih problemih.

Raziskovalne hipoteza: Iz formalnih analiz iger lahko sklepamo, kako optimalno razmišljati pri zahtevnih problemih realnega sveta – kdaj podrobno analizirati in kdaj se odločiti »strateško«.

Za primere iz realnosti smo izbrali iskanje oaze v puščavi, lovljenje antilope in določanje števila otrok. Tretji problem je verjetno eden najzahtevnejših in najpomembnejših vprašanj v življenju vsakega človeka. Torej - ali se je o številu otrok bolje odločati formalno ali strateško, na osnovi podrobne analize ali na osnovi načel oz. strateških vrednot.

2 PRIMER ENEGA IN DVEH IGRALCEV

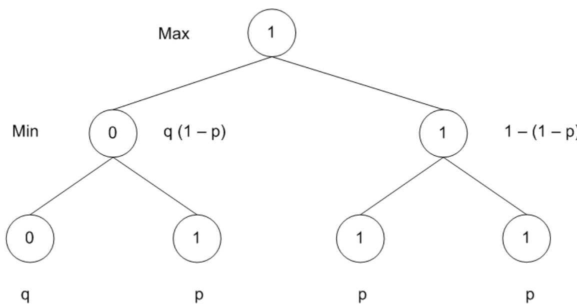
V tem razdelku bomo najprej prikazali dva preprosta primera patologije enega in dveh igralcev. Na slikah 1 (dva igralca) in 2 (eden igralec) sta dva primera binarnih dreves z globino preiskovanja 2. Korenina drevesa predstavlja osnovno pozicijo, v kateri imamo na izbiro levo ali desno poddrevo. Prave vrednosti vozlišč so navedene v vozlišču in so lahko le 0 (izgubljena) ali 1 (dobljena). Drevo na sliki 1 prikazuje drevo igre, kjer na vsako našo potezo odgovori nasprotni igralec. Nasprotnik nam skuša odvzeti zmago oz. jo spremeniti v naš poraz. Na sliki 2 pa je drevo, kjer vse poteze vlečemo sami.

Realna primera omenjenih dreves bi lahko bila: Za igro enega igralca bi lahko dejali, da izbira pot v puščavi proti oazi. Lahko se odloči glede na videz bližnjih sipin, lahko pa vzame daljnogled in se odloči glede na izgled daljnih sipin. Realni primer dveh igralcev je lahko takle: Lovec lovi antilopo in jo želi ujeti, kar zanj predstavlja 1. Antilopa lahko zbeži v travo ali v reko, vendar jo v reki lovec vedno ujame. V travi pa lahko steče v hrib ali v dolino in v hribu jo čaka drugi lovec, medtem ko po dolini uide. Torej jo lovec ulovi v treh od štirih primerov.

Vsi dogodki pa so verjetnostni in pračlovek lahko vnaprej samo z določeno verjetnostjo oceni situacijo, pa tudi ogled sipin samo z določeno verjetnostjo nakazuje, kje bi bila voda, npr. z nekaj več travnimi bilkami. V teh osnovnih modelih je vedno predpostavljeno, da je verjetnost, da bo 0 narobe ocenjena s »q« in verjetnost, da bo 1 narobe ocenjena s »p«. Pogosto predpostavimo še, da sta p in q enaki, včasih celo, da sta obe 0.1.

Formalna modela ustrezata modelom iger – minimin oziroma maxmax sta igri, kjer igralec na igri skuša doseči najmanjšo ali največjo možno vrednost in igra sam. Kadar igrata dva igralca, eden skuša minimizirati in drugi maksimizirati, od tod izraz minimax. Minimin se uporablja za preiskovanje dreves ALI, medtem ko se minimax uporablja za preiskovanje dreves IN/ALI, kjer se vozlišča zaporedoma izmenjujejo.

Sedaj lahko preračunamo verjetnosti v drevesih. Če se na nivoju 2, tj. najnižjem nivoju, saj označujemo korenino z nivojem 0, lahko zmotimo z verjetnostjo p ali q odvisno od vrednosti, potem so vrednosti na nivoju 1 označene desno od vozlišča. Za enega igralca sta možnosti $q(1\text{-}p)$ za levo vozlišče in za desno vozlišče $1\text{-}(1\text{-}p)^2$. Izračun je preprost: Da se zmotimo in levo vozlišče, ki ima pravo vrednost 0, ocenimo kot 1, se moramo zmotiti v levem nasledniku (q), medtem ko moramo desno podvozlišče pravilno oceniti $(1\text{-}p)$. Skupna verjetnost je produkt.

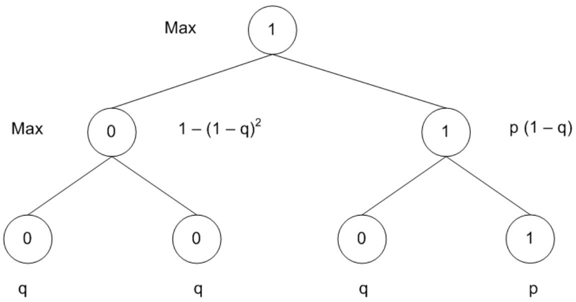


Slika 1: Binarno drevo dveh »igralcev«, kjer je preiskovanje v globino 2 slabše kot v globino 1. Prikazuje dilemo pračloveka, ki lovi antilopo.

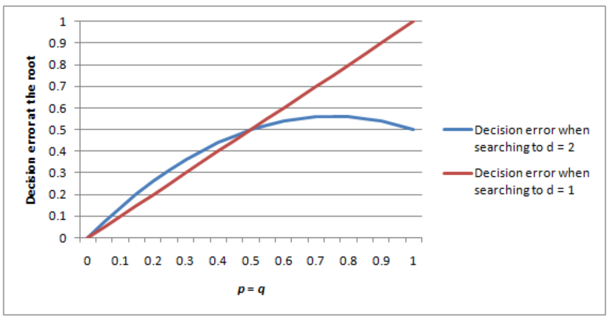
V desnem vozlišču prvega nivoja lahko naredimo napako le, kadar se ne zmotimo v oceni obeh vozlišč hkrati, od tod $1\text{-}(1\text{-}p)^2$. Ko izračunamo oceno napake v korenini ob dodatni predpostavki, da je $p=q$, dobimo $p(3–2\ p)/2$. To je torej ocena napake v korenini binarnega drevesa, kjer igra le en igralec.

Pri ocenjevanju verjetnosti napake drevesa, kjer igrata dva igralca, se izračun nekoliko spremeni, zato dobimo vrednosti napake, kot je označena desno pri vozliščih srednjega nivoja. Izračun napake v korenini pri pogoju $p=q$ da enak izraz kot pri prejšnjem drevesu: $p(3–2\ p)/2$.

Ker imamo sedaj le en parameter, to je $p=q$, lahko narišemo verjetnost uspešnega preiskovanja, oziroma verjetnost napake $p(3–2\ p)/2$, v globini 2, v primerjavi z verjetnostno napako v globini 1, tj. p. Slika 3 (Dacar, 2009) očitno pokaže, da se preiskovanje v bolj splača v globini 1 in manj v globini 2 za vse pričakovane vrednosti napak pri oceni pozicije, od 0 do 0,5.



Slika 2: Binarno drevo enega igralca, kjer je preiskovanje v globino 2 pogosto slabše kot v globino 1. Prikazuje dilemo izgubljenega v puščavi.



Slika 3: Primerjava izraza $p(3–2p)/2$, tj. vrednost napake pri preiskovanju v globino dva v primerjavi z globino 1, kjer je $p=q$. Obe drevesi na slikah 1 in 2 sta torej patološki za večino smotrnih vrednosti p.

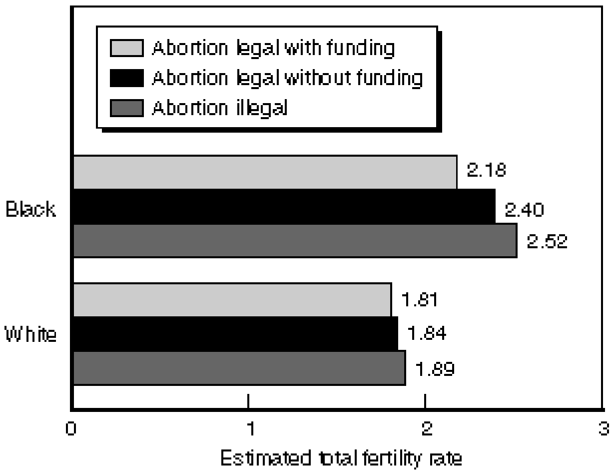
Za omenjena dva primera naredimo razlago za primer puščave in lovljenja antilope. Če pogledamo iz tekoče sipine do dveh najbližjih in se odločimo na osnovi teh dveh ocen, bomo naredili manjšo napako, kot če gremo pogledati do naslednjih dveh. In če pralovec razmišlja samo do prvega nivoja, bo bolj verjetno izbral pravo odločitev. Torej se preiskovanje v globino v obeh realnih primerih ne izplača!

Predpostavke formalnega modela se zdijo smotrne v realnem življenju, tj. da ocenjujemo le z uspehom (1) ali neuspehom (2) in da se ocena s povečano globino ne izboljšuje. Seveda pogosto preiščemo možnosti do konca ali vsaj do nekih končnih rešitev, kar je pogost primer v realnih igrah, samo prvotna predpostavka je bila, da imamo opravka z zapleteno situacijo, kjer se po nekaj potezah rešitev ne izkristalizira.

V (Nau et al., 2011) je pokazano, da se tudi za vse možne kombinacije do globine 5 v takih razmerah ne izplača preiskovati v večjo globino, ob tem da analiza uvaja še nove parametre. V (Dacar, 2009) pa je pokazano, da se za vse primere pri teh pogojih na izplača preiskovati v globino ne glede na obliko binarnega drevesa, oz. porazdelitev vrednosti na drugem nivoju.

3 KOLIKO OTROK?

Pri analizi števila otrok se zdi, da je odločitev globalno gledano patološka ne samo v nekaterih državah, ampak kar pri vseh belcih, ki imajo v povprečju premalo otrok, kar dolgoročno vodi do izumrtja (Gams & Krivec, 2011). Še posebej to velja za države kot Slovenijo, ki je z rodnostjo okoli 1.5 v zadnjih desetletjih med tistimi z najslabšimi projekcijami.



Slika 4: Koliko se poveča rodnost v primeru prepovedi splava? So odločitve o številu otrok patološke?

Formalni modeli v prvih sekcijah tega prispevka so zajemali preprosta drevesa. Za odločanje v primeru otrok bi sicer lahko zgradili drevesa, ki bi prikazovala, s kakšnimi ukrepi (npr. prepovedjo splava ali kontracepcije) bi lahko dosegli dolgoročno zdržnost, vendar za nivo cele države. Zdi se, da je razlika med temi formalnimi modeli in dejanskim kognitivnem odločanjem posameznika prevelika, saj se nihče ne odloča na tak način. Žal torej ne znamo zgraditi dreves, ki bi vsaj približno kazala razmišljanje ljudi pri odločanju o otrocih.

Pa vendar – nekaj mora biti narobe. Ko so se ljudje odločali na osnovi prevladujoče ideologije, npr. cerkvene, in niso premišljali v globino, so se odločali za večje število otrok, torej nepatološko. Celo sedaj imajo belke želje o številu otrok, ki omogoča dolgoročno zdržnost, realno število otrok na osnovi podrobnega premisleka pa je premajhno. Določene asociacije torej kažejo na analogijo s patologijo pri igrah, vendar je nujno še enkrat poudariti, da je ogromna razlika med formalnimi igrami in sklepanjem o otrocih.

4 ZAKLJUČEK

Formalni modeli kažejo, da je patologija možna v vsaki domeni, če le med preiskovanjem ne naletimo na rešitev in torej situacija ostane nejasna. Preproste primere realnega sveta lahko stresemo iz rokava, npr. iskanje oaze v puščavi ali razmišljanje pralovca, kako bi ujel antilopo. Tudi pri zelo zapletenem odločanju o številu otrok se zdi, da patološkost izvira iz preračunavanja in ne odločanja na osnovi strateških načel. V isti sapi pa je potrebno napisati, da je taka življenjska odločitev oz. odločanje zelo verjetno precej drugačna od formalnih modelov.

Vseeno se zdi povsem možno, da se ljudje v razvitem svetu odločamo na osnovi premišljanja v nekih formalnih shemah, ker menimo, da je tako preiskovanje najboljše. S tem prispevkom smo pokazali, da **je mogoče trditi, da je v določenih situacijah realnega sveta boljše odločanje glede na strategijo oz. čustva kot pa na osnovi podrobnega razčlenjevanja**.

Nekoliko bolj nepreverjena se zdi hipoteza, da so bili praljudje uspešnejši z ideologijami, verovanjem, pa celo praznoverjem, saj tedaj še niso znali dobro formalno razčlenjevati. Danes znamo podrobno sklepati, zato je npr. praznoverje slabše kot logični premislek, še vedno pa je v določenih razmerah pretirano logično razmišljanje patološko. Ko bomo znali dobbro ugotoviti, kdaj se splača razmišljati in kdaj strateško odločiti, bomo naredili še en civilizacijski korak naprej.

Zahvala:

Zahvaljujemo se ARRS za financiranje raziskovalne programske skupine Umetna inteligenca in inteligentni sistemi. Zahvaljujemo se naslednjim sodelavcem za sodelovanje pri raziskavah: Mitja Luštrek, Boštjan Kaluža, Rok Piltaver, Aleš Tavčar, Aleksander Sadikov, Matej Guid, Ivan Bratko, Dana Nau, Jana Krivec.

Literatura:

Beal DF. (1980) An Analysis of Minimax. In Advances in Computer Chess 2, ed Clarke MRB (Edinburgh University Press, Edinburgh), pp 103–109.

Bratko I., Gams M. (1982) Error Analysis of the Minimax Principle. In Advances in Computer Chess 3, ed Clarke MRB (Pergamon Press, Oxford), pp 1–15.

Dacar F. (2009), interno poročilo, dostopno pri avtorju.

Gams M., Krivec J. (2010) Is Fertility Rate Of The Country Related To The System Of Values?, Zbornik konference Informacijska družba.

Kaluža B. (2008) Analysis of pathological minimax models and Pearl's game (Slovene: Analiza patoloških modelov minimaksa in Pearlove igre). BSc thesis, University of Ljubljana.

Luštrek M., Bratko I., Gams M. (2005) Why minimax works: An alternative explanation. In Proc IJCAI, pp. 212–217.

Luštrek M., Gams M., Bratko I. (2006) Is real-valued minimax pathological. Artificial Intelligence 170(6–7): 620–642.

Nau D. (1982) An investigation of the causes of pathology in games. Artificial Intelligence 19(3): 257–278.

Nau D., Luštrek M., Parker A., Gams M. (2011) When Is it Better Not To Look Ahead?, Artificial Intelligence, accepted for publication.

Pearl J. (1983) On the nature of pathology in game searching. Artificial Intelligence 20(4): 472–453.

Piltaver R. (2008) Search pathology in eight-puzzle (Slovene: Patologija preiskovanja v igri osmih ploščic). BSc thesis, University of Ljubljana.

Sadikov A., Bratko I., Kononenko I. (2005) Bias and pathology in minimax search. Theoretical Computer Science 349(2): 261–281.

DUŠEVNOST KOT EMERGENCA IN SEMIOZA TER NEKATERE IMPLIKACIJE

Matej Plankar
Inštitut BION
Stegne 21, 1000 Ljubljana
Tel: 040 432 298; email: matej.plankar@bion.si

IZVLEČEK

Predstavljena sta dva pristopa, ki na dinamičen in holističen način obravnavata življenje in posledično nudita tudi sistemska orodja za obravnavo kognitivnega aparata kot emergentne in obenem semiotske strukture, ter nekatere možne aplikacije na kognitivne fenomene.

1 UVOD

Namen prispevka je predstaviti dve medsebojno dopolnjujoči se smeri za sistemsko obravnavo kognitivnega aparata kot dinamične in holistične strukture. Kot kognitivni aparat (KA) smatram celovit splet možgansko-duševnega kompleksa na materialni (fizikalni) in virtualni (fenomenološki) ravni, torej v najširšem možnem smislu. Prednost obeh pristopov oz. njune kombinacije je predvsem široko konceptualno ogrodje, ki ni omejeno zgolj na kognitivne procese, pač pa obravnava kognicijo kot eno (naj)višjih hierarhičnih ravni (samo)organizacije materije in na ta način ponuja sistemska interdisciplinarna orodja za razumevanje kognitivnih procesov.

2 DUŠEVNOST KOT EMERGENCA

Teorija kompleksnih sistemov in njene izpeljanke (nelinearna termodinamika, teorija kaosa, kibernetika drugega reda, sinergetika, sistemska biologija (organicizem) itd.) je univerzalno orodje za obravnavo kompleksnih pojavov v naravi, katerih osrednja značilnost je samoorganizacija. Koncept samoorganizacije predstavlja temeljni mehanizem kreativnosti v naravi, ki omogoča spontano porajanje novih oblik in funkcij ter evolucijo žive in nežive narave. Na ta način omogoča načelno nedeterminiranost (odprtost) in s tem historičnost sveta, saj v determiniranem svetu ne bi bilo prostora za nikakršno izbiro in posledično tudi svobodno voljo (Bruni 2008). Samoorganizacija je univerzalen pojav, kar pomeni, da je značilna tako za materialne (fizikalni, biološki) kot tudi virtualne (npr. umetna inteligenca) sisteme (Heudin 2006). Med slednje gotovo sodi tudi kognicija kot sploh eden najkompleksnejših pojavov v naravi (Varela 2006).

Samoorganizacije so zmožni sistemi z mnogo elementarnimi komponentami, ki imajo sposobnost lokalnih medsebojnih interakcij in so podvrženi stalnemu metastabilnemu toku energije, ki te interakcije omogoča. Slednje ne pomeni samo stalen dotok energije v sistem, pač pa tudi njeno disipacijo (razpršitev) v zunanje okolje. Zato v materialnih sistemih tovrstnim strukturam pravimo tudi disipativne (Nicolis in Prigogine 1977). Tok, ki omogoča samoorganizacijo elementov sistema, je v disipativnih strukturah v obliki fizikalne energije in vzdržuje sistem v termodinamičnem neravnovesju. Disipativne strukture si fizikalno gledano prizadevajo čim hitreje izničiti energijske gradiente, ki jih poganjajo, torej doseči termodinamično ravnovesje (s tem težijo k lastnemu propadu). Pri tem opravljajo fizikalno delo, katerega velikost je primarno odvisna od velikosti preprek v energijskem toku. Učinkovitost opravljenega dela pa je sorazmerna redukciji prostostnih stopenj sistema, ali drugače, bolj kot je sistem notranje uglašen oz. dinamika njegovih elementov sinhrona (ubrana, koherentna, kolektivna), več dela lahko opravi (Kauffman 2010). Funkcionalna organizacija sistema oz. koherenca je tako ena najpomembnejših posledic samoorganizacije, ki bistveno vpliva na delovanje sistema in je zelo pomembna tudi za kognitivne procese (sinhrone oscilacije nevronskih skupnosti, kvantni modeli zavesti, dinamika odločanja itd.).

Strukturam višjega reda, ki nastanejo kot posledica samoorganizacije, pravimo emergentne (emergirati = vzrasti, vzniknit). Ena ključnih lastnosti emergentnih struktur je načelna ireducibilnost struktur višjega reda na strukture nižjega reda in obratno, nededucibilnost struktur višjega reda iz struktur nižga reda (Heudin 2006). Višja kot je raven organizacije, manjša je njena neposredna odvisnost od struktur nižjega reda, od tod tudi srčika biološkega holizma o avtonomiji organizmov, na kar je med prvimi opozarjal že Aristotel in kasneje med drugimi Kant (Van de Vijver 2006).

Emergentne lastnosti lahko obstajajo na mnogih organizacijskih ravneh in bolj kot je sistem kompleksen, bogatejšo hierahično organizacijo lahko ima. Načeloma lahko sistem kreira nove strukture oz. organizacijske ravni, dokler ima na razpolago zunanje prostostne stopnje, torej dokler je informacijsko odprt in stalno dobiva nove informacije; če se informacijski tok prekine, postane determiniran. V tem smislu šum oz. izvorna kaotičnost sveta ni uničevalec, pač pa primarni vir novih informacij (Atlan in Cohen 2006, Goujon 2006).

Klasičen primer nežive disipativne emergentne strukture je vrtinec. Kemijska zgradba vrtinca se v ničemer ne razlikuje od svojih sestavnih elementov (voda, zrak…), pa vendar kot celota predstavlja strukturo višjega reda. Vrtinec obstaja, dokler ne uplakne tok energije, ki ga poganja, v obliki tlačnih gradientov. Molekule v vrtincu imajo manj prostostnih stopenj kot če bi se gibale kaotično, kar pomeni, da je njihova dinamika organizirana (koherentna), zato lahko opravlja fizikalno delo. To analogijo lahko dalje razširimo na sklopljene verige metabolnih ciklov v bioloških sistemih, ki delujejo kot hierarhične mreže avtokatalitičnih reakcij (vsaj del produktov se vrača v cikle kot reaktanti), le da je energija, ki omogoča samoorganizacijo, metabolična oz. elektrokemična (Igamberdiev 1999).

Podobno lahko smatramo kognitivni aparat kot funkcionalen sklop nadaljnih ravni samoorganizacije. Materialna (fizična) komponenta KA se na biokemijski oz. celični ravni ne razlikuje bistveno od ostalih sistemov, medtem ko ima raven medcelične komunikacije oz. nevronskih skupnosti že nekatere lastne značilnosti. Doživljanje kot virtualno oz. fenomenološko komponento KA lahko smatramo kot najvišji nivo samoorganizacije, ki presega neposredno vezanost na materialno komponento in je zato do določene mere avtonomno oz. neodvisno od specifičnosti posameznih elementov, iz katerih vzrašča (npr. atomov, molekul, celic). Obenem pa lahko po načelu krožne vzročnosti[1] navzdol usmerja njihovo dinamiko (Heudin 2006, Scott 2006).

Tovrsten pristop ima nekaj pomembnih implikacij za obravnavo kognitivnih procesov. Prvič, ponuja fizikalistični temelj za priznanje avtonomnosti duševnosti kot emergentne strukture (naj)višjega reda in s tem realnost svobodne volje (glej tudi razdelek 3). Drugič, doživljanje lahko obstaja le kot kontinuirani dinamični tok in torej ne more biti statično, zato ga je potrebno tudi obravnavati z dinamičnimi koncepti. Fizikalna struktura duševnosti domnevno obstaja kot nepredstavljivo kompleksen splet vzorcev-atraktorjev in njihovih trajektorij v faznem prostoru možganskega kompleksa, toda vprašanje je, ali kot taka že implicira konkretno fenomenalno (doživljajske) vsebino. Ta problem slikovito ponazarja Strawson: duševno vsebino lahko primerjamo z zapisom simfonije na plošči; le-to lahko namesto na predvajalnik glasbe vežemo npr. na napravo za svetlobne učinke, ki verno razpošilja svetlobne žarke v prostor glede na zapis na plošči. Plošča sama po sebi torej ni prav nič bolj glasbena kot je svetlobna, ali kaj tretjega. Njena eksplicitna fenomenalna kakovost se pokaže šele ob *reprezentaciji* njene vsebine; ravno tako implicitna duševna vsebina postane doživljajska oz. izkustvena šele ob reperzentaciji skozi zavest (Strawson 1994, prirejeno).

[1] Krožna vzročnost (cirkularna kavzalnost) pomeni paralelno in kontinuirano medsebojno součinkovanje oz. sodoločanje različnih organizacijskih ravni (Heudin 2006).

Sledeč tej interpretaciji se torej zdi, da za (fenomenalno) izkustvo oz. doživljanje, ki je dejansko vse, kar lahko vem o svetu in onstran česar ne morem ničesar spoznati, čutiti ali kako drugače vedeti, potrebujemo vsaj dve temeljno različni substanci – zavest ter implicitno duševno vsebino, ki predstavlja fizikalno strukturo duševnosti oz. fizičen "zapis na plošči", in ga nadalje enačim s sebstvom. Dodaten argument v prid tej predpostavki je, da nekateri avtorji (npr. Varela 2006, Strawson 1994) menijo, da je tovrstno implicitno polje v svoji doživljajski eksplikaciji primarno enotno, torej gre v osnovi za eno samo substanco, pa najsi gre za doživljanje čustev, racionalnih miselnih procesov, senzoričnih reprezentacij itd.; slednje delitve doživljanja so torej šele sekundarne. Na drugi strani pa zavest kot taka nima nikakršne notranje organizacije in/ali vsebine in je torej sama po sebi brezčasna, brezprostorska in brezpomenska, zato jo je tudi izredno težko kvalitativno opredeliti s kakršnimikoli zunanjimi koncepti izven nje same. Zavest je preprosto neskončna rekurzivna zanka zavedanja (zavedanje zavedanja zavedanja...), torej ni stvar ali objekt, pač pa process, oziroma čista izvorna dinamika, ki potrebuje objekt. V kontekstu teorije kompleksnih sistemov bi torej zavest bila analogna dinamičnemu toku, ki omogoča samoorganizacijo primarnih fenomenalnih reprezentacij kot osnovnih elementov duševnosti (kvalij[2]) v strukture oz. kategorizacije vedno višjega reda tekom duševnega razvoja subjekta.

Šele interakcija zavesti s tovrstnim holističnim poljem potencialnih objektov (=sebstvo) omogoča njihovo interpretacijo in s tem temeljno ločljivost (diskriminabilnost) ter organizacijo pojavnega sveta: zavesten subjekt doživlja različne dražljaje iz okolice, morda tudi sebe kot od okolice ločene avtonomne celote in morda celo lastne miselne procese, nezavesten subjekt pa ne doživlja ničesar od tega. Tovrstno eksplicitno doživljajsko polje, v katerega je subjekt neobhodno vržen, zato zahteva izredno kompleksno organizacijo kot avtonomna, mnogodimenzionalna, domnevno holografsko organizirana virtualna emergentna struktura, lahko bi ji rekli kar fenomenalni hiperprostor (FH). V tem smislu je pozornost (intencionalnost) pravzaprav uperjenost zavesti proti specifičnemu segmentu FH; zavest kontinuirano skenira sebstvo (v fizikalnem smislu bi rekli, da preletava energijsko-konfiguracijski prostor, kar imenujemo ergodičnost), in manifestira subjektivno doživljanje kot njuno dinamično interakcijo, ki je pravzaprav sama kognicija. Raznovrstne načine pozornostne dinamike je analiziral Arvidson (2006) v svoji teoriji sfere pozornosti.

3 DUŠEVNOST KOT SEMIOZA

Naslednji pristop k sistemski obravnavi kognitivnega aparata predstavlja biosemiotika kot relativno mlada znanstvena disciplina, ki pa ima globoke temelje tako v filozofiji kot v eksaktnih znanostih.

[2] Kvalije povzemam po Sanguineti (2007) kot osnovne enote subjektivnosti oz. kar kot »fenomenalne kvante«.

Biosemiotika ni poddisciplina klasične (Piercove) semiotike, pač pa njena temeljita multidisciplinarna nadgradnja. Charles S. Pierce je v 19.stol. osnoval triado osnovnih kategorij, prvotno kot osnovno orodje svoje teorije logike: znak, objekt in interpretant. Znak simbolizira nek objekt kot svoj pomen, vendar šele skozi interpretacijo. Objekt znaka ♪ je pomen kot ga interpretira interpretant, npr. glasba ali pa ton določene dolžine. Pomen ni nikoli povsem determiniran, saj je interpretacija nujno vsaj deloma odvisna od konteksta. Semioza je dinamični proces potencialno neskončne zanke interpretiranja znakov, kodiranja pomenov v nove znake, ponovne interpretacije itn. Brez interpretanta torej ni znaka in obratno, brez znaka ni interpretanta, zato se semiotska zanka nujno vzpostavi šele kot triada, ali pa ne moremo govoriti o semiozi (Jämsä 2008).

Po drugi liniji je Jakob von Uexküll, sicer eden prvih sistemskih biologov, ne da bi se tega eksplicitno zavedal, razvijal svoj sistem semiotike, uporaben za študij vedenja organizmov. Koncept umwelta je ena njegovih najpomembnejših idej in ga lahko opredelimo kot individualni nabor vseh znakov, ki jih lahko biološki subjekt interpretira, torej kot subjektivno znakovno polje. Thomas Sebeok je kasneje združil oba pristopa in utemeljil biosemiotiko kot samostojno znanstveno in filozofsko disciplino. Danes obstaja v biosemiotiki več struj mišljenja, izmed katerih sta najpomembnejši t.i. klasična "kopenhagenska šola" (Hoffmeyer, Emmeche, Kull idr.), ki je izrazito organicistično usmerjena in poudarja vlogo emergence in samoorganizacije v semiozi, in teorija organskih kodov Marcella Barbieria, ki je za mnoge raziskovalce revolucionarna (Favareau 2008).

Biosemiotika enotno pojmuje življenje kot semiozo, torej je prav semioza tisto, kar ločuje živo od neživega. Organizme vidi kot aktivne interpretatorje lastnega znakovnega polja (umwelta), ki na višjih interpretativnih ravneh postajajo vse bolj neodvisni od specifičnih fizikalno-kemijskih dejavnikov, torej avtonomni. Fizikalizem nasprotno vidi materialni svet in pojave v njem, vključno z organizmi, izključno kot mehanično diadno prepletanje vzrokov in posledic, ki jih posredujejo mehanske sile (4 osnovne fizikalne sile). Toda fizikalni zakoni sami po sebi so le potencialni in potrebujejo začetne in mejne pogoje, da se lahko izrazijo, ti pa so lahko le v obliki specifične, kvalitativne informacije, torej v Batesonovem (in ne Shannonovem) smislu. Fizikalni zakoni torej ne morejo determinirati življenja, pač pa ga zgolj dovoljujejo in v tem smislu deskriptorna moč biosemiotike seže onstran fizikalizma (Pattee 2008).

Barbieri (2003) v svoji knjigi *Organic Codes* eksaktno opiše, kako v evoluciji nastane semiotska zanka (s čemer obenem definira življenje), a je natančna razlaga izven konteksta te obravnave. Ključen koncept v njegovi teoriji je organski kod, determiniran nabor pravil, ki določajo, kako se specifične informacije (znaki!) nekega sistema preslikajo v nek drug sistem. V fizikalnem jeziku bi rekli, da kod definira relacije med dvema neodvisnima faznima prostoroma. Ključno pri tem je, da je kod poljuben, torej da je določen arbitrarno, kar zagotavlja načelno neodvisnost specifičnih relacij od fizikalnih dejavnikov in torej nerelevantnost specifične fizikalno-kemijske narave znaka za njegovo interpretacijo. Primer splošno uveljavljenega koda je genetski kod kot urejen nabor pravil, ki določajo relacijo med nukleotidnimi tripleti (kodoni) in aminokislinami (optimizacija koda je v tem primeru sekundarna, saj je selekcijsko pogojena). Določena aminokislina je torej pomen oz. interpretacija specifičnega kodona ali širše, fenotip kot organizem je interpretacija genotipa kot nabora kvalitativnih informacij v obliki nukleotidnega zaporedja. Toda v biologiji obstaja poleg genetskega še cela vrsta drugih kodov (ustrezajoč Barbierijevim pogojem), le da jih ne obravnavamo kot take – primeri so npr. histonski, spajalno-izrezovalni, signalno-transdukcijski, adhezijski, citoskeletni, lektinski kod itd.

Sistemski pomen organskega koda je v tem, da predstavlja most, ki povezuje dva različna svetova in na ta način odpira nove interpretativne ravni. Signalno-transkukcijski kod npr. kodira primarne sporočevalce (hormoni, nevrotransmiterji itd.) v sekundarne in tako omogoča urejeno medcelično komunikacijo, adhezijski kod pa določa pravila za prepoznavanje molekul na zunanji celični površini; oba koda skupaj sta omogočila prehod iz enoceličnega na mnogocelično življenje. V širšem smislu je kod pravzaprav korespondenčni sistem, ki je lahko bolj ali manj determiniran. Na višjih interpretativnih ravneh postajajo korespondenčni sistemi vse manj determinirani (torej kodi v ožjem pomenu) in posledično vse bolj podvrženi kontekstualno odvisni interpretaciji. Z biosemiotskega zornega kota bi torej življenje lahko opredelili v smeri konkretne materializacije semioze, ki na nek način »jezdi« na samoorganizacijskih lastnostih kompleksnih dinamičnih sistemov, tako da aktivno usmerja njihov razvoj, pri čemer biološki interpretant kot emergentni sistem sočasno deluje na mnogih hierarhičnih ravneh, ki se na najvišji ravni zlivajo v sam biološki subjekt (Bruni 2008). Na nižjih ravneh je interpretacija bolj determinirana, kodificirana, zaprta zaradi osnovnih generičnih relacij, ki sploh omogočajo osnovne življenjske funkcije, na višjih pa bolj odprta, interpretabilna, kontekstualno odvisna (npr. jezik). Biološki subjekt je tako inherentno uperjen navzven, izven samega sebe in torej izven lastne subjektivnosti, saj ima le na ta način lahko smoter – s tem, da stalno vzdržuje informacijski pretok in na ta način presega lastno determiniranost (razdelek 2).

V smislu zgoraj navedenih izhodišč lahko sebstvo kot individualno-specifično fizikalno strukturo duševnosti v luči biosemiotike razumemo kot primarni duševnostni kod, ki skozi zavest interpretira lastno znakovno polje (umwelt) kot fenomenalni hiperprostor. V tem smislu je vsakršno doživljanje a priori že interpretacija, sebstvo pa "most" med objektivnim in spoznavnim svetom.

Sebstvo tekom ontogenetskega razvoja stalno raste in se diferencira in analogno se FH kategorizira v vedno bolj abstraktne in hierarhično razvejane objekte. Že Pierce je opozarjal, da je vsaka misel oz. percept pravzaprav znak (od tod njegov pojem thought-sign) in kognicija potencialno neskončna semiotska zanka (infinite semiosis) (Vekhavaara 2008). Posledično postane nujno, da se oblikujejo novi – izključno fenomenološki, lahko bi jim rekli sekundarni – korespondenčni sistemi, ki usmerjajo kompleksno vedenjsko polje subjekta, kot ga omogoča zlasti človeški kognitivni aparat. Nadalje jih lahko delimo na intra- in intersubjektivne (slednji so npr. jezik, razni socialni "kodi" kot je subkulturni, itd.).

Do tu sem se namerno izognil vlogi afektivnosti kot "ključne gonilne sile človkovega vedenja od njegovih najzgodnejših časov, ki vključuje tako notranje miselne procese kot navzven usmerjena dejanja" (Sanguineti 2007). Kot afektivnost smatram nabor vseh stanj, ki se se skozi zavestno interpretacijo in v specifičnem izkustvenem kontekstu[3] reprezentirajo kot emocije, afekti, razpoloženja in druga sorodna stanja, torej kot čustvovanje v najširšem pomenu. Afektivnost v osnovi razumem kot mehanizem, ki usmerja vedenjsko polje v situacije, ki jih subjekt čuti kot ugodne, in igra zlasti pomembno vlogo v primerih, ko racionalno mišljenje ne zadošča za dovolj precizno odločanje. Na tem mestu zaradi omejenosti zgolj predpostavljam možnost semiotske obravnave afektivnosti. Sanguineti (2007) pojmuje vrednostni sistem (value system) kot integralni del sebstva, ki določa osnovne vzorce čustvovanja. Menim, da je del vrednostnega sistema že vgrajen v strukturo duševnosti kot del kolektivne arhetipske zapuščine (v obliki nagonskih in refleksnih vedenj), večji del pa si subjekt sam ustvarja skozi intersubjektivnost. Če je vrednostni sistem del sebstva, sledi, da je del primarnega duševnostnega koda, medtem ko je afektivnost oz. čustvovanje, analogno, "del" FH kot njegova izkustvena in kontekstualno odvisna interpretacija. Vrednostni sistem torej lahko smatramo kot nekakšen aktivni, interpretabilni, nedeterminirani, semiotsko odprti in nikoli dokončani "rob" oziroma "gradbišče" sebstva, kjer se stalno ustvarjajo nove kategorizacije, pravzaprav kot jedro subjektivnosti.

4 LITERATURA

[1] Arvidson P.S. 2006. The Sphere of Attention. Springer, Dordrecht.

[2] Atlan H. in Cohen I. 2006 (121-139). Self-Organization and Meaning in Immunology. V: Feltz B., Crommelinck M. in Goujon P. (ur.) 2006. Self-Organization and Emergence in Life Sciences. Springer, Dordrecht, The Netherlands.

[3] Barbieri M. 2003. The Organic Codes: An Introduction to Semantic Biology. Cambridge University Press.

[4] Bruni L. E. 2008 (365-407). Cellular Semiotics and Signal Transduction. V: Barbieri M. (ur.) 2008. Introduction to Biosemiotics: The New Biological Synthesis. Springer, Dordrecht, The Netherlands.

[5] Favareau D. 2008 (1-67). The Evolutionary History of Biosemiotics. V: Barbieri M. (ur.) 2008. Introduction to Biosemiotics: The New Biological Synthesis. Springer, Dordrecht, The Netherlands.

[6] Heudin J.-C. 2006 (227-247). V: Feltz B., Crommelinck M. in Goujon P. (ur.) 2006. Self-Organization and Emergence in Life Sciences. Springer, Dordrecht, The Netherlands.

[7] Jämsä T. 2008 (69-100). Semiosis in Evolution. V: Barbieri M. (ur.) 2008. Introduction to Biosemiotics: The New Biological Synthesis. Springer, Dordrecht, The Netherlands.

[8] Kauffman S. 2010. Reinventing the Sacred: A New View of Science, Reason, and Religion. Basic Books, New York.

[9] Nicolis G. in Prigogine I. 1977. Self-Organization in Non-Equilibrium Systems : From Dissipative Structures to Order Through Fluctuations. Willey New York.

[10] Pattee H. H. 2008 (115-132). The Necessity of Biosemiotics: Matter-Symbol Complementarity. V: Barbieri M. (ur.) 2008. Introduction to Biosemiotics: The New Biological Synthesis. Springer, Dordrecht, The Netherlands.

[11] Sanguineti V. 2007. The Rosetta Stone of the Human Mind: Three languages to integrate neurobiology and psychology. Springer Science+Business Media, LLC.

[12] Scott A. 2006 (171-191). Physicalism, Chaos and Reductionism. V: Tuszynski J. A. (ur.) 2006. The Emerging Physics of Consciousness. Springer-Verlag Berlin Heidelberg.

[13] Strawson G. (1994). Mental Reality. Cambridge USA: the MIT Press, Bradford Books.

[14] Reisenzein R. in Döring S.A. 2009. Ten Perspectives on Emotional Experience: Introduction to the Special Issue. Emotion Review 1(3):195-205.

[15] Van de Vijver G. 2006 (143-161). Kant and the Intuitions of Self-Organization. V: Feltz B., Crommelinck M. in Goujon P. (ur.) 2006. Self-Organization and Emergence in Life Sciences. Springer, Dordrecht, The Netherlands.

[16] Varela F. 2006 (95-108). Neuronal Synchrony and Cognitive Functions. V: Feltz B., Crommelinck M. in Goujon P. (ur.) 2006. Self-Organization and Emergence in Life Sciences. Springer, Dordrecht, The Netherlands.

[17] Vekhavaara T. 2008 (257-282). From the Logic of Science to the Logic of the Living. The Relevance of Charles Pierce to Biosemiotics. V: Barbieri M. (ur.) 2008. Introduction to Biosemiotics: The New Biological Synthesis. Springer, Dordrecht, The Netherlands.

[3] T.i. dvoparametrski modeli afektivnosti (Reisenzein in Döring 2009).

DOŽIVLJANJE ČUSTEV

Urban Kordeš
Univerza v Ljubljani, Pedagoška fakulteta
Kardeljeva ploščad 16
1000 Ljubljana
urban.kordes@guest.arnes.si

POVZETEK

Pri poskusih znanstvene opredelitve čustev je ta morda najbolj očitna raznolikost možnih pogledov na raziskovanje duševnosti. V večini razprav o tem pojavu/procesu/fenomenu ni jasno definirano, ali gre za vedenjsko kategorijo (pogled s tretjeosebne perspektive), za zbirko fizioloških procesov ali za doživljajska poročila. Dejstvo je, da je ravno pri razmišljanju o čustvih jasno, da ne moremo dobiti popolne slike brez poznavanja vseh treh nivojev. V pričujočem prispevku poročam o fenomenologiji čustev, torej o tem kar vemo o prvoosebnem doživljanju tega fenomena. Osredotočim se na rezultate raziskav skupine, zbrane okrog Russella Hurlburta, utemeljitelja metode t.i. deskriptivnega vzorčenja izkustva.

1. Tri glavne perspektive opazovanja duševnih pojavov

Kognitivna znanost je blizu točke, ko bomo o njej govorili kot o novi, samostojni disciplini in ne več samo kot skupnem imenu za vede, ki se dotikajo vprašanj duševnosti, kognicije, zavesti itd. Osamosvajanje kognitivne znanosti pa ni preprosto saj iskanje širših spoznanj o duševnih pojavih pomeni nujno upoštevanje različnih perspektiv. Konstitutivne discipline, ki sodijo pod dežnik kognitivnih znanosti, prihajajo z vseh delov znanstvenega spektra – od naravoslovnih ved, medicine, do družboslovja in humanistike. Preprosto dodajanje spoznanj različnih ved ni mogoče, kako bi izgledalo njihovo združevanje na novem nivoju pa zenkrat tudi še ni povsem jasno. Kognitivna znanost je trenutno na točki razvoja, ko se zaveda, da obstajajo različne perspektive opazovanja določenega pojava in da je upoštevanje vseh teh perspektiv nujno. Meje med njimi niso povsem jasne, vsekakor pa so.

Tretjeosebne perspektive so dobro uveljavljene, študije vedenja, še bolj pa v zadnjih časih študije t.i. "dogodkov v živčnem sistemu" (kot se večkrat izrazi A. Damasio) imajo trdno vlogo pri graditvi temeljev kognitivne znanosti. Veliko bolj izmuzljivo področje pa je raziskovanje prvoosebnega oz. doživljajskega sveta. Človeško izkustvo nam je intimno znano, po drugi strani pa smo pri njegovem (sistematičnem, znanstvenem) raziskovanju najbolj nerodni. Kljub temu, da je razumevanje doživljajskega dela določenega kognitvinega fenomena nujno, se njegovemu opisu pogosto poskušamo izogniti ali pa ga nekako neartikulirano pomešamo med ostale (tretjeosebene) opise.

Pri poskusih znanstvene opredelitve čustev je ta zadrega morda najbolj očitna. V večini razprav o tem pojavu/procesu/fenomenu ni jasno definirano, ali gre za vedenjsko kategorijo (pogled s tretjeosebne perspektive), za zbirko fizioloških procesov (ki bi se lahko zgodili tudi brez vpletanja zavesti) ali za doživljajska poročila. Dejstvo je, da je ravno pri razmišljanju o čustvih jasno, da ne moremo dobiti popolne slike brez poznavanja vseh treh nivojev (kot je ugotovila že psihologinja Elis-Izard, 1977).

V nadaljevanju bom poskusil predstaviti možnost raziskovanja doživljanja čustev in nekaj prvih rezultatov, ki jih je zbrala mlada veja kognitivne znanosti, ki se ukvarja s preučevanjem prvoosebnih poročil (t.i. fenomenološko raziskovanje).

2. Fenomenološko raziskovanje

Fenomenološko raziskovanje se ne sprašuje o vzrokih za določeno doživljanje (kot na primer psihoanaliza), niti po njegovih fizioloških korelatih (nevroznanost). Zanima ga predvsem kaj subjekt doživlja in – ker je subjekt edini, ki ima neomejen dostop do svojega doživljanja – je lahko le on sam je »ekspert«, ki odloča kaj je kaj. Fenomenološki raziskovalec lahko subjektu (poročevalcu) pomaga izostriti pogled na doživljanje, lahko mu pomaga raziskovati njegovo lastno doživljajsko pokrajino, ne more pa soditi katero poročilo je resnično in katero ne. Z vidika fenomenološkega raziskovanja lahko samo on oceni katere vrste izkustvo je doživel.

Iz zgornjega sledi, da je – s fenomenološkega glediišča – *čustvo* samo tisto doživljanje, ki ga subjekt spozna kot takega. Ko bom v nadaljevanju govoril o čustvih, bom torej mislil na posebno vrsto izkustva, ki so ga poročevalci opisali kot čustvo.

Na tej točki lahko vidimo, da se področje, ki ga raziskujemo ne pokriva povsem z nevrološkim in/ali psihološkim raziskovanjem čustev. Mnogo je znanih primerov, kjer ljudje kljub vedenju, ki očitno kaže na prisotnost določenega čustva, ne poročajo o doživljanju čustev. In obratno (ti primeri so seveda še veliko bolj pogosti): mnogokrat zabeležimo doživljajska poročila o doživljanju čustev, ki ni bilo pospremljeno z vedenjem. Menim, da bi lahko našli tudi primere, kjer je bil npr. limbični sistem aktiviran, subjekt pa ni doživljal čustvenega stanja.

Doživljanje čustev je kompleksen del človekove fenomenološke pokrajine. Fenomenologi že dolgo vemo, da modeli delovanja kognicijskega sistema, ukrojeni po računalniškem modelu input/output ne delujejo. Seveda nam lahko pri razmišljanju pomaga slika sistema, ki se odziva na »vhodne« signale (stimule), vendar mora vsak resen razmislek upoštevati tudi avtopoetsko naravo duševnosti – zgodovina sistema in njegova notranja dinamika sta tista, ki imata glavno besedo pri tem kaj bomo izkusili in kako.

Čustva lahko v evolucionarni psihologiji modeliramo kot telesne odzive na različne zunanje situacije, vsekakor pa nas – avtonomnih bitih – v njihovo doživljanje ne sili narava okoliščin. Čustva so laho, niso pa nujno, odzivi na zunanji dražljaj.

3. Poročila o doživljanju čustev

Pri pregledu fenomenoloških poročil in redkih tekstov, ki opisujejo fenomenologijo čustev (npr. Heavy in dr., 2010) pridemo do zanimivega spoznanja, da je doživljanje čustev (mnogokrat) odvisno od časovnega intervala v katerem opazujemo. Dobimo vtis, kot da bi se čustva – bolj kot se bližamu opazovanju doživljanja v trenutku – mnogokrat razgradijo. Kot bi bilo doživljanje nekaterih čustev *sestavljeno.*

Bolj kot se približujemo trenutku, manjša je verjetnost, da bo poročevalec poročal o doživljanju čustva in večja je verjetnost, da bo stanje opisoval z fizičnimi in nedoločenimi (mnogokrat jih imenujejo "energetske") občutki. Kar je bilo – gledano bolj od "daleč" žalost, se na primer mnogokrat spremeni v tiščanje v prsih. Veselje v "valovanje energije po vsem telesu" in/ali "prijeten nemir" in "hitro se izmenjujoče misli".

Ta ugotovitev se sklada s poročili raziskovalske skupine, ki jo je vodila Susana Siegel (2008), ki je raziskovala doživljanje najstnikov in presenečeno ugotovila, da med doživljajskimi poročili ni skoraj nobenih poročil o čustvih. Po drugi strani pa so bila pogosta poročila o tem, kako so opazovali "tiščanje v prsih" itd. in *razmišljali* "verjetno je to, kar doživljam žalost".

Seveda so rezultati vedno pogojeni z načinom zbiranja podatkov. Pri fenomenološkem raziskovanju to velja še posebej. Kot sem omenil v nekaterih drugih tekstih (npr. Kordeš, 2009) prevladujeta dva glavna pogleda na to kako se lahko dobimo najbolj kvalitetna doživljajska poročila. Del raziskovalcev zagovarja uporabo metod, ki izurijo subjekta, da lahko kar se da čisto podoživlja spomine na pretekla doživetja, druga skupina pa meni, da je potrebno zmanjšati retrospekcijo na minimum in da je vsako doživljajsko poročilo, ki je časovno preveč oddaljeno, neuporabno.

Metoda t.i. deskriptivnega vzorčenja izkustva (DVI oziroma DES – descriptive experience sampling) spada v drugo kategorijo. Njena prednost je, da lahko zajame doživljajska poročila o različnih, naključno izbranih, trenutkih. Njena slaba stran je, da lahko izpusti pomembne (doživljajske) dogodke, predvsem pa ne more zaznati bolj kompleksnih, dlje trajajočih doživljajskih vsebin. Po drugi strani pa z naključnim vzorčenjem dobimo presenetljivo točno sliko subjektove doživljajske pokrajine. O doživljanju čustev je zbranih največ podatkov ravno s to metodo, zato si jo je vredno nekoliko bolj podrobno pogledati.

4. Deskriptivno vzorčenje izkustva

Deskriptivno vzorčenje izkustva predstavlja relativno lahek uvod v fenomenološko opazovanje in poročanje. Metoda DVI je relativno preprosta (subjekt poroča o doživljanju v nekaj, naključno izbranih trenutkih), kljub temu pa lahko skozi njo pridemo do globokih vpogledov v doživljanje. Tako nabrani podatki so dali precej zanimivih vpogledov v doživljanja povezana s čustvi.

Russell Hurlburt (1992), utemeljitel tehnike deskriptivnega vzorčenja izkustva. Cilj DVI je sestaviti nekakšno enciklopedijo osnovnih doživljajskih elementov in povezav med njimi. DVI lahko primerjamo z geološkim sondiranjem tal: na naključnih mestih vzamemo vzorce, ki jih potem v laboratoriju prečistimo in analiziramo. Analogno, pri DVI v naključno izbranih trenutkih sondiramo (vzorčimo) izkustvo. Sondiranje je v praksi izvedeno tako, da raziskovalni subjekt s sabo nosi napravico, ki nežno zazvoni ob naključno izbranih trenutkih. Subjekt poskuša "zamrzniti" doživljanje tik pred piskom – to stori tako, da ga čim bolj natančno opiše v priročno beležko. Ne kasneje kot štiriindvajset ur po sondiranju, se subjekt sreča z raziskovalcem, ki poskuša skozi pogovor o vzorcih dobiti čim bolj jasne podatke o doživljanju. Hultburt uči, da je treba takoj odstopiti od vprašanja, če subjekt začne premlevati oziroma ko ni povsem prepričan v svoje odgovore (»hm...«, »ne vem...«, »zdi se mi...« - tako imenovanje »subjectifications«). Na tak način poskuša retrospekcijo zmanjšati na minimum.

5. DVI in čustva

Ker je DVI usmerjeno na "zajemanje" trenutkov izkutva, je manj verjetno, da bi lahko ujelo trenutek intenzivnega čustva. S tega vidika je treba interpretirati zgornjo ugotovitev, da mnoga doživljanja čustev izgledajo sestavljena iz telesnih senzacij. Če bi ujeli trenutek močnega, kratkotrajnega čustva, bi bilo bolj verjetno, da bi tudi v mikro-introspekciji zaznali čustvo in ne z njim povezanih telesnih in "energetskih" doživljanj.

V nadaljevanju bom naštel nekaj ugotovitev, povezanih z doživljanjem čustev, o katerih so poročali raziskovalci skupine Russella Hurlburta z univerze v Nevadi. Objavljeni so v kratkem povzetku (Heavey, Hurlburt, Lefforge, 2010), celoten članek pa je še v pripravi.

Prva zanimiva ugotovitev mnogih let zbiranja podatkov o doživljajskih pokrajinah ljudi je, da je čustev mnogo manj, kot bi pričakovali. Heavey in sodelavci (2010) pišejo:

"Prvič: čustva so prisotna. Ljudlje poročajo o neposrednem, trajajočem doživljanju čustev.
Drugič: čustva mnogokrat niso prisotna. ... Prepričani smo, da v mnogih, morda kar v večini trenutkov, ljudje v polju svoje zavesti nimajo jasno razpoznavnih čustev. Poročamo lahko o nekaj pomembnih variantah primerov, ko ni nobenega doživljanja čustev, kljub temu, da so prisotni (zunanji op. U. K.) znaki čustev."

Nadaljje Heavey in sodelavci (2010) poročajo o tem, da doživljanje čustev mnogokrat vsebuje telesne senzacije. Le-te so največkrat locirane v zgornjem delu torza, redkeje pa v glavi.

Zaznave senzacij – takrat ko se pojavijo – varirajo med ultra-jasnimi in komaj opaznimi. Kot sem že omenil zgoraj, so te senzacije večinoma *del* čustev, ne pa kakšen spremljajoč pojav. Zanimivo je še, da so čustva včasih povsem "razumska" (mental) in jih poročevalci doživljajo v obliki senzacij v glavi.

Zelo zanimiva je tudi ugotovitev, da obstajajo (morda je takšnih celo večina) doživljanja čustev, ki jih generirajo misli. Pri teh čustvih so raziskovalci ugotovil, da subjekti poročajo o časovnem intervalu, ko – če so dobro izurjeni pri opazovanju svojega doživljanja – lahko vidijo kako misel ustvari čustvo. Nekateri celo poročajo, da lahko v tem delčku sekunde preprečijo nastanek čustva (Bennet Goleman, 2002, govori o "magic quarter second" in poroča o nevroloških raziskavah, ki so potrdile tovrstna doživljanja).

Subjekti mnogokrat poročajo tudi o pozitivni povratni zanki med mislimi in čustvi s katero se obe komponenti ohranjata: določena miselna vsebina ohranja čustvo in obratno.

Na tej točki bom zaključil to kratko predstavitev. Raziskovalci skupine Russella Hurlburta so sestavili seznam najbolj pogostih doživljajskih oblik. Seznam je delno dostopen na http://faculty.unlv.edu/hurlburt/codebook.html

Viri

Ellis Izard, C. (1977). *Human Emotions*. Plenum Press, New York.

Siegel, S. (2008). Phenomenal Contrast and the High-level Contents of Consciousness. V: *Abstract of VIII conference Towards the Science of Consciousness, Center for Consciousness Studies*, University of Nevada, Tucson

Heavey, C., Hurlburt, R. T., Lefforge N. L. *Some Characteristics of Feelings* (v pripravi za objavo)

Kordeš, U. (2008). »Fenomenološko raziskovanje v psihoterapiji«. *Kairos, slovenska revija za psihoterapijo, 2 (3-4)*, str. 9-21.

Bennet Goleman, T. (2002). *Emotional Alchemy.* Harmony books, New York.

MISTIČNI MOŽGANI: Očrt in kritika nevroteologije[i]

Sebastjan Vörös
Oddelek za filozofijo, Filozofska fakulteta
Aškerčeva 2, 1000 Ljubljana, Slovenia
e-mail: sebastjan.voros@guest.arnes.si

POVZETEK

V prispevku obravnavam vprašanje, v kolikšni meri lahko sodobni nevro(teo)loški pristopi pripomorejo k razumevanju verskih (zlasti mističnih) izkustev kot podmnožice spremenjenih stanj zavesti. Najprej predstavim trenutno najodmevnejši in najcelovitejši primer nevroteološke razlage (Newberg-d'Aquilijev modela), nato pa prikažem nekaj ključnih pomanjkljivosti tako obravnavanega modela kot nevroteologije nasploh. V zaključku izpostavim, da je poleg vseh empiričnih in pojmovnih težav, ki pestijo tovrstne pristope, njihova glavna težava ravno v tem, da ne zmorejo premostiti razlagalne vrzeli (oziroma brezna), ki se razpira med fenomenološko in fizično ravnjo analize.

1. UVOD: NEVROTEOLOGIJA

Nevroteologija velja za eno od najmlajših potomk »nevroznanstvene revolucije«, do katere je prišlo v zadnjem desetletju 20. stoletja in za katero je značilno skokovito izboljšanje tehnik slikanja možganov (CT, MRI, fMRI, SPECT itd.). Te izboljšave so privedle do boljšega poznavanja zgradbe in delovanja možganov ter omogočile proučevanje možganskih procesov med izvajanjem različnih telesnih in duševnih dejavnosti. Na ta način so se odprla vrata za empirično proučevanje področij, za katera se je nekoč zdelo, da so onstran dometa znanosti, in eno od teh področij so tudi spremenjena stanja zavesti ter z njimi povezana verska in mistična izkustva.[ii]

Sam izraz *nevroteologija* se v strokovnem kontekstu prvič pojavi v 80. letih prejšnjega stoletja, v 90. letih in na prehodu v novo tisočletje pa postane splošna oznaka za vsa prizadevanja, ki skušajo tako ali drugače pokazati, da »verska fenomenologija izhaja iz nevropsihologije« (d' Aquili in Newberg, 1999: 4). Nevroteologije ne gre pojmovati v ozkih mejah »običajne« teologije, saj je ne zanima neko specifično pojmovanje boga (npr. v abrahamovskih religijah) in/ali temeljni nauki nekega specifičnega religioznega sistema, temveč v okviru »širše« teologije, ki jo zanimajo »vse prvine religije, celo tiste, ki so prej veljale za iracionalne in zato niso bile vključene v domeno teologije«. S svojimi razlagalnimi modeli skuša razložiti pojav *religije kot celote* – od njenih temeljnih mitov in verskih naukov do ritualov in verskih izkustev.

Ravno slednji – *izkustveni* – vidik pa je tisti, ki mu nevroteologija iz metodoloških razlogov posveča največ pozornosti. Čeprav d'Aquili in Newberg ponudita osnutke nevroloških razlag tudi za druge vidike religije (denimo, utemeljitvene mite in rituale), so na tem mestu njune domneve zelo spekulativne. Kar jih razločuje od njunih razlag izkustvenih vidikov, je to, da so slednja podkrepljena z empiričnimi raziskavami (slikanji možganov), kar jim daje večjo trdnost. Ker so se prizadevanja večine avtorjev s področja nevroteologije spontano omejila na proučevanje verskih izkustev, bomo to dejstvo upoštevali tudi v pričujočem sestavku in se v analizi omejili zgolj na nevroteološke razlage *izkustvenega* vidika religije.

2. NEWBERG-d'AQUILIJEV NEVROTEOLOŠKI MODEL ABSOLUTNE ENOTNE BITI (AEB)

Raziskave Eugena d'Aquilija, pokojnega profesorja psihiatrije na Univerzi v Pennsylvaniji, in Andrewa B. Newberga, profesorja za psihiatrijo in radiologijo na Univerzi v Pennsylvaniji, temeljijo na dveh podmenah: prvič, da so verska izkustva prisotna v vseh religijskih sistemih, in, drugič, da imajo ta izkustva skupno *fenomenološko* jedro, ki ni odvisno od »dogmatskih resnic« posameznega sistema in sestoji iz občutka »enosti z nečim, kar presega naš jaz« (Newberg, d'Aquili in Rause, 2002: 101). Različna verska izkustva lahko glede na prisotno »stopnjo poenotenja oziroma povezave z drugimi stvarmi« razporedimo v t.i. »kontinuum enotnosti« (d'Aquili in Newberg, 1999: 96). Na eni strani tega kontinuuma se nahaja »osnovno stanje duha, v katerem preživimo večino našega vsakdanjega življenja. Jemo, spimo, delamo, občujemo z drugimi, in čeprav se zavedamo, da smo na nek način povezani s svetom, ki nas obdaja […], zaznavamo ta svet kot nekaj, od česar smo jasno ločeni« (Newberg, d'Aquili in Rause, 2002: 115). Na drugi strani kontinuuma najdemo mistična izkustva oziroma občutja »absolutne enotne biti« (*Absolute Unitary Being*; AEB), v katerih »subjekt izgubi vsakršno zavedanje ločenega omejenega bivanja in toka časa ter doživi uničenje dihotomije jaz-drugi« (d'Aquili in Newberg, 1999: 110). Ko se pomikamo od ene skrajne točke (»osnovno stanje duha«) k drugi (»absolutna enotna bit«), postaja torej razvdojenost med nami in drugimi vse šibkejša. V tem vmesnem polju se nahajajo vsa ostala verska izkustva (občutki duhovnega zanosa, transi, ekstatična stanja ipd.).

Po Newbergu in d'Aquiliju verska izkustva nastopijo bodisi *spontano* bodisi jih spodbudimo s pomočjo *kolektivnih* (rituali) in/ali *individualnih* metod (meditacija/kontemplacija). Čeprav se kolektivni in individualni pristopi medsebojno razlikujejo, se obe kategoriji »poslužujeta istih ali podobnih nevrofizioloških mehanizmov« (d'Aquili in Newberg, 1999: 13), različni pa sta si v načinu, kako končni cilj (manjši ali večji »občutek enosti«) dosežeta. Kolektivni pristopi delujejo po načelu »od spodaj navzgor«, saj vzdražijo naš avtonomni živčni sistem in živčno aktivacijo postopoma prenesejo do centralnega živčnega sistema, medtem ko individualni načini delujejo po načelu »od zgoraj navzdol«, kjer se živčna dejavnost začne »v [možganski] skorji, čemur sledi aktivacija progresivno nižjih struktur, kar doseže vrhunec v aktivaciji avtonomnega živčnega sistema« (99). Ker so individualni pristopi zaradi svoje »umirjene« narave (močno omejena ali povsem odsotna fizična dejavnost na strani subjekta) primernejši za eksperimentalno proučevanje, poleg tega pa sta avtorja prepričana, da lahko le z njimi dosežemo najvišjo stopnjo verskih izkustev (občutek »absolutne enotne biti«), se bomo v nadaljevanju osredotočili samo nanje.

Avtorja razdelita meditativne/kontemplativne tehnike v dve splošni kategoriji. V prvo uvrstita *pasivne* metode (t.i. *via negativa*), katerih glavni namen je »izprazniti duha vseh zavestnih misli«, v drugo pa *aktivne* metode (t.i. *via positiva*), katerih cilj je »popolnoma osredotočiti duha na nek predmet pozornosti – denimo, na mantro, nek simbol ali vrstico iz svetega pisma« (Newberg, d'Aquili in Rause, 2002: 117). Nevrološka modela za obe kategoriji se sicer nekoliko razlikujeta, a vodita v isto smer – k doseganju »absolutne enotne biti«. Pa si oglejmo obe kategoriji zapovrstjo.

Po mnenju avtorjev se vse »duhovne poti« začnejo z »dejanjem volje«. Pasivna metoda se začne »z zavestno željo, da iz duha odstranimo vse misli, čustva in zaznave« (*prav tam*), kar na nevrološki ravni ustreza aktivaciji frontalnega (čelnega) možganskega režnja, natančneje *prefrontalne skorje*, ki jo avtorja poimenujeta »nevrološki sedež volje«, saj »igra pomembno vlogo pri usmerjanju kompleksnih, integriranih telesnih gibov in vedenj, povezanih z doseganjem ciljev« (29).

Aktivacija prefrontalne skorje povzroči upad dotoka informacij (deaferentacijo) v zadnji del *parietalnega* (temenskega) režnja. Zadnji del parietalnega režnja ima več funkcij: po eni strani sodeluje pri »analizi in integraciji vidnih, slušnih in somatosenzoričnih informacij višjega reda« (33) – je torej pomembno središče, kjer se stekajo informacije z različnih čutov in generirajo misli –, po drugi strani pa je močno povezan z izdelavo »jaza«, saj ustvarja »prostorski občutek jaza« (levi reženj) in »fizični prostor, v katerem lahko ta jaz obstaja« (desni reženj) (Newberg, d'Aquili in Rause, 2002: 28). Avtorja zato menita, da delni deaferentaciji parietalnega režnja na nevrološki ravni ustreza počasno odmiranje misli, čustev in zaznav, skupaj s krhanjem meje med jazom in zunanjostjo.

Deaferentacija parietalnega režnja sproži val aktivacije v limbičnem delu možganov, kar privede do vzburjenja *hipotalamusa*, »starodavne možganske strukture«, ki »povezuje dejavnosti višjih možganskih predelov z bazičnimi funkcijami avtonomnega živčnega sistema in nadzira zmožnost avtonomnega živčnega sistema za ustvarjanje tako pomirjajočih kot vzburjajočih občutkov« (118). Impulzi, ki prispejo v hipotalamus, vzdražijo njegove pomirjajoče dele, kar ima za posledico občutek sproščenosti in pomirjenosti. Z aktivacijo hipotalamusa pride do sprožitve novega niza živčnih impulzov, ki preko limbičnega sistema potujejo nazaj v parietalni reženj, to pa vzpostavi povratno zanko med prefrontalno skorjo, zadnjim delom parietalnega režnja in hipotalamusom. Bolj ko si meditator prizadeva izpraznitvi duha (večja aktivacija prefrontalne skorje), izrazitejša je deaferentacija zadnjega dela parietalnega režnja in posledično pomirjajoča dejavnost hipotalamusa.

Ko omenjena pomirjajoča dejavnost na neki točki trči ob svoje meje (»hiperpomirjenost«), pride do pojava, imenovanega »razlitje« oziroma »preboj«, v katerem sta sočasno maksimalno vzdražena tako vzburjajoči kot pomirjajoči sistem, kar v meditatorju sproži občutek ekstaze in blaženosti. Nadalje, sočasna maksimalna simulacija obeh sistemov v hipu privede do popolne deaferentacije zadnjega dela parietalnega režnja. To povzroči »absolutno subjektivno zaznavo čistega prostora« oziroma »enosti« (»polnosti«) ter »uničenje dihotomije jaz-drugi« (112 – 113), tj. stanje »absolutne enotne biti«:

> *V tem stanju popolne deaferentacije orientacijskega predela [parietalni reženj] duh zaznava nevrološko realnost, ki je skladna s številnimi mističnimi opisi najvišje duhovne enosti: [V tej realnosti] ni ločenih predmetov ali bitij, nobene zaznave prostora ali toka časa, nobene meje med jazom in ostalim vesoljem. Pravzaprav sploh ni subjektivnega jaza, temveč le absoluten občutek enosti – brez misli, besed in občutkov. Duh biva brez jaza v stanju čiste, nedeljene pozornosti.* (Newberg, d'Aquili in Rause, 2002: 120)

Do istega rezultata pridemo tudi s pomočjo *aktivnih metod*, z nekaj manjšimi razlikami. Osnovna – in najpomembnejša – razlika je ta, da se proces začne z intenzivno osredotočenostjo na neko misel ali predmet, kar namesto deaferentacije zadnjega dela parietalnega režnja privede do njegove stimulacije in – posledično – namesto aktivacije pomirjajočega dela hipotalamusa do aktivacije njegovega vzburjajočega dela. Maksimalna vzburjajoča dejavnost (»hipervzburjenost«) nato povzroči »razlitje« oziroma »preboj«, od tod naprej pa je razvoj dogodkov tako rekoč enak tistemu v prvemu modelu. Razlike v začetnih korakih na nevrofiziološki ravni imajo na fenomenološki ravni za posledico, da namesto postopnega ugašanja duševnih vsebin pride do vse izrazitejšega identificiranja s predmetom, na katerega smo osredotočeni, dokler naposled ta osredotočenost ne prerase v vsesplošen občutek enosti.

Newberg in d'Aquili sta svojo tezo tudi empirično podprla, in sicer s SPECT slikanjem možganov osmih tibetanskih budistov med meditacijo in treh frančiškanskih nun med kontemplativno molitvijo. SPECT (*Single Photon Emission Computer Tomography* – računalniška tomografija emisij enotnih fotonov) je posebna slikovna tehnika, ki »meri krvni pretok v različnih predelih možganov. Več je krvnega pretoka, večja je možganska dejavnost – in obratno.« (Newberg in Waldman, 2006: 172) Pri večini budistov in nun je prišlo do povišane živčne dejavnosti v prefrontalni skorji in zmanjšane dejavnosti v zadnjem delu parietalnega režnja, kar se sklada z napovedmi njunega modela. Avtorja sta mnenja, da njune analize in raziskave jasno potrjujejo, da poročila mistikov ne temeljijo na »varljivih idejah, temveč na izkustvih, ki so nevrološko realna« (Newberrg in d'Aquili, 2002: 126).

3. KRITIKA NEWBERG-d'AQUILIJEVEGA MODELA IN NEVROTEOLOGIJE NASPLOH

a. EMPIRIČNA KRITIKA

Videli smo, da v Newberg-d'Aquilijevem modelu ključno vlogo pri razlagi verskih in mističnih izkustev igrata parietalni in frontalni reženj, kar pa je v izrazitem neskladju s hipotezami nekaterih drugih nevroznanstvenikov. Persinger in Ramachandran, decimo, dokazujeta, da ključna vloga pripada *temporalnemu* (temenskemu) režnju (mistična izkustva naj bi bila posledica »mikroepileptičnih« napadov v imenovanem režnju). Obojim oporeka Austin, ki meni, da da odločilno vlogo pri nastanku mističnih izkustev ne igrajo višja (kortikalna) središča, temveč *nižji (subkortikalni) predeli* (retikularna formacija, talamus). Nesoglasja med nevroznanstveniki torej zaenkrat niso zamejena na nekaj manjših neskladij, temveč se pričnejo na najbolj rudimentarni ravni: pri poskusu določitve približnega nevrološkega substrata verskih in mističnih doživetij.

Newberg-d’Aquilijev model pa pesti tudi nekaj bolj specifičnih težav. Austina čudi, da so pri obdelavi podatkov, dobljenih pri SPECT slikanju možganov med meditacijo, iz nadaljnjih statističnih analiz izključili nekatere možganske predele, v katerih je prišlo do precejšnjega povečanja krvnega pretoka (orbitofrontalni korteks, cingulatni girus in talamus) (Austin, 2006: 220). Pravzaprav tudi samo Newbergovo eksperimentalno delo kasnejšega datuma do neke mere spodkopava njegove zgodnejše ugotovitve. Slikanje možganov članov binkoštne cerkve med t.i. »glosolalijo« oziroma »govorjenjem v jezikih« – obliko bogoslužja, ki vključuje »petje, vokalno oglašanje [govor, podoben eksotičnemu tujemu jeziku, ki pa nima lingvističnega pomena, op.av.] in ekstatična telesna doživetja« (Newberg in Waldman, 2006, str. 196) – je namreč pokazalo, da so aktivirane druge možganske strukture kot pri meditaciji in kontemplaciji: namesto povečane aktivnosti v frontalnem režnju in zmanjšane aktivnosti v parietalnem režnju (nune in budisti) se je zgodilo ravno obratno – aktivnost v frontalnem režnju se je zmanjšala, v parietalnem režnju pa povečala. Poleg tega je prišlo do povečanja aktivnosti v talamusu in temporalnem režnju, kar ruši idejo o »kontinuumu enotnosti« in skupnem nevrofiziološkem substratu za *vsa* verska izkustva (200 – 209).

b. POJMOVNA KRITIKA

Čeprav sodi Newberg-d’Aquilijev model med bolj dodelane nevroteološke modele (veliko bolj dodelane, kot sta npr. Persingerjev ali Ramachandranov), ga kazi nekaj precej izrazitih nekonsistentnosti. V postuliranem »kontinuumu enotnosti« je zlasti problematična opredelitev »vmesnih izkustev«, tj. izkustev, ki ležijo med vsakdanjim zavestnim stanjem in izkustvom »absolutne enotne biti«. V to ohlapno kategorijo naj bi med drugim sodila tudi razna ekstatična stanja in transi, vendar so za vsaj eno *ekstatično* stanje (prej omenjeno glosolalijo) kasnejše raziskave pokazale, da je fenomenološko (in tudi nevrobiološko) popolnoma različna od meditativno-kontemplativne dejavnosti. Težave pa so tudi s samim pojmom »absolutna enotna bit«, saj je po Austinu preveč splošen in lahko zajema tako višje kot nižje oblike mističnega »zlitja« oziroma »enosti«.

Nadaljnja težava, ki pesti Newberg-d'Aquilijev model, so pomanjkljiva fenomenološka poročila testirancev med SPECT slikanji. Avtorja namreč ne navajata poročil o tem, *kaj* pravzaprav so subjekti med slikanjem doživljali, zato je vprašljivo, v kolikšni meri lahko res trdimo, da dobljeni rezultati potrjujejo njun osnovni model, ki ponuja nevrofiziološko razlago dozdevno najvišjega mističnega izkustva (»absolutna enotna bit«). Austin poudarja, da so najvišje stopnje na mistični poti »redki« in »naključni dogodki«, ki jih »[n]e gre mešati z zavestno sproženimi, bolj površinskimi meditativnimi stanji« (2006: 320). Tega se zaveda tudi Newberg, saj je v pogovoru s Horganom priznal, »da je njegova študija morda zajela le možganska stanja, povezana z meditacijo in molitvijo […] in ne globokih mističnih stanj kot takih« (Horgan, 2003: 81). Ker se je za eno obliko verskega doživljanja (»glosolalija«) izkazalo, da je nevrofiziološko in fenomenološko različna od mističnih izkustev, je pri določanju odnosa med meditativno-kontemplativnimi stanji in višjimi duhovnimi stanji nepogrešljiva natančnejša fenomenološka analiza.

c. FILOZOFSKA KRITIKA IN SKLEP

Vse nevroteološke pristope (tako Newberg-d'Aquilijevega kot Persingerjevega, Ramachandranovega in Austinovega) pa pesti temeljni problem, ki – tudi če bi nevroznanstvenikom nekoč uspelo odpraviti vsa empirična in pojmovno-fenomenološka neskladja – meče senco dvoma na možnost, da bi nevroteologija bila sposobna razložiti verska izkustva na *načelni* ravni, tj. da bi nevrofiziološki modeli lahko dali *dokončno* oziroma *celovito* razlago tovrstnih izkustev. Gre za t.i. »problem kvalij« – za razkol med »prvoosebno« in »tretjeosebno vednostjo« oziroma za prepad med »zavestnimi« in »fizičnimi« pojavi, ki v primeru verskih izkustev dobijo dodatne razsežnosti. Poglejmo si, o čem je govora.

K problemu najlažje pristopimo s slavnim Nagelovim miselnim poskusom o tem, kako je biti netopir. Nagel se vpraša, kaj pomeni, da je nek organizem *zavesten*, in zaključi, da to pomeni, da obstaja nekaj takšnega, kot je »*biti ta organizem*«: »[O]rganizem [ima] zavedajoča se mentalna stanja tedaj in le tedaj, če obstaja nekaj takšnega, kot [je] *biti* ta organizem.« (Nagel, 1991: 401) Ta *subjektivni* značaj izkustva je specifičen za duševne pojave in ga ni moč zajeti v kakršnikoli analizi, sloneči na *objektivnem* fizičnem izrazju. Nagel to ponazori na primeru netopirja. Še tako dobro in nadrobno poznavanje zgradbe in delovanja netopirjevih možganov nam ne pove, *kako je biti netopir*. Namreč tudi če si skušamo predstavljati, kako bi bilo, če bi na rokah imeli krila, imeli zelo slab vid itd., nam to pove le to, kako bi bilo *nam*, če bi se vedli kot netopir; ne bi pa bili zmožni ugotoviti, kako je *netopirju* biti netopir (402). Doživeti subjektivni značaj izkustva pomeni torej nič manj kot *biti organizem, ki ta značaj doživlja*. Kot zavestna bitja smo omejeni z lastno subjektivnostjo: tako kot nihče drug ne more izkusiti, kako je biti jaz, tudi sam ne morem doživeti subjektivnosti nekoga ali nečesa drugega.

V podobni smeri je razmišljal Frank Jackson, ko je oblikoval znameniti miselni poskus o Mary, prvovrstni nevrologinji, ki do potankosti pozna nevrofiziologijo barvnega zaznavanja. Maryjina posebnost pa je, da je (iz takšnega ali drugačnega razloga) odraščala v črno-belem okolju, tako da kljub temu, da ve čisto vse o nevrofiziologiji barvnega zaznavanja, ni bila še nikoli izpostavljena dejanskim barvam. Kaj se zgodi, ko Mary zapusti svojo črno-belo sobo in vstopi v svet barv? Bo Mary izvedela kaj novega o svetu ali ne? Jackson je prepričan, da vsekakor *bo* izvedela nekaj novega: izvedela bo, kako je, ko *dejansko izkusimo neko barvo*, tj. kakšen je *subjektivni* oz. *kvalitativni* značaj te barve. Poznavanje fizičnih dejstev ne more zaobjeti subjektivnega izkustva, ki ga bo doživela, saj ne more opisati *kvalij* (občutka »kako je biti«), ki spremljajo zaznavo neke barve (Jackson, 2002: 275). Joseph Levine je na podlagi tega sklenil, da med zavestnimi in snovnimi stanji zeva *razlagalna vrzel*: zavestna stanja imajo nek presežek – kvalitativni značaj (»kako je biti« ali »kako se občuti«) –, ki ga ne moremo razložiti z objektivnim fizičnim besednjakom (Levine, 2002: 354 – 356).

Na podlagi povedanega si lahko zastavimo provokativno vprašanje: *kako je biti mistik*? Če namreč navedeni ugovori veljajo za običajna zavestna stanja, so v primeru izjemnih stanj, zlasti mističnih, potencirani do skrajnosti. Glede slednjih so si različne verske in filozofske tradicije edine, da je pri njih bistveno ravno to, da jih *izkusimo*. Daisetz Suzuki poudarja, da je osnova vseh verskih izkustev *»takšnost«* oziroma »videti stvari takšne, kot so« (brez primesi našega jaza); in če hočemo »takšnost« razmeti, jo moramo tudi osebno doživeti. Austin se s tem strinja in dodaja, da preden je sam doživel »razsvetljenje«, je bila »takšnost« zanj le »beseda na listu«, le »bledi posnetek dejanske stvari«. Prava »takšnost« pa ni pojem, temveč »izkustvo, ki presega meje razuma« (Austin, 1999: 549). Duhovnost poleg tega ni le spoznavna, temveč tudi *transformativna* dejavnost. Z besedami Kena Wilberja: »[Duhovnost] ne pomeni le drugačnega mišljenja o duhu, temveč misleca dejansko spremeni.« (Horgan, 2003: 60) Roberts je zato v pogovoru z Horganom strokovnjake, ki proučujejo mistična izkustva s tretjeosebnega (objektivnega vidika), šaljivo označil za »mistične evnuhe«, kajti »evnuh lahko opazuje spolno vedenje drugih in piše o njem, vendar je očitno, da nima nobenega vpogleda v subjektivno izkustvo seksa« (51).

»Razlagalna vrzel« se v primeru mističnih izkustev zdi še toliko bolj nepremostljiva, zaradi česar Horgan poudarja, da »[n]evroteologi niso soočeni z razlagalno vrzeljo, temveč *prepadom*« (137). Vprašljivo je, ali lahko ta grozeč prepad med objektivnostjo in subjektivnostjo premosti še tako kompleksen in natančen nevrofiziološki model. Austin sklene: »Ne domišljajte si, da nevrološke slikovne tehnike, opisane v teh zadnjih petih poglavjih, dajejo 'odgovor' o zenu [ali mistiki vobče]. Njihovim podatkom manjka zavest osebe, ki je v stroju. Zena [in mistike] ne bomo razumeli s pomočjo elektronske čarovnije, temveč preko zadnje plati naših hlač [meditacije] in nemara intuicij, ki jih občutimo na prostem ali v nekaterih vrsticah poezije« (Austin, 2008: 226).

LITERATURA

Austin, James H. (2006). *Zen-Brain Reflections*. Cambridge, Massachusetts, London, England: The MIT Press

Damasio, Antonio. (2008). *Iskanje Spinoze: Veselje, žalost in čuteči možgani*. Ljubljana: Krtina

d'Aquili, Eugene in Newberg, Andrew. (1999). The *Mystical Mind: Probing the Biology of Religious Experience*. Minneapolis: Fortress Press

Hick, John. (2006). *The Frontier of Religion and Science. Religious Experience, Neuroscience and the Transcendent*. New York: Palgrave Macmillan

Horgan, John. (2003). *Rational Mysticism: Spirituality Meets Science in the Search of Enlightenment*. Boston, New York: Mariner Books

Jackson, Frank (2002). »Epiphenomenal Qualia«. V Chalmers, David J. (ur.). *Philosophy of Mind: Classical and Contemporary Readings*. New York & Oxford: Oxford University Press, str. 273 – 280.

Levine, Joseph (2002). »Materialism and Qualia: Explanatory Gap«. V Chalmers, David J. (ur.). *Philosophy of Mind: Classical and Contemporary Readings*. New York & Oxford: Oxford University Press, str. 354 – 361.

Nagel, Thomas (1990). »Kako je biti netopir?«. V Hofstadter, Douglas R., Dennet, Daniel C. (ur.). *Oko duha: Fantazije in refleksije o jazu in duhu*. Ljubljana: Založba mladinska knjiga, str. 399 – 421.

Newberg, Andrew in Waldman, Mark Robert. (2006). *Why We Believe What We Believe*. New York, London, Toronto, Sydney: Free Press

Newberg, Andrew in Waldman, Mark Robert. (2009). *How God Changes Your Brain*. New York: Ballantine Books

Newberg, Andrew, d'Aquili Eugene in Rause, Vince. (2001). *Why God Won't Go Away*. New York: Ballantine Books

Ramachandran, V. S. in Blakeslee, Sandra. (1998). *Phantoms in the Brain*. New York, London, Toronto, Sydney: Harper Perennial.

Vörös, Sebastjan (2010). »Blišč in beda nevroteologije«, *Analiza*, let. 14, 1-2, str. 71-91..

---

[i] V pričujočemsestavku v strnjeni obliki predstavljam vsebine, ki sem jih nadrobneje predstavil v: Vörös, 2010.

[ii] Izraz »spremenjena stanja zavesti« zajema široko paleto zavestnih stanj, ki jih (zelo grobo rečeno) določajo netipične doživljajske komponente. Poleg »verskih« in »mističnih izkustev« sodijo sem tudi npr. psihopatološka zavestna stanja in stanja, ki nastopijo pod vplivom halucinogenih drog. »Mistična izkustva« pa obravnavam kot posebno podmnožico verskih stanj (čeprav bi jih bilo pravilneje obravnavati kot posebno množico, ki se le mestoma prekriva z verskimi izkustvi), katere cilj je doseči enotnost z zadnjo realnostjo (Absolutom, Bogom).

ZNAČILNOSTI NAGRAJEVALNEGA MOTIVACIJSKEGA SISTEMA, MOTIVACIJSKEGA SISTEMA UMIKA TER IMPULZIVNOSTI PRI OSEBAH Z BIPOLARNO MOTNJO RAZPOLOŽENJA

Barbara Dolenc, Lilijana Šprah
Družbenomedicinski inštitut, Znanstvenoraziskovalni center SAZU
Novi trg 2, Ljubljana
tel: 01 470 64 39; faks: 01 426 14 93
e-mail: dolenc.barbara@gmail.com

POVZETEK

V pričujoči raziskavi smo preučili značilnosti vedenjskega motivacijskega sistema približevanja (BAS) in umika (BIS) ter impulzivnosti pri osebah z bipolarno motnjo razpoloženja. Rezultati so pokazali pomembno bolj izraženo tako impulzivno vedenje, kot tudi BAS in BIS občutljivost v klinični skupini, pri čemer se omenjeni osebnostni konstrukti v večini niso medsebojno pomembno povezovali. Izsledki naše raziskave govorijo v prid hipotezi o možnih osebnostnih lastnostih, ki bi lahko predstavljale dejavnik tveganja za razvoj bipolarne motnje razpoloženja.

1. UVOD

Bipolarna motnja razpoloženja (v nadaljevanju BMR) je definirana s ponavljajočimi se periodami privzdignjenega razpoloženja in depresije, kar se odraža tudi v mišljenju in vedenjskih vzorcih posameznika (Benazzi, 2007). Med dejavnike tveganja za začetek razvoja BMR avtorji prištevajo več osebnostnih potez, med njimi tudi različne vidike motivacijskega vedenja ter impulzivnosti, ki so bili tudi predmet naše raziskave.
Grayev (1990) aktivacijski in inhibicijski vedenjski model sestoji iz dveh neodvisnih nevrobioloških sistemov, in sicer vedenjskega sistema umika (BIS, *behavioural inhibition system*) in nagrajevalnega vedenjskega sistema (BAS, *behavioural activation system*). Posamezniki se razlikujejo v odzivnosti njihovega BIS in BAS in različni vzorci občutljivosti sistemov lahko tako predstavljajo dejavnike tveganja za razvoj različnih psihiatričnih motenj (Johnson, Turner in Iwata, 2003). BIS je nevrofiziološki sistem, ki je občutljiv na signale kazni, odvzem nagrade in novosti in je tako osnova za izogibalno motivacijo in vedenje (Gray, 1990). Aktiviran BIS v odgovor na signale nevarnosti sproža doživljanje negativnih emocij in inhibira potekajoče vedenje, ekstremna BIS občutljivost pa naj bi tako bila osnova za individualne razlike v nagnjenosti k anksioznosti, depresivnosti in obsesivno-kompulzivnim motnjam (Carver in White, 1994; Meyer, Johnson in Winters, 2001). BAS pa je za razliko občutljiv na signale nagrade ter izogibanje kazni ter je tako osnova za približevalno motivacijo in vedenje (Gray, 1990). Aktiviran BAS torej usmerja vedenje k aktivnemu doseganju ciljev in iskanju novih, nagrajujočih izkušenj ter sproži doživljanje pozitivnih emocij ter kognitivno aktivnost, zaradi česar avtorji povezujejo BAS aktivnost z večjo impulzivnostjo posameznika (Carver in White, 1994).
Depue je s sodelavci (1987, 1989, v Alloy in dr., 2009) tako zasnoval BAS hipersenzibilno teorijo BMR. Osebe, pri katerih obstaja tveganje za razvoj BMR, naj bi namreč posedovale bolj občutljiv BAS, ki se hiperaktivno odzove na določene pomembne namige, kar pa se posledično odraža v maničnih ali hipomaničnih simptomih. Osebe z BMR naj bi imele tako bolj izraženo potrebo po iskanju nagrajujočih dražljajev, to pa vodi tudi v komorbidnost BMR z boleznimi odvisnosti (Alloy in dr., 2009). V nasprotju pa depresivni simptomi reflektirajo mirovanje oziroma deaktivacijo vedenjskega približevanja (Kasch, Rottenberg, Arnow in Gotlib, 2002; McFarland, Shankman, Tenke, Bruder in Klein, 2006). Vse to pomeni, da naj bi posamezniki s hiper-občutljivim BAS imeli okrnjeno sposobnost efektivno regulirati svoja čustva in vedenja, zato pri njih obstaja nagnjenost k hipomaničnim in depresivnim stanjem, torej k BMR (Alloy in dr., 2009; Johnson in dr., 2000).
Za razliko pa so rezultati povezanosti sistema BIS z BMR nekoliko dvoumni. Medtem ko nekateri poročajo o povezanosti BIS občutljivosti tako z unipolarno, kot tudi z bipolarno depresijo (Johnson, Turner in Iwata, 2003; Kasch in dr., 2002), rezultati drugih raziskav ne kažejo pomembne vloge BIS pri posameznikih z BMR (Alloy in dr., 2008).
Nekatere raziskave (Patton, Stanford in Barratt, 1995; Swann in dr., 2001) poročajo tudi o povečani izraženosti impulzivnosti pri posameznikih z BMR, ki pa naj bi bila prisotna med vsemi fazami bolezni, kar pomeni, da bi lahko tudi impulzivnost predstavljala dejavnik tveganja za razvoj BMR. Impulzivno vedenje definiramo kot hitro, spontano in nenačrtovano akcijo, kjer gre za nesposobnost odlaganja gratifikacije. Prav zato jo raziskovalci povezujejo z manijo in jo med drugim postavljajo tudi kot enega od kriterijev za njeno diagnozo (Swann in dr., 2001). Vendar pa sam odnos med maničnimi simptomi in specifičnimi aspekti impulzivnosti še ni čisto jasen, povezanost impulzivnosti s samo depresijo pa je še bolj dvoumna. Patton, Stanford in Barratt (1995) navajajo, da bi impulzivnost pri načrtovanju oziroma samo pomanjkanje občutka za prihodnost, lahko bila povezana z brezupom in depresijo kot tako. Tudi Stanford, Mathias, Dougherty, Lake, Anderson in Patton (2009) poročajo o povezanosti impulzivnosti z afektivnimi

stanji pri motnjah razpoloženja. Motorična impulzivnost naj bi namreč bila bolj povezana z maničnimi epizodami in nezmožnostjo odložitve gratifikacije, impulzivnost pri načrtovanju naj bi bila bolj povezana z depresivnimi epizodami, predvsem z anhedonijo in občutki brezupa (Swann, Steinberg, Lijffijt in Moeller, 2008), medtem ko naj bi podlestvica pozornosti bila povezana tako z maničnimi, kot tudi z depresivnimi epizodami.

2. NAMEN RAZISKAVE

Namen raziskave je bil preučiti značilnosti in povezanost obeh motivacijskih sistemov BIS in BAS ter impulzivnosti pri osebah z BMR v fazi remisije ter pri zdravih prostovoljcih.

3. METODA

3.1 Udeleženci

V raziskavi je sodelovalo 70 oseb z BMR, ki so bili v času testiranja v remisiji, ter 70 zdravih prostovoljcev, ki so se s klinično skupino ujemali po spolu, starosti in izobrazbi. Demografski podatki obeh skupin so prikazani v tabeli 1.

Tabela 1
Demografski podatki vzorcev, vključenih v raziskavo

		Skupina	
		Klinična	Kontrolna
N	Moški (%)	34 (48,6)	34 (48,6)
	Ženske (%)	36 (51,4)	36 (51,4)
Starost	*M* (*SD*)	42,19 (9,89)	39,47 (12,21)
Izobrazba	*M* (*SD*)	13,84 (2,23)	14,45 (2,33)

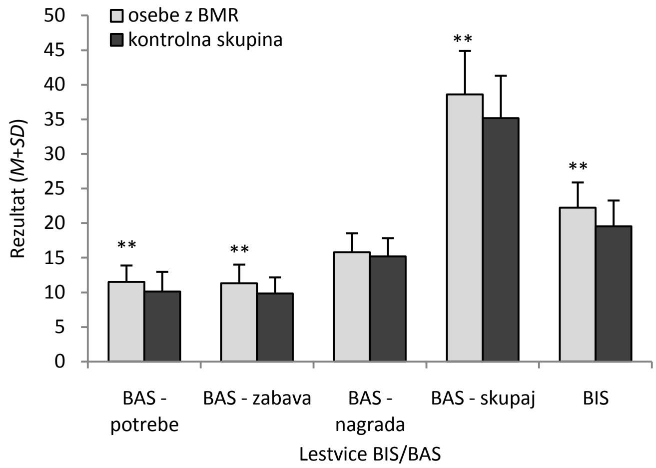
3.2 Pripomočki

Lestvica BIS/BAS (Carver in White, 1994) je samoocenjevalna lestvica, ki meri občutljivost BIS in BAS motivacijskega sistema. Vsebuje 20 postavk, na katere posameznik odgovarja na 4-stopenjski Likertovi lestvici. Lestvica BIS meri občutljivost na dogodke, ki lahko sprožijo anksioznost in so povezani s kaznovanjem, lestvica BAS pa vključuje tri podlestvice, in sicer *dovzetnost za nagrade* (stopnja, do katere nagrade vodijo k pozitivnim emocijam), *potrebe* (aktivno prizadevanje za želene cilje) in *iskanje zabave* (tendenca k iskanju in impulzivni vključitvi v potencialno nagrajujoče dražljaje). Višji rezultat na posamezni lestvici odraža večjo občutljivost motivacijskega sistema.
Barrattova Lestvica impulzivnosti BIS-11 (Patton, Stanford in Barratt, 1995) je 4-stopenjska samoocenjevalna lestvica namenjena merjenju impulzivnosti kot osebnostne poteze. Vsebuje 30 postavk, ki se nanašajo na tri aspekte impulzivnosti, in sicer *pozornost* (sposobnost osredotočiti se na trenutne naloge), *motorika* (spontano in nepremišljeno reagiranje) ter *načrtovanje* (trenutna orientacija oziroma pomanjkanje vpogleda v prihodnost).

3.3 Postopek

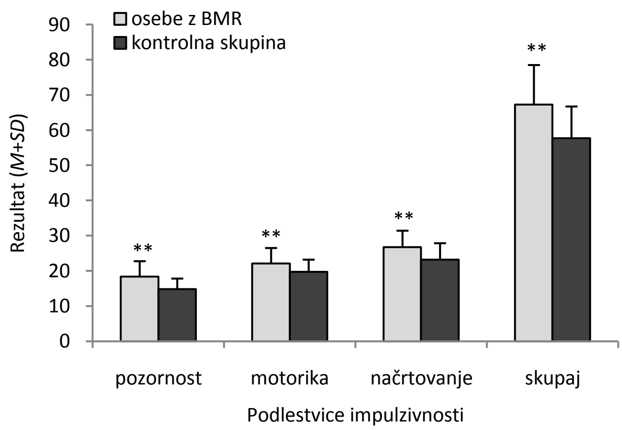
Tako klinična, kot tudi kontrolna skupina sta izpolnili oba vprašalnika, nato pa smo rezultate analizirali s statističnim paketom SPSS 16.0. Razlike med skupinami smo izračunali s pomočjo Mann-Whitneyevega U-testa. Izračunali smo tudi Cohenovo mero učinka *d*. Da bi ugotovili povezanost impulzivnosti in BIS/BAS motivacijskih sistemov, smo izračunali Spearmanove korelacijske koeficiente.

4. REZULTATI



Slika 1. Občutljivost nagrajevalnega motivacijskega sistema (BAS) in motivacijskega sistema umika (BIS) pri osebah z BMR in zdravih prostovoljcih.
*** statistično pomembna razlika na 1% nivoju tveganja*

Rezultati so pokazali statistično pomembno večjo občutljivost tako BIS (*Z*=-4,392, *p*=0,000, *d*=0,68) kot tudi BAS (*Z*=-3,133, *p*=0,002, *d*=0,53) v klinični skupini v primerjavi z zdravimi prostovoljci (glej sliko 1). Statistično pomembne razlike so se pokazale na dveh podlestvicah BAS občutljivosti, in sicer na podlestvici potrebe (*Z*=-2,960, *p*=0,003, *d*=0,51) in zabava (*Z*=-3,526, *p*=0,000, *d*=0,56), medtem ko razlika na podlestvici nagrada (*Z*=-1,514, *p*=0,130, *d*=0,22) ni bila statistično pomembna.



Slika 2. Izraženost dimenzij impulzivnosti, ocenjenih z BIS-11, pri osebah z BMR in zdravih prostovoljcih.
*** statistično pomembna razlika na 1% nivoju tveganja*

Poleg tega so rezultati pokazali tudi, da so osebe z BMR pomembno bolj impulzivne (Z=-4,781, p=0,000, d=0,84), ta impulzivnost pa je pomembno višja tako na področju pozornosti (Z=-4,872, p=0,000, d=0,84), kot tudi na področju motorike (Z=-3,462, p=0,001, d=0,58) in načrtovanja (Z=-4,042, p=0,000, d=0,71) (glej tudi sliko 2).

Tabela 2
Spearmanovi koeficienti korelacije med posameznimi lestvicami impulzivnosti in BIS/BAS občutljivostjo

	Impulzivnost			
	Pozornost	*Motorika*	*Načrtovanje*	*Skupaj*
Osebe z BMR				
BAS - potrebe	0,079	0,207	-0,002	0,119
BAS - zabava	-0,013	0,383**	0,111	0,219
BAS - nagrada	0,032	0,047	-0,141	-0,012
BAS - skupaj	0,059	0,288*	0,004	0,160
BIS	0,337**	0,058	0,012	0,163
Kontrolna skupina				
BAS - potrebe	-0,174	-0,051	-0,038	-0,093
BAS - zabava	0,246*	0,210	0,162	0,239*
BAS - nagrada	0,040	-0,066	-0,071	-0,071
BAS - skupaj	0,027	0,022	-0,020	-0,007
BIS	0,230	0,018	-0,065	0,059

** statistična pomembnost na 5% nivoju tveganja*
*** statistična pomembnost na 1% nivoju tveganja*

Analiza povezanosti med posameznimi lestvicami impulzivnosti in občutljivostjo BIS/BAS motivacijskih sistemov je prikazana v tabeli 2. Iz nje je razvidno, da se impulzivnost in občutljivost BIS/BAS motivacijskega sistema v večini primerov ne povezuje statistično pomembno, nekaj povezav obstaja le med BAS-zabavo in določenimi aspekti impulzivnosti, pri klinični skupini pa je statistično pomembna povezanost tudi med vedenjskim sistemom umika in impulzivnostjo na področju pozornosti.

5. RAZPRAVA

V pričujoči raziskavi smo preučevali značilnosti motivacijskega sistema približevanja in umika ter impulzivnosti pri osebah z BMR v primerjavi z zdravimi prostovoljci. Prav tako nas je zanimala povezanost obeh osebnostnih konstruktov pri obeh skupinah.

Rezultati so pokazali pomembno večjo občutljivost tako BAS kot BIS pri osebah z BMR. Večja BAS občutljivost se odraža na domala vseh podlestvicah, razen na podlestvici nagrada, kjer razlike niso statistično pomembne. Tudi mere učinka kažejo srednjo do visoko praktično pomembnost. Osebe z BMR torej kažejo večjo pripravljenost vstopanja v vedenje, usmerjeno k cilju, ti rezultati pa so tudi v skladu s teorijo o hipersenzibilnem BAS pri osebah z BMR. Ti posamezniki tako kažejo večjo tendenco k aktivnemu prizadevanju za želene cilje ter večjo nagnjenost k iskanju in impulzivni vključitvi v potencialno nagrajujoče dražljaje. Vendar pa klinična skupina ne kaže pomembnih razlik na podlestvici BAS-nagrada. Ta se osredotoča bolj na emocionalni vidik in meri stopnjo, do katere nagrade vodijo k pozitivnim emocijam. Kljub temu se pri osebah z BMR kaže trend hipersenzibilnega BAS, ki je očitno prisoten tudi v fazi remisije, ko torej oseba z BMR ne kaže ne depresivnih ne maničnih simptomov. Zanimivo je, da osebe z BMR hkrati kažejo tudi večjo nagnjenost k anksioznosti in vedenju umika, kar lahko sklepamo iz višjega BIS rezultata in torej hipersenzibilnosti BIS. Naši rezultati se skladajo z nekaterimi prejšnjimi raziskavami, ki sicer pomembno vlogo v BMR pripisujejo predvsem BAS občutljivosti (Alloy in dr., 2009; Johnson in dr., 2000), hkrati pa naši izsledki podpirajo tudi tiste raziskave, ki kot pomemben dejavnik tveganja za razvoj BMR navajajo tudi hipersenzibilen BIS (Johnson, Turner in Iwata, 2003; Kasch in dr., 2002). Rezultati naše raziskave so tako pokazali, da imajo osebe z BMR bolj občutljiv tako BAS kot tudi BIS motivacijski sistem, na osnovi česar bi lahko sklepali, da imajo te osebe hkrati bolj izraženo potrebo po iskanju nagrajujočih dražljajev in nagnjenost izogibanju kaznovalnim dražljajem. Verjetno se pri teh osebah kažejo tudi težave z regulacijo lastnih emocij in vedenja, kar bi lahko bil odraz impulzivnosti ter BIS in BAS hipersenzibilnosti.

Rezultati študije so pokazali tudi, da so osebe z BMR statistično pomembno bolj impulzivne, kar pa se odraža na vseh njenih aspektih. Tudi mere učinka kažejo na razmeroma velik učinek in visoko praktično pomembnost rezultatov. Osebe z BMR so tako v primerjavi z zdravimi prostovoljci bolj odkrenljive, imajo več težav z osredotočanjem na trenutne naloge, bolj spontano in nepremišljeno reagirajo ter imajo manj vpogleda v prihodnost, težave pa imajo tudi z odložitvijo gratifikacije. Ti rezultati so v skladu z nekaterimi prejšnjimi študijami (Patton, Stanford in Barratt, 1995; Stanford in dr., 2009; Swann in dr., 2001), ki kažejo na povečano izraženost impulzivnosti pri osebah z BMR, ki pa je, tako kot motivacijsko vedenje, prisotna tudi v evtimični fazi bolezni. To bi lahko bil pokazatelj relativno stabilne osebnostne lastnosti oziroma poteze, ki se pri osebah z BMR ohranja skozi vse epizode bolezni. Vse to pa bi lahko pomenilo, da tako impulzivnost, kot tudi motivacijsko vedenje posameznika, morda predstavljata dejavnika tveganja za razvoj BMR, čeprav bi morali v prid tej hipotezi odnos raziskati še s kavzalnega vidika.

Zanimivo pa je, da naši rezultati niso pokazali visoke povezanosti impulzivnosti z nobenim od motivacijskih sistemov. Pričakovali bi namreč visoko korelacijo med impulzivnostjo in BAS občutljivostjo, saj Gray (1990) za BAS uporablja kar termin »impulzivnost«. Impulzivni posamezniki naj bi namreč bili bolj občutljivi na samo nagrado v primerjavi z manj impulzivnimi posamezniki (Carver in White, 1994; Gray, 1990), čeprav novejše raziskave mečejo luč dvoma na tiste, ki enačijo občutljivost za nagrade z impulzivnostjo kot tako (Franken in Muris, 2006). Tudi v naših rezultatih tako lahko opazimo le nekaj pomembnih korelacij med BAS-zabavo in impulzivnostjo (motorično impulzivnostjo pri klinični skupini in impulzivnostjo v pozornosti pri zdravih prostovoljcih), pri osebah z BMR pa z impulzivnostjo v pozornosti korelira tudi BIS občutljivost. Povezanost impulzivnosti in obeh motivacijskih sistemov tako še ni čisto jasna, saj nekatere raziskave poročajo o visoki povezanosti predvsem BAS z impulzivnim vedenjem. Omenjena konstrukta naj bi si namreč bila podobna v svoji fenotipski ekspresiji (Quilty in Oakman, 2004), česar v našem vzorcu ni opaziti. Po drugi strani pa mnogo avtorjev opozarja na multidimenzionalnost obeh konstruktov, zato je verjetno napak govoriti o enoznačni povezanosti med impulzivnostjo in motivacijo približevanja oziroma umika. Quilty in Oakman (2004) tako poročata predvsem o povezanosti motorične impulzivnosti in impulzivnosti v pozornosti z BAS občutljivostjo, podobne povezave, četudi precej nizke, pa smo dobili tudi v naših dveh vzorcih. Sklepamo, da sta BAS in BIS na eni ter impulzivnost na drugi strani dva ločena konstrukta, ki med seboj ne kažeta velike povezanosti, njune lastnosti pa se na nekaterih mestih verjetno prekrivajo. To prekrivanje bi se lahko odražalo v podlestivci BAS-zabava, ki naj bi, kot menijo nekateri avtorji (Knyazev, Slobodskkaya in Wilson, 2004; Smillie, Jackson in Dalgleish, 2006) odslikavala predvsem impulzivnost, medtem ko naj bi bili podlestivici potrebe in dovzetnost za nagrade bolj ali manj ključni koncept BAS.

6. ZAKLJUČEK

Rezultati naše raziskave potrjujejo domnevo o impulzivnosti ter občutljivosti vedenjskih motivacijskih sistemov BIS/BAS kot pomembnih dejavnikov, ki so vpleteni v razvoj BMR. Nekateri avtorji sklepajo, da BAS občutljivost (še posebej skupni BAS rezultat ter rezultat na lestvici BAS-zabava) lahko predstavlja prediktor za diagnozo bipolarnega spektra. Podobno pa avtorji sklepajo tudi o impulzivnosti ter tako postavljajo hipotezo, da povečana BAS občutljivost in bolj izražena impulzivnost predstavljata povečano vulnerabilnost za razvoj bolezni bipolarnega spektra (Alloy in dr., 2009). Naši rezultati so tako v skladu s to hipotezo, vendar pa je ključno vprašanje, na katerega bo v prihodnosti potrebno odgovoriti, to, ali je takšno stanje vzrok za bolezen ali zgolj njena posledica. Ker iz preteklih raziskav (Alloy in dr., 2009) vemo, da povečana BAS občutljivost predstavlja tudi dejavnik tveganja za razvoj nekaterih oblik zasvojenosti (zasvojenost z alkoholom in drogami, kompulzivno nakupovanje, igre na srečo), pa bi bilo zanimivo raziskati tudi, če lahko govorimo o specifičnem profilu poteznih značilnosti, ki osebo z diagnozo BMR naredijo bolj ranljivo za tovrstna vedenja. Vemo namreč, da osebe v manični fazi kažejo kopico tovrstnih vedenjskih vzorcev (Benazzi, 2007) in morda so prav določene osebnostne poteze tiste, zaradi katerih se določeno vedenje še toliko bolj izrazi. Vsekakor bi lahko tudi s pomočjo slikovnih študij možganov lažje odgovorili na vprašanja, v kolikšni meri se spremenjeno delovanje nekaterih možganskih področij, ki vplivajo na kognitivno kontrolo in uravnavanje čustvenih procesov, pri osebah z BMR odraža tudi na povečani dovzetnosti za nekatera tvegana vedenja, povezana z zasvojenostjo.

*** Študija je bila izvedena v sklopu raziskovalnega projekta L3-9698-2485-06, ki sta ga sofinancirala ARRS in Astrazeneca.*

7. LITERATURA

[1] Alloy, L. B., Abramson, L. Y., Walshaw, P. D., Cogswell, A., Grandin, L. D., Hughes, M. E., Iacoviello, B. M., Whitehouse, W. G., Urosevic, S., Nusslock, R. in Hogan, M. E. (2008). Behavioral Approach System and Behavioral Inhibition System sensitivities and bipolar spectrum disorders: prospective prediction of bipolar mood episodes. *Bipolar Disorders, 10 (2)*, 310-322.

[2] Alloy, L. B., Bender, R. E., Wagner, C. A. in Whitegouse, W. G. (2009). Bipolar Spectrum – Substance Use Co-occurrence: Behavioral Approach System (BAS) Sensitivity and Impulsiveness as Shared Personality Vunrerabilities. *Journal of Personality and Social Psychology, 97 (3)*, 549 – 565.

[3] Benazzi, F. (2007). Bipolar disorder – focus on bipolar II disorder and mixed depression. *The Lancet, 369 (9565)*, 935 – 945.

[4] Carver, C. S. in White, T. L. (1994). Behavioral Inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology, 67 (2)*, 319-333.

[5] Franken, I. H. A. in Muris, P. (2006). Gray's impulsivity dimension: A distinction between Reward Sensitivity versus Rash Impulsiveness. *Personality and Individual Differences, 40 (7)*, 1337-1347.

[6] Gray, J. A. (1990). Brain systems that mediate both emotion and cognition. *Cognition and Emotion, 4 (3)*, 269-288.

[7] Johnson, S. L., Sandrow, D., Meyer, B., Winters, R., Miller, I, Solomon, D. in Keitner, G. (2000). Increases in manic sympoms after life events involving goal attainment. *Journal of Abnormal Psychology, 109 (4)*, 721-727.

[8] Johnson, S. L., Turner, R. J. in Iwata, N. (2003). BIS/BAS levels and psychiatric disorder: an epidemiological study. *Journal of Psychopathology and Behavioral Assessment, 25 (1)*, 25-36.

[9] Kasch, K. L., Rottenberg, J., Arnow, B. A. in Gotlib, I. H. (2002). Behavioral activation and inhibition systems and the severity and course of depression. *Journal of Abnormal Psychology, 111 (4)*, 589-597.

[10] Knyazev, G. G., Slobodskaya, H. R. in Wilson, G. D. (2004). Comparison of the construct validity of the Gray-Wilson Personality Questionnaire and the BIS/BAS Scales. *Personality and Individual Differences, 37 (8)*, 1565-1582.

[11] McFarland, B. R., Shankman, S. A., Tenke, C. E., Bruder, G. E. in Klein, D. N. (2006). Behavioral activation system deficits predict the six-month course of depression. *Journal of Affective Disorders, 91 (2-3)*, 229-234.

[12] Meyer, B., Johnson, S. L. in Winters, R. (2001). Responsiveness to Threat and Incentive in Bipolar Disorder: Relations of the BIS/BAS Scales With Sympoms. *Journal of Psychopathology and Behavioral Assessment, 23 (3)*, 133-143.

[13] Patton, J. H., Stanford, M. S. in Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology, 51 (6)*, 768 – 774.

[14] Quilty, L. C. in Oakman, J. M. (2004). The assessment of behavioural activation – the relationship between impulsivity and behavioural activation. *Personality and Individual Differences, 37 (2)*, 429-442.

[15] Smillie, L. D., Jackson, C. J. in Dalgleish, L. I. (2006). Conceptual distinctions among Carver and White's (1994) BAS scales: A reward-reactivity versus trait impulsivity perspective. *Personality and Individual Differences, 40 (5)*, 1039-1050.

[16] Stanford, M. S., Mathias, C. W., Dougherty, D. M., Lake, S. L., Anderson, N. E. in Patton, J. H. (2009). Fifty years of the Barratt Impulsiveness Scale: An update and review. *Personality and Individual Differences, 47 (5)*, 385 – 395.

[17] Swann, A. C., Janicak, P. L., Calabrese, J. R., Bowden, C. L., Dilsaver, S. C., Morris, D. D., Petty, F. in Davis, J. M. (2001). Structure of mania: depressive, irritable, and psychotic clusters with different retrospectively-assessed course patterns of illness in randomized clinical trial participants. *Journal of Affective Disorders, 67 (1-3)*, 123 – 132.

[18] Swann, A. C., Steinberg, J. L., Lijffijt, M. in Moeller, F. G. (2008). Impulsivity: Differential relationship to depression and mania in bipolar disorder. *Journal of Affective Disorders, 106 (3)*, 241 – 248.

SOCIODEMOGRAFSKI IN KLINIČNI DEJAVNIKI DOŽIVLJANJA KVALITETE ŽIVLJENJA PRI OSEBAH Z BIPOLARNO MOTNJO RAZPOLOŽENJA

Tatjana Novak, Lilijana Šprah
Družbenomedicinski inštitut, Znanstvenoraziskovalni center SAZU
Novi trg 2, 1000 Ljubljana, Slovenija
Tel: +386 1 470 6445; fax: +386 1 426 1493
e-pošta: tatjana.novak@zrc-sazu.si

POVZETEK

Bipolarna motnja razpoloženja (BMR) lahko pušča neugodne in dolgotrajne posledice na posameznikovem psihosocialnem in poklicnem delovanju in blagostanju. Izpopolnitev farmakološkega in psihosocialnega zdravljenja je preusmerila klinično pozornost od obvladovanja afektivnih simptomov na kvaliteto življenja kot ene izmed mer izida motnje. Raziskave povezujejo kvaliteto življenja pri BMR z vrsto sociodemografskih in kliničnih dejavnikov. Namen pričujoče raziskave je bil proučiti, kateri sociodemografski in klinični dejavniki se povezujejo s kvaliteto življenja pri osebah z BMR, in kakšne so razlike v kvaliteti življenja glede na kontrolno skupino. V raziskavi je sodelovalo 23 oseb z BMR in 9 svojcev, ki so izpolnili kratko obliko vprašalnika kvalitete življenja The World Health Organization Quality of Life(WHOQOL-BREF). Skladno z nekaterimi raziskavami se je okrnjena kvaliteta življenja pri osebah z BMR povezovala z višjo starostjo, nezaposlenostjo in številčnostjo gospodinjstva med sociodemografskimi dejavniki in s krajšim obdobjem zadnje remisije, prisotnostjo komorbidnosti in hitrociklično obliko tekom motnje med kliničnimi dejavniki. Pri pojasnjevanju dejavnikov izida motnje se čedalje bolj uveljavljajo mere kvalitete življenja, ki lahko služijo tudi kot pomemben indikator učinkov različnih oblik zdravljenja.

1. UVOD

Bipolarno motnjo razpoloženja (BMR) opredeljujejo izmenjajoče se faze manije, depresije in remisije (okrevanja), ki pomembno prizadanejo posameznikovo psihosocialno in poklicno delovanje in blagostanje (Michalak in sod., 2005a). BMR se povezuje z visoko stopnjo napačnih diagnoz, komorbidnostjo, povečano samomorilnostjo, prezgodnjo umrljivostjo in nezaposlenostjo oz. izobrazbi neustrezno zaposlitvijo in visokimi ekonomskimi stroški tudi v evropskih državah (Fajutrao in sod., 2009).

Zaradi dolgotrajne funkcionalne okrnjenosti, pogosto doživljenjske postavitve diagnoze BMR in epizodičnega, variabilnega poteka motnje s spremljajočimi zapleti, kot sta hitrociklična oblika in različne telesne in duševne komorbidnosti, je pomembno spremljati izid motnje (Brissos in sod., 2008). Enega od indikatorjev ocene izida motnje predstavljajo vprašalniki zaznane kvalitete življenja (Brieger in sod., 2007). Svetovna zdravstvena organizacija obravnava kvaliteto življenja celostno in večplastno, ne zgolj kot odsotnost bolezni, temveč kot subjektivno zaznavo telesnega, duševnega in socialnega blagostanja, ki je vpeta v posameznikovo kulturno in družbeno okolje in se povezuje z njegovimi cilji, pričakovanji, standardi in skrbmi (WHO, 1998).

Mere kvalitete življenja prinašajo vrsto prednosti pred objektivnimi kliničnimi informacijami, saj omogočajo opredelitev posameznikove preference do določenih področij življenja, odkrivanje sprememb v kvaliteti življenja v različnih fazah motnje in subtilnih razlik na zdravljenje (Michalak in sod., 2005a, Brissos in sod., 2008). Zaradi okrnjenega uvida v lastno duševno in zdravstveno stanje, ki se pojavlja predvsem v manični fazi, so se pojavili dvomi o veljavnosti samoporočanih ocen kvalitete življenja (Gazalle in sod., 2007). V eni od raziskav se je pokazalo, da se zaznana kvaliteta življenja ni bistveno razlikovala pri osebah z oškodovanim uvidom od zaznave oseb z ohranjenim uvidom v stanju okrevanja (Dias in sod., 2008).

Zaznana kvaliteta življenja pri osebah z BMR je slabša kot v splošni populaciji (Arnold in sod., 2000, Brissos in sod., 2008) in primerljiva z okrnjenostjo pri unipolarni depresiji (Yatham in sod., 2004) in shizofreniji (Brissos in sod., 2008). Doživljanje kvalitete življenja pri osebah z BMR je slabše predvsem na področjih telesnega zdravja in socialnih odnosov in na psihičnem področju v primerjavi s kontrolami (Brissos in sod., 2008).

Kvaliteta življenja pri BMR je večja v fazi okrevanja kot v afektivnih fazah (Michalak in sod., 2005a). Na doživljanje kvalitete življenja vpliva tudi vrsta afektivne simptomatike, saj depresivni simptomi v večji meri kot manični slabšajo kvaliteto življenja (Vojta in sod., 2001, Michalak in sod., 2005b). Tudi nedavna depresivna epizoda se povezuje s slabšo kvaliteto življenja kot nedavna manična epizoda (Michalak in sod., 2005b). Med ostale klinične dejavnike, ki se povezujejo s slabšo kvaliteto življenja pri BMR, se uvrščajo pozna diagnoza BMR (Gazalle in sod., 2005), trajanje motnje (Robb in sod., 1997), telesna komorbidnost (Pirraglia in sod., 2010), starost ob prvi hospitalizaciji in hospitalizacija v zadnjem letu (Gazalle in sod., 2007). Nasprotno pa nekatere raziskave ne odkrivajo pomembnega vpliva določenih kliničnih spremenljivk, kot sta trajanje motnje in število predhodnih afektivnih epizod, na kvaliteto življenja pri BMR (Ozer in sod., 2002, Brissos in sod., 2008). Izmed sociodemografskih značilnosti se ženski spol, nižja starost, višja izobrazba, zaposlenost in urejene življenjske razmere povezujejo z večjo kvaliteto življenja pri BMR (Gazalle in sod., 2007, Brissos in sod., 2008).

2. NAMEN RAZISKAVE

Namen preliminarne raziskave je bil proučiti, kateri sociodemografski in klinični dejavniki poteka in resnosti motnje se povezujejo z doživljanjem kvalitete življenja v zadnjem mesecu pri osebah z BMR, in razlike v izraženosti kvalitete življenja glede na skupino svojcev.

2.1 Udeleženci

V raziskavi je sodelovalo 23 oseb z bipolarno motnjo razpoloženja (12 moških in 11 žensk, starost: 43,81 ± 12,03 let, izobrazba: 15,04 ± 2,07 let, zaposlitveni status: 9 zaposlenih, 4 nezaposleni, stan: 8 poročenih, 5 samskih) in 9 svojcev s podobnimi sociodemografskimi značilnostmi (3 moški in 4 ženske, starost: 33,6 ± 15,44 let, izobrazba: 15,2 ± 0,45 let, zaposlitveni status: 2 zaposlena, 3 nezaposleni, stan: 2 poročena, 3 samski). Klinične značilnosti poteka motnje za osebe z BMR so prikazane v Tabeli 1.

Tabela 1. *Spremenljivke kliničnega poteka motnje pri osebah z BMR (N=23).*

Spremenljivke kliničnega poteka motnje		
trajanje bolezni (leta)	*M*	8,62 (6,4)
pozna diagnoza BMR (leta nediagnosticirane motnje)	*(SD)*	5,15 (5,05)
število (hipo)maničnih epizod		2,9 (1,85)
trajanje (hipo)maničnih epizod (meseci)		4,57 (2,74)
število depresivnih epizod		4,64 (2,20)
trajanje depresivnih epizod (meseci)		8,59 (21,06)
število hospitalizacij		1,77 (1,54)
prisotnost psihoze	*N*	4 (17%)
prisotnost hitrociklične oblike	*(%)*	6 (26%)
komorbidnost (telesna ali psihična)		9 (39%)

Uporabljeni instrumenti

Vprašalnik kvalitete življenja WHOQOL–BREF (WHO, 1998) meri doživljanje kvalitete življenja na področjih telesnega zdravja, socialnih odnosov, povezanosti z okoljem in na psihičnem področju (WHO, 1998). Na področju telesnega zdravja se postavke nanašajo na fizično bolečino, zmožnost opravljanja dnevnih aktivnosti, kvaliteto spanca, zmožnost za delo in energetsko opremljenost. Na psihičnem področju postavke merijo sprejemanje lastnega izgleda, zadovoljstvo s seboj, pogostost doživljanja negativnih čustev, veselje do življenja, osebna prepričanja in sposobnost koncentracije. Področje socialnih odnosov se nanaša na zadovoljstvo z medosebnimi odnosi, socialno podporo in zadovoljstvo s spolnim življenjem. Povezanost z okoljem preverja finančne vire, svobodo in fizično varnost, dostop do zdravstvene oskrbe, možnosti za pridobivanje novih informacij in spretnosti, možnosti za prostočasne aktivnosti, dostop do transporta in klimo. Vprašalnik sestavlja 26 postavk, od katerih sta dve namenjeni splošni oceni kvalitete življenja in splošnemu zadovoljstvu z zdravjem. Udeleženci ocenjujejo svojo kvaliteto življenja v zadnjih štirih tednov na petstopenjski lestvici. Več točk na vprašalniku pomeni višjo zaznano kvaliteto življenja.

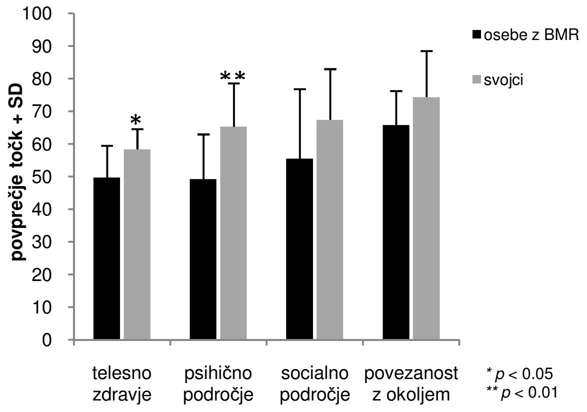
Postopek

Udeleženci so izpolnili vprašalnik WHOQOL-BREF v skupini na Centru za izvenbolnišnično psihiatrijo v Ljubljani. Sociodemografski in klinični podatki o bolezni so bili pridobljeni z intervjujem oz. retrogradno iz kartotečnih popisov bolezni, ki jih je priskrbel lečeči psihiater.

S t-testi smo merili razlike v samooceni posameznih področij kvalitete življenja med osebami z BMR in svojci. S Pearsonovimi in Spearmanovimi korelacijskimi koeficienti smo proučevali povezanost posameznih področij kvalitete življenja in sociodemografskih in kliničnih dejavnikov pri osebah z BMR in pri svojcih. Kot mero statistične pomembnosti smo uporabili $p < 0.05$. Zaradi medsebojne primerljivosti področij kvalitete življenja in primerljivosti z dolgo verzijo WHOQOL vprašalnika smo surove točke pretvorili pretvorili v transformirani skor v razponu od 0 do 100 (WHO, 1998).

3. REZULTATI

Osebe z BMR so slabše ocenile svojo kvaliteto življenja ($t(30)= -2.27$, $p= 0.03$) in so bile manj zadovoljne s svojim zdravjem ($t(30)= -3,28$, p= 0.00) v primerjavi s svojci. Glede na svojce so tudi slabše ocenile posamezne vidike kvalitete življenja, in sicer na psihičnem področju ($t(30)= -2.99$, $p= 0.01$) in področju telesnega zdravja ($t(30)= -2.49$, $p= 0.02$). Če primerjamo posamezna področja kvalitete življenja med sabo, sta bili obe skupini najbolj zadovoljni na področju povezanosti z okoljem v primerjavi s psihičnim in socialnim področjem, medtem ko so svojci za razliko od oseb z BMR bolje ocenili svojo kvaliteto življenja na psihičnem področju glede na področje telesnega zdravja ($t(8)= -2,44$, p= 0.04) (Slika 1).



Slika 1. Primerjava ocenjevanih področij kvalitete življenja med skupinama oseb z BMR in njihovih svojcev.

Med sociodemografskimi dejavniki sta zaposlenost in večja številčnost gospodinjstva najbolj pozitivno korelirali s kvaliteto življenja, medtem ko je višja starost negativno korelirala s kvaliteto življenja v skupini oseb z BMR (Tabela 2).

Med kliničnimi spremenljivkami je bilo največ pomembnih negativnih korelacij med kvaliteto življenja in prisotnostjo hitrociklične oblike tekom motnje, trajanjem depresivnih epizod in prisotnostjo komorbidnosti, medtem ko je daljše obdobje zadnje remisije pozitivno koreliralo s kvaliteto življenja (Tabela 2).

Tabela 2. *Korelacije med področji kvalitete življenja in sociodemografskimi in kliničnimi spremenljivkami pri osebah z BMR.*

SPREMENLJIVKE	PODROČJA KVALITETE ŽIVLJENJA					
	Kvaliteta življenja splošno	*Zadovoljstvo z zdravjem splošno*	*Telesno zdravje*	*Psihično področje*	*Področje socialnih odnosov*	*Povezanost z okoljem*
sociodemografske						
spol	-0.32	-0.31	-0.03	0.02	0.25	0.12
starost	**-0.61***	-0.48	-0.24	**-0.55***	-0.40	-0.26
izobrazba (leta)	0.16	-0.24	-0.26	-0.19	0.18	-0.10
stan	-.007	-0.08	0.09	0.12	0.38	0.32
zaposlitveni status	**0.62***	**0.74****	**0.59***	**0.75****	0.13	0.30
številčnost gospodinjstva	0.17	**0.44***	0.38	0.12	0.09	0.27
klinične						
trajanje motnje (leta)	-0.26	-0.47	-0.31	-0.37	-0.16	-0.05
pozna diagnoza BMR (leta nediagnosticirane BMR)	-0.26	-0.47	-0.31	-0.37	-0.16	-0.05
vrsta zadnje afektivne epizode (depresivna, manična)	0.32	0.26	0.38	0.30	0.19	0.40
trajanje remisije (meseci)	0.15	0.15	0.35	**0.59****	0.25	**0.42***
število (hipo)maničnih epizod	0.66	0.53	0.00	0.27	0.57	0.00
trajanje (hipo)maničnih epizod (meseci)	0.54	0.24	-0.33	-0.36	-0.26	-0.37
število depresivnih epizod	0.16	-0.37	-0.13	0.17	0.28	0.03
trajanje depresivnih epizod (meseci)	-0.31	-0.27	-0.26	-0.14	**-0.72***	-0.53
število hospitalizacij	-0.18	-0.25	-0.42	-0.26	0.32	0.01
prisotnost psihoze	-0.28	-0.13	0.02	0.05	-0.40	-0.01
prisotnost hitrociklične oblike	-0.15	-0.50	**-0.79****	**-0.63***	-0.29	**-0.74****
komorbidnost	-0.43	**-0.62***	0.03	-0.26	0.24	-0.13

* $p < 0.05$
** $p < 0.01$

4. DISKUSIJA

V preliminarni raziskavi smo proučevali razlike in vzorce povezanosti med področji kvalitete življenja in sociodemografskimi ter kliničnimi dejavniki pri osebah z BMR in njihovih svojcih. Skladno z raziskavami o okrnjenosti kvalitete življenja na splošno in na posameznih področjih (Yatham in sod., 2004, Michalak in sod., 2005a, Brissos in sod., 2008) so naši rezultati pokazali, da so osebe z BMR glede na svojce slabše ocenile svojo kvaliteto življenja na področju telesnega zdravja in splošnega zadovoljstva z zdravjem in na psihičnem področju.

Okrnjena kvaliteta življenja pri osebah z BMR se je povezovala z nekaterimi sociodemografskimi dejavniki, o katerih poročajo tudi druge raziskave; in sicer z višjo starostjo, nezaposlenostjo in številčnostjo gospodinjstva (Gazzalle in sod., 2007). V nasprotju z nekaterimi izsledki pa nismo odkrili povezav med kvaliteto življenja in spolom ter izobrazbo (Gazalle in sod., 2007, Brissos in sod., 2008).

V naši in v drugih raziskavah so se kot pomembni potrdili naslednji klinični dejavniki okrnjene kvalitete življenja: trajanje zadnje remisije, ki je indikator okrevanja (Michalak in sod., 2005a) in oteževalna dejavnika poteka motnje, tj. prisotnost komorbidnosti (Pirraglia in sod., 2009) in pojav hitrociklične oblike tekom BMR, ki je sicer na tem področju slabše raziskan. V nasprotju z nekaterimi izsledki nismo našli pomembnih korelacij kvalitete življenja s pozno diagnozo BMR (Gazalle in sod., 2005). Naše ugotovitve glede šibkih in nepomembnih povezav med trajanjem BMR in zgodovino depresivnih faz na eni strani in kvaliteto življenja na drugi strani podpirajo nekateri izsledki (Brissos in sod., 2008, Ozer in sod., 2002). Prav tako nismo uspeli potrditi povezanosti med zadnjo manično oz. depresivno epizodo in višjo oz. nižjo kvaliteto življenja (Vojta in sod., 2001, Michalak in sod., 2005b).

Posplošljivost rezultatov omejujejo metodološke pomanjkljivosti raziskave, ki je sicer še v preliminarni fazi, in se nanašajo na majhna in neuravnotežena numerusa primerjanih skupin, vzorčenje na podlagi motiviranosti za reševanje vprašalnika WHOQOL-BREF, pomanjkanje ocen trenutne afektivne simptomatike, ki je močan prediktor kvalitete življenja, pomanjkanje kontrole nad vplivom prejete farmakoterapije, ki vpliva na doživljanje kvalitete življenja, in na odsotnost psihiatrične ocene kliničnega statusa oseb z BMR v času reševanja vprašalnika.

Zaznana kvaliteta življenja odraža posameznikovo vrednotenje življenjskih okoliščin in predstavlja pomemben vidik pri pojasnjevanju dejavnikov izida motnje (Brieger in sod., 2007). Kljub temu, da obstaja več splošnih vprašalnikov kvalitete življenja, pa do nedavnega ni bilo vprašalnika, prilagojenega posebnim značilnostim BMR, kot je menjavanje afektivnih faz, kjer se sočasno lahko pojavljajo depresivni in manični simptomi, z obdobji okrevanja. Redkokateri vprašalnik kvalitete življenja je prilagojen specifičnim vsebinam BMR, kot so rutina, neodvisnost, duhovnost in stigma (Michalak in sod., 2005). Potrebo po specifičnem vprašalniku za BMR bi morda lahko izpolnil vprašalnik QOL-BD, ki je sicer še v fazi razvoja (Murray in Michalak, 2010).

Mere kvalitete življenja so lahko v veliko pomoč pri oceni simptomatskega in funkcionalnega okrevanja (Gazalle in sod., 2007) in nudijo širok spekter uporabe, med drugim omogočajo tudi spremljanje učinkov farmakoloških in psihosocialnih intervenc. Ugotovitve farmakoloških raziskav tako na primer kažejo, da se je po uvedbi podpornega zdravila zaznana kvaliteta življenja na večini merjenih področij dvignila (Namjoshi in sod., 2004), podobne obetavne rezultate pa dajajo tudi raziskave psiholoških intervenc, ki poročajo o večji kvaliteti življenja na področjih telesnega zdravja (Michalak in sod., 2005b), socialnih odnosov in splošnega zadovoljstva z zdravjem (Dogan in sod., 2003) po končani psihoedukaciji pri osebah z BMR.

5. ZAKLJUČEK

Kljub zelo omejeni posplošljivosti naši rezultati kažejo, da je zaznana splošna kvaliteta življenja in njena bolj specifična ocena na dveh merjenih področjih slabša pri osebah z BMR v stanju relativne remisije glede na njihove svojce. Zaposlitveni status in prisotnost zapletov tekom motnje (hitrociklična oblika) sta se izkazala za dejavnika, ki korelirata z največ področji kvalitete življenja pri osebah z BMR. Mere kvalitete življenja obravnavajo izid motnje tudi s subjektivne perspektive, ki posamezniku omogoča vpogled v različna področja njegovega življenja in vpliva na njegovo sodelovanje v procesu zdravljenja. Pokazala se je tudi potreba po vprašalniku kvalitete življenja, prilagojenem specifičnim značilnostim BMR.

** Študija je bila opravljena v sklopu raziskovalnega projekta L3-9698-2485-06, ki sta ga sofinancirala ARRS in Astrazeneca.*

6. LITERATURA

[1] Arnold LM, Witzeman KA, Swank ML, McElroy SL, Keck PE (2000). Health-related quality of life using the SF-36 in patients with bipolar disorder compared with patients with chronic back pain and the general population. Journal of Affective Disorders, 57; 235-239.

[2] Brieger P, Roetting S, Roettig D, Marneros A, Priebe S (2007). Dimensions underlying outcome criteria in bipolar I disorder. Journal of Affective Disorders, 99; 1-7.

[3] Brissos S, Dias VV, Carita AI, Martinez-Aran A (2008). Quality of life in bipolar type I disorder and schizophrenia in remission: Clinical and neurocognitive correlates. Psychiatry Research, 160; 55-62.

[4] Dias VV, Brissos S, Carita AI (2008). Clinical and neurocognitive correlates of insight in patients with bipolar disorder I in remission. Acta Psychiatrica Scandinavica, 117; 28-34.

[5] Dogan S, Sabanciogullari, S (2003). The effects of patient education in lithium therapy on quality of life and compliance. Archives of Psychiatric Nursing, 17; 270-275.

[6] Fajutrao L, Locklear J, Priaulx J, Heyes A (2009). A systematic review of the evidence of the burden of bipolar disorder in Europe. Clinical Practice and Epidemiology in Mental Health, 5:3; 1-19.

[7] Gazalle FK, Frey BN, Hallal PC, Andreazza AC, Cunha ABM, Santin A, Kapczinski F (2007). Mismatch between self-reported quality of life and functional assessment in acute mania: A matter of unawareness of illness? Journal of Affective Disorders,103; 247-252.

8] Gazalle FK, Andreazza AC, Cereser KM, Hallal PC, Santin A, Kapczinski F (2005). Clinical impact of late diagnose on bipolar disorder. Journal of Affective Disorders, 86; 313-316.

[9] Michalak EE, Yatham L, Lam, RW (2005a). Quality of life in bipolar disorder: A review of the literature. Health and quality of life outcomes, 3: 72; 1-17.

[10] Michalak EE, Yatham LN, Wan DC, Lam RW (2005b). Perceived quality of life in patients with bipolar disorder. Does group psychoeducation have an impact? Canadian Journal of Psychiatry, 5:2; 95-100.

[11] Murray G, Michalak EE (2010). The quality of life in bipolar disorder (QoL.BD) scale: Development and psychometric properties.(abstract). Journal of Affective Disorders, 122; S36-S37.

[12] Namjoshi MA, Risser R, Shi L, Tohen M, Breier A (2004). Quality of life assessment in patients with bipolar disorder treated with olanzapine added to lithium or valproic acid. Journal of Affective Disorders, 81; 223-229.

[13] Ozer S, Ullusahin A, Batur S, Kabakci E, Saka MC (2002). Outcome measures of interepisode bipolar patients in Turkish sample. Social Psychiatry and Psychiatric Epidemiology, 37; 31-37.

[14] Pirraglia PA, Biswas K, Kilbourne AM, Fenn H, Bauer MS (2009). A prospective study of the impact of comorbid medical disease on bipolar disorder outcomes. Journal of Affective Disorders, 115; 355-359.

[15] Vojta C, Kinosian B, Glick H, Altshuler L, Bauer MS (2001). Self-reported quality of life across mood states in bipolar disorder. Comprehensive Psychiatry, 42; 190-195.

[16] WHO (1998). WHOQOL User Manual. Ženeva, Svetovna zdravstvena organizacija (WHO).

[17] Yatham LN, Lecrubier Y, Fieve RR, Davis KH, Harris SD, Krishnan AA (2004). Qouality of life in patients with bipolar I depression: data from 920 patients. Bipolar Disorders, 6; 379-385.

INSTITUTIONALISING OUT-OF-SCHOOL EDUCATION

Josip Stepanić[1], *Kornelija Špoler Ćanić*[2], *Dubravka Rasol*[2]
[1]Faculty of Mech.Eng. & Naval Arch. – University of Zagreb
I. Lučića 1, HR-10000 Zagreb, Croatia
Tel: +385 1 6168592; fax: +385 1 6168 466
e-mail: josip.j.stepanic@fsb.hr
[2]Meteorological and Hydrological Service
Grič 3, HR-10000 Zagreb, Croatia
Tel: +385 1 4565666; fax: +385 1 4851901
e-mails: spoler@cirus.dhz.hr

ABSTRACT

Number of fields for which additional, out-of-class education activities for children and youth is useful and needed, is rather large. These activities differ in content, structure, space and time considerations, prerequisites imposed to participants etc. In order to enlarge the availability of underlying resources on the one hand, and to keep their sustainability at the sufficient level, combining of these activities is needed.

This paper describes educational activities in advanced technologies and robotics for middle-school pupils in Croatia. Activities are structured, their perspectives discussed. As an additional topic, in particular, the needs for and possibilities of their institutionalisation in a science museum are described and compared.

1 INTRODUCTION

Educational activities for children and youth contribute to their optimal personal development, to appropriate realisation of their interests and with that related preparatory phase of their future professional development. Examples of such activities are found within a broad range of out-of-class activities, ranging from additional classes in schools to international meetings and competitions.

Educational activities for children and youth are institutionalised in different amount. Some are physically conducted within schools and universities, others are conducted in specially formed institutions. Examples of later type include museums which, along with their cultural and social dimension contribute to education. In particular, the science museums are sustainable institutions which contribute in a particular way to education.

Since number of topics of interest to children and youth is rather large, every child and young person has to choose one or few of the educational activities (of course, if locally more of them are available).

Activities related to technology and several disciplines of science usually require a definite quantity of available equipment, rather specialised and well-educated trainers or educators, which finally enlarge their financial dimension to a value above average for educational activities.

Furthermore, participants' costs and efforts involved rise if there are several places in which activities could be performed in different and uncorrelated times.

Overall, enlarging of costs lessen possibility for further development and broadening of number and types of educational activities related to technology and science.

In the second section a scientific museum is described in more details, and in third section we present two case studies and argue that they converge toward a science museum. Fourth section analyses elements of games and contribution of games in functioning of a science museum. Fifth section briefly summarises previous statements.

2 SCIENCE MUSEUM

Science museum is well-known answer to questions and problems listed in introduction. On the one hand it collects diverse, yet related exercises, optimises availability and expenses of mentorship, maintenance, and time use of participants. Furthermore, science museums throughout the world usually have thorough multimedia and interactive components. Their content is rather dynamic and spans several levels of knowledge required for understanding. Their functioning is on commercial base which influences list of their functions. However in most cases, at least partially science museum obtain local, regional or national funds predicted to education.

Science museums therefore represent institutionalised form for out-of-school education. Institutionalisation brings about their constant presence in society and thereby related possibility for close relating their content with school programme.

In other words, science museum contain critical mass of experience, possibilities and knowledge for successful education of groups of participants in diverse, elementary or advanced topics.

Along with that facts regarding their content and education in a given set of disciplines, science museum contribute toward social cohesion of interested participants, especially in a form of coordinated group activities of participants.

How did any science museum emerge? Broadly, one way is top-down, i.e. their establishing by local, regional or national governances. Another way is gradual development of activities held on smaller scales, such as temporary or non-constant periodic activities, for smaller regions with smaller quantity of available equipment and smaller accompanied multimedia service.

3 EXAMPLES OF EDUCATIONAL ACTIVITIES

Let us briefly describe two educational, not-for-profit activities currently held in Croatia. Their differences will be clearly stated. Their combination will be addressed to as a separate entity and its, somewhat emergent, characteristics, will be emphasised in this section.

3.1 SMALL METEOROLOGICAL WORKSHOP

Small Meteorological Workshop (SMW) is educational activity for elementary school pupils. In SMW children learn about elements of our atmosphere and our climate, using notions and experiments appropriate for their age, developed skills and previous experience about the weather. The workshop serves furthermore as a starting point for description of sciences and development of an active relation toward observed facts.

The SMW has been held for the last four years in different parts of Croatia. The plan is to broaden its activities. The plan is based on existing experience gathered from previous realisations, existing human and material resources and expected collecting the funds necessary for transport and maintenance. However, broadening of the SMW scope is accompanied with riding the costs, which is treated as a rather serious issue.

3.2 LEAGUE OF TEAMS

The League of Teams (LoT) is educational activity for middle-school youth. In it, participants work with mobile robots throughout the school year. Topic of that activity is mobile robotics and underlying mehatronics. Participants are grouped into teams, each of which represents one school. Groups gather on several regional and national meetings and tournaments. Groups are led throughout the school year in their preparation for meetings using problems which teams have to solve for meetings. Interactive communication with organisers is achieved using online resources between meetings, and with public presentations of teams' solutions and moderated discussions during meetings. That activity is planned to be broadened with inclusion of additional schools, and with enabling youth to form several or more teams within one school.

3.3 COMPARISON

Let us compare SMW and LoT in some detail. Many of their differences are manifest:

- SMW is predicted for children in elementary schools and LoT for middle-school youth
- SMW contains elementary level exercises while LoT contains elementary and advanced level exercises,
- SMW subject belongs to applied and fundamental sciences, while LoT subject belongs to technology, etc

However, they have substantial similarities:

- they are out-of-class, yet in-school educational activities,
- they assume longer-in-time involvement of participants and their teachers,
- they contain topics which are related to school programme, yet which is not regularly included in it,
- they are applicable to same pupils in their different ages: while SMW serves as an introduction toward advanced scientifically and technologically related programmed, LoT represents similarly introductory step toward advanced technologies and expressions of technical creativity.

Last but not least, both these activities are held in a formal way by a not-for-profit, non-governmental organisation, rather non-institutionalised, and without long-term funding agreed upon.

All stated significantly influences scopes and ranges of activities held, as well as of activities planned.

In one approach, these activities can be viewed as initial base for further development, which could lead to formation of science museum, if development of existing activities, their broadening and linking to other educational activities is performed in a long-term.

4. INVOLVED ELEMENTS OF GAMES

Games are type of structured activities which can be used in education. We concentrate here on the fact that games are characterised with the set of rules, interaction among interested parties and challenge. Games can be individual or team activities. If properly learned and repeated, games can contribute to psycho-physical development of involved participants. Development includes rise in level of specific individual skills as well as development of level of social skills.

Science museums contain educational character which is rather useful and rather easily combined with the notion of challenges, intrinsic part of games. In particular, if prepared exercises are put to participants (youth and children) as challenges, we obtain one element of individual games. Furthermore, if rules are defined as facts from sciences which are learned in regular school programme, then all participants in science museums' challenges know (at least most of the) rules in advance. At last, during individual challenges, there is constant interaction with the equipment and other resources of the science museum. If records are held, then we obtain asymmetric, asynchronous interaction of a participant with the previous and future participants through the list of records and other related lists.

Overall, interactive experiments in science museums are specific type of educational games.

Similar considerations can be applied onto team activities. They similarly as individual educational games, represent educational games. The difference is that team activities more thoroughly develop social activities and contribute to early phase of networking

Finally, owing to the media activities of science museums and their relations with other educational institutions – presumably schools – the spatial extent of challenges and interaction can be broadened to include whole region which gravitates to science museums. Similarly, duration of challenges and interaction can be significantly extended, thereby.

The extensions of educational games from science museums onto broader region for a longer period of time contributes significantly to the level of motivation of children and youth as well as to their achieved level of understanding of "rules of the game", i.e. class topics from regular school programmes.

The larger the number of participants, one the one hand the smaller the time and funding requirements per participant and, on the other hand, the larger requirements for organisation.

5. SUMMARY AND CONCLUSIONS

Constant and intense development of science and technology asks for constantly held, yet constantly modified, out-of-class and out-of-school educational activities. It is opportune to institutionalise them in a science museum, an institution which combines cultural and educational function. Such an approach is illustrated in the realistic situation of educational activities organised by a not-for-profit, non-governmental society.

Games are important part of successful attracting of children and youth to the science museums and their programmes, as well as to the optimal rise of skills and specific education of participating children and youth.

6. ACKNOWLEDGMENT

Authors acknowledge the resources of the Croatian Interdisciplinary Society they were provided with during preparation of this paper.

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

Robotika

Robotics

Uredili / Edited by

Jadran Lenarčič, Leon Žlajpah, Andrej Gams

http://is.ijs.si

15. oktober 2010 / October 15th, 2010
Ljubljana, Slovenia

PREDGOVOR

Po okriljem multikonference »Informacijska družba« drugič zapored organiziramo tudi konferenco Robotika, s katero nadaljujemo dolgo tradicijo raziskovalne robotike v Sloveniji. Kljub finančno in raziskovalno zahtevnem področju raziskovanja se je uveljavilo več raziskovalnih skupin, ki ne samo, da sledijo trendom sodobne robotike, temveč jih so-ustvarjajo.

Robotika se je v zadnjih desetletjih utrdila v številnih industrijskih procesih kot nepogrešljiv del moderne, ekonomične in človeku prijazne tehnologije. Brez robotskih manipulatorjev si ne znamo več predstavljati varjenja avtomobilskih ohišij, vstavljanja obdelovancev v stiskalnice, razpršilnega barvanja. Ne presenečajo nas kirurški roboti ali servisni mobilni roboti, ki bodo kmalu čistili naša stanovanja. Vse večje število kvalitetnih mednarodnih robotskih strokovnih revij pa dokazuje, da postaja robotika priznana tudi kot samostojna znanost.

Sodobna robotika izrazito presega okvire industrijskih robotskih manipulatorjev. Označili bi jo lahko kot »inteligentno gibanje mehanizmov«. Sem sodijo tako mobilni roboti, hodeči roboti, roboti v medicini, servisnih robotov, ki nam bodo v pomoč v vsakdanjem življenju in pa raznovrstne industrijske naprave, ki so kakorkoli povezana z gibanjem: na primer dvigala, podajalniki. Eden izmed glavnih ciljev moderne robotike je razvoj metod in znanj, ki bodo omogočila prenos robotskih tehnologij iz tovarniških v domača okolja. V našem domačem okolju lahko roboti postanejo izvrstni pomočniki na več področjih, od pomoči ostarelim do pomočnikov pri zahtevnih in vsakodnevnih opravilih. V ta namen moramo usposobiti robote za avtonomno delovanje v nepričakovanih situacijah.

Prvi prispevek predstavlja izrazito industrijsko aplikacijo, ki nakazuje na interdisciplinarnost uporabe robotike, v tem primeru povezane s tehnološkim procesom varjenja.

Temu sledi drugi sklop prispevkov, kjer je tematika humanoidna robotika, ki je eno izmed prednostnih raziskovalnih področij sodobne robotike. Prispevki pokrivajo teme od posploševanja gibanja, kot na primer seganje povezano z robotsko hojo, izvajanja metov, kjer je pokazano posploševanje hitrih gibov, pa do vključevanja človeka v zaprtozančne sisteme vodenja robotov.

Posebno pozornost pa v zadnjem času izkazujejo rehabilitacijski roboti, ki se vse več uporabljajo kot »pametni« funkcionalni rehabilitacijski pripomoček. Rehabilitacijski robot lahko razbremeni fizioterapevte težkega fizičnega dela, hkrati pa poskrbi za učinkovito terapijo poškodovanca. To je tudi tematika zadnjega prispevka, ki opisuje uporabo robotike za rehabilitacijo pacientov po možganski kapi.

Eden izmed namenov te robotske konference pa je, da vzpodbudi interdisciplinarno debato med znanstveniki, ki se ukvarjajo z različnimi področji robotike. Letošnji prispevki predstavljajo različne inovativne raziskave sodelavcev treh skupin na katerih temelji prihodnost raziskovalnega dela na
področju robotike pri nas.

Jadran Lenarčič
Leon Žlajpah
Andrej Gams

Problematika robotskega laserskega kaljenja pri conah prekrivanja

Matej Babič
Emo-Orodjarna d.o.o.
Bežigrajska 10, 3000 Celje, Slovenija
e-mail: matej.babic@emo-orodjarna.si

Povzetek

Prispevek obravnava robotsko lasersko kaljenje kovin s prekrivanjem kaljene cone. Opisuje rezultate dosedanjega dela, raziskav in pridobljenih izkušenj na robotskem laserskem kaljenju kovin. Zanima nas trdota kovine po kaljenju že kaljene kovine. Na koncu so zbrani rezultati.

Ključne besede: razvoj, kaljenje, robot, laser,

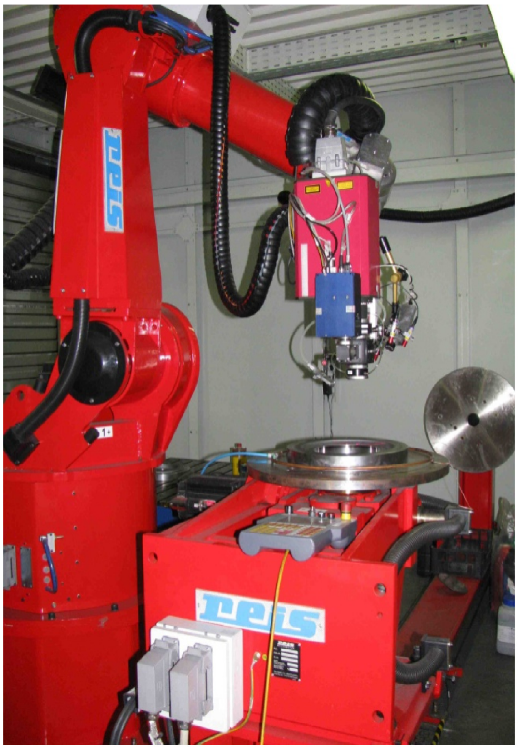
Abstract

This paper deals with robotic laser hardening of metals by coating the hardened zone. It describes the results of previous work, research and lessons learned from the robotic laser quenching metal. We are interested in quenching hardness of metal under the already hardened metal. Finally, the results are gathered.

Key words: development, hardening, robot, laser,

1 UVOD

V prispevku je opisan postopek robotskega laserskega kaljenja kovin. Predstavljena je problematika robotskega laserskega kaljenja pri conah prekrivanja. Zanimalo nas je, kakšna je trdota kaljenega materiala med conami prekrivanja pri različnih širinah. Eksperimentalno delo je potekalo na materialu OCR in na orodnem jeklu standardne oznake po ISO standardu 1.7225. Narejeni so bili poizkusi z optiko 23 mm. Meritve trdote so zbrane v tabelah. Na koncu je podana primerjava med kovinama OCR in orodnem jeklu standardne oznake po ISO standardu 1.7225. V zaključku je predstavljen še nov odprti problem robotskega laserskega kaljenja.



Slika 1: Robotska laserska celica

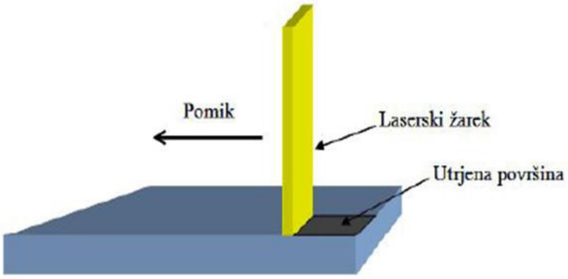
2. PREDSTAVITEV ROBOTSKEGA LASERSKEGA KALJENJA

Robotsko lasersko kaljenje je površinska toplotna obdelava. Laserski žarek je vir energije pri robotskem laserskem kaljenju, kateri izredno hitro segreje površino kovine in s tem poveča njeno trdoto. Robotsko lasersko kaljenje se uporablja pri strojnih delih za torzijske vzmeti, zobnike, obdelovalne matrice, rezilne robove v strojni, vojaški in tudi v letalski industriji. Robotsko lasersko kaljenje in precej prednosti pred induktivnim kaljenjem. Te so:

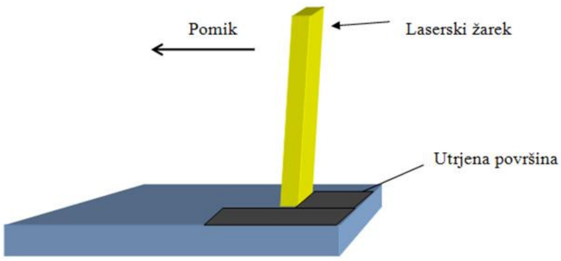
- manjši stroški obdelave,
- visoka natančnost in prilagodljivost,
- kaljenje kompliciranih geometrijskih oblik,
- izjemna obrabna obstojnost kaljene površine,
- visoka fleksibilnost,…

3 PREDSTAVITEV PROBLEMATIKE

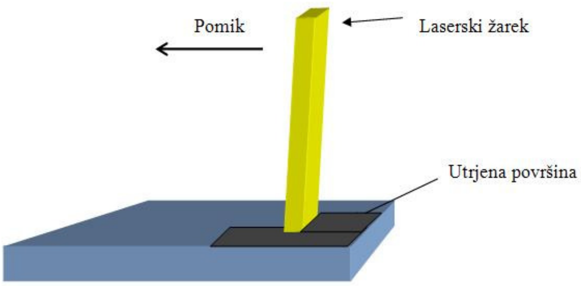
Pri robotskem laserskem kaljenju hitro opazimo najrazličnejše težave. V tem primeru nas je zanimala trdota materiala (OCR). Najprej smo material kalili. Nato pa smo isti material ponovno kalili tako, da smo se najprej približali kaljeni površini, potem pa smo kalili preko kaljenega dela materiala. Napravili smo več testov in meritev. Na sliki 2 je prikazan postopek kaljenja. Na sliki 3 vidimo kaljenje točno zraven kaljene cone. Na sliki 4 pa vidimo kako laserski žarek prekriva kaljeni del materiala. Tukaj smo upoštevali več možnosti. Testirali smo prekrivanje kaljene cone za 2 mm, 3 mm, 4 mm, 6 mm in 10 mm.



Slika 2: Kaljenje pravokotno na material



Slika 3: Kaljenje materiala zraven kaljene cone



Slika 4: Kaljenje materiala preko klaljene cone

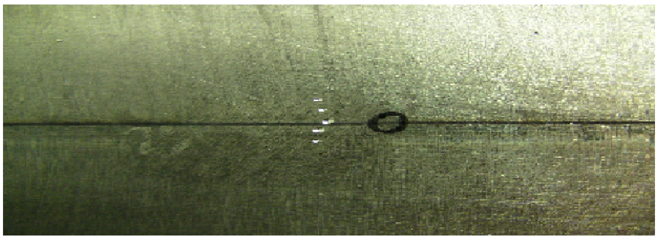
4 EKSPERIMENTALNO DELO

4.1 Material OCR

Zanima nas kako trdoto lahko dosežemo pri kaljenju materiala, če kalimo do kaljene cone in čez njo. Izbrali smo material OCR z osnovno trdoto 7 HRC. Vsi podatki so zbrani v spodnjih tabelah. Trdnost materiala pred in po kaljenju smo izmerili z namiznim merilcem za merjenje trdote WPM. V tabeli 1 so zbrani podatki ko smo laserski žarek približali natančno zraven kaljene cone. V tabeli 1 so zbrani rezultati meritev trdote, ko se kaljeni coni stikata. V tabelah 2, 3, 4, 5 in 6 so zbrani podatki, ko laserski žarek prekriva kaljeno cono.

Tabela 1: Trdnost materiala, ko se coni kaljenja dotikata

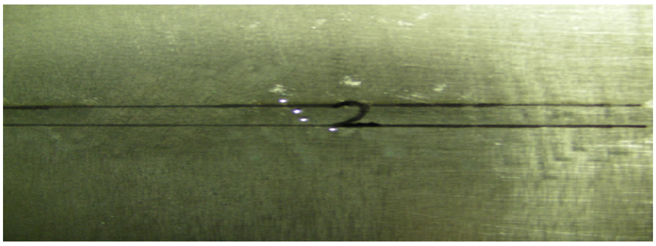
Prekrivanje (mm)	Trdnost (HRc)	smer
0	59,59,57,55,57,50,42,15	desno
0	58,58,55,54,53,52,44,15	levo



Slika 5: Kaljeni coni se dotikata

Tabela 2: Trdnost materiala, ko se coni kaljenja prekrivata 2 mm

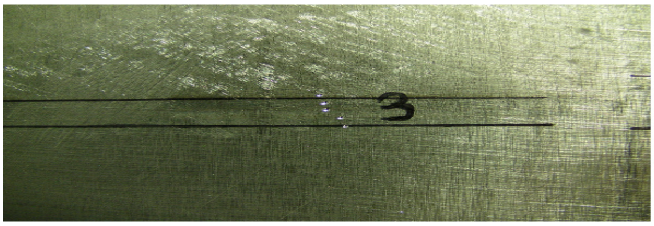
Prekrivanje (mm)	Trdnost (HRc)	smer
2	58, 55, 50, 45, 31, 12	desno
2	57, 56, 54, 51, 46, 36	levo



Slika 6: Kaljeni coni se prekrivata 2mm

Tabela 3: Trdnost materiala, ko se coni kaljenja prekrivata 3 mm

Prekrivanje (mm)	Trdnost (HRc)	smer
3	56, 54, 52, 50, 43	desno
3	52, 48, 36, 14, 18	levo



Slika 7: Kaljeni coni se prekrivata 3mm

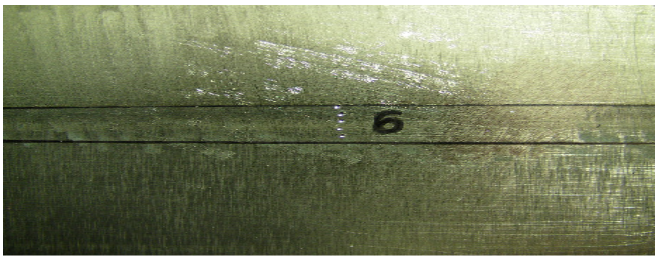
Tabela 4: Trdnost materiala, ko se coni kaljenja prekrivata 4 mm

Prekrivanje (mm)	Trdnost (HRc)	smer
4	55, 52, 50, 47, 40, 30	desno
4	57, 54, 47, 41, 29	levo

Slika 8: Kaljeni coni se prekrivata 4mm

Tabela 5: Trdnost materiala, ko se coni kaljenja prekrivata 6 mm

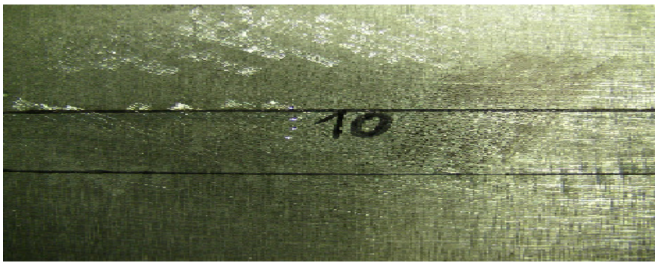
Prekrivanje (mm)	Trdnost (HRc)	smer
6	57, 53, 48, 42, 39	desno
6	56, 54, 48, 43	levo



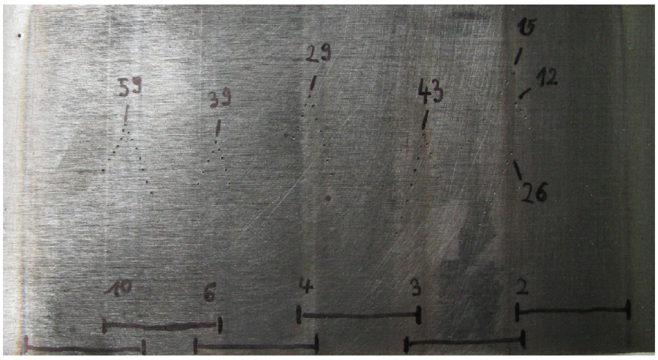
Slika 9: Kaljeni coni se prekrivata 6mm

Tabela 6: Trdnost materiala, ko se coni kaljenja prekrivata 10 mm

Prekrivanje (mm)	Trdnost (HRc)	smer
10	55, 52, 50, 53, 59, 60, 60	desno
10	61, 61, 62, 62, 62, 61, 59	levo



Slika 10: Kaljeni coni se prekrivata 10mm



Slika 11: Kaljenje materiala preko kaljene cone

Na sliki 12 je prikazana še globina kaljenega materiala OCR. S slike je razvidno, da je kaljena površina materiala OCR kar precej globoka, kar omogoča boljšo obrabno obstojnost kaljene površine.

Slika 12: Globina kaljenega materiala OCR

4.2 Material 1.7225

Teste smo napravili še na orodnem jeklu standardne oznake po ISO standardu 1.7225. Trdota materiala 1.7225 je 6,5 HRc. Meritve smo napravili na mehki coni. V tabelah so prikazani rezultati merjenja. Najprej v tabeli

7 meritve trdote po kaljenju brez prekrivanja. Meritve smo izvajali od mehke cone proti sredini kaljene cone.

Tabela 7: Trdnost materiala brez prekrivanja

Trdnost (HRc)	smer
7, 10, 49, 57, 57	Od mehke cone proti sredini kaljene cone

V tabeli 8 so predstavljeni podatki merjenja trdote, ko se kaljeni coni dotikata (prekrivanje 0mm) in ko se kaljeni coni prekrivata 1, 2, 3, 4, 6, 10 mm.

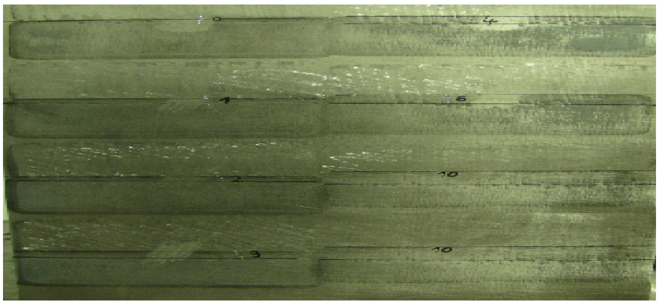
Tabela 8: Trdnost materiala, ko se coni kaljenja prekrivata 0, 1, 2, 3, 4, 6, 8, 10 mm

Prekrivanje (mm)	Trdnost (HRc)
0	55, 43, 10, 9, 38
1	53, 30, 10, 45, 57
2	51, 14, 26, 57
3	52, 28, 22, 53, 57
4	41, 22, 52, 56
6	45, 32, 43, 54, 57
10	52, 33, 51, 56

5 REZULTATI

Iz podatkov, ki so zbrani v tabelah lahko opazimo optimalno trdoto kaljenega materiala pri različnih conah prekrivanja. Pri vsakem vzorcu smo uporabili optiko 5×23 F200. Temperatura laserskega robotskega kaljenja je bila 1100°C. Pri hitrosti 2mm/s smo vsak vzorec lasersko kalili 20s. Lahko primerjamo podatke materiala OCR in 1.7225. Iz tabel lahko vidimo, da so podatki pri obeh materialih OCR in 1.7225 dokaj podobni. Če pogledamo podatke pri laserskem kaljenju, ko se coni kaljenja dotikata, opazimo trdoto med 15 HRc in 59 HRc. Dosežena trdota pri prekrivanju kaljene cone med 2 HRc, 3 HRc, 4 HRc in 6 Hrc opazimo podobne rezultate. Trdot se giblje med 30 HRc in 58 HRc. Zanimivi so podatki pri prekrivanju kaljene cone 10 mm. Iz tabele 6 lahko opazimo, da dobimo ravno pri tej opciji najboljše rezultate. Tukaj smo izmerili trdoto med 59 Hrc in 62 HRc.

Na sliki 13 so zbrani vsi vzorci OCR robotskega laserskega kaljenja pri coni dotika (prekrivanje 0mm) in pri conah prekrivanja 1mm, 2mm, 3mm, 4mm, 6mm in 10mm.



Slika 13: Vsi vzorci robotskega laserskega kaljenja pri conah prekrivanja

6 ZAKLJUČEK

Robotsko lasersko kaljenje je zelo zanimivo in uporabno. V prispevku smo videli, da pa tudi pri robotskem laserskem kaljenju nastopajo težave in problematika, ki ima praktično uporabo. Ugotovili smo kako lahko širina kaljene cone prekrivanja laserskega kaljenja vpliva na trdoto kaljenega materiala. V prihodnosti nas še zanimajo različni problemi s katerimi se srečamo pri robotskem laserskem kaljenju. Ena izmed problematik je dvožarkovno robotsko lasersko kaljenje, kjer lahko laserski žarek razdelimo na dva dela. Pri tem pa nastopijo določene težave. Tako nas v nadalje čaka še nekaj odprtih problemov.

7 LITERATURA

[1] George E., Totten, Ph.D., FASM, Steel heat treatment, second edition, 2006.

[2] J. Grum, P. Žerovnik, R. [turm: Measurement and Analysis of Residual Stresses after Laser Hardening and Laser Surface Melt Hardening on Flat Specimens; *Proceedings of the Conference "Quenching '96"*, Ohio, Cleveland, 1996.

[3] H. W. Bergman: Current Status of Laser Surface Melting of Cast Iron, *Surface Engineering*, 1, 1985.

[4] Daley, W.W. Laser processing and analysis of material. Chapter 1: Laser and Laser radiation Plenum Press: New Yourk, 1983, 158-162.

[5] Gregson, V.G. Chapter 4: Laser and heat treatment. In Laser Materials Procesing. Bass, M., Ed.; Materials Procesing Theory and Practices; North-Holland Publishing Company: Amsterdam, 1983, Vol. 3, 201-234.

Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad

Operacijo je delno financirana iz evropskega socialnega sklada.

UPORABA GAUSSOVE REGRESIJE PRI STROJNEM UČENJU PRIJEMANJA

Denis Forte, Aleš Ude
Odsek za avtomatiko, biokibernetiko in robotiko
Institut »Jožef Stefan«
Jamova 39, 1000 Ljubljana, Slovenija
Tel: +386 1 4773419; fax: +386 1 4251038
E-pošta: denis.forte@ijs.si

USING GAUSSION PROCESS REGRESSION FOR GRASPING WITH MACHINE LEARNING

Intelligent robots cannot be programmed in advance for all possible situations, but they should be able to generalize based on the acquired knowledge. In robot learning based on imitation of human activity we often use statistical methods that generalize observed (learned) movements. The acquired data is used to generate useful robots responses in situations for which the robot has not been specifically instructed how to respond.

The paper describes robot learning with Gaussian process regression that creates the model and estimates the parameters for generalization of the acquired motor knowledge, which is accumulated as a database of example movements. New actions are synthesized by applying Gaussian process regression, where the goal and other characteristics of an action are utilized as queries to create an optimal control policy with respect to the previously acquired knowledge.

The paper demonstrates that the proposed methodology can be integrated with an active vision system of a humanoid robot. 3D vision data is used to provide query points for statistical generalization. We also include robot walking in our method.

1 UVOD

Strojno učenje je področje umetne inteligence za razvoj tehnik, ki omogočajo računalnikom oz. strojem (robotom) pridobivanje raznih znanj. Strojno učenje se močno opira na statistiko, saj se tudi statistika ukvarja z obdelavo in s posploševanjem podatkov.

Nadzorovano učenje je princip strojnega učenja, ki modelira funkcije na podlagi učne množice vzorcev. Naloga učečega sistema je, da posploši znanje, ki ga dobi iz učne množice, na pravilen način. Rezultat nadzorovanega učenja je v večini primerov nek globalen model funkcije, ki preslika vhodne podatke v nek želeni izhod.

Med različnimi metodami, ki jih lahko uporabimo za posploševanje pridobljenih znanj, se je Gaussova regresija (Gaussian process regression – GPR) izkazala za zelo učinkovito. GPR je Bayesova regresijska metoda, ki napove razpršenost rezultatov. Predvidene porazdelitve se lahko uporabi za merjenje negotovosti ocenjene funkcije. Dokazano je, da ta tehnika prekaša druge metode regresije pri problemih, kot so ocenjevanje inverzne dinamike robotske roke s sedmimi prostostnimi stopnjami [1].

Naša naloga je bila programirati humanoidnega robota HOAP-3, da s pomočjo Gaussove regresije posplošuje gibanje desne in leve roke od začetne lege do katerekoli točke v njegovem delovnem območju in tako seže za predmetom ter ga zgrabi. Če se predmet nahaja izven delovnega območja robota, bo HOAP-3 uporabil robotsko hojo, da se mu približa.

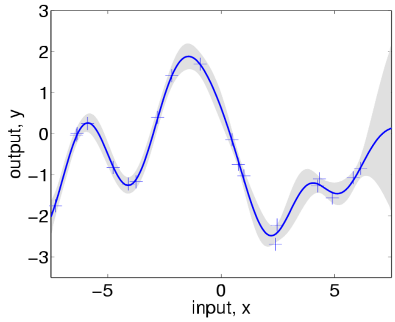
Algoritmu smo podali 140 predhodno naučenih trajektorij seganja s končnimi točkami, enakomerno porazdeljenimi po celotnem delovnem območju desne roke, ki znaša približno 30x30x10 cm. Končne točke so bile med seboj oddaljene okrog 5 cm. Vrh roke je moral biti v vidnem polju robotovih kamer. Izkazalo se je, da ponekod tudi vidno področje robota omejuje delovni prostor. Trajektorije so bile ustvarjene s fizičnim pomikom desne roke robota in merjenjem kotov štirih sklepov na desni roki, trajektorije za levo roko pa smo priskrbeli tako, da smo kotom dene roke le spremenili predznake, saj to omogoča simetrija sklepov robota HOAP-3. S posploševanjem lahko tako robot izvede gib desne ali leve roke (ki ga predhodno nismo naučili) z enako obliko trajektorije, kot jo imajo naučeni gibi.

Vsako posplošeno trajektorijo smo predstavili z DMP-ji, ki imajo to lastnost, da konvergirajo proti trenutni ciljni točki, tudi če med gibanjem robotske roke cilj spremenimo [2]. Tako lahko robot z dvema kamerama locira 3D-pozicijo objekta in seže proti njemu, če se predmet nahaja v njegovem delovnem območju rok. Če med gibanjem roke predmet spremeni pozicijo, bo dlan zapustila osnovno posplošeno trajektorijo in začela slediti predmetu.

2 GAUSSOV REGRESIJSKI PROCES

Gaussovi procesi so primer verjetnostnega modeliranja z verjetnostnim napovedovanjem izhodov. Uporaba Gaussovih procesov izvira iz področja statistike in temelji na Bayesovem verjetnostnem modeliranju. Modeli na osnovi Gaussovih procesov imajo zanimivo in za vodenje uporabno lastnost, da model poleg vrednosti izhoda napoveduje tudi zaupanje v to vrednost. To je pomembno, saj pove v kolikšni meri lahko zaupamo predikciji modela za določen vhodni signal.

Če izberemo nekaj poljubnih vhodov in jim definiramo izhode (slika 1), bo Gaussova regresija posplošila te podatke, ustvarila funkcijo med njimi in ocenila varianco te funkcije. Največja varianca je v območjih z manj testnimi točkami in tam, kjer so testne točke bolj oddaljene od aproksimacije (slika 1). Ponavljanje eksperimenta z več testnimi točkami razporejenimi po širšem območju vodi do natančnejšega ocenjevanja vmesnih vrednosti pri posploševanju.



Slika 1: *Največja varianca je v območjih z manj testnimi točkami in tam, kjer so testne točke bolj oddaljene od aproksimacije.*

Gaussov proces je naključen večdimenzionalen proces, katerega naključne spremenljivke x_i so določene z vektorjem srednjih vrednosti $\mu(X)$ in kovariančno matriko K ter medsebojno porazdeljene po normalnem porazdelitvenem zakonu. K modeliramo s kovariančno funkcijo $C(x_i,x_j)$, ki izraža korelacijo med posameznima izhodoma $f(x_i)$ in $f(x_j)$, glede na vhoda x_i in x_j. Kovariančna funkcija je lahko katerakoli funkcija, ki tvori nenegativno definitno kovariančno matriko K za katerikoli nabor vhodnih vektorjev X. Ponavadi se izbere taka kovariančna funkcija, da so točke, ki so bližje skupaj v vhodnem prostoru, tudi močneje korelirane. Pri predpostavki o stacionarnosti procesa je najpogosteje uporabljena Gaussova kovariančna funkcija, kjer element K_{ij} kovariančne matrike K izračunamo kot [1]:

$$K_{ij} = C(x_i, x_j) = \sigma_f^2 \sum_{d=1}^{D} \exp\left(-\frac{1}{2l_d^2}\left(x_i - x_j\right)^2\right) + \sigma_n^2 \delta_{ij} \tag{1}$$

Enakovredna vektorska oblika zapisa je:

$$C(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{1}{2}(x_i - x_j)^T W^{-1}(x_i - x_j)\right) + \sigma_n^2 \delta_{ij},$$

kjer je
$$W^{-1} = \begin{bmatrix} l_1 \ddots & 0 \\ 0 & \ddots l_D \end{bmatrix}. \tag{2}$$

Parametri kovariančne funkcije (σ_f, σ_n, l_1, … l_D), imenovani hiperparametri, so določeni (naučeni) na podatkih učne množice. D je dolžina vhodnega vektorja x oz. njegova dimenzija. Da dana kovariančna matrika ustreza pogojem nenegativne definitnosti, morajo biti vsi hiperparametri večji od nič. Parameter σ_f skalira velikost variance, parametri l_d odražajo relativno pomembnost posamezne komponente vhodnega signala, izraz $\sigma_n^2\delta_{ij}$ pa modelira beli šum na izhodu sistema in je različen od nič le, ko velja $x_i = x_j$ [1].

Predikcija izhoda pri novi vhodni točki je normalna verjetnostna porazdelitev s srednjo vrednostjo $\mu(X^*)$ in varianco izhoda:

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X)+\sigma_n^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}\right) \tag{3}$$

$$\mu(X^*) = \bar{y}^* = K(X^*,X)\left[K(X,X)+\sigma_n^2 I\right]^{-1} y \tag{4}$$

$$\mathrm{cov}(y^*) = K(X^*,X^*) - K(X^*,X)\left[K(X,X)+\sigma_n^2 I\right]^{-1} K(X,X^*) \tag{5}$$

Velikost variance podaja mero zaupanja v napovedano vrednost izhoda modela. Vektor X^* je vhodni testni vektor z dimenzijo D, pri katerem nas zanima predikcija izhoda $\bar{y}^*$. Pričakovana vrednost izhoda $\bar{y}^*$ je tako enaka srednji vrednosti porazdelitve $\mu(X^*)$. Na vektor $K(X^*,X)[K(X,X)+\sigma_n^2]^{-1}$ lahko gledamo kot na vektor uteži, ki določa uteženost izhodov. Če se nov vhod precej razlikuje od učnih, bo izraz $K(X^*,X)[K(X,X)+$ $+\sigma_n^2]^{-1}K(X,X^*)$ majhen, varianca $cov(y^*)$ pa velika [1].

3 DMP (DYNAMIC MOVEMENT PRIMITIVE)

DMP (Dynamic Movement Primitive) je metoda za predstavitev gibanj, ki temelji na nelinearnem dinamičnem sistemu. DMP-ji temeljijo na sistemu nelinearnih diferencialnih enačb drugega reda, ki opisujejo lastnosti želenega gibanja. Ena najpomembnejših prednosti DMP-jev je ta, da niso direktno odvisni od časa. Kot je razloženo v [2], je eksplicitno izražanje časa nadležno, ker poveča kompleksnost prekinitve, ustavitev ali ponovnega zagona ure ob pojavu nepredvidljivih motenj med izvajanjem gibanja. Ugodna lastnost DMP-jev je, da lahko analitično prilagajamo dinamični sistem, pri čemer ključne diferencialne enačbe ostajajo nespremenjene. Takšno prilagajanje je lahko zelo uporabno v primeru različnih nepredvidljivih motenj med izvajanjem, ki niso del naučenih podatkov. Takšen pristop je primeren za sodelovanje s strojnim vidom humanoidnega robota.

Predstavimo teoretično ozadje DMP-jev. Za eno prostostno stopnjo označeno z y, ki je lahko ena od notranjih sklepov ali zunanjih prostorskih koordinat, je bil predlagan naslednji sistem linearnih diferencialnih enačb s konstantnimi koeficienti kot osnova za opis gibanja

$$\tau\dot{z} = \alpha_z\left(\beta_z(g-y) - z\right) \tag{6}$$

$$\tau\dot{y} = z \tag{7}$$

Ob pravilni izbiri parametrov ($\alpha_z = 4\beta_z$ in $\tau > 0$) sistem (6)-(7) konvergira proti enolično določeni točki $[y\ z]^T = [g$

$0]^T$. Diferencialni enačbi (6) in (7) zagotovita, da y konvergira proti cilju g in se ju kot taki lahko uporabi za realizacijo diskretnega giba od točke do točke. Da povečamo omejeno število trajektorij, ki jih enačbi (6) in (7) lahko opišeta in tako omogočimo aproksimacijo splošnega »point-to-point« giba, moramo enačbo (6) nekoliko spremeniti. V primeru diskretnih gibov lahko enačbi (6) dodamo linearno kombinacijo radialnih baznih funkcij:

$$f(x) = \frac{\sum_{i=1}^{N} w_i \psi_i(x)}{\sum_{i=1}^{N} \psi_i(x)} x, \quad \psi_i(x) = \exp\left(-h_i (x - c_i)^2\right), \tag{8}$$

kjer je $h_i > 0$, c_i pa so središča radialnih baznih funkcij razporejena po trajektoriji. V enačbi (8) je namesto časa uporabljena fazna spremenljivka x, da se izognemo direktni časovni odvisnosti funkcije f. Dinamika fazne spremenljivke x je definirana kot:

$$\tau \dot{x} = -\alpha_x x, \tag{9}$$

z začetno vrednostjo $x(0) = 1$. Rešitev enačbe (9) je $x = \exp(-\alpha_x t / \tau)$ in tako ob povečevanju časa t, x pada proti 0. Rezultat je naslednji sistem diferencialnih enačb:

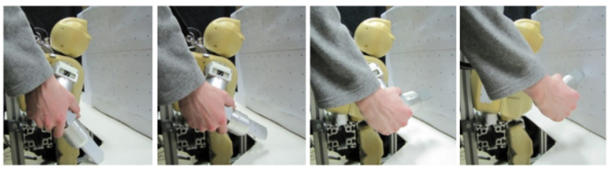
$$\tau \dot{z} = \alpha_z \left(\beta_z (g - y) - z\right) + f(x), \tag{10}$$

$$\tau \dot{y} = z, \tag{11}$$

ki ga lahko uporabimo za aproksimacijo diskretnih gibov različnih oblik. Ker x teži k nič, se vpliv nelinearne funkcije $f(x)$ s časom manjša in sistem enačb (10)-(11) konvergira proti $[g\ 0]^T$ kot sistem enačb (6)-(7). Naslednja vloga x-a je lokaliziranje radialnih baznih funkcij po trajektoriji, ki jo aproksimiramo. Rezultirajoča nadzorna taktika povezana s spremenljivko y definira tako imenovani enostavni opis dinamičnega gibanja (DMP) [3].

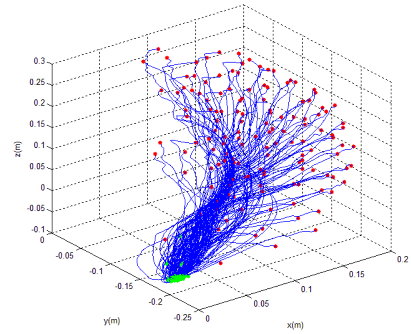
4 POSPLOŠEVANJE Z GAUSSOVO REGRESIJO

Naš cilj je bil robota HOAP-3 naučiti, kako naj seže za predmetom in ga prime. Vedenje seganja smo učili tako, da smo robotu pokazali nekaj primerov (učenje s posnemanjem), vse vmesne gibe pa bo HOAP-3 izvedel s posploševanjem naučenih gibov na podlagi Gaussove regresije.



Slika 2: *Zaporedje slik prikazuje učenje seganja s fizičnim vodenjem roke robota HOAP-3.*

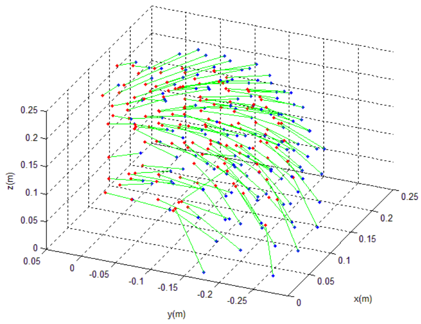
Robotu je bilo najprej potrebno pokazati, kako naj z roko sega po prostoru. To smo storili tako, da smo ga prijeli v bližini sklepa, ki predstavlja desni komolec, in roko fizično vodili od izbrane začetne lege do želenih končnih točk pred robotom (slika 2). Ves čas smo merili vse štiri kote v sklepih desne roke robota. Roko smo pomikali do ploščadi pred robotom, na kateri so bile narisane točke na medsebojni razdalji 5 cm. Ploščad smo postopoma odmikali stran od robota in tako na štirih razdaljah napravili serije gibov. Nekaj gibov je bilo napravljenih tudi ob robotu brez ploščadi. Tako so bile končne točke trajektorij enakomerno porazdeljene znotraj delovnega območja desne roke. Vse skupaj smo izvedli 148 gibov (slika 3), a se je kasneje izkazalo, da 8 končnih točk robot s svojimi kamerami ne vidi. Te gibe smo morali odstraniti iz baze predhodnega znanja.



Slika 3: *Modre trajektorije predstavljajo naučene gibe desne roke, pri čemer zelene pike predstavljajo začetne točke, rdeče pike pa končne točke teh trajektorij.*

S spremembo predznakov vseh štirih sklepov desne roke, smo ustvarili še trajektorije seganja leve roke robota. Takšna poenostavitev je bila mogoča, zaradi primerne konstrukcije robota HOAP-3.

Poleg kotov v sklepih smo merili tudi končne točke trajektorij, kot jih oceni strojni vid robota. Izkazalo se je, da strojni vid končnih točk ne zaznava tam, kjer je njihova dejanska pozicija v baznem koordinatnem sistemu robota (slika 4). Pozicijo predmeta se namreč oceni s pomočjo zamika slik obeh kamer. Tako robotski kameri predmet v povprečju vidita kar 63 mm stran od njegove dejanske pozicije.



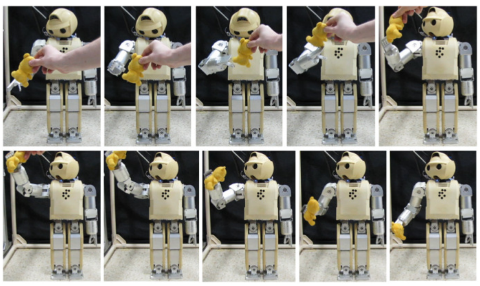
Slika 4: *Modre pike predstavljajo dejanske končne točke vseh trajektorij, rdeče pike pa so te iste točke, kot jih oceni strojni vid robota. Zelene črte prikazujejo zamik končnih točk.*

Ta napaka se verjetno pojavi zaradi nenatančno kalibriranih kamer, čeprav je bil postopek kalibracije izveden korektno. Napako zaradi kalibracije smo odpravili tako, da smo na koncu vsake trajektorije v robotsko dlan položili majhen sferičen predmet in določili njegovo pozicijo v vsaki končni točki, kot jo vidijo kamere in ker bo strojni vid robota ocenjeval pozicijo predmeta tudi med samostojnim izvajanjem naloge seganja, se bodo napake pri naučenih pozicijah in napake pri trenutnih pozicijah predmeta med sabo izničile.

Shranjene trajektorije smo zakodirali z nelinearnimi dinamičnimi sistemi (DMP-ji) in jih tako aproksimirali s 25 primerno uteženimi radialnimi baznimi funkcijami. Eksperimentalno smo določili, da je 25 uteži optimalno število, glede na kompleksnost in dolžino učnih trajektorij. Omenjeno polje uteži se nato uporabi v Gaussovi regresiji pri izračunu novih uteži za trajektorije, ki sežejo v katerokoli točko v delovnem območju robotovih rok. Regresija prav tako posploši funkcijo, ki preslika končne vrednosti trajektorij v zunanjih koordinatah (kot so bile ocenjene s strojnim vidom) v polje končnih vrednosti rok robota v notranjih koordinatah (enačba (4)). Gaussovemu procesu podamo tudi polje časovnih trajanj posameznih učnih gibov in tako se lahko za vse nove trajektorije s posploševanjem določi ustrezen čas izvedbe posameznega giba.

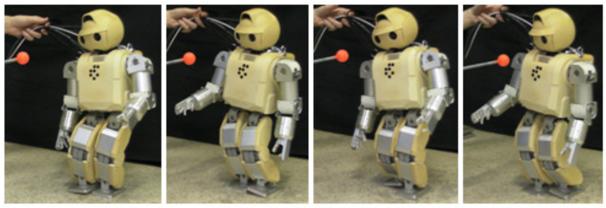
Ko se s pomočjo kamer določi pozicija predmeta v zunanjih koordinatah, Gaussova regresija na podlagi posploševanja izračuna končno pozicijo robotske dlani v notranjih koordinatah (g_j), 25 uteži za želeni gib seganja (w_k) in primeren čas trajanja tega giba (t). Dobimo tako imenovane DMP-parametre (w_k, g_j, t), ki opisujejo želeno trajektorijo (enačbi (10) in (11)), glede na novo ciljno točko. Te parametre posredujemo nelinearnemu dinamičnemu sistemu (DMP), ga integriramo z Eulerjevo integracijo in izhod (vrednosti sklepov vzdolž trajektorije) uporabimo za vodenje roke robota.

Če se predmet pojavi kjerkoli v delovnem območju robotovih rok, bo robot segel za predmetom s podobno obliko giba, kot smo mu prej pokazali in ga zgrabil. V primeru, da se med izvajanjem seganja pozicija predmeta spreminja, robot uporabi aktivni strojni vid za oceno nove pozicije predmeta in prilagodi DMP, da doseže nov cilj. Robot bo tako sledil predmetu s kamerami in roko ter prav tako prišel do cilja (slika 5). Oblika giba bo ostala podobna naučenim gibom vse do konca ocenjenega trajanja giba, nato bo zavzela obliko, kot jo narekuje kritično dušen sistem. Točen prihod v končno točko ovira napaka Gaussove regresije, ki v povprečju znaša 7,6 mm, kar je za predstavljeno nalogo sprejemljivo.

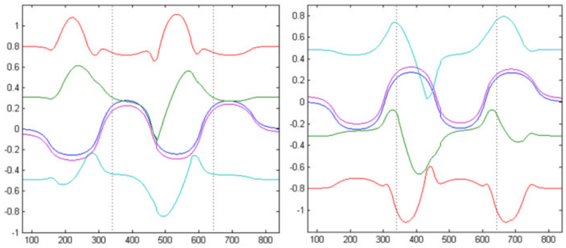


Slika 5: *Robot HOAP-3 zazna predmet in mu sledi. Ko se predmet ustavi, ga robot lahko prime in pomakne roko nazaj v začetno pozicijo.*

Če se predmet nahaja izven dosega robotovih rok, HOPA-3 s kamerami oceni razdaljo med predmetom in baznim koordinatnim sistemom ter uporabi robotsko hojo (slika 6), da se predmetu približa. Med hojo robot sproti sledi predmetu z glavo in kamerami ter ves čas ocenjuje razdaljo. Robotska hoja je bila matematično določena od proizvajalca robota Fujitsu. Tako smo iz niza sledečih si kotov v sklepih, ki pripomorejo k robotski hoji, izrezali tri delčke: začetni, vmesni in končni korak (slika 7).



Slika 6: *Če se predmet nahaja izven dosega rok, bo robot uporabil robotsko hojo in se predmetu približal.*



Slika 7: *Prvi graf prikazuje spreminjanje sklepov desne noge, drugi graf pa leve noge robota. Vsak od grafov je razdeljen na 3 dele, ki predstavljajo začetni, vmesni in končni korak, pri čemer so vrednosti kotov na mejnih področjih enake, kar omogoča dodajanje ali odstranjevanje vmesnih korakov.*

Robot med hojo sproti ocenjuje koliko vmesnih korakov je še potrebnih do cilja. Če kamere opazijo, da je predmet več kot 15° pod kotom glede na bazni koordinatni sistem, se bo robot na mestu za 15° obrnil proti predmetu in hojo nadaljeval.

5 ZAKLJUČEK

V članku je predstavljeno, kako je dobre rezultate mogoče doseči s predlaganim pristopom, vse dokler je na voljo zadostna količina naučenih podatkov. Če se želimo izogniti velikim količinam začetnega znanja, lahko zmanjšamo število učnih trajektorij pridobljenih s fizičnim vodenjem. Robot lahko nato postopno pridobiva vse več podatkov z avtomatskim seganjem v različne pozicije z majhnim sferičnim predmetom v dlani in sprotnim ocenjevanjem končnih točk s strojnim vidom.

Literatura

[1] C. E. Rasmussen, C. Williams, *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press, 2006

[2] S. Schaal, P. Mohajerian, A. Ijspeert, *Dynamics systems vs. optimal control – a unifying view*, Brain Research, vol. 165, no. 6, pp. 425–445, 2007

[3] A. Gams, A. Ude, *Generalization of Example Movements with Dynamic Systems*, 9th IEEE-RAS International Conference on Humanoid Robots, 2009

Posploševnje robotskega gibanja - met

Andrej Gams, Tadej Petrič, Leon Žlajpah, Aleš Ude
Department for Automation, Biocybernetics and Robotics
Jozef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: andrej.gams@ijs.si

ABSTRACT

Controlling robots with imitation can only be useful if the robots can adapt their actions to the current state of the external world. This paper presents a generalization algorithm which generates new actions from a library of learned actions. The new action has the properties of nearby demonstrated actions. The paper focuses on the execution of robotic throwing at a random target within the area of demonstrated and learned actions. The new movement is not one of the demonstrated ones but a newly generated one. Throwing has no explicit connection between the final position of the arm and the target and this application shows that the generalization algorithm works for such tasks as well.

1 UVOD

Eden od problemov moderne robotike je še zmeraj generiranje trajektorij ter njihova modulacija [1]. Manipulacija z objekti v tri-dimenzionalnem svetu zahteva, da se robot nauči prav vseh možnih gibov, kar pa ni izvedljivo, kajti prostor učenja je enostavno prevelik. Druga opcija je spreminjanje oz. modulacija naučenih trajektorij iz obstoječega znanja [2, 3].

Obstajajo različni pristopi moduliranja naučenih trajektorij, ki uporabljajo tudi različne matematične mehanizme za shranjevanje trajektorij. Zapis trajektorije je ključnega pomena za učinkovit algoritem [4]. Med najpreprostejše spada shranjevanje časovno-indeksiranih vektorjev [5]. Alternativa so kompaktnejši zapisi z uporabo npr. zlepkov (ang. splines). HMM (ang. Hidden Markov Models) se lahko učinkovito uporabijo tako za modulacijo kot tudi generacijo trajektorij [6].

Eden od pristopov je z uporabo nelinearnih diferencialnih enačb oz. nelinearnih dinamičnih sistemov. Pristop, kjer z nelinearnimi enačbami drugega reda opišemo poljubne trajektorije se imenuje pristop z dinamičnimi primitivi gibanja (DMP) [7, 8]. Pokazano je bilo, da lahko s spreminjanjem diferencialnih enačb, ki sestavljajo DMP, moduliramo naučene trajektorije, kot je to prikazano v npr. [9,l0]. Te modifikacije mora implementirati strokovnjak in jih ne moremo izvajati avtonomno.

Na drugi strani lahko posploševanje iz knjižnice naučenih gibov avtonomno generira trajektorije, s katerimi se prilagajamo dejanskemu stanju okolice. Obstajajo različni pristopi generalizacije gibanj, npr. s HMM [6] ali pa z DMP-ji [3]. V našem pristopu uporabljamo pristop z DMP-ji.

Glavni namen tega prispevka je pokazati, da lahko algoritem generalizacije uporabimo tudi za met na poljubno tarčo. Met nima neposredne povezave med končno lego roke ter tarčo, kajti roka se po trenutku, ko jo projektil zapusti, lahko prosto premika dalje oz. izvede še kakšen dodaten gib. Za razliko od seganja do poljubne točke tako nimamo neposredne povezave med zaključkom gibanja ter točko povpraševanja iz knjižnice naučenih gibov.

V naslednjih poglavjih najprej predstavimo pristop z uporabo dinamičnih primitivov gibanja oz. DMP-jev. V tretjem poglavju razložimo algoritem posploševanja gibanja iz knjižnice naučenih gibov. Četrto poglavje opisuje izvedbo eksperimenta na realnem robotskem mehanizmu. Na kratko je predstavljen tudi sistem zajemanja pozicije tarče in rezultati. V zadnjem poglavju so podani zaključki.

2 DINAMIČNI PRIMITIVI GIBANJA - DMP

V tem poglavju je na kratko predstavljen pristop posnemanja periodičnega gibanja z nelinearnimi dinamični sistemi oz. DMPji. Ijspeert in drugi [7] so predlagali zapis gibanja s sistemom nelinearnih enačb, ki imajo dobro definirano dinamiko atraktorja.

Posamezni stopnji prostoti gibanja v notranjih koordinatah ali pa v prostoru nalog naslednji sistem diferencialnih enačb poda osnovo za določitev specifik gibanja.

$$\tau \dot{z} = \alpha_z(\beta_z(g-y)-z) \quad (1)$$

$$\tau \dot{y} = z \quad (2)$$

Pri tem modajo biti parametri α_z, β_z in $\tau > 0$ primerno določeni recimo $\alpha_z = 4\beta_z$ (v tem primeru je sistem celo kritično dušen) in sistem ima samo en atraktor pri $y = g, z = 0$.

Enačbam (1) – (2), ki opisujeta zelo omejen nabor gibanja, dodamo nelinearn člen, s katerim preoblikujemo rezultirajočo trajektorijo. Za zaključeno oz. diskretno gibanje dodamo

$$f(x)=\frac{\sum_{i=1}^{N}\omega_i\Psi_i(x)}{\sum_{i=1}^{N}\Psi_i(x)}x, \tag{3}$$

$$\Psi_i=\exp(-h_i(x-c_i)^2). \tag{4}$$

Z vpeljavo fazne spremenljivke x se znebimo direktne odvisnosti od časa. Pridobimo ga z integracijo enačbe

$$\tau\dot{x}=-\alpha_x x. \tag{5}$$

Enačba (5) podaja eksponencialno približevanje ničli, ki ga prekinemo, ko se gib izteče. S tem se enačbi (1) – (2) spremenita v

$$\tau\dot{z}=\alpha_z(\beta_z(g-y)-z)+f(x) \tag{6}$$

$$\tau\dot{y}=z \tag{7}$$

s katerima pa lahko zapišemo poljubno diskretno oz. aperiopdočno trajektorijo. Enačbi (5) in (6) opisujeta en primitive gibanja, torej en DMP.

3 POSPLOŠEVANJE GIBANJ

V tem poglavju je opisan algoritem posploševanja diskretnih gibanj iz knjižnice naučenih gibov. Najprej je predstavljena metoda učenja ene same trajektorije, kasneje pa posploševanje iz več trajektorij.

3.1 Učenje ene trajektorije

Poljubno gladko trajektorijo lahko ponovimo s spreminjanme parametrov w_i iz enačbe (3). Pri tem začnemo iz posnetih demonstracijskih pozicij, hitrosti in pospeškov $\{y_d(t_j),\dot{y}_d(t_j),\ddot{y}_d(t_j)\}$, kjer je t_j čas vzorčenja. Podatke dobimo s kinestetičnim vodenjem robota ali pa človeško demonstracijo. Enačbi (5) – (6) zapišemo v eno samo enačbo drugega reda

$$\tau^2\ddot{y}+\alpha_z\tau\dot{y}-\alpha_z\beta_z(g-y)=f(x), \tag{8}$$

kjer je $f(x)$ definiran z enačbo (3). Z zapisom

$$F(t_j)=\tau^2\ddot{y}_d(t_j)+\alpha_z\tau\dot{y}_d(t_j)-\alpha_z\beta_z(g-y_d(t_j)), \tag{9}$$

$$f=\begin{bmatrix}F(t_1)\\ \ldots\\ F(t_T)\end{bmatrix},\quad \mathrm{w}=\begin{bmatrix}w1\\ \ldots\\ w_N\end{bmatrix}, \tag{10}$$

dobimo naslednji set linearnih enačb:

$$\mathbf{Xw}=\mathbf{f} \tag{11}$$

ki ga moramo rešiti, da lahko ocenimo DMP, ki opisuje želeno gibanje. X podaja

$$\mathrm{X}=\begin{bmatrix}\frac{\Psi_1(x_1)}{\sum_{i=1}^{N}\Psi_1(x_1)}x_1 & \ldots & \frac{\Psi_N(x_1)}{\sum_{i=1}^{N}\Psi_N(x_1)}x_1\\ \ldots & \ldots & \ldots\\ \frac{\Psi_1(x_T)}{\sum_{i=1}^{N}\Psi_1(x_T)}x_T & \ldots & \frac{\Psi_N(x_T)}{\sum_{i=1}^{N}\Psi_N(x_T)}x_T\end{bmatrix}. \tag{12}$$

x dobimo z integracijo enačbe (4).

3.2 Posploševanje iz več trajektorij

Za posploševanje iz več trajektorij lahko uporabimo vse točke, ki so vključene v knjižnici naučenih trajektorij. The točk je zelo veliko, zato je pomembno, da izberemo učinkovito metodo za estimacijo posplošenih gibanj. Lokalnoa utežena regresija (*ang*. Locally Weighterd Regression – LWR) [11] je neparametrična regresijska metoda, ki prilagodi lokalni model bližnjim točkam. Ima manjšo kompleksnost računanja kot podobne neparametrične metode, npr. Gaussova regresija [12]. Z LWR in ob izbrani točki povpraševanja q, optimalne parameter izračunamo z utežitvijo funkcije

$$C(\mathrm{q})=\sum_k L(\Xi(\mathrm{q}_k,\mathrm{w}),\mathrm{f}_k)K(d(\mathrm{q},\mathrm{q}_k)). \tag{13}$$

Na osnovi enačbe (11) lokalne modele opisuje

$$L(\Xi(\mathrm{q}_k,\mathrm{w}),\mathrm{f}_k)=\|\mathrm{X}_k\mathrm{w}-\mathrm{f}_k\|^2 \tag{14}$$

in glede na $\mathbf{w}$ moramo minimizirati funkcijo

$$\sum_{k=1}^{M}\|\mathbf{X}_k\mathbf{w}-\mathbf{f}_k\|^2K(d(\mathbf{q},\mathbf{q}_k)), \tag{15}$$

kjer je K jedrna funkcija in d metrika v prostoru točke povpraševanja. Izbrali smo naslednje utežitveno jedro

$$K(d) = \begin{cases} (1-|d|^3)^3 & \text{če } |d|<1 \\ 0 & \text{sicer} \end{cases}, \qquad (16)$$

ki je končno ter ima zvezen prvi in drugi odvod. Ker podatke pridobimo inkrementalno, tako tudi posplošujemo, kar pomeni, da moramo pred posploševanjem poskrebeti, da so vsi podatki v primernem vrstnem redu. K in d določata vpliv posameznih naučenih gibov na končni posplošeni gib. Njihov vpliv se zmanjša z razdaljo od točke povpraševanja. Če so točke povpraševanja v evklidskem prostoru, potem lahko uporabimo standardno evklidsko razdaljo

$$d(\mathbf{q},\mathbf{q}_k) = \|\mathbf{D}(\mathbf{q}-\mathbf{q}_k)\|, \ \mathbf{D} = \mathrm{diag}(a_i), \ a_i > 0 .$$

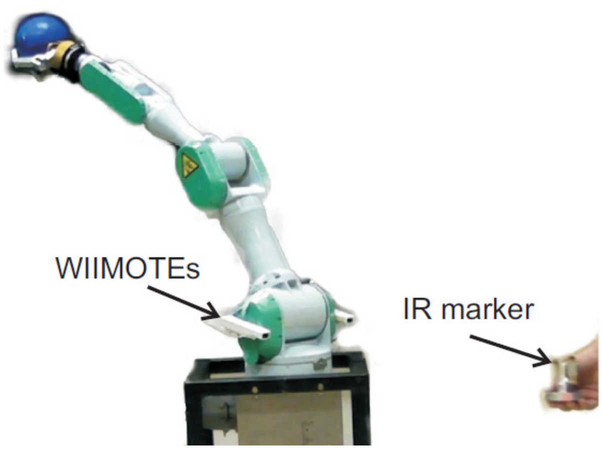
d smo določili tako, da sta bila vsaj dva primera v vsaki smeri od točke povpraševanja ob računanju novega DMP. Opisani pristop je primeren samo, če si demonstracijske trajektorije smiselno zvezno sledijo kot funkcija točk povpraševanja. V nasprotnemprimeru podatki ne nudijo zadostnih informacij o gibanju, ki bi ga povezali z novo točko povpraševanja. To pomeni, da mora biti funkcija

$$\mathbf{G} : \mathbf{q} \to \left[\mathbf{w}^T, \tau, g^T\right], \qquad (17)$$

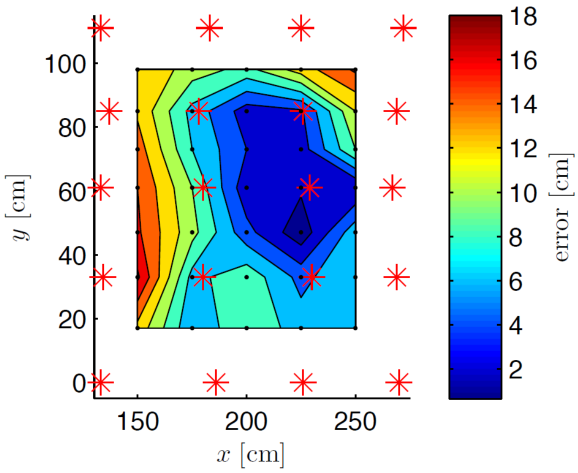
ki je v splošnem neznana, gladka. Ponavadi je tako, če robot uporablja enako strategijo za reševanje naloge v različnih situacijah.

4 POSPLOŠEVANJE GIBANJ - MET

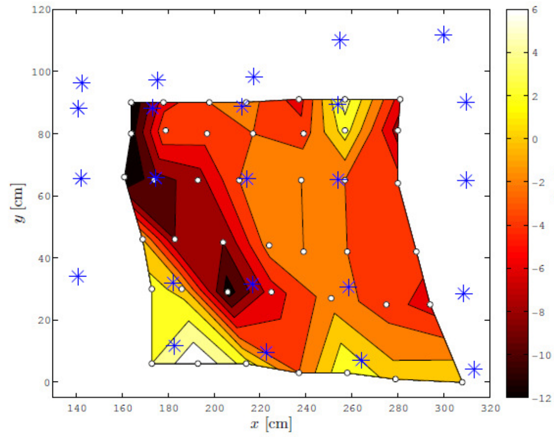
Robotski met je že bil predmet raziskav [13]. Met na tarčo znotraj nekega območja je gib, ki nima eksplicitne povezave med končno lego roke ter tarčo. Eksplicitno povezavo ima na primer seganje, pri metu pa se lahko po izmetu roka poljubno premika. Eksperiment posploševanja meta znotraj območja neučenih gibov prikazuje uporabnost opisanega algoritma tudi za tako nalogo.



Slika 1: *Eksperimentalni sestav.*



Slika 2: *Rezultati meritev ob merjenju z metrom. Zvezdice prikazujejo naučene gibe, točke pa točke povpraševanja oz. tarče. Napako podaja barva.*

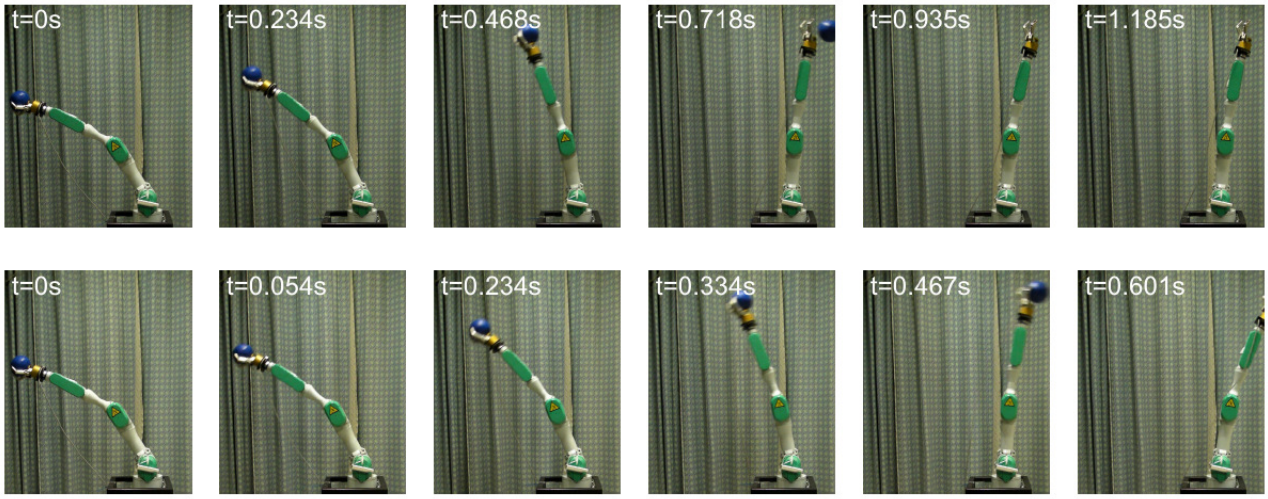


Slika 3: *Rezultati meritev ob merjenju z uporbo dveh naprav wiimote. Napaka je znotraj območja naučenih gibov dokaj majhna.*

Nalogo smo izvedli z robotom Mitsubishi Pa-10, na katerega smo pritdili prijemalko. Sestav prikazuje slika 1. Izmet se je zgodil zaradi premikanja robotske robe ter ni bil pogojen s prijemalko. Izvedli smo dva poskusa, najprej z merjenjem z navadnim metrom Rezultate prikazuje slika 2. Namestili smo tudi sistem stereo vida za sledenje IR markerjem, ki je bil sestavljen iz dveh Nintendo Wiimote igralnih palica [14]. S tem je bilo merjenje izvedeno avtomatsko. Rezultate prikazuje slika 3. Sliak 4 prikazuje izvedbo dveh metov na tarči blizu in daleč od robota.

5 ZAKLJUČEK

Z izvedbo posploševanja meta na tarčo znotraj območja naučenih gibov smo pokazali, da algoritem posploševanja



Slika 4: *Dve trajektoriji meta na tarči blizu in daleč od robota*

deluje tudi za naloge, kjer ni eksplicitne povezave med končno lego robotske roke ter točko povpraševanja.
Met, kot naloga, ki zahteva hitro, dinamično gibanje ter hitre pospeške, je zahteven in kaže na robustnost posploševanja, ki deluje ne samo po poziciji temveč tudi pri posnemanju doseženih pospeškov in hitrosti.

References

[1] S. Degallier, C. Santos, L. Righetti, and A. Ijspeert, "Movement Generation Using Dynamical Systems: a Humanoid Robot Performing a Drumming Task," in IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS06), 2006.

[2] S. Schaal, "Is Imitation Learning The Route to Humanoid Robots?" Trends in cognitive sciences, no. 3, pp. 233–242, 1999.

[3] A. Gams and A. Ude, "Generalization of Example movements with Dynamical Systems," in IEEE-RAS International Conference on Humanoid Robots, 2009, pp. 28 – 33.

[4] A. Gams, A. J. Ijspeert, S. Schaal, and J. Lenarˇciˇc, "On-line learning and modulation of periodic movements with nonlinear dynamical systems," Autonomous Robots, vol. 27, no. 1, pp. 3–23, 2009.

[5] S. Kawamura and N. Fukao, "Interpolation for input torque patterns obtained through learning control," in Proceedings of The Third International Conference on Automation, Robotics and Computer Vision (ICARCV'94), 1994.

[6] S. Calinon, F. D'halluin, D. Caldwell, and A. Billard, "Handling of multiple constraints and motion alternatives in a robot programming by demonstration framework," in Proc. IEEE-RAS Intl Conf. on Humanoid Robots (Humanoids), December 2009.

[7] S. Calinon, F. Guenter, and A. Billard, "On learning, representing and generalizing a task in a humanoid robot," IEEE Transactions on Systems, Man and Cybernetics, Part B. Special issue on robot learning by observation, demonstration and imitation, vol. 37, no. 2, pp. 286–298, 2007.

[8] A. J. Ijspeert, J. Nakanishi, and S. Schaal, "Movement imitation with nonlinear dynamical systems in humanoid robots," in Proc. IEEE Int. Conf. Robotics and Automation, Washington, DC, 2002, pp. 1398–1403.

[9] ——, "Learning rhythmic movements by demonstration using nonlinear oscillators," in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems, Lausanne, Switzerland, 2002, pp. 958–963.

[10] S. Schaal, P. Mohajerian, and A. Ijspeert, "Dynamics systems vs. optimal control – a unifying view," Progress in Brain Research, vol. 165, no. 6, pp. 425–445, 2007.

[11] C. G. Atkeson, A.W. Moore, and S. Schaal, "Locally weighted learning," AI Review, vol. 11, pp. 11–73, 1997.

[12] D. Nguyen-Tuong, M. Seeger, and J. Peters, "Model learning with local gaussian process regression," Advanced Robotics, vol. 23, no. 15, pp. 2015 – 2034, 2009.

[13] B. Hove and J. Slotine, "Experiments in robotic catching," in Proceedings of the 1991 American Control Conference, 1991, pp. 380 – 385.

[14] T. Petrič, A. Gams, A. Ude, and L. Žlajpah, "Real-time 3d marker tracking with a WIIMOTE stereo vision system: application to robotic throwing," in Submited to RAAD 2010, 2010.

ROBOTSKI PARALELNI MEHANIZEM ZA ANALIZO ČLOVEKOVEGA RAVNOTEŽJA

Goran Škorja in Jan Babič
Odsek za Avtomatiko, Biokibernetiko in Robotiko
Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana, Slovenija
e-mail: goran.skorja@ijs.si

POVZETEK

Namen te študije je bil konstruirati in uporabiti paralelni robotski mehanizem za povzročanje motnje v ravnotežju oseb pri pokončni stoji. Nagibni paralelni mehanizem je generiral naključne rotacijske motnje v šestnajstih različnih smereh okoli navpične osi. Z merjenjem središča pritiska oseb na podporno podlago s pomočjo pritiskovne plošče, pritrjene na vrh platforme paralelnega mehanizma, smo imeli možnost voditi orientacijo paralelnega robotskega mehanizma v realnem času. Poleg merjenja odmika središča pritiska smo merili tudi mišične aktivnosti. Rezultate raziskave smo predstavili s polarnimi diagrami.

1 UVOD

Običajni metodološki pristop za študij vzdrževanja človekovega ravnotežja je analiza fizioloških in biomehanskih odzivov na generirane mehanske motnje v podporni površini oseb [1]. Najpogostejši način realizacije mehanske motnje je s spodrsavanjem, nagibanjem, pospeševanjem ali zaviranjem podporne površine. Vzdrževanje ravnotežja je v primeru nepredvidljive motnje podporne površine odziv na zunanje sile, ki posredno delujejo na težišče telesa [2].
Meritev gibanja središča pritiska s pritiskovno ploščo je eden od standardnih pristopov določanja stabilnosti vzdrževanja ravnotežja. Središče pritiska na podplatih se izraža kot sila reakcije podlage in je rezultat nadzora ravnotežja. S primernim orodjem in ustrezno programsko opremo lahko opazujemo časovno odvisnost gibanja središča pritiska [3].
Namen tega prispevka je predstavitev paralelne nagibne platforme in analiza mišičnih odzivov na motnje v podporni površini. Merili in analizirali smo parametre za vzdrževanje ravnotežja desetih zdravih oseb. Osebe so stale na paralelni nagibni platformi, ki se je med eksperimentom naključno nagibala v šestnajstih različnih smereh okoli navpične osi. V primerjavi s tradicionalnimi metodami generiranja motnje [4], kjer se podporna površina rotira okoli poljubne osi smo mi uporabili napreden eksperimentalni pristop. Uporabili smo Stewartovo paralelno platformo s pritrjeno pritiskovno ploščo s katero smo med eksperimentom merili položaj središča pritiska. Podatek o položaju je omogočal, da so preizkušanci z odmikom težišča telesa aktivno vodili orientacijo Stewartove platforme.
Rezultate raziskave smo grafično predstavili s polarnimi diagrami aktivnosti posameznih mišičnih skupin v odvisnosti od smeri motnje podporne površine.

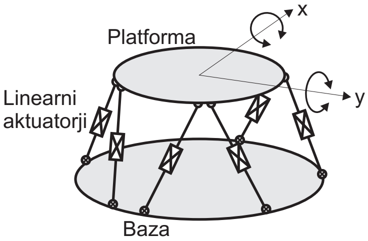
2 METODOLOGIJA

2.1 Preizkušanci

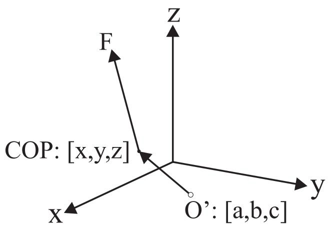
V raziskavi je sodelovalo deset oseb moškega spola brez okvar gibalnega ali mišičnega sistema. Povprečna starost udeležencev je bila 27 let, s povprečno maso 82 kg in povprečno višino 175 cm. Vsi preiskovanci so bili informirani o poteku eksperimenta, nalogah, ki so jih morali opraviti ter pripravah in načinu merjenja. Raziskavo je odobrila Državna Komisija za medicinsko etiko.

2.2 Eksperimentalno okolje

Za generiranje motnje v ravnotežju osebe smo uporabili paralelno nagibno ploščad (slika 1). Nagibna ploščad je Stewartov paralelni robotski mehanizem s šestimi prostostnimi stopnjami. Mehanizem označuje kinematično strukturo, pri katerem dva toga segmenta povezuje več kinematičnih verig [5]. Eden od segmentov predstavlja podlago, ki se ji reče baza, drugi segment pa je gibajoč in mu pravimo platforma ali ploščad. Kinematične verige, ki povezujejo bazo s platformo pa so noge oziroma linearni aktuatorji. Za naš eksperiment nam Stewartov robotski mehanizem omogoča rotacijo podporne površine okoli treh osi koordinatnega sistema na vrhu ploščadi.



Slika 1: *Stewartov paralelni mehanizem. Primer rotacij okoli x in y osi na vrhu platforme.*



Slika 2: *Koordinatni sistem pritiskovne plošče. O' je referenčna točka pritiskovne plošče, COP je točka središča pritiska na podlago (vrh pritiskovne plošče).*

Motnjo v ravnotežju oseb je predstavljala rotacija podporne površine v antero-posteriorni smeri (rotacija okoli y osi), medio-lateralni smeri (rotacija okoli x osi) in kombinaciji obeh. Motnja je imela konstanten nagib 8.5° s kotno hitrostjo 50°/s in se je naključno generirala v 16 različnih smereh okoli navpične smeri.

Spremembo mišičnih aktivnosti zaradi vzdrževanja ravnotežja med eksperimentom smo merili z elektromiografom. Površinske EMG elektrode smo namestili na 4 mišične skupine spodnjih ekstremitet, ki so relevantne pri vzdrževanju ravnotežja med pokončno stojo.

2.3 Vodenje paralelnega mehanizma

Na vrhu nagibne ploščadi smo pritrdili Kistler pritiskovno ploščo s katero smo med eksperimentom merili položaj središča pritiska oseb na podlago. Podatek o položaju središča pritiska smo uporabili za povratno-zančno vodenje orientacije Stewartove paralelne ploščadi oziroma za vodenje orientacije podporne površine v realnem času.

Pritiskovno ploščo smo uporabili na podlagi proizvajalčevih specifikacij in merili ortogonalne komponente sil vzdolž *x*, *y* in *z* osi. Neobdelane signale iz posameznih sprejemnikov smo zajemali z osebnim računalnikom s frekvenco vzorčenja 500 Hz, preko 16 bitnega analogno-digitalnega pretvornika. Kot prikazuje slika 2 lahko vse sile, ki delujejo med stopali in podlago seštejemo v en sam vektor sile na podlago **F** ter določimo navor v prijemališču te sile. Pozicijo središča pritiska **COP**(x,y,0) izračunamo iz navora, ki ga povzroča sila na podlago okoli referenčne točke **O'**(a,b,c) pritiskovne plošče, ki je določena s strani proizvajalca

$$\mathbf{COP} = (x, y, 0)\,, \tag{1}$$

$$x = -\frac{M_y + cF_x}{F_z} + a, \quad \mathrm{y} = \frac{\mathrm{M}_x - cF_y}{F_z} + b\,. \tag{2}$$

M_y predstavlja navor sile F_y vzdolž frontalne ravnine, M_x navor sile F_x vzdolž sagitalne ravnine in F_z vertikalno silo. Iz enačbe 2 je razvidno, da je pozicija središča pritiska zelo občutljiva na majhne vrednosti vertikalne komponente sile F_z. Problem omilimo s postavitvijo minimalne vrednosti vertikalne sile, ki odpravi singularnost, če sila ni prisotna.

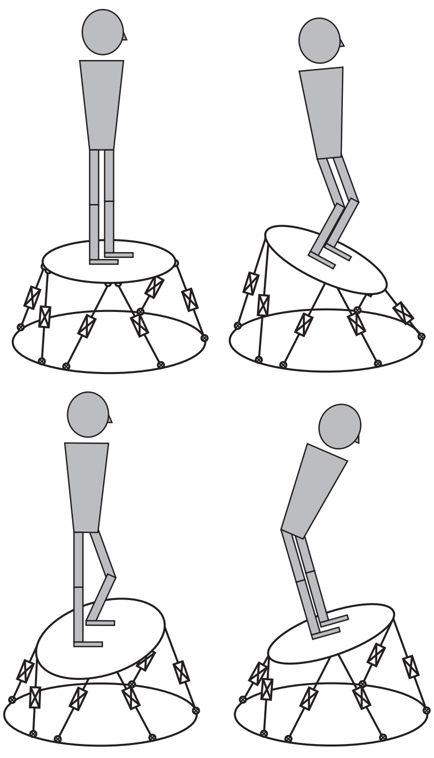
Vodenje Stewartovega paralelnega mehanizma smo zasnovali tako, da je odmik središča pritiska v pozitivno *x* smer koordinatnega sistema platforme rotiral platformo okoli koordinatne osi *y* v nasprotni smeri ure in odmik središča pritiska v pozitivno *y* smer koordinatnega sistema platforme rotiral platformo okoli *x* osi v smeri ure. Podobno velja za negativne odmike središča pritiska.

Uporabili smo PD regulator, kjer je izhod hitrostno reguliral Stewartov paralelni mehanizem. Vhod v regulator je napaka med referenčno dolžino nog in dejansko dolžino nog mehanizma. Referenčno dolžino nog dobimo z računanjem inverzne kinematike mehanizma. Odmik središča pritiska je sorazmeren z želeno oz. referenčno orientacijo, ki je vhod v inverzno kinematiko.

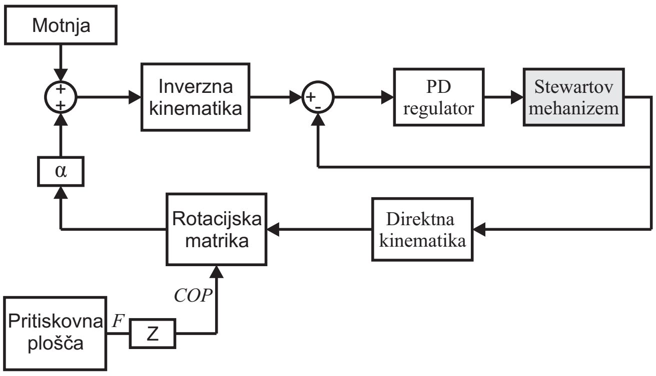
Ko se platforma zarotira za želen kot, se zarotira tudi koordinatni sistem postavljen na vrh platforme, kar pomeni, da moramo za korektno rotacijo odmika središča pritiska upoštevati tudi trenutno rotacijo koordinatnega sistema vrha platforme. V ta namen uporabimo rotacijsko matriko (slika 4) in njen izhod upoštevamo kot želen kot rotacije platforme. Vhod v rotacijsko matriko je podatek o odmiku središča pritiska, ki ga dobimo iz pritiskovne plošče.

Ker so preiskovanci stali na pritiskovni plošči na vrhu platforme (slika 3), so z odmikom središča pritiska na ploščo lahko aktivno vodili orientacijo paralelnega mehanizma. V skladu s superponirano motnjo so poskušali motnjo kompenzirati tako, da so se motnji s svojimi senzoričnimi sistemi za posturalni nadzor uprli in podporno površino poskušali vrniti v izhodiščni položaj.

Čas eksperimenta za posameznega preiskovanca je bil določen s časom motnje podporne površine in je trajal 240s.



Slika 3: *Odziv telesa na motnjo v različnih smereh.*



Slika 4: *Bločni diagram vodenja robotskega paralelnega mehanizma.*

2.4 Potek eksperimenta

Preizkušancem smo najprej pritrdili EMG elektrode na štiri mišične skupine (vastus lateralis, tibialis anterior, soleus in paraspinalis). Nato smo določili maksimalno hoteno kontrakcijo posamezne mišične skupine, ki so jo preizkušanci izvršili z izometričnim upiranjem posamezne mišice. Eksperiment je potekal tako, da so preizkušanci s prekrižanimi rokami pokončno stali na pritiskovni plošči, pritrjeni na platformi paralelnega mehanizma. Z odmikom svojega težišča so preko središča pritiska vplivali na vodenje paralelnega mehanizma. Aktivnemu vodenju orientacije mehanizma se je v naključni smeri prištela motnja s periodo 15 s. Preizkušancem smo naročili naj ohranijo svoje ravnotežje kar se da najbolje.

2.5 Obdelava podatkov

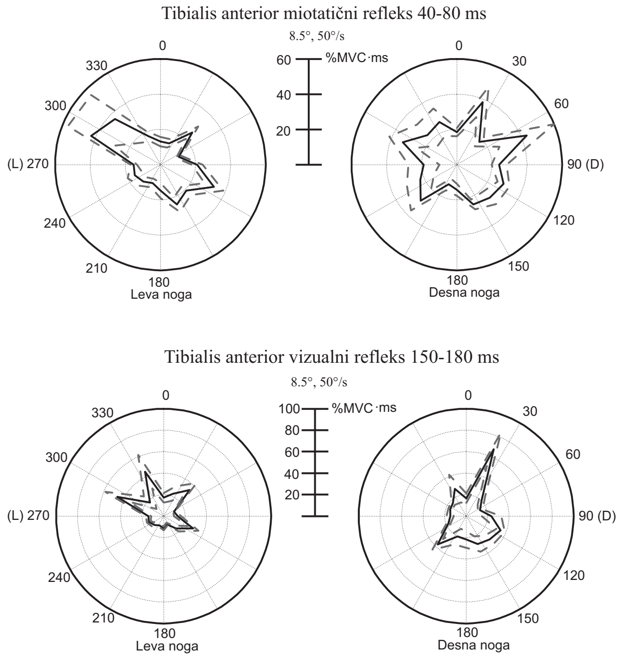
Izmerjene EMG signale smo najprej usmerili, razdelili na 16 delov od začetka posamezne motnje do njenega upada in sortirali glede na smer motnje od 0° do 360° v smeri ure. V času, ko motnja ni bila prisotna, smo za vsak del razdeljenega signala izračunali povprečno vrednost in to vrednost odšteli od preostalega signala. Dobljenemu signalu smo izvedli normalizacijo z vrednostjo maksimalne hotene kontrakcije. EMG normalizirano površino mišične aktivnosti v odvisnosti od smeri motnje, smo izračunali z uporabo trapeznega integrala na določenem časovnem intervalu povezanim z značilnimi refleksi mišic (miotatični refleks 45 ms – 80 ms, proprioceptivni refleks 80 ms – 120 ms, vizualno zaznavanje 150 ms – 180 ms) [6]. Za 10 serij obdelanih meritev smo izračunali povprečno vrednost in standardno deviacijo EMG normalizirane površine mišičnih aktivnosti v smeri posamezne motnje.

Končno obdelane signale smo predstavili s polarnimi diagrami, ki predstavljajo odvisnost EMG normalizirane površine od smeri motnje.

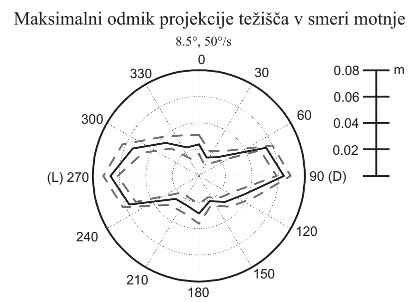
Podatek o središču pritiska smo dobili iz pritiskovne plošče. Filtriran signal o poziciji središča pritiska smo razdelili na 16 delov od trenutka motnje do njenega upada ter ga sortirali glede na smer motnje. V času, ko motnja ni bila prisotna, smo od vseh razdeljenih delov signala izračunali srednjo vrednost in jo od preostalega signala odšteli. Tako dobljen podatek je predstavljal maksimalen odmik središča pritiska v smeri motnje.

3 REZULTATI

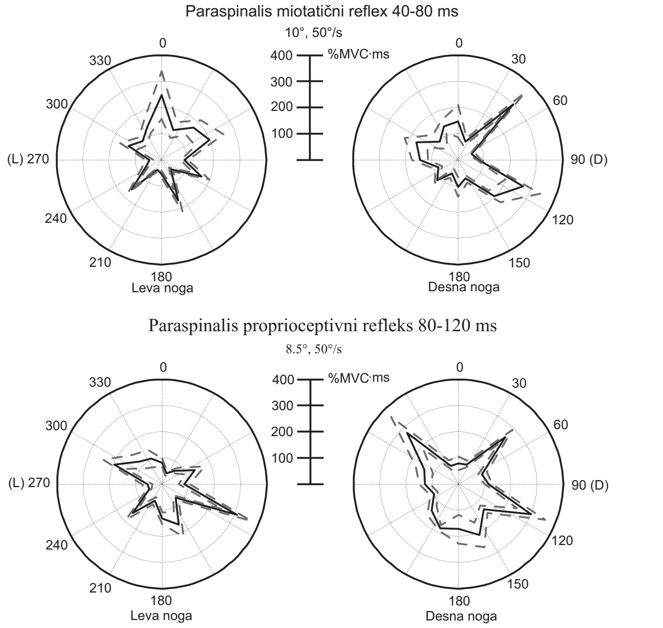
S polarnimi diagrami (slika 5, slika 7 in slika 8) smo predstavili srednjo vrednost in standardno deviacijo normalizirane površine mišičnih aktivnosti štirih mišičnih skupin, za leve in desne mišice v odvisnosti od smeri motnje. Diagrami so predstavljeni za različne časovne intervale, ki so ključni pri mišičnih aktivnostih med vzdrževanjem ravnotežja. Polarni diagram smo uporabili tudi za prikaz maksimalnega odmika središča pritiska v smeri motnje (slika 6).



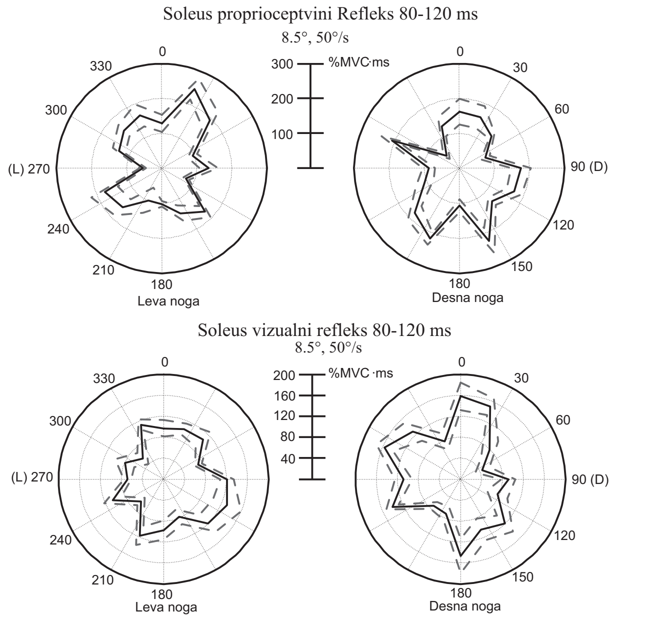
Slika 5: *Polarni diagrami povprečne vrednosti površine pod normalizirano EMG aktivnostjo in standardna deviacija (črtkana črta) mišice tibialis anterior za 10 preizkušancev. Odzivi prikazujejo meritve pri motnji podporne površine v 16 različnih smereh med miotatičnim (45 ms) in med vizualnim zaznavanjem (150-180 ms).*



Slika 6: *Maksimalni odmik središča pritiska v smeri motnje.*



Slika 7: *Polarni diagrami za mišico paraspinalis. Za detajlni opis diagrama glej Sliko 5.*



Slika 8: *Polarni diagrami za mišico soleus. Za detajlni opis diagrama glej Sliko 5.*

4 ZAKLJUČEK

Predhodne študije z uporabo motenj podporne površine v različnih smereh so pokazale smerno občutljivost mišic, ki sodelujejo pri ohranjanju ravnotežja in mirne pokončne stoje [7], vendar v teh študijah niso upoštevali in ločili reakcijskih časov mišic za določitev, kako motnja podporne površine v različnih smereh vpliva na posamezen reakcijski čas mišice. Leta 1998 so Carpenter, Allum et al. [8] raziskali tudi takšen primer. Rezultat njihovega eksperimenta je pokazal, da so zgodnejši reakcijski časi mišic občutljivi na smer motnje podporne površine.

Z našim eksperimentom smo nadgradili raziskave omenjenih avtorjev. Uporabili smo podoben koncept, vendar upoštevali dejstvo, da je oseba s svojimi senzoričnimi sistemi bila zmožna kompenzirati naključno motnjo podporne površine na takšen način, da je vrnila podporno površino v okolico izhodiščnega položaja. Aktivnost mišic je bila v takšnem primeru različna v primerjavi z dobljenimi rezultati raziskave, ki so jo izvedli Carpenter et al. Predvsem se razlika kaže v večji aktivnosti mišic v vseh smereh motnje podporne površine.

S tem eksperimentom smo zgradili eksperimentalni pristop, kjer je človek v interakciji s strojem. Na eni strani študiramo vpliv motnje robotskega mehanizma na človekov senzorični sistem pri katerem je oseba sposobna vzdrževati ravnotežje, na drugi strani pa smo realizirali admitančno vodenje paralelnega robotskega mehanizma, kjer je regulacija robotskega manipulatorja v dotiku z okolico, pri katerem je referenčni vhod regulatorja sila.

Reference

[1] D. A. Winter. Human balance and posture control during standing and walking. *Gait & Posture*, 3:193-214, 1995.

[2] R. Balasubramaniam and A. M. Wing. The dynamics of standing balance. *Trends in Cognitive Sciences*, 6(12):531-536, 2002.

[3] D. Rugelj and F. Sevšek. Postural sway area of elderly subjects. *WSEAS transactions on signal processing*, 3(2):213-219, 2007.

[4] M. G. Carpenter, J. H. J. Allum, and F. Honegger. Directional sensitivity of stretch refexes and balance corrections for normal subjects in the roll and pitch planes. *Experimental Brain Research*, 129(1):93-113, 1999.

[5] J. Lenarčič and T. Bajd. *Robotski mehanizmi*. Založba FE in FRI, Ljubljana, 2003.

[6] E. R. Kandel. *Principles of neural science*. McGrew-Hill, New York, 4 edition, 2000.

[7] S. P. Moore et al. Human automatic postural responses: responses to horizontal perturbations of stance in multiple directions. *Experimental Brain Research*, 73(3):648-658, 2004.

[8] M. G. Carpenter et al. Contributions od proprioceptive and vestibular inputs to postural control in the roll and pitch planes. In: J. H. J. Allum et al. Proprioceptive control of posture: a review of new concepts, *Gait & Posture*, 8(3):214-242, 1998

RITMIČNO VODENJE ROBOTA Z UPORABO NELINEARNIH OSCILATORJEV IN ČLOVEŠKE MIŠIČNE AKTIVNOSTI

Tadej Petrič, Andrej Gams, Martin Tomšič, Leon Žlajpah
Department for automation, biocybernetics and robotics
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773461; fax: +386 1 4773154
e-mail: tadej.petric@ijs.si

ABSTRACT

In the paper we address the problem of controlling the periodic robot motion by surface electromiography (EMG), a technique by which muscle action potentials are gathered by electrodes placed on the patient's forearm skin. These potentials can be used to track the muscles activation of the patient. Surface EMG is therefore a relatively simple way of detecting what the patient is doing with his limbs. However, since the EMG signal is quasi periodic with a lot of frequency components and noise, it is difficult to determine the frequency and phase of the measured limb motion. Therefore we propose to use nonlinear dynamical system capable of extracting the frequency and the phase from an unknown periodic signal with an arbitrary waveform. The method uses a whole Fourier series representation in the feedback loop. In this way it is capable of extracting the frequency and phase from an unknown periodic signal in real-time and without any additional signal processing or preprocessing. Combining this method with an output dynamic system based on dynamic motion primitives, which generates the desired trajectory, enables synchronization between human muscles and robots.

1 UVOD

Ljudje stalno izvajamo različne periodične naloge, med katere spadajo npr. rokovanje [2], hoja [3, 11, 12], bobnanje [4] ali igranje z različnimi igrali, kot so jojo [5, 10] ali Powerball [1, 6, 14]. Skupno tem nalogam je, da se jih ljudje hitro naučijo s posnemanjem drugih ali s pomočjo navodil [1]. Po drugi strani pa je izvajanje takšnih nalog z roboti zahteven problem, ki zahteva uporabo naprednih algoritmov in senzorskih sistemov [1]. Vodenje ritmičnih nalog z roboti pa poleg preciznega generiranja želene trajektorije zahteva tudi natančno določevanje frekvence in faze gibanja. Natančno določevanje faze je še posebej pomembno, kadar je potrebno sistem sinhronizirani na določene naloge, kot je to npr. pri vodenju igrače jojo ali pri enostavni sinhronizaciji robota z gibanjem človeka.

Izvajanje periodičnih oziroma ritmičnih nalog z robotom lahko izvedemo na različne načine [3,4]. Eden od teh je, da s pomočjo ustreznega vodenja ponovimo vnaprej določen vzorec giba, ki je lahko posnetek demonstracije ali rezultat matematične izpeljave. Drug način je, da se trajektorija določa v odvisnosti od trenutnega stanja naprave [5]. Pri uporabi te metode mora imitacijski sistem poleg generiranja trajektorije ujeti tudi pravilno frekvenco in fazo.

Premike človeških okončin lahko zelo natančno izmerimo s pomočjo laboratorijskih merilnih naprav, kot je npr. optični sistem The Optotrak Certus®, ki ga sestavljajo kamere in markerji, pritrjeni na pacientu. Takšen sistem je statičen in zato ni primeren za širšo uporabo, kot je to npr. uporaba v umetnih udih ali exoskeletonih. Za slednje aplikacije se pogosto uporablja metoda EMG, ki meri mišično aktivnost pacienta.

Sinhronizacija in vodenje robotov s pomočjo merjenja mišične aktivnosti (EMG) je že bilo predmet raziskav [15], kjer so avtorji pokazali, da se s pomočjo EMG meritev in naprednih samo-učečih sistemov, kot so npr. adaptivne nevronske mreže, lahko uspešno izvaja diskretne gibe robota oz umetnega uda. V tem članku pa želimo pokazati, da lahko s pomočjo dvo-nivojskega imitacijskega sistema [13] in merjenja EMG signala robotsko gibanje sinhroniziramo v realnem času z gibanjem pacientovega uda. To je zahtevna naloga, saj je EMG signal kvazi-periodičen z velikim številom frekvenčnih komponent in slabim razmerjem signal-šum.

Predlagamo uporabo dvo-nivojskega imitacijskega sistema [13], kjer prvi nivo izloči frekvenco in fazo vhodnega signala, drugi nivo pa generira želeno trajektorijo. Prvi nivo, imenovan kanonični dinamični sistem, temelji na nelinearnem faznem oscilatorju združenem z adaptivno Fourierevo vrsto. Ta sistem nam omogoča določevanje frekvence in faze v realnem času brez procesiranja vhodnega signala prej in potem. Praktična uporabnost bo preliminarno prikazana na primeru sinhronizacije med bicepsi na pacientu in komolcem humanoidnega robota. Takšna struktura sistema se lahko uporablja tudi za imitacijo človeškega gibanja na robotu, vodenje exoskeletonov ali npr. rehabilitaciji s stimulatorji.

V članku je opisano ritmično vodenje robota z uporabo nelinearnega oscilatorja in merjenja mišične aktivnosti. V drugem poglavju je opisan sistem za določevanje frekvence in faze vhodnega signala. V tretjem poglavju je predstavljen preliminarni eksperiment z rezultati. Zaključki so podani v četrtem poglavju.

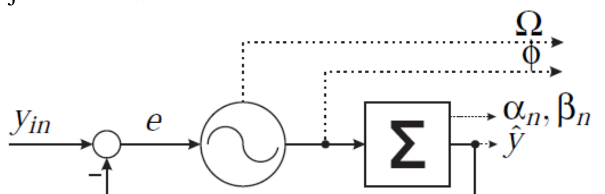
2 KANONIČNI DINAMIČNI SISTEM Z ADAPTIVNO FOURIERJEVO VRSTO

V tem poglavju je predstavljena arhitektura kanoničnega dinamičnega sistema z nelinearnim oscilatorjem in adaptivno Fourierevo vrsto. Osnova kanoničnega dinamičnega sistem je en frekvenčno prilagodljiv fazni oscilator [7, 8, 9]. Ta je združen z povratno-začno strukturo, ki vsebuje adaptivno Fourierevo vrsto. Povratna zanka z adaptivno Fourierevo vrsto je prikazana na sliki 1, podana pa je z enačbami

$$\dot{\phi} = \Omega - Ke\sin\phi, \tag{1}$$

$$\dot{\Omega} = -Ke\sin\phi, \tag{2}$$

kjer je K konstanta medsebojne povezanosti, ϕ je faza in Ω je frekvenca.



Slika 1: *Struktura adaptivnega oscilatorja, združenega v povratno zanko z adaptivno Fourierevo vrsto.*

Vhod e v oscilator je definiran z

$$e = y_{in} - \hat{y}, \tag{3}$$

kjer je y_{in} vhodni signal. Pri tem je $\hat{y}$ signal povratne zanke definiran z

$$\hat{y} = \alpha_0 + \sum_{i=1}^{M}\left(\alpha_i\cos(i\phi) + \beta_i\sin(i\phi)\right), \tag{4}$$

kjer je M velikost Fouriereve vrste, α_0, α_i in β_i pa so amplitude posameznih členov v Fourierevi vrsti. Algoritem učenja posameznih amplitud podaja

$$\dot{\alpha}_0 = \eta \cdot e, \tag{5}$$

$$\dot{\alpha}_i = \eta\cos(i\phi)\cdot e, \tag{6}$$

$$\dot{\beta}_i = \eta\sin(i\phi)\cdot e, \tag{7}$$

kjer je $i = 1...M$. Kot je razvidno iz slike 1, je vhod v oscilator razlika med vhodnim signalom in Fourierevo vrsto. Ker je uporabljena negativna povratna zanka, se bo napa približevala ničli, ko se bo Fouriereva vrsta približevala vhodnemu signalu. Takšna povratnozančna struktura izvede tako imenovano adaptivno Fourierevo analizo, kjer je fazna razlika med posameznimi harmoniki enaka $n\pi/2$, kjer je $n = 0,1,2,3$.

Predlagani pristop se tako lahko prilagodi na poljuben periodičen vhodni signal. Število harmonikov, ki jih lahko natančno določi, je odvisno od velikosti uporabljene Fouriereve vrste. Obenem pa je potrebno poudariti, da sta frekvenca in faza nedvoumno določeni, ker je uporabljen samo en oscilator. To pa je posebej uporabno kadar imamo opravka z robotskimi periodičnimi nalogami.

Poleg uvodoma omenjene povezave kanoničnega dinamičnega sistema v dvo-nivojski imitacijski sistem, pa lahko kanonični dinamični sistem uporabimo kot samostojen imitacijski sistem, ki lahko posnema poljuben periodičen signal. To lahko storimo, kadar e konvergira k ničli in dokler je e enak 0, bo signal zakodiran v Fourierevi vrsti. Natančnost konvergence signala je odvisna od števila uporabljenih členov v Fourierevi vrsti.

Proces učenja je vgrajen v algoritem in se izvaja v realnem času, zato ni potrebe po zunanjem optimizacijskem procesu ali dodatnemu algoritmu učenja. Z dodanim izhodnim dinamičnim sistemu [8] skupaj sistema tvorita dvo-nivojski imitacijski sistem, kar nam omogoča sinhronizacijo gibov robota glede na merjeno periodično veličino. V tem primeru je merjena periodična veličina sedaj zakodirana v Fourierevi vrsti, želena trajektorija pa v izhodnem dinamičnem sistemu. Ker lahko izločanje frekvence in učenje želene izhodne trajektorije storimo hkrati, bodo vse sistemske zakasnitve že avtomatsko zajete. Poleg tega pa lahko zakasnitve oziroma fazni zamik enostavno popravimo s prištevanjem želenega premika k fazi. To nam nato omogoča, da lahko uporabimo bodisi pred-definirano obliko želene trajektorije ali pa se sistem nauči nove.

Kljub temu, da se lahko kanonični dinamični sistem uporablja samostojno, pa ima povezava v dvo-nivojski sistem še dodatne prednosti, kajti omogoča nam enostavno moduliranje amplitude, frekvence in učenja zapletenih periodičnih vzorcev brez izločanja vseh frekvenčnih komponent. Obenem je dvo-nivojski sistem tudi bolj robusten [8], kar pa je pomembna lastnost kadar imamo opravka s hitrimi periodičnimi nalogami ali nalogami, kjer imamo opravka s šumnimi meritvami, kot je to npr. merjenje mišične aktivnosti.

Poleg že naštetih prednosti dvo-nivojskega sistema, ima ta sistem še dodatno prednost, kadar potrebujemo več izhodnih signalov s poljubnimi trajektorijami in faznimi zamiki med njimi, saj za sinhronizacijo potrebujemo le en kanonični dinamični sistem. Izhodi pa so zakodirani v poljubnem številu izhodnih dinamičnih sistemov. Število teh sistemov je enako številu aktuiranih prostostnih stopenj. Na ta način bi lahko npr. samo z merjenjem ene mišične aktivnosti na eni nogi vodili hkrati poljubno število aktuatorjev na drugi nogi. Trajektorije na drugi nogi bi bile med seboj pravilno fazno zamaknjene in sinhronizirane z mišično aktivnostjo na prvi nogi. To pa nakazuje na uporabnost takšnega pristopa tudi pri exoskeletonih ali pri rehabilitaciji, kjer bi npr. ena okončina sledila drugi (ena noga sledi drugi).

Izhodni dinamični sistem dvo-nivojskega imitacijskega sistema je podan z

$$\dot{z} = \Omega\left(\alpha_z \left(\beta_z \left(g - y \right) - z \right) + \frac{\sum_{i=1}^{N} \Psi_i w_i r}{\sum_{i=1}^{N} \Psi_i} \right), \tag{6}$$

$$\dot{y} = \Omega z. \tag{7}$$

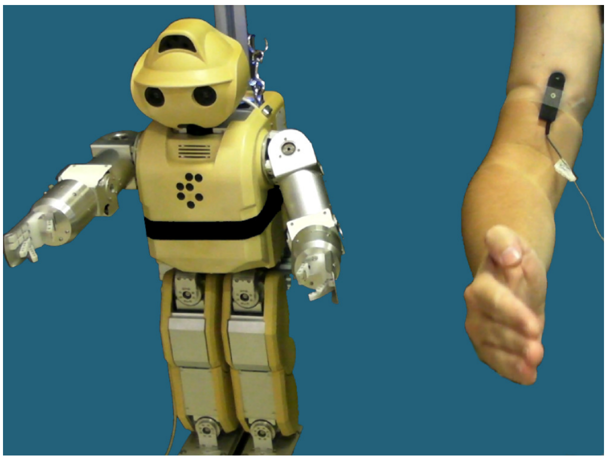
Kjer je Ω frekvenca, dobljena iz kanoničnega dinamičnega sistema, $\alpha_z = 8$ in $\beta_z = 2$ sta pozitivni konstanti, N je število Gaussovih jedrnih funkcij. Amplituda trajektorije je podana s parametrom r. Nelinearna odvisnost določa želeno obliko trajektorije z množenjem jedrnih funkcij in pred-definiranega vektorja uteži. Jedrne funkcije so podane z

$$\Psi_i = \exp\left(h\left(\cos\left(\phi - c_i \right) - 1 \right) \right), \tag{8}$$

in imajo Gaussovo obliko. Parametra jedrnih funkcij sta h, ki določa njihovo širino in c_i, ki določa njihovo porazdeljenost preko ene periode. Parameter c_i je enakomerno porazdeljen med 0 in 2π v N korakih.

3 APLIKACIJA SINHRONIZACIJE ROBOTA NA MIŠIČNO AKTIVNOST

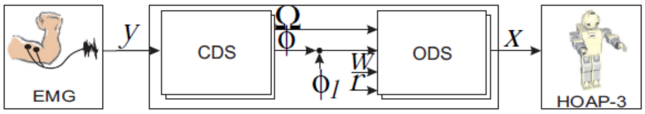
V tem poglavju so predstavljeni rezultati sinhronizacije robota na EMG signal, izmerjen na bicepsih človeške roke. Eksperimentalni sestav prikazuje slika 2. Takšen tip aplikacije je uporaben predvsem za vodenje protez ali exoskeletonov.



Slika 2: *Eksperimentalen sestav.*

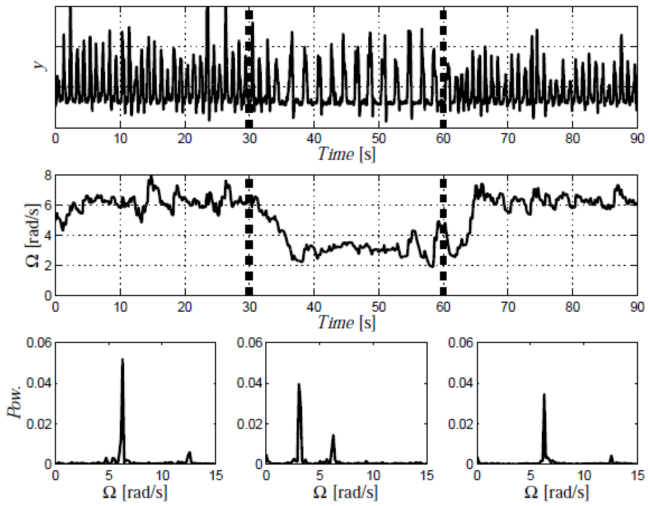
Za merjenje mišične aktivnosti smo uporabili elektrode EMG in signal monitor Myolab II MA-110 podjetja Motion Control Inc. Elektrode so površinske in so integrirane s predojačevalniki, ki imajo ojačanje od 270x do 380x. Elektrode smo pritrdili na kožo nad biceps mišico, kot je to prikazano na sliki 2. Eksperimentalni osebi smo naročili, naj izvaja fleksijo komolca (bicepsa), ko zasliši zvočni signal. Zvočni signal (pisk) je imel frekvenco ponavljanja posameznih piskov 1 Hz na začetku, 0.5 Hz po 30 s in nazaj na 1 Hz po 60 s od začetka izvajanja eksperimenta.

Vodenje je bilo izvedeno v Matlab/Simulinku programskem okolju. Blokovna shema je podana na sliki 3, kjer opazimo, da je vhod v sistem ovojnica izmerjenega EMG signala. Dvo-nivojski imitacijski sistem je sestavljen iz nelinearnega oscilatorja, združenega z adaptivno Fourierevo vrsto (CDS). Naloga tega dela sistema je izločiti frekvenco in fazo vhodnega signala. Izhodni dinamični sistem (ODS) dvo-nivojskega sistema pa poskrbi za želeno trajektorijo robota.



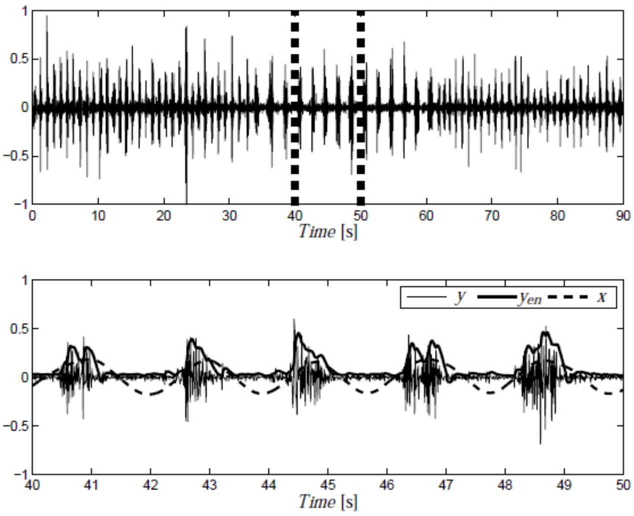
Slika 3: *Blokovna shema predlaganega algoritma vodenja.*

Na sliki 4 so prikazani rezultati frekvence vhodnega signala (vmesni graf). Vhodni signal je ovojnica izmerjenega EMG signala (zgornji graf). Spodnji trije grafi prikazujejo močnostni spekter vhodnega signala v časovnih intervalih od 0 do 30 s, od 30 s do 60 s in od 60 do 90 s. S primerjavo izločene frekvence (vmesni graf) in izračunanimi močnostnimi spektri (spodnji grafi) vidimo, da je sistem uspešno konvergiral k osnovni frekvenci vhodnega signala tudi ob spremembi ritma.



Slika 4: *Ovojnica izmerjenega EMG signala na zgornjem grafu. Izločena frekvenca na vmesnem grafu in močnostni spektri v spodnjih grafih.*

Na sliki 5 je na zgornjem grafu predstavljena meritev EMG signala. Na spodnjem grafu pa je podana primerjava med izmerjenim EMG signalom, njegovo ovojnico in generirano trajektorijo robota na odseku od 40 s do 50 s. Na spodnjem grafu na sliki 5 lahko opazimo, da je faza med vhodnim EMG signalom in generirano trajektorijo ves čas enaka. Ta eksperiment je pokazal, da je z uporabo dvo-nivojskega sistema mogoče iz ovojnice EMG signala izločit frekvenco in fazo v realnem času brez dodatnega pred- in po- procesiranja vhodnega signala.



Slika 5: *EMG signal na zgornjem grafu in primerjava med EMG signalom y, ovojnico EMG signala $\hat{y}$ in generirano trajektorijo robota x.*

4 ZAKLJUČEK

V članku smo pokazali sinhronizacijo robota na EMG signal s pomočjo dvo-nivojskega imitacijskega sistema. Ta dinamični sistem, ki je uporabljen za določanje frekvence in faze vhodnega signala, temelji na nelinearnem oscilatorju, združenim z Fourierevo vrsto. Sistem lahko izloči frekvenco in koeficiente Fouriereve vrste poljubnega periodičnega signala v realnem času brez dodatnega pred- ali po-procesiranja signala. Integracija sistema z izhodnim dinamičnim sistemom, pa omogoča enostavno in računsko učinkovito vodenje različnih ritmičnih periodičnih nalog, ki imajo vsaj eno periodično merljivo veličino. Med te naloge spada tudi sinhronizacija človeškega gibanja in robota.

Pokazali smo, da se lahko takšen sistem samostojno sinhronizira na ovojnico EMG signala (usmerjenega in integriranega) in generira trajektorijo robota s pravilno frekvenco in fazo. To pa ima potencial v vodenjih protez in eksoskeletonov.

Literatura

[1] T. Petric, A. Gams, and L. Zlajpah. Modeling and control strategy for robotic powerball. In *18th International Workshop on Robotics in Alpe-Adria-Danube Region, Brasov, Romania, May 25-27, 2009, RAAD 2009 : [proceedings]*, 2009.

[2] Tomofumi Kasuga and Minoru Hashimoto. Human-robot handshaking using neural oscillators. In *ICRA*, pages 3802–3807, 2005.

[3] Auke Jan Ijspeert. Central pattern generators for locomotion control in animals and robots: A review. *Neural Networks*, 21(4):642–653, 2008.

[4] Sarah Degallier, Ludovic Righetti, Lorenzo Natale, Francesco Nori, Giorgio Metta, and Auke Ijspeert. Amodular bio-inspired architecture formovement generation for the infantlike robot iCub. In *Proceedings of the 2nd IEEE RAS / EMBS International Conferenceon Biomedical Robotics and Biomechatronics (BioRob)*, 2008. 20

[5] Leon Žlajpah. Robotic yo-yo: modelling and control strategies. *Robotica*, 24(2):211–220, 2006.

[6] Andrej Gams, Leon Žlajpah, and Jadran Lenarčič. Imitating human acceleration of a gyroscopic device. *Robotica*, 25(4):501–509, 2007.

[7] Jonas Buchli, Ludovic Righetti, and Auke Jan Ijspeert. Frequency Analysis with coupled nonlinear scillators. *Physica D: Nonlinear Phenomena*, 237:1705–1718, 2008.

[8] Andrej Gams, Auke Jan Ijspeert, Stefan Schaal, and Jadran Lenarčič. On-line learning and modulation of periodic movements with nonlinear dynamical systems. *Auton. Robots*, 27(1):3–23, 2009.

[9] Ludovic Righetti, Jonas Buchli, and Auke Jan Ijspeert. Dynamic hebbian learning in adaptive frequency oscillators. *Physica D*, 216(2):269–281, 2006.

[10] Hui-Liang Jin, Qian Ye, and Miriam Zacksenhouse. Return maps, parameterization, and cycle-wise planning of yo-yo playing. *Trans. Rob.*, 25(2):438–445, 2009.

[11] J. Buchli and A.J. Ijspeert. A simple, adaptive locomotion toy-system. In S. Schaal, A.J. Ijspeert, A. Billard, S. Vijayakumar, J. Hallam, and J.A. Meyer, editors, *From Animals to Animats 8. Proceedings of the Eighth International Conference on the Simulation of Adaptive Behavior (SAB'04)*, pages 153–162. MIT Press, 2004.

[12] L. Righetti and A. J. Ijspeert. Programmable Central Pattern Generators: an application to biped locomotion control. In *Proceedings of the 2006 IEEE International Conference on Robotics and Automation*, 2006. 21

[13] Andrej Gams, Tadej Petrič, and Leon Žlajpah. Controlling yo-yo and gyroscopic device with nonlinear dynamic systems. In *18th International Workshop on Robotics in Alpe-Adria-Danube Region, Brasov, Romania, May 25-27, 2009, RAAD 2009 : [proceedings]*, 2009.

[14] Tadej Petrič, Boris Curk, Peter Cafuta, and Leon Žlajpah. Modeling of the robotic powerball: a nonholonomic, underactuated, and variable structure-type system. *Mathematical and Computer Modelling of Dynamical Systems*, 2010.

[15] C. Castellini and P. van der Smagt. Surface emg in advanced hand prosthetics. *Biological Cybernetics*, 100(1):35–47, 2009.

ADAPTIVNA SENZORNA INTEGRACIJA NA OSNOVI BIOMEHANSKIH IN FIZIOLOŠKIH MERITEV V REHABILITACIJSKI ROBOTIKI

Domen Novak, Matjaž Mihelj, Jaka Ziherl, Andrej Olenšek, Marko Munih
Laboratorij za robotiko in biomedicinsko tehniko
Fakulteta za elektrotehniko
Tržaška 25, 1000 Ljubljana, Slovenia
Tel: +386 1 4768196; fax: +386 1 4768239
e-mail: domen.novak@robo.fe.uni-lj.si

POVZETEK

V prispevku je predstavljena metoda za senzorno integracijo na osnovi biomehanskih in fizioloških meritev v sistemu za rehabilitacijo roke. Sistem temelji na haptičnem robotu HapticMaster, ki je povezan z navideznim okoljem, v katerem lahko uporabnik vadi gibanje roke in prijemanje predmetov. Senzorji na robotu merijo gibe, sile in navore, poleg tega pa psihofiziološki senzorji merijo elektrokardiogram, prevodnost kože, dihanje in temperaturo kože uporabnika. S pomočjo adaptivne diskriminantne analize vse te meritve združimo v oceno, kakšna težavnost naloge bi bila primerna za uporabnika. Metoda je bila preizkušena na 24 zdravih osebah in 11 hemiparetičnih pacientih. V navzkrižni validaciji je bila pravilnost sistema (odstotek pravilno napovedanih merjenčevih izbir) 84,7 % za zdrave osebe in 89,4 % za paciente.

ABSTRACT

This paper presents a method of integrating biomechanical and physiological measurements in a hand and arm rehabilitation system. The system is based on the HapticMaster haptic robot, which is connected to a virtual environment that allows the user to practice arm movement and grasping. The robot's sensors measure movement, forces and torques. Additionally, psychophysiological sensors measure the user's electrocardiogram, skin conductance, respiration and skin temperature. Using adaptive discriminant analysis, all these measurements are integrated into an estimate of the difficulty level that would be appropriate for the user. The method was tested with 24 healthy subjects and 11 hemiparetic patients. In cross-validation, the system accuracy (percentage of correctly predicted subjects' choices) was 84.7 % for healthy subjects and 89.4 % for patients.

1 UVOD

V današnjem času roboti sodelujejo z ljudmi na najrazličnejših področjih. Eno od teh področij je rehabilitacija, kjer lahko roboti nadomestijo terapevta in pacientu nudijo intenzivno vadbo oslabljenih okončin [1]. S pomočjo senzorjev sil in premikov lahko pacientove gibe razpoznajo v začetni fazi giba in pacientu pomagajo le toliko, kot je nujno potrebno [2]. Žal pa so raziskave pokazale, da je težko ovrednotiti in interpretirati subjektivne občutke človeka do robota [3], saj senzorji sil in premikov ne nudijo informacij o človekovih občutkih: stresu, zadovoljstvu, motivaciji … Ravno ti občutki pa so pri rehabilitaciji zelo pomembni, saj motivirani pacienti intenzivneje vadijo in hitreje okrevajo [4].

Informacije o pacientovem psihološkem stanju bi lahko robot pridobil s pomočjo takoimenovanih psihofizioloških meritev: meritev fizioloških odzivov na psihološke dražljaje. Te so se v interakciji človek-stroj že izkazale za zelo uspešne. Čustveni odzivi igralcev računalniških iger so naprimer vidni v njihovem srčnem utripu, prevodnosti kože in temperaturi kože [5]. Nedavne raziskave psihofizioloških odzivov pacientov med rehabilitacijo roke so prav tako pokazale povezavo med fiziološkimi odzivi in psihološkimi stanji [6]. Tako bi v rehabilitacijski robotiki lahko združili biomehanske meritve s psihofiziološkimi meritvami in pridobili bolj popolno, celostno sliko stanja pacienta.

V prispevku je predstavljena integracija uspešnosti v nalogi, biomehanskih meritev (sile in premiki) ter psihofizioloških meritev (srčni utrip, prevodnost kože, frekvenca dihanja in temperatura kože) v rehabilitacijski robotiki s pomočjo diskriminantne analize. Ker se psihofiziološke meritve močno razlikujejo od osebe do osebe, je predstavljena tudi metoda prilagajanja sistema trenutnemu uporabniku s pomočjo adaptivne različice diskriminantne analize.

2 MATERIALI IN METODE

Cilj raziskave je bil združiti meritve, pridobljene med rehabilitacijo roke v navidezni resničnosti, in iz njih določiti, kakšna težavnost vaje bi bila primerna za uporabnika. Strojna oprema sistema je predstavljena v poglavju 2.1, navidezno okolje pa v poglavju 2.2. Postopek merjenja je opisan v poglavju 2.4. Iz surovih signalov smo po postopkih, opisanih v poglavju 2.5, izluščili več spremenljivk in jih združili v oceno primerne težavnosti po metodi, opisani v poglavju 2.6. Metoda je bila preverjena z navzkrižno validacijo, ki je podrobneje predstavljena v poglavju 2.7.

2.1 Strojna oprema

Strojna oprema sestoji iz treh komponent: haptičnega vmesnika, prikazovalnika navideznega okolja, in psihofiziološkega merilnega sistema.

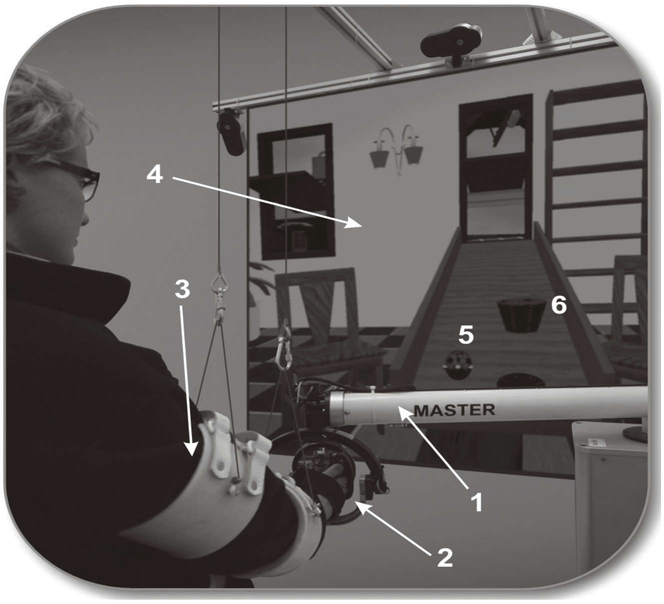
Haptični vmesnik je osnovan na haptičnem robotu HapticMaster [7] s tremi prostostnimi stopnjami. Na vrhu robota se nahaja tudi senzor za merjenje sile v treh oseh kartezičnega koordinatnega sistema. Na senzor za merjenje sile je pritrjen modul za prijemanje predmetov v navideznih okoljih. Robot HapticMaster z navideznim okoljem je prikazan na sliki 1. Poleg robota HapticMaster z modulom za prijemanje ima haptični vmesnik še sistem za kompenzacijo teže roke. Na roko uporabnika sta pritrjeni dve manšeti (ena nad in ena pod komolcem), ki sta preko vrvi povezani z motorjema na stropu. Tako na uporabnikovo roko stalno deluje sila, ki kompenzira silo teže.

Navidezno okolje je prikazano na ekranu velikosti 1,4 x 1,4 metra. Uporabnik sedi približno 1,25 metra pred ekranom, haptični vmesnik pa je postavljen med sedež in ekran.

Med nalogo lahko s pomočjo ojačevalnika g.USBamp (proizvajalec g.tec Medical Engineering GmbH, Avstrija) in pripadajočih senzorjev merimo štiri različne fiziološke signale: elektrokardiogram, prevodnost kože, frekvenco dihanja in temperaturo kože. Elektrode za elektrokardiogram so nameščene na trup. Senzor za prevodnost kože je realiziran v obliki dveh elektrod, nameščenih na kazalec in sredinec. Senzor za temperaturo kože je termistor, nameščen na konico mezinca. Tudi senzor dihanja je realiziran v obliki termistorja, ki je nameščen pod nos merjenca in meri razliko med temperature vdihanega ter izdihanega zraka.

2.2 Navidezno okolje

V raziskavi smo uporabili navidezno okolje, ki je bilo že uporabljeno v raziskavi psihofizioloških odzivov v rehabilitacijski robotiki [6]. Prikazano je na sliki 1 skupaj s haptičnim vmesnikom. V sredini navideznega okolja je miza, nagnjena navzdol proti gledalcu. Vrh robota HapticMaster je predstavljen z rumenimi oznakami. Na vrhu mize se pojavi žoga, ki se začne kotaliti navzdol. Naloga uporabnika je, da s pomočjo haptičnega vmesnika prime žogo, preden le-ta doseže spodnji rob mize. Ko je žoga prijeta, se nad mizo pojavi koš. Uporabnik mora žogo držati in jo postaviti v koš. Ko je žoga v košu ali pa pade z mize, se na vrhu mize pojavi nova žoga in naloga se nadaljuje. Okolje ima sedem različnih težavnostnih stopenj. Z višanjem stopnje žoga postaja vedno hitrejša in manjša.



Slika 1: Merjenec opravlja nalogo z robotom HapticMaster (1) in modulom za prijemanje (2), medtem ko je roka podprta z manšetama (3). Ekran (4) prikazuje nagnjeno mizo, žogo (5) in koš (6).

2.3 Merjenci

V raziskavi je sodelovalo štiriindvajset zdravih oseb (20 moških, 4 ženske, starost 31.1 ± 10.9 let) ter enajst hemiparetičnih pacientov po možganski kapi oziroma operaciji neoplazme možganov (8 moških, 3 ženske, starost 43.2 ± 13.5 let). Šest pacientov je imelo levostransko, pet pa desnostransko hemiparezo.

2.4 Merilni postopek

Ob prihodu je vodja raziskave vsakemu merjencu predstavil robota in raziskave. Po podpisu soglasja o sodelovanju v raziskavi se je merjenec usedel pred robota, terapevt pa je nanj namestil manšeti za kompenzacijo sile teže ter psihofiziološke merilne senzorje. Merjenec je nato opravil kratko vajo naloge pri tretji težavnostni stopnji. Po vaji je merjenec dve minuti počival. Ta interval je bil namenjen merjenju psihofizioloških signalov v mirovanju. Nato je pričel z opravljanjem naloge na težavnostni stopnji 3, 4 ali 5 (naključno izbrano). Po dveh minutah opravljanja naloge je bila ta za kratek čas prekinjena, merjenec pa je bil vprašan, ali bi raje nadaljeval z višjo ali z nižjo težavnostno stopnjo. Težavnost je bila spremenjena za eno ali dve stopnji v izbrano smer. Naloga se je nato nadaljevala na novi težavnostni stopnji. Merjenec je nalogo skupaj opravljal dvanajst minut (šestkrat po dve minuti, s spremembo težavnosti na vsaki dve minuti).

2.5 Obdelava signalov

Med počitkom in opravljanjem naloge so bili signali zajemani z vzorčno frekvenco 100 Hz (sile in premiki) oziroma 1200 Hz (fiziologija). Po zaključku meritev smo za vsak dvominutni interval iz surovih signalov izluščili več relevantnih spremenljivk. Te so razdeljene v tri kategorije: uspešnost, biomehanika in psihofiziologija.

Spremenljivke uspešnosti opisujejo, kako dobro je merjenec opravil nalogo in kako dolgo jo je že opravljal. Štiri spremenljivke uspešnosti so *trenutna težavnost* (1-7), *trenutni interval naloge* (1 – prvi, 6 – zadnji), *odstotek ujetih žog* ter *odstotek žog, postavljenih v koš*.

Biomehanske spremenljivke opisujejo sile in gibe, ki jih izvajajo merjenci. Te spremenljivke so *povprečna sila*, *povprečna hitrost*, *povprečni pospešek*, *skupno delo*, *povprečna frekvenca signala pozicije*, *povprečna frekvenca signala hitrosti*, *povprečna frekvenca signala pospeška*, in *povprečna frekvenca signala sile*. Izračunane so le za gibe v vodoravni ravnini med ujemanjem žoge.

Psihofiziološke spremenljivke so izračunane iz štirih fizioloških signalov: srčnega utripa (elektrokardiograma), prevodnosti kože, dihanja in temperature kože. Spremenljivke, pridobljene iz srčnega utripa, so *povprečni srčni utrip* ter pet ocen variabilnosti srčnega utripa: *SDNN*, *RMSSD*, *pNN50*, *skupna moč v nizkofrekvenčnem pasu srčnega utripa*, ter *skupna moč v visokofrekvenčnem pasu srčnega utripa*. Te ocene so podrobno definirane v [8]. Signal prevodnosti kože delimo na nivo prevodnosti kože (nizkofrekvenčna komponenta) in odzive prevodnosti kože (hitra, kratkotrajna povečanja prevodnosti kože). Spremenljivke, pridobljene iz prevodnosti kože, so *povprečni nivo prevodnosti*, *povprečni odvod nivoja prevodnosti*, *povprečna frekvenca odzivov prevodnosti*, in *povprečna amplituda odzivov prevodnosti*. Spremenljivki, pridobljeni iz dihanja, sta *povprečna frekvenca dihanja* in *standardna deviacija frekvence dihanja*. Iz temperature kože je izluščena *temperatura ob koncu intervala*.

2.6 Senzorna integracija

Izračunane spremenljivke je potrebno združiti v oceno, kakšna težavnost naloge bi bila primerna za uporabnika. Ker v primeru rehabilitacijske robotike povezave med gibi, psihofiziologijo in težavnostjo naloge niso dobro znane, smo se odločili za senzorno integracijo na podlagi učne množice in metode diskriminantne analize.

Med meritvijo smo merjence večkrat vprašali, ali bi raje nadaljevali z višjo ali z nižjo težavnostno stopnjo. Tako smo pridobili učno množico z znanimi vhodi v sistem (spremenljivke, izračunane v poglavju 2.5) ter znanimi izhodi iz sistema (merjenec izbral višjo/nižjo težavnost). Klasifikacijo večih vhodnih spremenljivk v dva možna razreda (višja / nižja težavnost) lahko opravimo s statistično metodo linearne diskriminantne analize [9]. Enačbe za klasifikacijo vhodnega vektorja $\boldsymbol{x}$ v enega od dveh razredov s pomočjo linearne diskriminantne analize so:

$$D(\boldsymbol{x}) = b + \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x}$$
$$b = -\boldsymbol{w}^{\mathrm{T}} \cdot \tfrac{1}{2} \cdot (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$
$$\boldsymbol{w} = (\boldsymbol{S}_1 + \boldsymbol{S}_2)^{-1} \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$
$$C(\boldsymbol{x}) = \begin{cases} 1; D(\boldsymbol{x}) < 0 \\ 2; D(\boldsymbol{x}) \geq 0 \end{cases}$$

kjer so $D(\boldsymbol{x})$ diskriminantna funkcija, b in $\boldsymbol{w}$ uteži te funkcije, $\boldsymbol{S}_k$ kovariančna matrika razreda, $\boldsymbol{\mu}_k$ povprečna vrednost za razred k, $C(\boldsymbol{x})$ pa izhodni razred za vektor $\boldsymbol{x}$.

Obstaja tudi adaptivna verzija linearne diskriminantne analize, ki s pomočjo Kalmanovega filtra rekurzivno posodablja uteži diskriminantne funkcije, ko dobi dostop do novega podatka [10]. V našem primeru jo uporabimo tako, da diskriminantno funkcijo najprej zgradimo s podatki iz učne množice, ko pa merjenec opravlja nalogo, se po vsakem dveminutnem intervalu uteži diskriminantne funkcije posodobijo s podatki iz preteklega intervala (vektor vhodnih spremenljivk in merjenčeva izbira). Tako se diskriminantna funkcija vedno bolj prilagaja trenutnemu merjencu. Enačbe za adaptivno diskriminantne analizo so:

$$\boldsymbol{H}_k = [1, \boldsymbol{x}_k^{\mathrm{T}}]$$
$$e_k = y_k - \boldsymbol{H}_k \cdot \hat{\boldsymbol{w}}_{k-1}$$
$$v_k = 1 - UC$$
$$Q_k = \boldsymbol{H}_k \cdot \boldsymbol{A}_{k-1} \cdot \boldsymbol{H}_k^{\mathrm{T}} + v_k$$
$$\boldsymbol{k}_k = \frac{\boldsymbol{A}_{k-1} \cdot \boldsymbol{H}_k^{\mathrm{T}}}{Q_k}$$
$$\hat{\boldsymbol{w}}_k = \hat{\boldsymbol{w}}_{k-1} + \boldsymbol{k}_k \cdot e_k$$
$$\tilde{\boldsymbol{A}}_k = \boldsymbol{A}_{k-1} - \boldsymbol{k}_k \cdot \boldsymbol{H}_k \cdot \boldsymbol{A}_{k-1}$$
$$\boldsymbol{A}_k = \frac{\mathrm{trace}(\tilde{\boldsymbol{A}}_k) \cdot UC}{p} + \tilde{\boldsymbol{A}}_k$$

kjer so e_k napaka predikcije Kalmanovega filtra, y_k trenutni razred, $\boldsymbol{x}_k$ trenutni vhodni vektor, $\hat{\boldsymbol{w}}_k$ vektor stanj ($\hat{\boldsymbol{w}}_k = [b, \boldsymbol{w}^T]$, uteži diskriminantne funkcije), Q_k ocenjena variance predikcije, $\boldsymbol{A}_k$ takoimenovana "a priori state error correlation matrix", $\tilde{\boldsymbol{A}}_k$ vmesna vrednost, potrebna za izračun $\boldsymbol{A}_k$, v_k varianca inovacijskega procesa, $\boldsymbol{k}_k$ ojačanje Kalmanovega filtra, UC posodobitveni koeficient in p število elementov vektorja $\hat{\boldsymbol{w}}_k$.

2.7 Navzkrižna validacija

Uspešnost diskriminantne analize pri integraciji meritev in napovedovanju merjenčevih želja smo ocenili s postopkom navzkrižne validacije. Diskriminantno funkcijo, ki iz signalov napove merjenčevo izbiro (višjo/nižjo težavnost) smo zgradili z učno množico, ki je vsebovala meritve vseh merjencev razen enega. To diskriminantno funkcijo smo nato preizkusili na preostalem merjencu. Postopek je bil ponovljen tolikokrat, kolikor je bilo merjencev. Zgradili smo več diskriminantnih funkcij: samo s spremenljivkami uspešnosti, samo z biomehanskimi spremenljivkami, samo s psihofiziološkimi spremenljivkami in z vsemi spremenljivkami. Preizkusili smo tako neadaptivno kot adaptivno diskriminantno analizo. Zdrave osebe in pacienti so bili obravnavani ločeno.

Diskriminantne funkcije smo ocenili glede na zmožnost napovedovanja merjenčevih izbir (izbral višjo/nižjo težavnost). Pravilnost diskriminantne funkcije je bila definirana kot odstotek pravilno napovedanih izbir.

3 REZULTATI

Tabela 1 prikazuje pravilnost diskriminantne funkcije za obe skupini merjencev (zdravi in pacienti), za neadaptivne in adaptivne funkcije, ter za različne tipe vhodnih podatkov.

Tabela 1: Pravilnost diskriminantnih funkcij v odstotkih za različne tipe vhodnih podatkov. Okrajšave: usp. = uspešnost, meh. = biomehanika, psih. = psihofiziologija.

	usp.	meh.	psih.	vse
zdravi neadaptivno	81,9	75,0	60,4	84,7
zdravi adaptivno	82,6	80,6	76,4	84,7
pacienti neadaptivno	81,8	75,8	60,6	89,4
pacienti adaptivno	81,8	75,8	68,2	89,4

4 RAZPRAVA

Za najpomembnejše so se izkazale spremenljivke uspešnosti, ki so tako pri zdravih kot pacientih vrnile najvišjo pravilnost diskriminantne funkcije (cca. 82 %). Pravilnost biomehanskih spremenljivk je bila slabša, psihofiziološke spremenljivke pa so se v neadaptivni diskriminantni analizi izkazale za relativno nezanesljive (pravilnost cca. 60 %). Adaptivna diskriminantna analiza lahko izboljša pravilnost psihofizioloških spremenljivk (za 16 % pri zdravih in 8 % pri pacientih), kar pa je še vedno slabše od ostalih tipov spremenljivk. Zaključimo lahko, da psihofiziološke meritve same po sebi niso zanesljive v rehabilitacijski robotiki.

Združitev različnih tipov podatkov poveča skupno pravilnost diskriminantne funkcije. Nadaljnja analiza je pokazala, da je povečanje rezultat kombinacije uspešnosti in psihofiziologije. Tako psihofiziologija lahko izboljša sistem, vprašljivo pa je, ali je izboljšava dovolj velika, da upraviči povečano kompleksnost sistema. Če imamo na voljo ocene uspešnosti, so psihofiziološke meritve verjetno nepotrebne, lahko pa se izkažejo za uporabne v primerih, ko ni možno objektivno oceniti uspešnosti pacienta v nalogi.

Adaptivni pristop izboljša pravilnost psihofizioloških spremenljivk, manj uspešen pa je pri drugih tipih podatkov. Vseeno pa vidimo, da se lahko sistem prilagodi trenutnemu merjencu, če ima na voljo podatke o merjenčevih izbirah.

5 ZAKLJUČEK

V prispevku smo predstavili metodo integracije različnih tipov meritev. S predstavljeno metodo lahko napovemo, kakšna težavnost naloge bi bila primerna za uporabnika, adaptivna komponenta pa se lahko prilagodi trenutnemu uporabniku in tako poveča uporabnost sistema. Metodo bi lahko uporabili tudi na drugih področjih, kjer je zahtevana integracija fizioloških meritev z drugimi viri informacij.

Zahvala

Raziskavo sta podprla Javna agencija za raziskovalno dejavnost Republike Slovenije ter Sedmi okvirni program Evropske skupnosti (projekt MIMICS, grant 215756).

Viri

[1] Lo, A. C., Guarino, P. D., Richards, L. G., Haselkorn, J. K., Wittenberg, G. F., Federman, D. G., et al. (2010). Robot-assisted therapy for long-term upper-limb impairment after stroke. *New England Journal of Medicine, 362*, 1772-1783.

[2] Riener, R., Lünenburger, L., Jezernik, S., Anderschitz, M., Colombo, G., & Dietz, V. (2005). Patient-cooperative strategies for robot-aided treadmill training: First experimental results. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 13*, 380-394.

[3] Bethel, C. L., Salomon, K., Murphy, R. R. and Burke, J. L. (2007) Survey of psychophysiology measurements applied to human-robot interaction. *16th IEEE International Conference on Robot and Human Interactive Communication*, 732-737.

[4] N. Maclean, P. Pound, C.Wulfe, and A. Rudd, “The concept of patient motivation: A qualitative analysis of stroke professionals’ attitudes,” *Stroke*, vol. 33, pp. 444–448, Feb. 2002.

[5] van Reekum, C. M. and Johnstone, T. Psychophysiological responses to appraisal dimensions in a computer game (2004). *Cognition and Emotion, 18*, 663-688.

[6] Novak, D., Ziherl, J., Olenšek, A., Milavec, M., Podobnik, J., Mihelj, M., et al. (2010). Psychophysiological responses to robotic rehabilitation tasks in stroke. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 18*, 351-361.

[7] R. Q. van der Linde, P. Lammertse. HapticMaster – a generic force controlled robot for human interaction. *Industrial Robot: An International Journal*, 30:515-524, 2003.

[8] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17, 354-381, 1996.

[9] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

[10] Vidaurre, C., Schlögl, A., Cabeza, R., Scherer, R., & Pfurtscheller, G. (2007). Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces. *IEEE Transactions on Biomedical Engineering, 54*, 550-556.

Zbornik 13. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2010

Proceedings of the 13th International Multiconference

INFORMATION SOCIETY – IS 2010

MATCOS 2010
(Mini konferenca iz uporabnega teoretičnega računalništva)

MATCOS 2010
(Mini-conference on Applied Theoretical Computer Science)

Uredil / Edited by

Andrej Brodnik

http://is.ijs.si

13.–14. oktober 2010 / October 13th–14th, 2010
Koper, Slovenia

PREDGOVOR / PREFACE

Draga bralka, dragi bralec!

MATCOS-10. Minikonferenca iz uporabnega teoretičnega računalništva in informatike (*Miniconference on Applied Theoretical Computer Science*) 2010. Kadarkoli organiziramo konferenco imamo seveda določene cilje. Najprej bi želeli dati možnost tistim raziskovalcem, ki so prišli do novih in pomembnih rezultatov na področju uporabnega teoretičnega računalništva in informatike, da jih predstavijo. Poleg tega menimo, da je pomembno, da se o tistih rezultatih, ki rešujejo praktične probleme z novimi in zanimivimi metodami, razvije razprava med, po in ob predstavitvi. Ker imajo organizatorji stike z vrsto mladimi študenti računalništva in informatike ter matematike, ki se ukvarjajo s področji kot sta teoretično računalništvo in informatika ter kombinatorične optimizacije, smo tem študentom želeli omogočiti, da občutijo »atmosfero resnične mednarodne konference«. Seveda si želimo, da si ob tem obrusijo svojo tehniko in način predstavitve.

Programski odbor je izbral prispevke, ki kažejo »barvitost« raziskovalnih interesov udeležencev in hkrati kažejo na trdno vpetost v praktične probleme. Upamo, da bodo letošnji Minikonferenci sledile naslednje ter smo hkrati trdno prepričani, da bo njihov nivo še višji in višji.

Na koncu a ne nazadnje: konferenca poteka v mestu Koper, ki je eden najlepših biserov na ogrlici Jadranskega morja. To še dodatno poudarja naše upanje, da se vidimo tudi naslednje leto na MATCOS-11!

Gábor Galambos and Andrej Brodnik

Dear reader :

MATCOS-10. Miniconference on Applied Theoretical Computer Science 2010. When organizing the conference, we had certain aims. Firstly, we would like to give possibility to those researchers who have new, significant results on the applied theoretical computer science to present them. For us it is also important that those results, which based on a practical problem and are solved by a remarkable method, get a panel discussion to show these efforts. Since the organizers have contact with a lot of young computer scientists and mathematicians who are students and are working in the field of the theoretical computer science and the combinatorial optimization, we wanted to give them to have a possibility to "feel the atmosphere" of an international conference. It is also preferable that they practice techniques and style, of course.

The programme committee selected papers, which show the "varicoloured" research interest of the participants, and also present a strong practical background. We hope that this year's Miniconference will be followed by others and we strongly believe that the level of the forthcoming conferences will be higher and higher.

Last but not least: the conference takes place in Koper, which is one of the most beautiful bead of the Adriatic. This also emphasizes our hope: see you next year in Koper on the MATCOS-11!

Gábor Galambos and Andrej Brodnik

Applications of combinatorics in statics*

András RECSKI
Budapest University of Technology and Economics
Email: recski@cs.bme.hu

Rigidity of frameworks, composed of rigid rods and rotatable joints, is an important question in civil engineering and, at the same time, a rich area for the applications of combinatorics.

In the first part of the talk

- we introduce the necessary concepts (rigidity and infinitesimal rigidity of the frameworks, the classical results of Maxwell etc),
- present results in the planar case, like Laman's theorem on the characterization of generic rigidity in the plane, the algorithm of Lovász and Yemini for checking this condition,
- briefly review the analogous results and open questions for tensegrity frameworks (where not only rigid rods but cables and struts are also permitted).

Then we concentrate on a special case: how to make a square grid rigid in the plane (or a cubic grid rigid in the space) by using diagonal bars. Here we shall see that seemingly innocent engineering questions can lead to very interesting (and sometimes quite deep) results in graph theory.

The lecture does not require any preliminary knowledge in engineering. Some basic concepts in graph theory are needed only.

References:

L. Lovász and A. Recski, 1982. Selected topics of matroid theory and its applications, Rendiconti del Circolo Matematico di Palermo II. 2, 171-185.

A. Recski, 1984. Applications of combinatorics to statics - a survey, Rendiconti del Circolo Matematico di Palermo II. 3, 237-247.

A. Recski, 1989. Matroid theory and its applications in electric network theory and in statics, Springer, Berlin.

N. Radics and A. Recski, 2002. Applications of combinatorics to statics -- rigidity of grids, Discrete Applied Ma[*]thematics, 123, 473-485.

[*] Research partially supported by the Hungarian National Research Fund and by the National Office for Research and Technology (Grant # OTKA 67651)

COMPUTING A LONGEST COMON SUBSEQUENCE OF TWO STRINGS WHEN ONE OF THEM IS RUN LENGTH ENCODED

Shegufta Bakht Ahsan, Tanaeem M. Moosa, M.Sohel Rahman and Shampa Shahriyar
AℓEDA Group
Department of CSE, BUET, Dhaka-1000, Bangladesh
e-mail: plaban777@yahoo.com ,tanaeem.moosa@csebuet.org, msrahman@cse.buet.ac.bd, shampa077@gmail.com

ABSTRACT

Given two strings, the longest common subsequence problem computes a common subsequence that has the maximum length. In this paper, we present new and efficient algorithms for solving the LCS problem for two strings one of which is run length encoded (RLE).

1 INTRODUCTION

Suppose we are given two strings $X[1..N] = X[1]X[2] \ldots X[N]$ and $Y[1..P] = Y[1]Y[2] \ldots Y[P]$. Without the loss of generality, we can assume that $N \leq P$. A subsequence $S[1..Q] = S[1]S[2] \ldots S[Q]$, $0 < Q \leq N$ of X is obtained by deleting $N - Q$ symbols from X. A common subsequence of two strings X and Y is a subsequence common to both X and Y. The longest common subsequence problem for two strings, is to find a common subsequence in both strings, having maximum possible length.

An interesting parameter for LCS problem is R, which is the total number of ordered pairs of positions at which the two strings match. More formally, we say a pair (i, j), $1 \leq i \leq N$, $1 \leq j \leq P$, defines a match, if $X[i] = Y[j]$. The set of all matches, M, is defined as follows:

$$M = \{(i,j) \mid X[i] = Y[j],\ 1 \leq i \leq N,\ 1 \leq j \leq P\}.$$

Observe that $|M| = R$. We use the notation M_i to denote the set of matches in Row i. Also, for the sake of better exposition we impose a numbering on the matches of a particular row from left to right as follows. If we have $M_i = \{(i, j_1),(i, j_2), \ldots, (i, j_l)\}$, such that $1 \leq j_1 < j_2 < \ldots < j_l$, then, we say that $number((i, j_q)) = q$ and may refer to the match (i, j_q) as the qth match in Row i. Note that $number((i, j_q))$ may or may not be equal to j_q.

In this paper, we are interested to compute a longest common subsequence (LCS) of two strings when one of them is run length encoded (RLE) [11]. The motivation for using compressed strings as input comes from the huge size of biological sequences. In a string, the maximal repeated string of characters is called a run and the number of repetitions is called the run-length. Thus, a string can be encoded more compactly by replacing a run by a single instance of the repeated character along with its run-length. Compressing a string in this way is called run-length encoding and a run-length encoding string is abbreviated as an RLE string.

In what follows, we use the following convention: if X is a (uncompressed) string, then the run length encoding of X will be denoted by $\tilde{X}$. For example, the RLE string of X=*bdcccaaaaaa* is $\tilde{X} = b^1d^1c^3a^6$. Note that for $\tilde{X}$, we define $\tilde{X}[1] = b^1$, $\tilde{X}[4] = a^6$ and so on. The notation $|X|$ is used to denote its usual meaning, i.e., the length of X; the length of the corresponding RLE string $\tilde{X}$ is denoted by $|\tilde{X}|$. We will use small letters to denote the length of an RLE string; whereas capital letters will be used to denote the length of an uncompressed string. For example, if $|X| = N$, then we shall use n to denote the length of $\tilde{X}$. Also, note that, the notion of a match and hence the definition of M can be extended in a natural way when one or both the strings involved are RLE. For example, the notion of a match $(i, j) \in M$, is extended when one input is an RLE string as follows: if $\tilde{Y}[i] = a^q$ and $X[j] = a$ then we say $(i, j) \in M$ and $run((i,j)) = q$.

We will use $LCS\_RLE(X, \tilde{Y})$ to denote an LCS of X and $\tilde{Y}$. There has been significant research on solving the LCS problem involving RLE strings in the literature. Mitchell proposed an algorithm [15] capable of computing an LCS when both the input are RLE strings. Given two RLE strings $\tilde{X}[1..n]$ and $\tilde{Y}[1..p]$, Mitchell's algorithm runs in $O((R+p+n)\log(R+p+n))$ time. Apostolico et al. [2] gave another algorithm for solving the same problem in $O(pn\log(pn))$ time whereas the algorithm of Freschi and Bogliolo [5] runs in $O(pN + Pn - pn)$ time. Ann et al. also proposed an algorithm to compute an LCS of two run length encoded strings [1] in $O(pn + \min\{p_1,p_2\})$ where p_1, p_2 denote the number of elements in the bottom and right boundaries of the matched blocks respectively. The version of the problem where only one string is run length encoded was handled recently by Liua et al. in [13]. Here, the authors proposed an $O(nP)$ time algorithm to solve the problem.

In this paper, we present two novel algorithms to compute $LCS\_RLE(X, \tilde{Y})$. In particular, we first present a novel and interesting idea to solve the problem and present an algorithm that runs in $O(nP)$ time. This matches the best algorithm in the literature [13] for the same problem. Subsequently, based on the ideas of our above algorithm, we present another algorithm that runs in $O(R \log\log p + N)$ time. Clearly, for $R < pN/\log\log p$, our second algorithm outperforms the best algorithms in the literature. Notably, in our setting, Mitchell's algorithm would run in $O((R+P+n)\log(R+P+n))$ time, which clearly is worse than ours[1].

2 A NEW ALGORITHM

In this section, we present Algorithm LCS_RLE-I which works in $O(pN)$ time. Since our algorithm depends on some ideas of the algorithm LCS-I of [9,10] we give a very brief overview of LCS-I in the following subsection.

2.1 Review of LCS-I

Note that LCS-I solves the classic LCS problem for two given strings X and Y. For the ease of exposition, and to remain in line with the description of [9,10], while reviewing LCS-I (in this section) we will assume that $|X| = |Y| = N$. From the definition of LCS it is clear that, if $(i,j) \in M$, then we can calculate $T[i,j], 1 \leq i,j \leq N$ by employing the following equation [18]:

$$T[i,j] = \begin{cases} Undefined & f(i,j) \notin M \\ 1 & if (i = 1 \;\; or \;\; j = 1) and \;\; (i,j) \in M \\ \max_{\substack{1 \leq l_i < i \\ 1 \leq l_j < j \\ (l_i,l_j \in M)}} \{(T[l_i,l_j])\} + 1 & if (i,j \neq 1) \;\; and \;\; (i,j) \in M \end{cases} \quad (1)$$

Here we have used the tabular notion $T[i,j]$ to denote $r(Y[1..i],X[1..j])$. In what follows, we assume that we are given the set of M in the prescribed order assuming a row by row operation. LCS-I depends on the following facts, problem and results.

Fact 1. ([8,18]) Suppose $(i,j) \in M$. Then for all $(i',j) \in M$, $i' > i$, (resp. $(i,j') \in M$, $j' > j$), we must have $T[i',j] \geq T[i,j]$ (resp. $T[i,j'] \geq T[i,j]$). □

Fact 2. ([8,18]) The calculation of the entry $T[i,j]$, $(i,j) \in M$, $1 \leq i, j \leq n$, is independent of any $T[l,q]$, $(l,q) \in M$, $l = i, 1 \leq q \leq N$. □

Problem 1. Range Maxima Query Problem. We are given an array $A = a_1a_2 \ldots a_n$ of numbers. We need to preprocess A to answer the following form of queries:

Query: Given an interval $I = [i_s \ldots i_e]$, $1 \leq i_s \leq i_e \leq n$, the goal is to find the index k (or the value $A[k]$ itself) with maximum value[2] $A[k]$ for $k \in I$. The query is denoted by $RMQ_A(i_s, i_e)$.

Theorem 1. ([4,6]) Range Maxima Query Problem can be solved in $O(n)$ preprocessing time and $O(1)$ time per query. □

Now, assume that we are computing the match (i,j). LCS-I maintains an array H of length N, where, for the current value of $i \in [1..N]$ we have, $H[l] = \max_{1 \leq k < i, (k,l) \in M}(T[k,l])$, $1 \leq l \leq N$. The 'max' operation, here, returns 0, if there does not exist any $(k,l) \in M$ within the range. Now, given the updated array H, LCS-I computes $T[i,j]$ by using the constant time range maxima query as follows: $T[i,j] = RMQ_H(1, j-1) + 1$. Because of Fact 1, LCS-I is able to maintain the array H on the fly using another array S, of length N, as a temporary storage. After calculating $T[i,j]$, such that $(i,j) \in M_i$, LCS-I stores $S[j] = T[i,j]$. It continues to update S (and not H) as long as the computation continues in the same row. As soon as the processing of a new row begins, it updates H with new values from S. Due to Fact 1, LCS-I does not need to reset S after H is updated (by it) for the next row. Now, for the constant time range maxima query, an $O(N)$ time preprocessing is required as soon as H is updated. But due to Fact 2, it is sufficient to perform this preprocessing once per row. So, the computational effort added for this preprocessing is $O(N^2)$ in total. Therefore, LCS-I runs in $O(N^2)$ time.

2.2 LCS_RLE-I

In this section we present our first algorithm, namely LCS_RLE-I, to solve the LCS problem when one of the strings is an RLE string. Recall that, the notion of a match $(i,j) \in M$, is extended when one input is an RLE string as follows: if $\widetilde{Y}[i] = a^q$ and $X[j] = a$ then we say $(i,j) \in M$ and $run((i,j)) = q$. Following the idea of [9,10] in the LCS_RLE-I algorithm, we maintain the arrays H and S and use them exactly the same way as they are used in LCS-I. We will be using another array K for the efficient implementation of our algorithm and its use will be clear as we proceed.

Now consider that we have completed the computation for the matches belonging to Row $i-1$ (i.e., M_{i-1}) and we start Row i. Given the updated array H, assume that we are processing the match (i,j). Also assume that when the the computation of the match (i,j) would be complete, i.e. $T[i,j]$ is completely computed, we would have the result of $LCS\_RLE(\widetilde{Y}'a^q, X')$, where $\widetilde{Y}'a^q$ and X' are prefixes of $\widetilde{Y}$ and X respectively. Then, clearly, the match is due to the letter a. Now, if $q = 1$, then to compute $T[i,j]$, we simply need to perform: $T[i,j] = RMQ_H(1, j-1)+1$. We also need to update S array to store the new value of $T[i,j]$ as the current highest value of the j column, i.e. we perform $S[j] = T[i,j]$.

[1] Notably, Mitchell's algorithm could also be used in our setting with an extra preprocessing step to compress the uncompressed string. In this case, the cost of compression must be taken into account.

[2] Ties can be broken arbitrarily, e.g. by taking the one with larger (smaller) index.

In what follows, we refer to the above operations as the *baseOperation*.

If $q>1$, then, we require two steps. Firstly, we perform the *baseOperation*. Then, in the second step (referred to as the *weightOperation*) we consider q previous matches[3] in Row i, including the current one. Now, note carefully that T-values for these matches have already been computed and reflected in the S array. We copy S to K and S array is never changed by any *weightOperation*. For Row i, we call $K[k]$ to be a *match position* if $(i,k) \in M_i$ and $K[k]$ and k are referred to as the corresponding K-value and K-index (similar notations are also defined for the arrays H and S). Now, we add a *weight* to each of the corresponding K-values: the *weight* is 0 for the current match, 1 for the previous match, 2 for the match before it and so on. Now observe that $T[i,j]$ will be the maximum of these values. This is because the kth element of this "window" from right, corresponds to matching k a's from the run a^q with rightmost k a's from X' and then matching the remaining substring with $\widetilde{Y}'$.

We will use array K to do this computation efficiently. Now recall that we are handling the match $(i, j) \in M_i$. Clearly, we can implement the *weightOperation* by adding the appropriate weights at the corresponding *match positions* of K and then performing the query $RMQ_K(u,j)$, such that $K[u]$ is a *match position* (due to (i,u)) and $q = number((i,j)) - number((i,u) +1$. In what follows we will refer to the above range (i.e. the range $[u..j]$) as the *weighted query window*. However, in this strategy, we may need to adjust the weights every time we compute a new match since for each match the *weighted query window* may change. Undoubtedly, this would be costly. In what follows, we discuss how to do this more efficiently.

Rather than adding the appropriate *weight*, for a particular row, we will add *relative weight* to all the *match positions* of K. This would ensure that the position of the maximum value remains the same, although the value may not. To get the correct value, we will finally deduct the appropriate difference from the value. We do it as follows. After the *baseOperation*, we copy the array S to array K. Then, to a match $(i,l), 1 \le l \le |M_i|$ we add $|M_i|$ - $number((i,j))+1$ as the *relative weight*. In other words we give weight 1 to the rightmost match, 2 to the next one and so on and finally, M_i to the first match.

Now recall that we are considering the match $(i, j) \in Mi$, i.e., we are computing $T[i,j]$. Assume that $number((i,j)) = |Mi|-k+1$, i.e., this is the kth *match position* from right. As before, we execute the query $RMQ_K(q,j)$, such that $K[u]$ is a *match position* (due to (i,u)) and $q = number((i,j)) - number((i,u)) +1$. However, this time we need to do some adjustment as follows. It is easy to realize that each of the values of the matched positions in $K[u..j]$, is k higher than the actual value. So, to correct the computation we perform $T[i,j]= RMQ_K(q,j)-k$.

[3] Ties can be broken arbitrarily, e.g. by taking the one with larger (smaller) index.

The analysis of the algorithm is similar to that of LCS-I algorithm of [9,10]. As we need to do at most two *RMQ* preprocessing per row, overall it will cost $O(Np)$ time (using $O(N)$ time preprocessing algorithm). We need two *RMQ* queries per match which amounts to $O(R)$ (using constant time *RMQ* query) time. Note that, in the worst case $R = O(Np)$. Finally, it is easy to see that, the set M in the prescribed order can be computed easily in $O(Np)$ time. Therefore, LCS_RLE-I solves our problem in $O(Np)$ time.

3 LCS_RLE-II

In this section, we use the ideas of LCS_RLE-I to present our second algorithm, LCS_RLE-II, which runs in $O(R\log\log p + N)$ time. To achieve this running time, we will use an elegant data structure (referred to as the vEB tree henceforth) invented by van Emde Boas [19] that allows us to maintain a sorted list of integers in the range $[1..n]$ in $O(\log\log n)$ time per insertion and deletion. In addition to that it can return *next(i)* (successor element of i in the list) and *prev(i)* (predecessor element of i in the list) in constant time.

We follow the same terminology and assume the same settings of Section 2 to describe LCS_RLE-II. So, assume that we are considering the match $(i, j) \in M_i$ and recall that when the the computation of the match (i,j) would be complete, i.e. $T[i,j]$ is completely computed, we would have the result of $LCS\_RLE(\widetilde{Y}'a^p,X')$. Note carefully that the *baseOperation* is basically the operation required to compute a normal LCS. We can use the LCS algorithm of [9] or [10] just to do the *baseOperation* for each match. Then, per match we would only need $O(\log\log p)$ time [9,10], requiring a total of $O(R \log\log p)$ time to perform all the *baseOperations*.

Now, we focus on the *weightOperations*. Now, our goal is to completely avoid any RMQ preprocessing. We need to modify the *weightOperation* as follows. We will use the vEB tree for this purpose. Now, recall that we want to find the maximum value of K in the *weighted query window*. Furthermore, note that, only the matched positions of K in the *weighted query window* are important in the calculation. So instead of maintaining the array K, we maintain a vEB tree where always the appropriate number (q in this case) of matches (with values after the addition of the relative weights) are kept. And as the computation moves from one match to the next, to maintain the appropriate *weighted query window*, only one element (corresponding to a match) is added to the vEB tree and at most one element is deleted. When we need the maximum value of the *weighted query window*, we just find the maximum from the vEB tree which can also be found in $O(\log\log p)$ time (by inserting a fictitious element having infinite value and then deleting it after computing its predecessor). As we need to insert and delete constant number of elements from the vEB tree for each match, this can be done in $O(R\log\log p)$ time on the whole. Like before, we would need to deduct the appropriate value $((|M_i| + 1 - number((i, j))$ in this case) from the returned maximum to do the proper adjustment.

Finally, the computation of the set M in the prescribed order can be done following the preprocessing algorithm of [9,10] which runs in $O(R\log\log p + N)$ time. So, we have an algorithm solving our problem in $O(R\log\log p + N))$ time.

4 CONCLUSION

In this paper, we have studied the longest common subsequence problem for two strings, where one of the input strings is run length encoded. We have presented two novel algorithms, namely LCS_RLE-I and LCS_RLE-II to solve the problem. We have first presented LCS_RLE-I combining some new ideas with the techniques used in [9, 10]. LCS_RLE-I runs in $O(pN)$ time, which matches the best algorithm in the literature. Then we present LCS_RLE-II which runs in $O(R\log\log p + N)$ time. Observe that in the worst case, $R = O(pN)$ and hence the worst case running time of LCS_RLE-II is slightly worse than the best algorithm in the literature. However, in many cases $R = o(pN)$, and our algorithm would show superior behavior in these cases. In particular, if $R < pN/\log\log p$, LCS_RLE-II will outperform the best algorithm in the literature. Additionally, if we run Mitchell's algorithm (the best algorithm in the literature for two RLE strings) in our setting, the running time would be $O((R+P+n)\log(R+P+n))$, which clearly is worse than ours. Also, employing some of the insights of Mitchel [15], we believe, our work can be extended to the version where both the input are RLE strings. Another research direction could be to implement the algorithms in the literature along the new ones proposed here and to compare them against each other from a practical point of view.

References

[1] Hsing-Yen Ann, Chang-Biau Yang, Chiou-Ting Tseng, and Chiou-Yi Hor. A fast and simple algorithm for computing the longest common subsequence of run-length encoded strings. Inf. Process. Lett., 108(6):360-364, 2008.

[2] Alberto Apostolico, Gad M. Landau, and Steven Skiena. Matching for run-length encoded strings. J. Complexity, 15(1):4-16, 1999.

[3] V.L. Arlazarov, E.A. Dinic, M.A. Kronrod, and I.A. Faradzev. On economic construction of the transitive closure of a directed graph (english translation). Soviet Math. Dokl., 11:1209-1210, 1975.

[4] Michael A. Bender and Martin Farach-Colton. The lca problem revisited. In LATIN, pages 88-94, 2000.

[5] Valerio Freschi and Alessandro Bogliolo. Longest common subsequence between run-length-encoded strings: a new algorithm with improved parallelism. Inf. Process. Lett., 90(4):167-173, 2004.

[6] H. Gabow, J. Bentley, and R. Tarjan. Scaling and related techniques for geometry problems. In STOC, pages 135-143, 1984.

[7] James W. Hunt and Thomas G. Szymanski. A fast algorithm for computing longest subsequences Commun. ACM, 20(5):350-353, 1977.

[8] Costas S. Iliopoulos and M. Sohel Rahman. Algorithms for computing variants of the longest common subsequence problem. Theor. Comput. Sci., 395(2-3):255-267, 2008.

[9] Costas S. Iliopoulos and Mohammad Sohel Rahman. New efficient algorithms for the lcs and constrained lcs problems. Information Processing Letters, 106(1):13-18, 2008.

[10] Costas S. Iliopoulos and Mohammad Sohel Rahman. A new efficient algorithm for computing the longest common subsequence. Theory Comput. Syst., 45(2):355-371, 2009.

[11] E.Fow (Eds.) K. Sayoood. Introduction to Data Compression. Morgan Kaufmann Publishers Inc, 2000.

[12] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Problems in Information Transmission, 1:8-17, 1965.

[13] Jia Jie Liu, Yue-Li Wang, and Richard C. T. Lee. Finding a longest common subsequence between a run-length-encoded string and an uncompressed string. J. Complexity, 24(2):173-184, 2008.

[14] William J. Masek and Mike Paterson. A faster algorithm computing string edit distances. J. Comput. Syst. Sci., 20(1):18-31, 1980.

[15] J.S.B Mitchell. A geometric shortest path problem, with application to computing a longest common subsequence in run-length encoded strings. Technical Report Department of Applied Mathematics, SUNY Stony Brook, 1997.

[16] Eugene W. Myers. An o(nd) di®erence algorithm and its variations. Algorithmica, 1(2):251-266, 1986.

[17] Narao Nakatsu, Yahiko Kambayashi, and Shuzo Yajima. A longest common subsequence algorithm suitable for similar text strings. Acta Inf., 18:171-179, 1982.

[18] M. Sohel Rahman and Costas S. Iliopoulos. Algorithms for computing variants of the longest common subsequence problem. In ISAAC, pages 399-408, 2006.

[19] P. van Emde Boas. Preserving order in a forest in less than logarithmic time and linear space. Information Processing Letters, 6:80-82, 1977.

[20] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. J. ACM, 21(1):168-173, 1974.

HOW TO EVALUATE CO-OCCURRENCES OF PRODUCTS IN MARKET-BASKETS FROM REAL-WORLD APPLICATIONS

Katharina A. Zweig and *Emőke-Ágnes Horvát*
Interdisciplinary Center of Scientific Computing (IWR)
University of Heidelberg
Speyerer Straße 6, 69120 Heidelberg, Germany
email: katharina.zweig@iwr.uni-heidelberg.de

ABSTRACT

The structural analysis of bipartite graphs, e.g., of products in market baskets, is very important in various applications such as recommendation systems. One aspect that is especially interesting is the number of market baskets in which two products both occur, i.e., their *co-occurrence*. To evaluate its significance, more than a dozen different *interestingness measures* have been proposed [5] and applied to questions from different disciplines. Many of these interestingness measures implicitly refer to a very simple independence model to determine the *expected co-occurrence*. In this article we will show theoretically and on one example that this simple independence model must not be applied to most real-world data sets since it will discover too many false-positive correlations between products. We show that a statistically more detailed model, e.g. one with fixed degree sequences (FDSM), needs to be used instead. We thus propose to generalize the classical interestingness measures to allow for arbitrary independence models that suit each data set the best.

1 INTRODUCTION

Some products are often bought together and their placement in a supermarket might be a crucial element for steering a customer's walk through the shop. Similarly, if many customers like film A and film B, it is reasonable to recommend film B to all customers who already watched and liked film A. The information about products that are frequently bought together can be computed by a so-called *market basket analysis* where each basket and each product is represented by a node, and a basket is connected by an edge with all the products it contains.

To understand whether the co-occurrence of any pair of products, i.e., the number of users that bought both products, is significant, it is necessary to define an *interestingness measure* like the *leverage* [4], *lift* and *conviction*. Although it is not immediately visible, most of the proposed interestingness measures rely on a term like $P(X)P(Y)$, where $P(X)$ is the frequency of market baskets containing X. This term is essentially a closed formula for the *expected co-occurrence* in a very simple independence model (SIM), in which every user buys every product X with the same probability $P(X)$. It is clear that in this null-model, the expected frequency of the pair X, Y is then given by $P(X)P(Y)$. As an example, if 20% of all users buy beer and 40% buy diapers, we expect in this simple model that 8% of all customers buy beer and diapers if the two products are unrelated.

In statistics, defining the best null-model against which to test a hypothesis is not easy. The more detailed a model is, the smaller its predictive power because if finally all events are in their own categorization group we can not predict much. If however too many different things are put into one category, false positive (and false negative) correlations might seem to be significant. We want to illustrate this problem by the following example: let's assume that in sum 1000 females and 1000 males apply for two different studies A and B with a total of 1000 admissions, but only 180 females are admitted. This can be the result of a fair admission procedure, if e.g. 900 females and 100 males apply for study A which grants only 100 of the 1000 admissions while the respective rest applies for study B with a total of 900 admissions. The effect that aggregation can lead to a (seemingly) significant difference from expectation is called the *Yule-Simpson effect* [1]. SIM is only conditional on the degree sequence of the side of interest. In this article we discuss the *fixed degree sequences model (FDSM)* as an alternative, which is conditional on both degree sequences. As a statistical model it is clear that FDSM models reality more closely. It can be likened to the example above where additional information in the model leads to new evaluations of the observed features. However, until now the interestingness measures suggested that SIM can be used as an approximative model instead of the more involved FDSM. In this abstract we will show both theoretically on a worst-case example and practically on a real-world data set that SIM makes implicit assumptions that lead to a high number of seemingly significant correlations between products.

2 DEFINITIONS

Let $U = \{u_1, u_2, \cdots, u_r\}$ denote a set of *users* or customers, and $P = \{p_1, p_2, \cdots, p_l\}$ a set of *products*. Let $E \subseteq U \times P$ be a set of pairs of users and products u_x, p_i, denoting that user u_x has bought product p_i. Note that E is a set, not a multi–set, i.e., we assume that each user buys each product at most once. The sets U, P, E can also be represented by a bipartite graph $G = (U \cup P, E)$ where U is a set of r and P a set of l vertices that are connected by an edge iff $(u_x, p_i) \in E$. We will denote the vertex and the represented object by the same label as long as there is no ambiguity. Let then $m := |E|$ denote the cardinality of this set that is at the same time the number of edges in the bipartite graph. By $deg(u_x)$ ($deg(p_i)$) we denote the *degree* of user u_x (product p_i), i.e., the cardinality of the set of pairs in E that contain u_x (p_i). If the data is represented in a 0-1 table where products are in rows and users in columns, then $deg(p_i)$ is equal to the i-th row sum, and $deg(u_x)$ is equal to the x-th column sum.

Given a bipartite graph $G = (U \cup P, E)$, for any pair of products $p_i, p_j,\ i \neq j$ we define as their *co-occurrence* $coocc_G(p_i, p_j)$ the cardinality of the set of users that bought both products, i.e., of all users u_x where (u_x, p_i) and $(u_x, p_j) \in E$.

Let now $R := (r_1 = deg(u_1), r_2 = deg(u_2), \cdots, r_r = deg(u_r))$ and $L := (l_1 = deg(p_1), l_2 = deg(p_2), \cdots, l_l = deg(p_l))$ denote the *degree–sequences* of the bipartite graph defined by U, P, and E. By $\mathcal{G}(L, R)$ we denote the *set of all bipartite graphs* $G' = (U \cup P, E')$ that obey these degree sequences. Note that we allow neither multi-edges nor self–loops and that the vertices are labeled.

To state whether the observed co-occurrence of two products in a given set $E \subseteq U \times P$ is significant or not, the value has to be compared to some expected value. This value depends on an appropriate independence model. As we have discussed in the introduction, a large part of all known interestingness measures implicitly relies on the simple independence model (SIM). As an alternative model FDSM can be used:

Fixed Degree Sequence Model (FDSM) Although SIM seems to be reasonable and is classically used, Gionis et al. argued [2], that it is more appropriate to use a model conditional on both degree sequences. This model defines a set $\mathcal{G}(L, R)$ which contains all possible bipartite graphs with exactly the same degree sequences as the original one. Thus, the *expected co-occurrence* $E[coocc_{\mathrm{FDSM}}(p_i, p_j)]$ of p_i and p_j is defined as:

$$\begin{aligned} & E[coocc_{\mathrm{FDSM}}(p_i, p_j)] \\ = \; & \frac{1}{|\mathcal{G}(L, R)|} \sum_{G \in \mathcal{G}(L,R)} coocc_G(p_i, p_j), \end{aligned}$$

i.e., we sum over the real co-occurrences of p_i and p_j in all feasible graphs in $\mathcal{G}(L, R)$ and divide by the number of graphs in this set.

Although Gionis et al. gave some anecdotal example of why the FDSM is more appropriate [2], there is so far no theoretical result that shows why and when the SIM is not suitable. In the following, we extend their work by presenting a theoretical observation of both models that clearly shows why and when SIM fails.

2.1 Expected Total Number of Co-occurrence Events

We first determine the total number of co-occurrences of vertices in P for any given data set: each vertex u_x in U on the right-hand side of the graph induces $\binom{deg(u_x)}{2}$ co-occurrences because any two products in the user's market basket co-occur together. We denote the sum of all co-occurrences in graph G by $Coocc(G)$:

$$Coocc(G) = \sum_{u_x \in R} \binom{deg(u_x)}{2}. \quad (1)$$

Since $Coocc(G)$ is sufficiently defined by the degree sequence R alone, we can also speak of $Coocc(R)$.

Let now L, R be given, and let $G = (L, R, E)$ be some graph of $\mathcal{G}(L, R)$, then the following observation is necessarily true:

Observation 1 *For any graph $G = (L, R, E)$, the sum*

$$\sum_{p_i \in L} \sum_{p_j \in L, j > i} coocc(p_i, p_j) = Coocc(R) \quad (2)$$

i.e., the sum of all co-occurrence values on the left hand side has to equal the number $Coocc(G)$ of co-occurrences induced by the right side.

Of course we expect that any reasonable independence model that tries to assess the significance of the single co-occurrence values, at least models the correct total number of co-occurrence events. I.e., we require that the sum of all **expected** co-occurrences of pairs of vertices in L for some given degree sequences L, R equals $Coocc(R)$. This is obviously the case for the FDSM since it keeps the right hand side degree sequence unchanged. SIM's expectation on the total number of co-occurrence events is in most cases different as the following lemma (without proof) shows:

Lemma 1 *The total number of co-occurrences of pairs p_i, p_j predicted by SIM is only depending on the degree sequence of L.*

The more skewed R is, the larger is $Coocc(R)$, but in SIM it is totally independent of R. For large deviations between $E[Coocc_{\mathrm{SIM}}(G)]$ and $Coocc(R)$, it is thus clear that the predictions by SIM must fail. We will now describe a worst-case family of graphs where the difference in the expected number of co-occurrence events is in $\Omega(n^3)$.

Theorem 1 *There is a family of bipartite graphs $G(n)$, representing n users and n products, such that the total number of co-occurrences expected by SIM differs by $\Omega(n^3)$ from $coocc(G(n))$.*

$G(n)$ is a bipartite graph between n products and n users, with $L = R = \{1, 2, \ldots, n\}$. A close inspection reveals that for every vertex there is exactly one subset of vertices that can be neighbours of it due to the constraint that no multi-edges are allowed. More generic, the only bipartite graph satisfying L, R is given by $E = \{\{u_x, p_i\} \mid x \leq i\}$. It follows that the observed co-occurrence $coocc(p_i, p_j)$ in $G(n)$ and the expected co-occurrence $E[coocc_{\mathrm{FDSM}}(p_i, p_j)]$ in FDSM is given by $\min\{i, j\}$ and thus $coocc(G) = n^3/6 - O(n^2)$.

In SIM, the expected co-occurence of two vertices i, j is given by ij/n. Thus, the expected total number of co-occurrence events is in $n^3/8 - O(n^2)$. This shows that there exists a highly skewed family of bipartite graphs in which SIM underestimates the total number of co-occurrence events by $\Omega(n^3)$. We now want to show that this is not merely a theoretical worst-case scenario but that it is also relevant for real-world data sets.

3 EXPERIMENTAL RESULTS

The Netflix prize [3] came with a data set of $100 \cdot 10^6$ ratings of 17770 films by 480,000 users. From this, we extracted all ratings of the first 10,000 users. For all pairs of films we calculated the difference between observed and expected co-occurrence, once in the SIM, denoted by lev_{SIM}, and once in the FDSM (lev_{FDSM}). Note that it is in general not possible to enumerate all graphs in $\mathcal{G}(L, R)$,

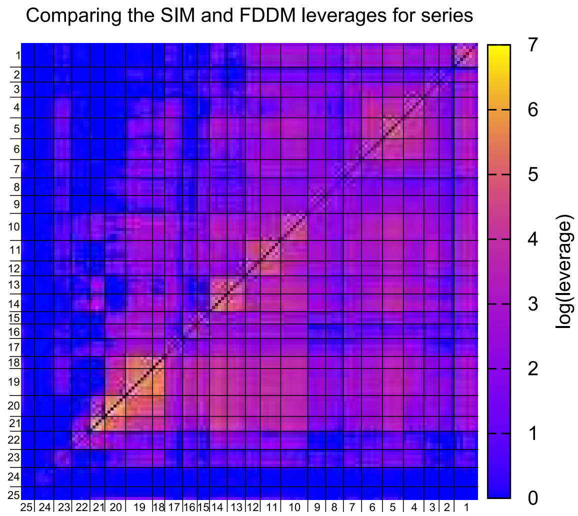


Figure 1: Log values of lev_{FDSM} (upper left triangle) and lev_{SIM} (lower right triangle) for all pairs of films which are part of a series with at least five parts; negative values are set to zero. It is clearly visible that SIM overestimates the significance of many co-occurrences. Assignments of numbers to films can be found in Table 1.

so instead we sampled uniformly at random from it as described in [2] and used the average co-occurrence in this sample as an approximation for the expected value. The sampling method is based on a Markov chain Monte Carlo simulation; we computed 5000 samples after simulating 35000 steps in the Markov chain. The main computational complexity arises from computing the co-occurrences for each sample, which is in $O(n(\sigma^2 + \mu^2))$ where σ is the variance and μ is the mean of the right-hand side degree sequence. Out of the possibly $157.9 \cdot 10^6$ pairs of films, only $52.9 \cdot 10^6$ pairs are at least once co-rated, so we only computed the expected values for these pairs in SIM and FDSM: The first result is that SIM expects in total $220 \cdot 10^6$ co-occurrence events while the graph produces $655 \cdot 10^6$ co-occurrence events. In general, SIM expects too few co-occurrence events. Although this does not directly imply that this is the case for each single pair of films, in this case 99.96% of pairs with a co-occurrence of at least 1 have a higher lev_{SIM} than lev_{FDSM} value, i.e., SIM overestimates the significance of most co-occurrences and thus produces false-positive correlations. Moreover, an astonishing 33% of all pairs of films have a positive lev_{SIM} value while its counterpart lev_{FDSM} is negative! Fig. 1 shows this on an automatically generated sample of films that contains all series with the keyword 'Season' in their title which had at least five parts. While SIM sees a positive correlation between almost all series (to various degrees), FDSM clearly differentiates between the films

Number	Title
25	Saved by the Bell
24	A Touch of Frost
23	Dr. Quinn, Medicine Woman
22	Homicide: Life on the Street
21	The Sopranos
20	Sex and the City
19	Friends
18	The Best of Friends
17	Frasier
16	Dawson's Creek
15	Gilmore Girls
14	The Simpsons
13	South Park
12	Angel
11	Buffy the Vampire Slayer
10	The X-Files
9	Xena: Warrior Princess
8	Hercules
7	Highlander
6	Star Trek: Deep Space Nine
5	Star Trek: The Next Generation
4	Star Trek: Voyager
3	Babylon 5
2	Andromeda
1	Stargate SG-1

Table 1: All film series with at least five parts with the keyword 'Season' in the title with the assigned number as shown in Fig. 1.

13-23 (e.g., Sex and the City, Friends, Dr. Quinn) and the films 1-9 (e.g., Star Gate, Star Trek, Andromeda). It thus seems that with the FDSM the 'beer and diaper' paradoxon, the case that highly popular items seem to have a significant co-occurrence without any explanation, can be diminished because it gives a better expectation value.

4 CONCLUSION

In summary we have first reported that most interestingness measures are based on a very simple independence model. We have then shown theoretically and practically that SIM fails when the degree sequence on the right hand side is strongly skewed. For the data set at hand, FDSM produced reasonable results. However, for other data sets, other random graph models might be even more suitable to assess the significance of co-occurrence values. We thus propose that interestingness measures should be more generally formulated, e.g., the improved leverage should be defined as the difference between the observed and the expected co-occurrence where the expected value is computed by a suitable independence model.

Acknowledgement We thank Netflix for publishing the dataset.

References

[1] David Freedman, Robert Pisani, and Roger Purves. *Introduction to the Practice of Statistics*. W.H. Freeman & Company, 5th edition edition, 2005.

[2] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. Assessing data mining results via swap randomization. In *KDD'06*, 2006.

[3] http://www.netflixprize.com. Netflix prize.

[4] G. Piatetsky-Shapiro. *Knowledge Discovery in Databases*, chapter Discovery, analysis, and presentation of strong rules. AAAI/MIT Press (Menlo Park, CA, USA), 1991.

[5] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right object measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

EMBEDDING OF COMPLETE AND NEARLY COMPLETE BINARY TREES INTO HYPERCUBES

Aleksander Vesel
University of Maribor, Faculty of Natural Sciences and Mathematics
Koroška c. 160, Maribor, Slovenia
vesel@uni-mb.si

Abstract

A new simple algorithm for optimal embedding of complete binary trees into hypercubes as well as a node-by-node algorithm for embedding of nearly complete binary trees into hypercubes are presented.

1 Introduction

Hypercubes and binary trees are omnipresent in computer science. In particular, hypercubes are very popular models for parallel computation because of their regularity, recursive structure and the ease of routing. On the other hand, binary trees can represent the basic computational structure of divide-and-conquer or branch-and-bound algorithms. In many cases however, it is more suitable that the internal structure of an algorithm is modeled by a more general structure - a nearly complete binary tree.

In this paper we consider the problem of embedding a (nearly) complete binary tree in a hypercube. This problem occurs during the implementation of divide-and-conquer algorithms in a hypercube network, e.g., see [3, 5]. An embedding is a mapping from the guest graph, representing the communication structure of the processes, into the the host graph, representing the communication network of the processors. Therefore, the problem of allocating processes to processors in a multiprocessor system is also known as the *mapping problem.*

A *tree* is a connected acyclic graph. One vertex is distinguished and called the *root.* A vertex of degree one is called a *leaf* of the tree if it is not the root. The *level* of a vertex v in a tree is the number of vertices on the simple path from the root to v. Note that the level of the root is one. The height of a tree T is the largest level of a vertex in T. A vertex u is called a *child* of v if u is adjacent to v and the level of u is greater than the level of v. If u is a child of v, then v is called the *parent* of u.

A *full binary tree* is a tree in which every node other than the leaves has two children. A full binary tree is ordered, i.e. we distinguish between left and right children. A *complete binary tree* is a full binary tree in which all leaves are at the same level. A *nearly complete binary tree* of height h is composed of a complete binary tree of height $h-1$ and with some nodes at level h (not necessary positioned to the left).

The *hypercube* of order d and denoted Q_d is the graph G where the vertex set $V(G)$ is the set of all binary strings $u_1u_2\dots u_d$, $u_i \in \{0,1\}$. Two vertices $x, y \in V(G)$ are adjacent in Q_d if and only if x and y differ in precisely one place.

An *embedding* of a graph G into a graph H is an injection $f : V(G) \to V(H)$ such that if (u,v) is an edge in $E(G)$ then $(f(u), f(v))$ is an edge in $E(H)$.

For binary vectors $s, t \in \{0,1\}^n$ let $s \oplus t$ denote the coordinate-wise addition modulo two, e.g. $100011 \oplus 000001 = 100010$. Let e_i^n be the binary vector $u_1, u_2, \dots, u_n$ with $u_i = 1$ and $u_j = 0$, $j \neq i$. If s is a binary vector of length n, then we will call the operation $e_i^n \oplus s$ a *reflection.* We will also use the "+" symbol as the concatenation operator, i.e. $s+t$ joins two binary vectors s and t end to end. If we concatenate a binary vector with a single bit (0 or 1), than we call this operation a *projection.* For a binary vector s of length n, $s+0$ and $s+1$ are projections of s into two disjoint subcubes of order n of Q_{n+1}.

The minimum h required for an embedding of a graph G into Q_h is called the *cubical dimension* of G. Deciding whether there exists an embedding of a given tree into a hypercube of a given dimension is known to be NP-complete [4]. Moreover, even in case of trees with bounded degrees, their cubical dimensions are unknown in most cases.

Obviously, if G is a graph such that $2^h \geq |V(G)| > 2^{h-1}$, then the cubical dimension of G is at least h. However, it is well known that the complete binary tree on $2^h - 1$ vertices cannot be embedded into Q_h for $h \geq 3$.

Havel in [1] conjectured that every binary tree T such that $2^h \geq |V(T)| > 2^{h-1}$ has an embedding into Q_{h+1}. The conjecture is still open, but there are many partial results supporting this assertion. It has been shown, for example, that a complete binary tree of height h can be embedded into the hypercube of order $h+1$, e.g. [3, 5].

2 Complete binary tree

Let C_h denote the complete binary tree of height h. Let also r_h denote the root of C_h and let C_h^l and C_h^r denote the left and the right subtree of C_h, respectively. Obviously, C_h^l and C_h^r are both complete binary trees of height $h-1$.

Let $\sigma_l : V(C_h^l) \to V(C_{h-1})$ be the mapping that in the natural way maps each vertex $v \in V(C_h^l)$ to the corresponding vertex of C_{h-1}, e.g. if v is the root of C_h^l, then $\sigma_l(v) = r_{h-1}$. Analogously we define the mapping $\sigma_r : V(C_h^r) \to V(C_{h-1})$.

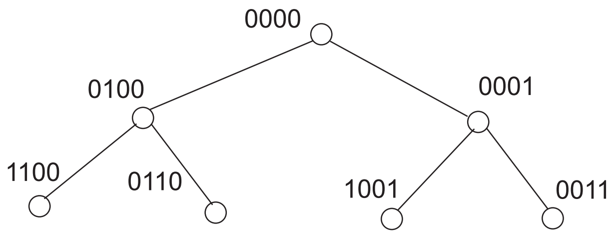


Figure 1: β_3 - the mapping of C_3

For each $h \geq 3$, we define a mapping $\beta_h : V(C_h) \to \{0,1\}^{h+1}$ as follows. The mapping $\beta_3 : V(C_3) \to \{0,1\}^4$ is depicted in Fig. 1, while for $h > 3$ the mapping is given by

$$\beta_h(v) = \begin{cases} 0^{h+1}, & v = r_h \\ \beta_{h-1}(\sigma_r(v)) + 1, & v \in V(C_h^r) \\ \beta_{h-1}(\sigma_l(v)) \oplus 10^{h-1}+0, & v \in V(C_h^l), h \text{ even} \\ \beta_{h-1}(\sigma_l(v)) \oplus 0^{h-3}100+0, & v \in V(C_h^l), h \text{ odd}. \end{cases}$$

For any v of C_h, the string $\beta_h(v)$ will be also called a *code word* defined by β_h in the sequel.

Theorem 1 *Let $h \geq 3$. Then β_h defines an embedding of the complete binary tree of height h in the hypercube of order $h+1$.*

The basis of the proof is the following lemma.

Lemma 2 *Let $h \geq 3$. Then $\beta_h(v) = 0^{h+1}$ if and only if v is the root of C_h.*

Proof. If v is the root of C_h, then $\beta_h(v) = 0^{h+1}$ by definition. On the other hand, the recursive definition of β_h implies, that if v is not the root of C_h, then $\beta_h(v)$ is obtained as a sequence of projections and reflections either from the root of the C_{i+1}, i.e. $\beta_{i+1}(r_{i+1}) = 00^i0$, $h > i \geq 3$, or from a code word defined by β_3.

Suppose first that $\beta_h(v)$ derives from 00^i0. Suppose also that i is even. Some of the code words derived from 00^i0 can be seen in the tree depicted in Fig. 2. The root of the tree is 00^i0, while the left and the right child of 00^i0 are obtained as code words derived from it in the left and the right subtree of C_{i+2}, respectively. Analogously, the left and the right children of 10^i000 and 00^i001 are code words in C_{i+3}, etc.

Note that the definition of β_j for $j \geq 4$ implies that $\beta_j(v)$ is derived from a code word s of C_{j-1} such that either the first or the $(j-3)$-th bit of s is reversed. It follows that from a code word s of C_{j-1} with a 1 in at least one of the positions: $2, 3, \ldots, j-4$, a code word of the form 0^{h+1} cannot be derived. Since every leaf of the tree in Fig. 2 in at least one of those positions possesses entry 1, it follows that 0^{h+1} cannot derive from 00^i0 if i is even. It is not difficult to see that a similar tree (having leaves of length j with a 1 in at least one of the positions $2, 3, \ldots, j-4$) can be derived for 00^i0, where i is odd.

For a $\beta_h(v)$ derived from a code word defined by β_3 observe first that all code words of the left subtree of C_3 depicted in Fig. 1 posses 1 at the second place, which implies that 0^{h+1} cannot be derived from any of them. For $\beta_h(v)$ derived from a code word of the right subtree of β_3, observe for example the tree depicted for code word 0001 the in Fig. 3. From the same arguments as above we can conclude that 0^{h+1} cannot derive from any code word defined by β_3 and the proof is complete. $\square$

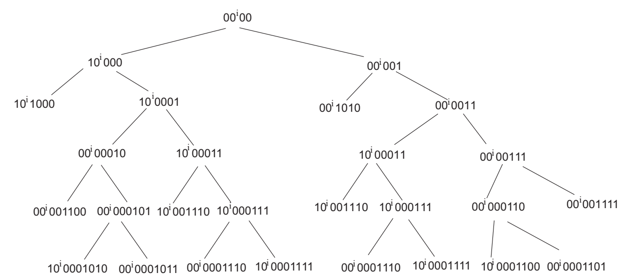


Figure 2: Code words derived from 00^i0 when i is even

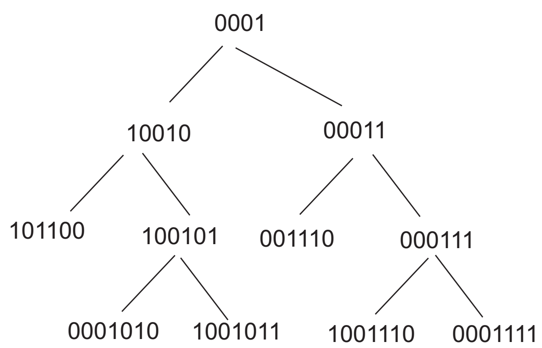


Figure 3: Some code words derived from 0001

Proof. [Proof (of Theorem 1)] The proof is by induction on h. The claim obviously holds for β_3. Suppose also that the claim holds for h. Note first that projections maps the vertices of C_{h+1}^l and C_{h+1}^r into two disjoint hypercubes. Moreover, reflections of an embedding in a hypercube preserve Hamming distance. Therefore β_{h+1} is an embedding of C^l_{h+1} and C^r_{h+1} in the hypercube of order $h+1$. Finally, since the root of C_{h+1} and the root of C^l_{h+1} (as well as C^r_{h+1}) differ in precisely one bit, we conclude that β_{h+1} embeds C_{h+1} into Q_{h+1} and the proof is complete. □

Theorem 1 is the basis for the algorithm to compute an optimal embedding of the complete binary tree of height h in Q_{h+1}.

We first present the algorithm to calculate an embedding of the complete binary tree of height h from an embedding of the complete binary tree of height $h-1$. It is assumed in the algorithm, that if v is an arbitrary node of a complete binary tree T, then $b(v)$ is the code word of v and $p(v)$ a parent of v in T. Furthermore, r denotes the root of T and T_l and T_r denote the left and the right subtree of T, respectively.

Procedure NEW TREE
input: h, T, b { $h > 3$, b an embedding of $T = C_{h-1}$ into Q_{h+1} }
output: b {An embedding of $T = C_h$ into Q_{h+1}}
  begin
   traverse T_r from level h to level 2
    for every $v \in T_r$ **do** $b(v) := b(p(v))) + 1$;
   if h mod $2 = 0$ **then**
      traverse T_l from level h to level 2
        for every $v \in T_l$ **do**
          $b(v) := b(p(v)) \oplus 10^{h-1} + 0$;
   else
      traverse T_l from level h to level 2
        for every $v \in T_l$ **do**
          $b(v) := b(p(v))) \oplus 0^{h-3}100 + 0$;
   $r := 0^{h+1}$; { The new root of T }
  end.

We next describe the algorithm to compute an optimal embedding of the complete binary tree of height h in a hypercube.

Procedure CODES
input: h {height of a tree, $h \geq 3$}
output: T, b {T is of height h with an embedding b}
  begin
   $T := C_3$.
   Set $b(v)$ for every $v \in T$ as in Fig. 1;
   for $i := 4$ to h **do begin**
     Augment T with new level of nodes to
     obtain C_i;
     NEW TREE(i, T, b);
  end.

Theorem 3 *For any $h \geq 3$, CODES embeds C_h into Q_{h+1} in linear time and space.*

Proof. The correctness of the algorithm is proved by induction on h. If $h = 3$, then the embedding is given in Step 2, the correctness of which can be verified by Fig. 1.

Assume now that for $i = h-1$ the algorithm correctly compute an embedding of T. In other words, when NEW TREE is called in Step 3 for $i = h$, the vector b corresponds to β_{h-1}. Moreover, for a node $v \in T_l$ ($v \in T_r$), the old value of $b(p(v)))$ corresponds to $\beta_{h-1}(\sigma_r(v))$ ($\beta_{h-1}(\sigma_l(v))$). Therefore, since the nodes of T are traversed from the last level to the roots of the subtrees and since NEW TREE accurately follows the definition of β_h, we can conclude that the embedding is correct.

T, p, and b can obviously be represented in linear space, therefore we only consider the time complexity. Let $n := 2^h - 1$ denote the number of nodes of a complete binary tree of height h. Note first that NEW TREE computes an embedding b in time which is linear in the size of T. Since the number of vertices of T in i-th iteration of the **for** loop equals $2^i - 1$, the total number of steps of the algorithm is given by

$$\sum_{i=4}^{h} 2^i - 1 = 2^{h+1} - 12 = O(n)\ .$$

This argument completes the proof. □

3 Nearly complete binary tree

In this section we present a simple node-by-node algorithm for constructing an embedding of a nearly complete binary tree into a hypercube. We assume that in each time step a nearly complete binary tree can grow by one node inserted at the last level of a tree. Note, that in [2] a somewhat similar approach has been studied, where the complete binary tree grows by a complete level of its leaves.

The algorithm presented herein computes the map of the nodes of a new tree using the map of their parent node. Moreover, if a new node does not change the height of a tree, the old nodes need not to be remapped.

The algorithm of the previous section implies that the embedding of C_h can be performed by using an embedding of C_{h-1} in such a way that the embedding of a node v is computed from the "old" embedding of its parent node.

This observation leads to a node-by-node algorithm for embedding of nearly complete binary trees into hypercubes. The algorithm augments a given nearly complete binary tree with one node, which is inserted at the level h and computes the mapping of the augmented tree. We will show that in the majority of cases the algorithm is able to determine the embedding of the augmented tree simply by expanding an embedding with the map of the new node. Moreover, the map of the new node can be computed with ease from the map of its parent node.

If a nearly complete binary tree T of height h with an embedding into Q_{h+1} is augmented with one new node, then for the resulting nearly complete binary tree T' the embedding into Q_{h+1} (or Q_{h+2}, if the height of T' is $h+1$) is computed. The new node can be

(i) a leaf at level h, if T is not complete or

(ii) a leaf at level $h+1$, if T is complete.

In order to obtain an embedding for T' we first show the following lemma.

Lemma 4 *Let for $h \geq 3$, v be a leaf of C_h and u the parent of v in C_h. Then*

$$\beta_h(v)=\begin{cases}\beta_h(u) \oplus 10^h, & v \text{ is the left child of } u\\ \beta_h(u) \oplus 001^{h-3}0, & v \text{ is the right child of } u\end{cases}$$

Proof. The proof is by induction with respect to h. The claim obviously holds if $h = 3$ as can be seen in Fig 1. Let us denote v a leaf of C_{h+1} and u the parent of v. If v (and u) is in the right subtree of C_{h+1}, then by inductive hypothesis $\beta_h(\sigma_r(v))$ and $\beta_h(\sigma_r(u))$ differ either in the first bit, if v is the left child of u, or in the third bit, if v is the right child of u. It is straightforward to see now that $\beta_{h+1}(v) = \beta_h(\sigma_r(v)) + 1$ and $\beta_{h+1}(u) = \beta_h(\sigma_r(u)) + 1$ differ either in the first or in the third bit. If v is in the left subtree of C_{h+1}, the proof is analogous. $\square$

In the following algorithm, let for an arbitrary vertex u of a nearly complete binary tree T, $b(u)$ and $p(u)$ denote the code word of u and the parent of u in T, respectively.

Procedure NEW NODE
input: h, T, b, v { b is the embedding of T, $h \geq 3$, v is a new node }
output: T, h, b { An augmented tree of height h with an embedding b }

```
begin
  if T is complete then begin
      NEW TREE (h, T, b);
      h := h + 1;
   end;
  Insert v at the level h in T;
  if v is the left child of p(v) then
        b(v) := b(p(v))) ⊕ 10^h;
     else  b(v) := b(p(v))) ⊕ 0010^(h−2);
end.
```

In order to obtain an embedding of a nearly complete binary tree with NEW NODE, before the algorithms is first called, Step 1 and Step 2 of CODES have to be executed. T is then the complete binary tree of height 3 with the embedding b.

Theorem 5 *If T is a nearly complete binary tree of height $h > 3$ and b an embedding of T into Q_{h+1}, then NEW NODE correctly embeds T' either*

(i) into Q_{h+2}, if T is complete or

(ii) into Q_{h+1}, if T is not complete.

Moreover, if T is C_3 and b an embedding of C_3 into Q_4, then NEW NODE correctly embeds T' into Q_5.

Proof.

Assume that T is either C_3 or an arbitrary nearly complete binary tree of height $h > 3$ and b its embedding into Q_{h+1}. If T is not complete, then the correctness of the algorithm follows from Lemma 4.

Let then T be a complete binary tree. When NEW TREE is called in Step 1, the value of h is not yet incremented, i.e. the output of the procedure is the complete tree of height h with an embedding into Q_{h+2}. However, in Steps 2 and 3, T is first augmented with a new node at level $h + 1$ and then an embedding into Q_{h+2} of the resulting nearly complete tree of height $h + 1$ is computed. $\square$

The following concluding comment concerning the time complexity of the algorithm is in order. The algorithm remaps the nodes of T only if T is a complete binary tree. In other cases a remapping is not performed. Clearly, the embedding of a new node depends only on the map of its parent node and can be performed in constant time. However, even in the case when remapping is needed, the computation of the new embedding can be done independently in each node v such that only the code word of a parent node of v is used. It follows that the remapping can be computed on the hypercube in parallel in constant time.

References

[1] I. Havel, On hamiltonian circuits and spanning trees of hypercubes, Časopis pro Pěstováni Matematiky, 109 (1984) 135 –152.

[2] V. Heun and E.W. Mayr, Optimal dynamic embeddings of complete binary trees into hypercubes, *J. Parallel Distrib. Comput.* **6** (2001) 11110–1125.

[3] A.S. Wagner, Embedding the complete tree in the hypercube, *J. Parallel Distrib. Comput.* **20** (1994) 241–247.

[4] A.S. Wagner, D.G. Corneil, Embedding trees in a hypercube is NP-complete, *SIAM J. Comput.* **7** (1990) 570–590.

[5] A. Y. Wu, Embedding of Tree Networks into Hypercubes, *J. Parallel Distrib. Comput.* **2** (1985) 238–249.

Better bounds for the bin packing problem with LIB constraint

György Dósa* Zsolt Tuza† Deshi Ye‡

September 24, 2010

Abstract

The (online) bin packing problem with LIB constraint is stated as follows: The items arrive one by one, and must be packed into unit capacity bins, but a bigger item can not be packed into a bin which already contains a smaller item. The number of used bins is minimized as usually.

We show that the performance bound of algorithm First Fit is not worse than $2 + 1/6$ for the problem, improving the previous best upper bound 2.5. If all items are not bigger than $1/d$, then we improve the previous best result $2 + 1/d$ to $2 + 1/d(d + 2)$, for any $d \geq 2$.

Moreover we define a problem with the generalized LIB constraint, where incoming items can not be packed into the bins of some already packed items. The incompatibility of the incoming item with some already packed items becomes to be known only at the arrival of the actual item, and is given by an undirected graph. We show that 3 is an upper bound for this general problem if some natural transitivity constraint is satisfied, and without this transitivity constraint no online algorithm can be c-competitive with any constant c.

1 Introduction

The classical bin packing problem has been extensively studied, where the problem is to pack a set of items, whose sizes are in $(0, 1]$, into the minimum number of bins with capacity 1. In this paper, we consider the bin packing problem with LIB (largest item in bottom) constraint. It is the classical bin packing problem with the additional requirement that in every bin larger (or longer) items must be placed below smaller (or shorter) items. In other words, items $p_1, p_2, ..., p_n$ arrive one by one (where $0 < p_i \leq 1$) and must be packed into bins of size 1, and the total size of items packed into a bin can not be bigger than 1 . With the additional LIB constraint, a later item p_i can not be packed into a bin already containing p_j, if $p_j < p_i$. (If all the information regarding the items is known in advance, then the problem is called *off-line problem.* The problem is called *online* if the items arrive one by one and we must make a decision upon the arrival of an item without knowing any information on future items.)

The LIB version of Bin Packing arises in the transportation applications where it is requested to have safe and stable packing (bigger item is not allowed to be packed on the top of some smaller item). The classical bin packing problem has many applications in computer science. With LIB constraint, the applications in computer science can be easily extended to the applications with priority. For example, in a multi-core system the CPU scheduler needs to assign the task with larger priority before the smaller priority in each core. However, each core has a fixed capacity of priority and the question is to minimize the number of cores used for serving all the tasks.

While applying to bin packing with LIB constraint, we describe the First Fit algorithm: FF places any incoming item i into the first bin B into which it fits, and no item j has already been packed into B such that $p_j < p_i$. If no such bin exists, a new bin will be opened and item i is packed into it.

1.1 Previous results regarding the LIB constraint

The bin packing problem with LIB constraint was introduced by Manyem, (see references in [1]). It was shown that NF is not constant approximated (algorithm NF or Next Fit at any time keeps open only one bin, and packs the next item into the open bin if fits, otherwise the bin is closed and the next item is packed into a new bin), but the competitive ratio of First Fit (FF) is at most 3. (The competitive ratio of an online algorithm A means the maximum of ratio $A(I)/OPT(I)$, where OPT is an optimal offline algorithm, and I is arbitrary input for the problem.) Epstein [1] improved the upper bound to 2.5, furthermore, Epstein [1] proved that the parametric upper bound of FF (where the size of each item is at most $1/d$) is at most $2 + 1/d$, for any integer $d \geq 2$. It was also shown that the competitive ratio of any online algorithm

*Department of Mathematics, University of Pannonia,Veszprem, Hungary, dosagy@almos.vein.hu.

†Computer and Automation Institute, Hungar. Acad. Sci., Budapest, Hungary, and Department of Computer Science and Systems Engineering, University of Pannonia, Hungary, tuza@sztaki.hu. Supported in part by OTKA grant 81493.

‡College of Computer Science, Zhejiang University, Hangzhou, China. Email: yedeshi@zju.edu.cn

for bin packing with LIB constraint is at least 2, for any parameter d.

1.2 Our contribution

In this paper, we revisit algorithm FF, and give a more careful analysis. We show that the competitive ratio of algorithm FF is not worse than $2 + 1/6$, improving the previous best upper bound 2.5 of Epstein [1]. Moreover, we show that the parametric competitive ratio of FF is at most $2 + 1/d(d + 2)$, again improving the previous upper bound $2 + 1/d$ of [1]. Finally, we also treat the problem in a generalized version, and prove upper and lower bounds.

1.3 Notation and preliminaries

Let a be any item. The bin where a is packed will be denoted by $B(a)$. The level of a bin B will be denoted as $l(B)$. After making a packing, we say that a bin B_{i_1} is earlier than bin B_{i_2}, or B_{i_2} is latter than bin B_{i_1}, if $i_1 < i_2$ holds.

We say that items x_v,x_{v-1}, ..., x_1 form a *chain*, if for any $1 \leq k < v$, x_{k+1} arrives before x_k, and $x_{k+1} < x_k$. Then from the LIB constraint follows that each element of the chain must be packed into different bins.

There are three lower bounds in the literature for the problem, namely: The total sum of the items and the length v of any chain are natural lower bounds on OPT, as easily seen. These lower bounds (total sum and length of longest chain) will be denoted by LB_1 and LB_2, respectively. The third lower bound [1] can be stated as follows:

Lemma 1 *Let x_v,x_{v-1}, ..., x_1 form a chain, $x_1 \leq 1/2$, and suppose that there are u items arriving after x_1, each bigger than a half. Then $LB_3 = v + u$ is a lower bound for the problem.*

2 The improved bounds

Due to space limitation, here we can give only an outline of the proof of the statement that $FF/OPT \leq 2 + \frac{1}{d(d+2)}$ holds for all $d \geq 2$.

Let us run algorithm FF, the created bins will be called as FF-bins, furthermore we denote the number of used bins simply by FF, while let OPT denote the optimum number of bins in case of an offline solution. Let a be an arbitrary item, being packed not into the first FF-bin, and let B be an FF-bin with smaller index than $B(a)$. Then we say that a is **not** packed into bin B by overflowing, if at the moment when item a arrives, $l(B) + a > 1$ holds where $l(B)$ means the actual level of bin B, and at this time there is no item packed into bin B smaller than a. Furthermore, we say that a is **not** packed into bin B by the LIB constraint, if there is an item $b < a$, already packed into bin B, just at the moment when item a is revealed (no matter that $l(B) + a$ is bigger than one or not). Now we classify the FF-bins as follows. Let k be an arbitrary integer.

Class H (bins with high level) will be the set of those bins where the level of the bin is at least $\frac{k(d+1)-1}{k(d+1)}$, the other bins form Class R (remaining bins). Bins of these classes will be called as H-bins and R-bins, respectively.

Next we define a special chain. Let x_1 be an arbitrary item being packed into the **last** R-bin. Suppose that $x_1, x_2, ..., x_k$ are already defined. Then let x_{k+1} be last item being packed into some R-bin B before the arrival of item x_k, such that B is the last R-bin where x_k is not packed by the LIB constraint, if there exists such item. Let the length of the chain be $v \geq 1$; we define the chain in the reverse order x_v,x_{v-1}, ..., x_1. We call items x_v,x_{v-1}, ..., x_1 as chain-items, the bins of these items as chain-bins, let this set of bins be Class C (chain bins). Let the class of the other R-bins be Class M (bins of medium level). The bins of these classes will be called as C-bins and M-bins, respectively. Then it can be shown that the level of any M-bin is at least $\frac{d}{d+1}$.

Now we divide the M-bins into k subclasses, according to their levels, as follows: subclass M_i is defined as the set of M-bins having level between $\frac{(d+1)k-i}{(d+1)k}$ and $\frac{(d+1)k-i+1}{(d+1)k}$, for $1 \leq i \leq k$. Since the level of any M-bin is at least $\frac{d}{d+1}$, these are all the M-bins. We also classify the chain bins, let C_i be the set of C-bins with level between $\frac{i-1}{(d+1)k}$ and $\frac{i}{(d+1)k}$, for $1 \leq i \leq k$. Since the level of a chain-bin can be $\frac{k}{(d+1)k}$ or bigger, let simply C_+ denote the C-bins with level at least $\frac{k}{(d+1)k}$. In order to prove $FF/OPT \leq 2 + \frac{1}{d(d+2)}$, let h, c_i and m_i be the number of H-bins, C_i-bins, and M_i-bins, respectively, for $1 \leq i \leq k$ (then $m_1 = 0$), and let c_+ be the number of C_+-bins. Then we gain the next lower bounds for the optimum value:

$$\begin{aligned} OPT \geq\ & \frac{(d+1)\,k-1}{(d+1)\,k}h \\ & + \sum_{i=1}^{k}\left(\frac{i-1}{(d+1)\,k}c_i + \frac{(d+1)\,k-i}{(d+1)\,k}m_i\right) \\ & + \frac{1}{d+1}c_+, \end{aligned} \tag{1}$$

$$OPT \geq c_1 + c_2 + ... + c_k + c_+ \tag{2}$$

$$\begin{aligned} OPT \geq\ & c_1 + ... + c_j \\ & + \sum_{i=j+1}^{k}\frac{(d+1)\,k-i}{(d+1)\,k}m_i, \end{aligned} \tag{3}$$

for any $1 \leq j \leq k - 1$.

After this treatment, from inequalities (1)–(3), multiplying them by some coefficients, and then summing them up, we obtain the desired result. The coefficients are the optimal solutions of the linear program, constructed by the lower bounds on OPT, while the objective is $z = h + \sum_{i=1}^{k} (c_i + m_i) + c_+$. Then we observe that the sum of the coefficients can be arbitrarily close (from above) to $2 + \frac{1}{d(d+2)}$ if $k \to \infty$. We only need to show that the sum of the coefficients of any variables h, c_i, m_i and c_+ are at least 1. We gain the next theorem:

Theorem *The parametric competitive ratio of algorithm FF is not bigger than* $2 + \frac{1}{d(d+2)}$*, for any* $d \geq 2$.

2.1 If items bigger than a half can occur

Now let us consider the case $d = 1$, i.e. there can be items bigger than a half. Note that the previous best upper bound is 2.5. With a little bit more careful analysis than that we made for cases $d \geq 2$, we can derive a better bound, namely we show that if items bigger than a half can occur, then $FF/OPT \leq 2 + 1/6$ still holds. The proof is quite similar to that in the previous section but we need one more type of lower bound, used never before, on the optimum value.

References

[1] L. Epstein, On online bin packing with LIB constraints, *Naval Research Logistics*, 56(8):780–786, 2009.

A REVIEW ON SEMI-ON-LINE BIN PACKING RESULTS

János Balogh, József Békési
Department of Applied Informatics
Institute of Applied Sciences, Gyula Juhász Faculty of Education, University of Szeged
Boldogasszony sgt. 6., 6725. Szeged, Hungary
Tel: +36 62 546080; fax: +36 62 420953
e-mail: {balogh,bekesi}@jgypk.u-szeged.hu

ABSTRACT

The present paper reviews some results in the field of semi on-line bin packing.
Semi-on-line algorithms for the bin-packing problem allow, in contrast to pure on-line algorithms, to execute one of certain types of additional operations in each step as repacking, reordering or buffering some elements before packing them. We describe our new lower and upper bounds for semi-on-line bin packing problems based on our papers [1-4].

1 INTRODUCTION

The classical one-dimensional bin packing problem is among the most frequently studied combinatorial optimization problems. In its traditional definition a list $L=x_1, x_2, \ldots, x_n$ of elements (also called items) with sizes in the interval *(0,1]* and an infinite list of unit capacity bins are given. Each element x_i from the list L has to be assigned to a unique bin such that the sum of the sizes of the elements in a bin does not exceed the bin capacity. The size of an element is also denoted by x_i. The *bin packing problem* consists of packing the items to the bins in such a way that as few bins as possible are used.
It is well-known that finding an optimal packing is *NP-hard* [17]. Consequently, a large number of papers have been published which look for polynomial time algorithms that find feasible solutions with an acceptable approximation quality.
One class of the algorithms is the so called on-line algorithms. An online algorithm packs the elements in order of their arrival, without knowing the subsequent elements. (Neither the size nor the number of the elements are known.) The packed elements cannot be removed. Off-line algorithms have full information about the list. They can take into consideration this information during their operation. Many times the list is sorted by decreasing order, and then an on-line technique is applied.

2 Definitions and earlier results

For measuring the efficiency of algorithms there are two general methods: the investigation of the worst-case behaviour or – assuming some probability distribution of the elements – a probabilistic analysis. In this paper we will concentrate on the asymptotic worst-case ratio of an algorithm. For a given list L we denote by $A(L)$ and $OPT(L)$ the number of bins used by algorithm A and the number of bins used in an optimal packing, respectively. Then the *asymptotic competitive ratio (ACR)* of algorithm A is

$$ACR(A) := \limsup_{k\to\infty}\left\{\max_{L}\left\{\frac{A(L)}{k} \middle| OPT(L)=k\right\}\right\}. \quad (1)$$

If an algorithm has an ACR, we also say that it is *ACR-competitive*.
For off-line algorithms Fernandez de la Vega and Lueker [12] provided an APTAS (Asymptotic Polynomial Time Approximation Scheme), while Karmakar and Karp [25] developed the first AFPTAS (Asymptotic Fully Polynomial Time Approximation Scheme). In [12] for every $\varepsilon>0$ an algorithm A_ε is given such that each A_ε runs in time polynomial in the length of the input list L (but exponential in $1/\varepsilon$) and $ACR(A_\varepsilon)= 1+\varepsilon$. In [25] a more complex algorithm is given. The running time of this is but depends on n and $1/\varepsilon$ polinomially and $A_\varepsilon(L)\le OPT(L)+\log^2(OPT(L))$ holds for this.
The best known on-line algorithm was defined by Seiden [31] after 2000. Its name is *Harmonic++*. Seiden (improving Richey's earlier method and analysis [30]) proved that the ACR of his algorithm was at most *1.58889*. Harmonic++ belongs to the *Super Harmonic* algorithm class defined in the same paper. For all algorithms from this class a lower bound of *1.58333* is valid [29], so $1.58333\le ACR(Harmonic{+}{+})\le 1.58889$. Note that all of the actual best algorithms are Super Harmonic type since 1985 [26,29,31]. For the ACR of an algorithm for the on-line problem the best known lower bound was *1.5401*, which was proved by van Vliet [32]. One of our new results improves this to *1.5403* [1]. This means that if we want to give a better on-line algorithm with better ACR than *1.58333*, we have to find a new idea. It is an open question, whether this is possible. Another possibility to give an algorithm with better ACR than *1.5403* is to slacken the on-line criterion.
In this paper we deal with semi-on-line (SOL) bin packing problems. The so called *semi-on-line algorithms* [5,8] are between the well-known on-line and off-line ones. For such algorithms at least one of the following operations is allowed: repacking of some items (e.g. [16,21]), lookahead of the next several elements [18], or some kind of preordering.

The most famous algorithms based on preordering are FFD and BFD. They sort the elements in decreasing order and pack them using FF (BF) strategy. Johnson proved in [24] that *ACR(FFD)=ACR(BFD)= 11/9=1.22222....* More exactly Johnson proved that *ACR(FFD) ≤(11/9)OPT(L)+4.* The tight value *6/9* of the additive constant was given by Dósa [9]. For the on-line algorithms running on preordered lists we improved the previous lower bound *8/7* [7] to *54/47* in [1].

The first semi-on-line bin packing algorithm with repacking was given by Galambos [13] for the case when a restricted number of bins can be open. A bin is called closed, if we cannot pack elements into it later. This algorithm uses two buffer bins for storing the elements temporarily. Improving this Galambos and Woeginger [14] defined a semi-on-line repacking algorithm using *3* buffer bins. Its ACR is *1.69103...*, which is optimal among the algorithms with restricted number of open bins.

To discuss the considered subclasses of semi-on-line bin packing, we need to introduce the role of the "scheduler" and the role of the "packer". The role of the scheduler is to produce the input list, while the role of the packer is to pack the items (that is, to realize the packing algorithm).

In the above mentioned problem classes the rule of the scheduler is trivial: to give the elements one by one (let say this is the *Insert* operation) and to mark the end of the (whole) list.

Gutin et al. introduced [19] the so called *batched bin packing problem* (BBPP) in 2005. This is a semi-on-line bin packing problem. The classical problem is modified in such a way, that the input is splitted into parts – so called batches – by the scheduler. In every step the scheduler either gives a new element, or marks the end of the current batch. Every batch is available only after the processing of the previous one. In other words, lookahead is not allowed outside the current batch. Otherwise the algorithm has to pack each batch as an offline one (that is, a lookahead is possible within the current batch). However, each batch has to be packed in an on-line manner: during the packing of a new batch the elements of the earlier batches cannot be moved. The algorithm has to pack each batch as an offline one (that is, a lookahead is possible within the current batch). One batch can consist of more elements or can be empty. If every batch has exactly one element, then we get the classical bin packing problem as a special case. If an input consists of exactly *m* batches, then we call this BBPP as *m*-BBPP. The asymptotic competitive ratio of the algorithms for this problem version can be defined the same way as before.

Gutin et al. study the *2*-BBPP problem and its variants in [19] in details. For the *2*-BBPP problem they proved an *1.3871...*lower bound here and they derived bounds for those special cases of the *2*-BBPP problem, when the sizes of the elements of the list are bounded from above by a given *p≥2* positive number. They proved – using the results of their paper [20] – that if *p=2* then *r(p)* is an optimal bound. They raise an open question, that for different *p(>2)* values their bound is optimal or not.

There is a relationship between bin packing and memory allocation scheduling task of computer programs (jobs). The size of a job corresponds to the size of an element, while a memory partition corresponds to the capacity of a bin. The most important difference is that a memory area ordered to a job can be freed after finishing the job. To model this new class of bin packing algorithms can be defined; the algorithms for *dynamic bin packing problem* (DBP), where the scheduler can specify *Delete* operation. This means that the removal of the elements is also allowed in addition to their arrival (Insert operation). This specification is part of the input, i.e. an input list is finite series of Insert and Delete operations. Thus the next input operation is always either an insertion or a removal of a previously inserted element. We underline, that a removal can be only part of the input, i.e. the algorithm, or the packer cannot delete an element, only the input provider (the scheduler) can do it. The number of used bins by a dynamic bin packing algorithm *A* can be defined as the maximum of the nonempty used bins during the steps. Then the asymptotic competitive ratio *ACR*(*A)* of a dynamic bin packing algorithm *A* can be defined similarly by formula (1). It is easy to see, that the original bin packing problem is the special case of the dynamic problem, where the list consists of inserts only. The dynamic bin packing problem has been defined and analyzed by Coffman, Garey and Johnson in [6]. They gave approximate algorithm for the problem and analyzed it.

When the scheduler can apply only Insert operation (no Delete) with repacking, Ivkovič és Lloyd proved in [23], that for for every *ε>0* there is a *(1+ε)*-competitive approximation scheme A_ε that requires *O(log n)* amortized time per Insert operation and there is a (*1+ε)-competitive* fully polylogarithmic approximation scheme A_ε that requires $O(\log^2 n)$ amortized time per Insert operation.

For the same bin packing problem Epstein and Levin [10] gave an APTAS. In their model the total size of the moved elements per step (insert operation) is bounded by *β* times the size of the arriving element. Their algorithm *Algorithm Dynamic APTAS* uses at most *(1+ε)OPT(L)+1* bins.

A special case of the dynamic bin packing problem is the so-called *fully dynamic bin packing* (FDBP, [21,22]) problem. The difference from the dynamic bin packing problem is that the packer is allowed to perform repacking. If the repacking is restricted, i.e. in every step the packer can repack at most *c* number of elements, then we speak about a *c-repacking fully dynamic bin packing problem* (*c*-repacking FDBP). Here *c* is a fixed positive integer. Obviously, the case *c=0* gives back the pure on-line bin packing problem.

The classical on-line bin packing can be also relaxed by allowing the repacking of at most *c* elements in each step. This version of the problem is called *c-repacking semi-on-line bin packing* (*c*-repacking SOL). Obviously, the case *c=0* gives back the pure on-line bin packing problem.

Some years after Galambos's first paper on repacking algorithms Gambosi et al. [15,16] returned to the analysis of certain semi-on-line algorithms. Their algorithm used repacking, but the application and the cost of the algorithm was defined in a special way. Their method packs the "big enough" elements one by one, while it composes bundles of the small items. Then one group is moved in one step, and the cost of such a movement is 1, i.e. this kind of movement counts 1-repacking. In this sense these algorithms may move even $O(n)$ piece of elements in one step. In paper [16] the authors analysed two algorithms. The faster, linear time algorithm A_1 has an ACR of *3/2*, while the other, $O(n \log n)$ running time algorithm A_2 has an ACR of *4/*3.

Ivkovič and Lloyd investigated the FDBP problem. Similarly to the technique presented in [16], their algorithm uses the bundle technique for the small items. Their asymptotic competitive ratio was *5/4* [22].

Until the last decades no lower bounds have been given for the efficiency of the semi-on-line algorithms (except the trivial ones). From this point of view the first paper is published by Ivkovič and Lloyd in 1996 [21]. They prove that there is no *c*-repacking FDBP algorithm, which has a better asymptotic competitive ratio than *4/3*. With a small modification the construction can be applied for the c-repacking SOL problem as well. This means that the lower bound *4/3* becomes valid for this subproblem, as it is mentioned in the paper of Csirik and Woeginger [8]. It is important that the bound is valid for both problems and for every *c*. The same holds for our improvement to *1.3871*.

4 Our results

In [3] we improved the best known lower bound *4/3* [21] to *1.3871* for the c-repacking semi-on-line problem. We presented our proof for the *c*-repacking SOL problem, but we proved that it remains valid for the *c*-repacking FDBP problem as well. The results obtained are valid for every *c*. We proved the lower bound by analyzing and solving a specific optimization problem. For the analysis of the construction we used different instruments: LP-techniques were combined with results from linear algebra, and finally we solved and analysed a special non-linear optimization problem. We expressed the exact value of the lower bound by the Lambert W function. Note that our construction generalizes the constructions in [19] and [21].

We proved some lower bound results for the special case of the problem as well. In [2] we dealt with the special case of the problem, when there is another condition: the maximal number of the different elements is restricted by a given p ($p \geq 2$) constant. The lower bounds are valid for these special cases of both above mentioned semi-on-line problems. Further they are valid for the *2*-BBPP problem defined in [19]. The bounds improve the lower bounds for the case $p \geq 3$ in [19] (they are valid either for the c-repacking or the classical version of the problem). This way we answered the open question in [19] for the *2*-batched bin packing problem allowing at most p different item sizes. In [2] we proved that our construction can improve these bounds, giving a negative answer for their optimality. The construction works for specific p ($p \geq 2$), and the lower bounds are valid for the above mentioned two problems, the *c*-repacking SOL problem and the *c*-repacking FDBP. In the above paper we prove, that although our bounds are presented in one dimension, the above mentioned lower bounds are valid in every *d*-dimension. To our best knowledge this is the first multidimensional semi-on-line bin packing result.

The curiosity of the lower bounds in [2] and [3] is that although the problem is clearly discrete, the solution requires solving specific nonlinear (continuous) optimization problems. To do this we used global optimization methods. So the construction presents a nice relationship between combinatorial and global optimization. In [2] we solved the nonlinear optimization problem by a reliable Branch and Bound method based on interval arithmetic [27,28]. Reliability means that the proof is produced by a computer, but the results are checked and the inaccuracy coming from roundings is eliminated.

In another paper [4] we improved our previous results and gave new upper bounds for the *c*-repacking SOL problem. We analysed our algorithm using classical methods (for example weighting function technique). We gave a series of algorithms *HFR-c* for every positive integer value *c*. We proved that if *c* goes to infinity, then *ACR(HFR-c)* goes to *3/2*. More exactly we proved that the asymptotic competitive ratio for a given *c* is not larger than $3/2+b_c/(1-b_c)$ where b_c is in the interval $(0, 1/(6c)]$. The given upper bounds prove that repacking really can help, i.e. this “allowance” can be used well. To see this we emphasize two cases. The ACR of our algorithm *HFR-2* for the case of $c=2$ is smaller or equal to *1.5728...*, which is below the best known ACR for on-line algorithms. [31]. Another curiosity of the case $c=2$ is that the given ACR is smaller than the lower bound *1.58333* proved for *Harmonic Fit* type on-line algorithms [29]. In the case of $c=1$ the result of our algorithm is irrelevant from that point of view that the best on-line algorithm is more competitive [31]. However our *1*-repacking algorithm uses much less bin classes. Another important issue, that for the case of $c=4$ our algorithm has better ACR than the best lower bound for on-line algorithms [32]. This means that it is more competitive than any other on-line algorithm. Of course this is due to the semi-on-line property, which enables us to do repacking. This proves that it is worth to investigate such problems.

Lot of open questions can be raised, we mention three of them. The first is to give a 1-repacking algorithm with an ACR better than *1.58333*! The second is to give a semi-on-line *c*-repacking algorithm with an ACR better than *1.5403*-for $c<4$! Further open problem is to improve the lower bounds for small *c* values ($c=1,2$).

Finally we mention a conjection. Epstein and Levin considered a subproblem in [11], when the bin packing algorithms know the value of the optimum in advance. For this case they proved a lower bound *1.30556*. We suppose that this can be improved to at least *1.32312*.

Acknowledgement. The research has been supported by the MÖB-DAAD German-Hungarian Research Project (number P-MÖB/837).

References

[1] J. Balogh, J. Békési, G. Galambos, New Lower Bounds for Certain Bin Packing Algorithms. *Accepted to 8th WAOA.* 2010.

[2] J. Balogh, J. Békési, G. Galambos, M.Cs. Markót. Improved lower bounds for semi-on-line bin packing problems. *Computing, 84.* pp. 139-148. 2009.

[3] J. Balogh, J. Békési, G. Galambos, G. Reinelt. Lower bound for bin packing problem with restricted repacking. *SIAM Journal on Computing 38.* pp. 398-410. 2008.

[4] J. Balogh, J. Békési, G. Galambos, G. Reinelt. On-line Bin Packing with Restricted Repacking. *Manuscipt.* 2010.

[5] E.G. Coffman, G. Galambos, S. Martello, D. Vigo. Bin packing approximation algorithms: Combinatorial analysis. In *Handbook of Combinatorial Optimization Supplement Volume*. Kluwer Academic Publishers. pp. 151–208. 1999.

[6] E.G. Coffmann, M.R. Garey, D.S. Johnson. Dynamic bin packing. *SIAM Journal on Computing 12*. pp. 227–260. 1983.

[7] J. Csirik, G. Galambos, Gy. Turán. Some results on bin packing. *Proceedings of EURO VI.* Vienna. 1983.

[8] J. Csirik, G.J. Woeginger, On-line packing and covering problems. In *On-line algorithms. Lecture Notes in Computer Science.* Vol. 1442. Berlin. pp. 147–177. 1998.

[9] Gy. Dósa: The Tight Bound of First Fit Decreasing Bin-Packing Algorithm Is FFD(I) <= 11/9OPT(I) + 6/9. *ESCAPE.* pp. 1-11. 2007.

[10] L. Epstein, A. Levin. A robust APTAS for the classical bin packing problem. *Mathematical Programming 119.* pp. 33-49. 2009.

[11] L. Epstein, A. Levin. On bin packing with conflicts. *SIAM Journal on Optimization* 19. pp. 1270-1298. 2008.

[12] W. Fernandez de la Vega, G.S. Lueker. Bin packing can be solved within 1+ε in linear time. *Combinatorica 1.* pp. 349–355. 1981.

[13] G. Galambos. A new heurisic for the classical bin packing problem. *Technical Report 82*. Institute fuer Mathematik. Augsburg. 1985.

[14] G. Galambos, G.J. Woeginger, Repacking helps in bounded space on-line bin packing. *Computing 49.* pp. 329–338. 1993.

[15] G. Gambosi, A. Postiglione, M. Talamo. New algorithms for on-line bin packing. In A*lgorithms and Complexity, Proceedings of the First Italian Conference.* World Scientific. Singapore. pp. 44–59. 1990.

[16] G. Gambosi, A. Postiglione, M. Talamo. Algorithms for the relaxed on-line bin-Packing model. *SIAM Journal on Computing 30*. pp. 1532–1551. 2000.

[17] M.R. Garey, D.S. Johnson. Computers and Intractability (A Guide to the theory of NP-Completeness). W.H. Freeman and Company. San Francisco. 1979.

[18] E.F. Grove. On-line bin packing with lookahead. *Proc. SODA.* pp. 430-436. 1995.

[19] G. Gutin, T. Jensen, A. Yeo. Batched bin packing. *Discrete Optimization 2.* pp. 71–82. 2005.

[20] G. Gutin, T. Jensen, and A. Yeo. Optimal on-line bin packing with two item sizes. *Algorithmic Operations Research 1.* pp. 72-78. 2006.

[21] Z. Ivkovič, E.L. Lloyd. A fundamental restriction on fully dynamic maintenance of bin packing. *Information Processing Letters 59.* pp. 229–232. 1996.

[22] Z. Ivkovič, E.L. Lloyd. Fully dynamic algorithms for bin packing: being (mostly) myopic helps. *SIAM Journal on Computing 28.* pp. 574-611. 1998.

[23] Z. Ivkovič and E. Lloyd. Partially dynamic bin packing can be solved within 1+ε in (amortized) polylogarithmic time. *Information Processing Letters 63.* pp. 45-50. 1997.

[24] D.S. Johsnon. Near-optimal bin packing algorithms. *PhD thesis.* MIT. Cambridge. MA. 1973.

[25] N. Karmarkar, R. Karp. An efficient approximation scheme for the one-dimensional bin packing problem. *Proc. 23rd Annual Symposium Foundation Computer Science (FOCS).* pp. 312-320. 1982.

[26] C.C. Lee, D.T. Lee. A simple on-line bin packing algorithm. *Journal of the ACM 32.* pp. 562-572. 1985.

[27] M.Cs. Markót, T. Csendes. A new verified optimization technique for the "packing circles in a unit square" problems. *SIAM Journal on Optimization 16.* pp. 193–219. 2005.

[28] M.Cs. Markót, T. Csendes, A.E. Csallner. Multisection in interval branch-and-bound methods for global optimization. II. Numerical tests. *Journal of Global Optimization 16.* pp. 219–228. 2000.

[29] P. Ramanan, D.J. Brown, C.C. Lee, D.T. Lee. On-line bin packing in linear time. *Journal of Algorithms 10.* pp. 305–326. 1989.

[30] M.B. Richey. Improved bounds for harmonic-based bin packing algorithms. *Discrete Applied Mathematics 34.* pp. 203-227. 1991

[31] S.S. Seiden. On the on-line bin packing problem. *Journal of the ACM 49.* pp. 640-671. 2002.

[32] A. van Vliet. An improved lower bound for on-line bin packing algorithms. *Information Processing Letters 43.* pp. 277–284. 1992.

Minimizing Total Weighted Earliness-Tardiness on a Single Machine Around a Small Common Due Date: An FPTAS Using Quadratic Knapsack

Hans Kellerer [a] and *Vitaly Strusevich* [b]

[a] Department of Operations Research and Statistics, University of Graz,
Universitätsstraße 15, A-8010 Graz, Austria, hans.kellerer@uni-graz.at.

[b] School of Computing and Mathematical Sciences, University of Greenwich,
Old Royal Naval College, Park Row, Greenwich, London SE10 9LS, U.K., V.Strusevich@greenwich.ac.uk.

Abstract

We design a fully polynomial-time approximation scheme (FPTAS) for a single machine scheduling problem to minimize the total weighted earliness and tardiness with respect to a common restrictive due date. Notice for this problem no constant-ratio approximation algorithm has been known so far. Our approach is based on adopting an FPTAS for a special version of the knapsack problem to minimize a convex quadratic non-separable function. For the continuous relaxation of such a knapsack problem we give an algorithm of a quadratic time complexity. The running time of each presented FPTAS is strongly polynomial.

1 Introduction

In this paper, we consider a single machine scheduling problem to minimize the sum of weighted earliness and tardiness computed with respect to a common due date. We show how to adapt a fully polynomial approximation scheme (FPTAS) earlier developed for a quadratic knapsack problem (QKP) of a special structure to handle the problem under consideration, provided that the due date is smaller than the total processing times of the jobs.

In the scheduling problem that we study here, the jobs of set $N = \{1, 2, \ldots, n\}$ have to be processed without preemption on a single machine. The processing of job $j \in N$ takes p_j time units. There is a positive weight w_j associated with job j, which indicates its relative importance. The completion time of job $j \in N$ in a feasible schedule S is denoted by $C_j(S)$ or just C_j. The jobs have a common due date d. In a schedule S, a job is said to be *early* if $C_j(S) - d \leq 0$, and its *earliness* is defined as $E_j(S) = d - C_j(S)$. On the other hand, a job is said to be *late* if $C_j(S) - d > 0$, and its *tardiness* is defined as $T_j(S) = C_j(S) - d$. The aim is to find a schedule that minimizes the function $\sum_{j \in N} w_j (E_j(S) + T_j(S))$.

Problems with an earliness-tardiness criterion are important in just-in-time manufacturing, where the earliness generates holding cost and the tardiness incurs a penalty for a late delivery. Notice that the weights are symmetric, i.e., for job j the same weight w_j is applied, no matter the job is late or early. In this paper, we concentrate on the problem of minimizing the total weighted earliness and tardiness $\sum w_j(E_j + T_j)$ with respect to a *small* or *restrictive* due date d, so that the sum of all processing times is larger than d. The problem is NP-hard in the ordinary sense as proved in [2, 3], and we design an algorithm that delivers its approximate solution.

Recall that for a problem of minimizing a function $Z(x)$, where x is a collection of decision variables, a polynomial-time algorithm that finds a feasible solution x^H such that $Z(x^H)$ is at most $\rho \geq 1$ times the optimal value $Z(x^*)$ is called a $\rho-$*approximation* algorithm; the value of ρ is called a *worst-case ratio bound*. A family of $\rho-$approximation algorithms is called a *fully polynomial-time approximation scheme (FPTAS)* if $\rho = 1 + \varepsilon$ for any $\varepsilon > 0$ and the running time is polynomial with respect to both the length of the problem input and $1/\varepsilon$.

The goal of this paper is to demonstrate that the scheduling problem to minimize the total weighted ealiness-tardiness with a small common due date admits an FPTAS, which is strongly polynomial. A similar approach (with different technical details) has been used in [6, 7] for obtaining FPTASs for various scheduling problems.

2 Preliminaries

Given an instance of the problem with the set $N = \{1, 2, \ldots, n\}$ of jobs and a common due date d, define $p(N)$ as the sum of all processing times, i.e., $p(N) = \sum_{j \in N} p_j$. If $p(N) \leq d$, the due date is called *large* or *nonrestrictive*;

otherwise, for $p(N) > d$, the due date is called *small* or *restrictive*.

As far as the problem with a large due date is concerned, here we only mention that the problem of minimizing the total unweighted earliness-tardiness $\sum_{j\in N}(E_j + T_j)$ with equal weights is solvable in $O(n \log n)$ time, and is NP-hard if the weights are not equal; see [1]. Kovalyov and Kubiak [8] present an FPTAS for the problem of minimizing the total weighted earliness-tardiness $\sum_{j\in N} w_j(E_j + T_j)$.

Following a widely accepted three-field notation for scheduling problems, we refer to this problem as $1|d_j = d|\sum w_j(E_j + T_j)$. The study on this problem has been independently initiated by Hall et al. [2] and Hoogeveen and van de Velde [3]. The problem is NP-hard in the ordinary sense even if the weights are equal. To the best of our knowledge, for problem $1|d_j = d|\sum w_j(E_j + T_j)$ no approximation algorithm with a constant worst-case ratio has been reported. If the weights are equal, a $\frac{4}{3}$−approximation algorithm by Hoogeveen et al. [4] is available.

Given an instance of problem $1|d_j = d|\sum w_j(E_j + T_j)$, assume that the jobs are numbered according to the *WSPT* (Weighted Shortest Processing Time) rule, i.e., in such a way that

$$\frac{p_1}{w_1} \le \frac{p_2}{w_2} \le \ldots \le \frac{p_n}{w_n}. \tag{1}$$

Recall that in an optimal schedule for the classical single machine problem of minimizing the weighted sum of the completion times, the jobs are processed according to the WSPT sequence, as shown by Smith [10].

3 Results

For problem $1|d_j = d|\sum w_j(E_j+T_j)$, an optimal schedule can be sought for in two classes of schedules. First, in an optimal schedule some job may complete exactly at time d, i.e., has neither earliness nor tardiness. There is no intermediate idle time in job processing, but some idle time may occur before the first early job, we call this class of schedules *Class 1.* Second, an optimal schedule may belong to the class of schedules in which the early jobs are processed starting at time zero and are followed by the *straddling* job that starts before time d and is completed after time d; in turn, the straddling job is followed by the block of late jobs. We call this class of schedules *Class 2*.

In a schedule of either class the early jobs are processed in the order opposite to their numbering by the WSPT rule, while the jobs that start after the due date are processed in the order of their numbering.

For each of the two classes that may contain an optimal schedule, we relate the task of finding a schedule in the corresponding class to a quadratic knapsack problem with the following Boolean decision variables

$$x_j = \begin{cases} 1, & \text{if job } j \text{ completes before the due date } d \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Assume that for problem $1|d_j = d, |\sum w_j(E_j+T_j)$ there exists an optimal schedule that belongs to Class 1. Considering the jobs in the order of their numbering defined by (1), we compute the completion time and the earliness of a job j that completes before time d as

$$C_j = d - \sum_{i=1}^{j-1} p_i x_i,\ E_j = \sum_{i=1}^{j-1} p_i x_i.$$

If job j is starts after the due date, then its completion time and tardiness are given by

$$C_j = d + \sum_{i=1}^{j} p_i(1 - x_i),\ T_j = \sum_{i=1}^{j} p_i(1 - x_i).$$

Thus, we obtain that

$$\begin{aligned} \sum_{j=1}^{n} w_j(E_j + T_j) = & \sum_{1\le i<j\le n}^{n} p_i w_j x_i x_j \\ & + \sum_{1\le i<j\le n}^{n} w_j \left(\sum_{i=1}^{j} p_i(1 - x_i) \right) (1 - x_j) \\ & + \sum_{j=1}^{n} p_j w_j(1 - x_j). \end{aligned}$$

Hence, for problem $1|d_j = d|\sum w_j(E_j + T_j)$, finding the best schedule in Class 1 (of schedules with no straddling job) reduces the following Boolean quadratic programming problem:

$$\begin{array}{ll} \text{Minimize} & Z = \sum_{1\le i<j\le n} p_i w_j x_i x_j \\ & + \sum_{1\le i<j\le n} p_i w_j (1 - x_i)(1 - x_j) \\ & + \sum_{j=1}^{n} p_j w_j (1 - x_j) \\ \text{Subject to} & \sum_{j=1}^{n} p_j x_j \le d \\ & x_j \in \{0, 1\},\ j = 1, 2, \ldots, n. \end{array} \tag{3}$$

Consider a Boolean quadratic programming problem of the following structure

$$\begin{array}{ll} \text{Minimize} & Z = \sum_{1\le i<j\le n} \alpha_i \beta_j x_i x_j \\ & + \sum_{1\le i<j\le n} \alpha_i \beta_j (1 - x_i)(1 - x_j) \\ & + \sum_{j=1}^{n} \mu_j x_j + \sum_{j=1}^{n} \nu_j (1 - x_j) + \Gamma \\ \text{Subject to} & \sum_{j=1}^{n} \alpha_j x_j \le A \\ & x_j \in \{0, 1\},\ j = 1, 2, \ldots, n. \end{array} \tag{4}$$

We call this problem the *Symmetric Quadratic Knapsack Problem*, or problem (SQKP). Here all coefficients $\alpha_j, \beta_j, \mu_j, \nu_j,\ j = 1, 2, \ldots, n$, and Γ are non-negative integers. We call the problem *symmetric* because both the quadratic and the linear parts of the objective functions are separated into two terms, one depending on the variables x_j, and the other depending on the variables $(1-x_j)$. An important feature of our problem is that the coefficients α_j in the linear constraint are the same as in the quadratic terms of the objective function. We can view the value α_j as the weight of item j, $1 \leq j \leq n$, i.e., $x_j = 1$ means that item j is placed into the knapsack of a total weight at most A, while $x_j = 0$ means that the corresponding item is not placed into the knapsack.

To see why this problem is relevant to our study on problem $1|d_j = d|\sum w_j (E_j + T_j)$, compare (3) and (4) and observe the former problem is a special case of the latter with

$$\alpha_j = p_j,\ \beta_j = w_j,\ \mu_j = 0,\ \nu_j = w_j p_j,\ j = 1, 2, \ldots, n, \\ A = d,\ \Gamma = 0. \tag{5}$$

In general, the quadratic knapsack problem (QKP) is NP-hard in the strong sense. See Chapter 12 of the book by Kellerer et al. [5] and a recent survey [9] by Pisinger for an overview of principal results on the QKP.

In [7], it has been proved that problem (SQKP) under certain conditions admits an FPTAS, which can be adopted for several single machine scheduling problems with machine availability constraints. One of the conditions of the existence of an FPTAS is the existence of a constant-ratio approximation algorithm that delivers an upper bound on the value of the objective function. No such an algorithm is given in [7] for problem (SKQP), because for the corresponding scheduling applications the required constant-ratio approximation algorithms can be obtained using pure scheduling reasoning, not related to the quadratic knapsack problems.

It is not known whether a $\rho-$approximation algorithm exists for all instances of problem (SQKP). We give such an algorithm for problem (SQKP) under specific additional conditions. In particular, we require that the items are numbered in such a way that

$$\frac{\alpha_1}{\beta_1} \leq \frac{\alpha_2}{\beta_2} \leq \ldots \leq \frac{\alpha_n}{\beta_n}. \tag{6}$$

Our algorithm is based on rounding a solution to a continuous relaxation of problem (SQKP). Informally, it works as follows. First, we show that the continuous relaxation of problem (4) for which (6) holds can be solved in $O(n^2)$ time. We start with a solution to a continuous relaxation of problem (SQKP) and round down to zero those components that are less than a specially chosen value. To determine the other variables, we solve a continuous linear knapsack problem to obtain a solution with at most one fractional component. Then we fix the value for that component to 1, thereby reducing the dimension of the problem, and repeat the process again.

Assume now that an optimal schedule belongs to Class 2, i.e., one of the jobs is straddling. To find a feasible (but not necessarily the best) schedule that belongs to this class, we select each job as a possible straddling job, find a Class 1 schedule S_m of the remaining $m = n - 1$ jobs. Thus, for each choice of the straddling job we need to find the best way of inserting the job into the schedule for the remaining jobs.

Theorem 1 *There is an FPTAS for the single machine scheduling problem to minimize the total weighted earliness and tardiness with respect to a common restrictive due date with strongly polynomial running time.*

4 Conclusion

Further research in this direction may include extending conditions under which the QKP admits an FPTAS or a constant-ratio approximation algorithm. The search for possible applications of the corresponding knapsack problems is also of interest.

References

[1] N.G. Hall and M.E. Posner, "Earliness-tardiness scheduling problems, I: weighted deviation of completion times about a common due date," *Oper. Res.* **39** (1991) 836-846.

[2] N.G. Hall, W. Kubiak and S.P. Sethi, "Earliness-tardiness scheduling problems, II: deviation of completion times about a restrictive common due date," *Oper. Res.* **39** (1991) 847-856.

[3] J.A. Hoogeveen and S.L. van de Velde, "Scheduling around a small common due date," *Europ. J. Oper. Res.* **55** (1991) 237-242.

[4] J.A. Hoogeveen, H. Oosterhout and S.L. van de Velde, "New lower and upper bounds for scheduling around a small common due date," *Oper. Res.* **42** (1994) 102-110.

[5] H. Kellerer, U. Pferschy, and D. Pisinger, *Knapsack Problems* (Springer, Berlin, 2004).

[6] H. Kellerer and V.A. Strusevich, "A fully polynomial approximation scheme for the single machine weighted total tardiness problem with a common due date," *Theor. Comp. Sci.* **369** (2006) 230-238.

[7] H. Kellerer and V.A. Strusevich, "Fully polynomial approximation schemes for a symmetric quadratic knapsack problem and its scheduling applications," *Algorithmica* **57** (2010) 769–795.

[8] M.Y. Kovalyov and W. Kubiak, "A fully polynomial approximation scheme for weighted earliness-tardiness problem," *Oper. Res.* **47** (1999) 757-761.

[9] D. Pisinger, "The quadratic knapsack problem - a survey," *Discr. Appl. Math.* **155** (2007) 623-648.

[10] W.E. Smith, "Various optimizers for single stage production," *Naval Res. Logist. Quart.* **3** (1956) 59-66.

Determining the expected runtime of exact graph coloring

Zoltán Ádám Mann and Anikó Szajkó
Budapest University of Technology and Economics
Department of Computer Science and Information Theory
Magyar tudósok körútja 2., 1117 Budapest, Hungary
e-mail: zoltan.mann@cs.bme.hu, szajko.aniko@gmail.com

September 4, 2010

Abstract

Exact algorithms for graph coloring tend to have high variance in their runtime, posing a significant obstacle to their practical application. The problem could be mitigated by appropriate prediction of the runtime. For this purpose, we devise an algorithm to efficiently compute the expected runtime of an exact graph coloring algorithm as a function of the graph's size, density, and the number of available colors.

1 Introduction and previous work

Graph coloring is one of the most fundamental problems in algorithmic graph theory, with many practical applications such as register allocation, frequency assignment, pattern matching, and scheduling [16, 5, 15]. Unfortunately, graph coloring is NP-complete [8]. Moreover, if $P \neq NP$, then no polynomial-time approximation algorithm with an approximation factor smaller than 2 can exist for graph coloring [7].

Exact graph coloring algorithms are often variants of the usual backtrack algorithm. The backtrack algorithm has the advantage that, by pruning large parts of the search tree, it can be significantly more efficient than checking the whole search space exhaustively. Although in the worst case the backtrack algorithm requires an exponential number of steps, its average-case complexity is $O(1)$ [19].

The probabilistic analysis of the coloring of random graphs was first suggested in the seminal paper of Erdős and Rényi [6]. Through subsequent work of several researchers, the coloring and, in particular, the chromatic number of random graphs is well understood [10, 4, 11, 17, 12, 2, 1]. In terms of the performance of backtracking on random graphs, only some lower and upper bounds are known on the moments of the distribution of the algorithm's runtime [3]. However, as the difference between the known lower and upper bounds is quite high, it is not possible to predict even the order of magnitude of the runtime of backtracking on a problem instance.

Predicting the runtime of the algorithm would greatly improve its practical usability, by informing the user in advance about the estimated runtime. This would let the user decide if the exact solution of the problem is realistic in the available time frame, or a heuristic solution should be used instead. More generally, it allows the manual or automated selection of the most suitable algorithm from an algorithm portfolio [9]. It also enhances load balancing when several problem instances are solved in parallel on multiple machines.

Hence, our aim is to obtain accurate results on the expected runtime of the backtrack algorithm in coloring random graphs. We restrict ourselves to the non-colorable case; extension of our model to the colorable case remains as future work. We use the size of the search tree as a measure of complexity and analyze the expected size of the search tree as a function of input parameters. Our contribution is an algorithm for determining the expected size of the search tree exactly. The algorithm uses dynamic programming, and its runtime is polynomial in the size of the graph. We also present our empirical findings on how the complexity of the problem depends on the input parameters.

2 Preliminaries

We consider the decision version of the graph coloring problem, in which the input consists of an undirected graph $G = (V, E)$ and a number k, and the task is to decide whether the vertices of G can be colored with k colors such that adjacent vertices are not assigned the same color. The input graph is a random graph from $G_{n,p}$, i.e. it has n vertices and each pair of vertices is connected by an edge with probability p independently from each other. The vertices of the graph will be denoted by $v_1, \ldots, v_n$, the colors by

$1, \dots, k$. A *coloring* assigns a color to each vertex; a *partial coloring* assigns a color to some of the vertices. A (partial) coloring is *invalid* if there is a pair of adjacent vertices with the same color, otherwise the (partial) coloring is *valid*.

The backtrack algorithm considers partial colorings. It starts with the empty partial coloring, in which no vertex has a color. This is the root – that is, the single node on level 0 – of the search tree. Level t of the search tree contains the k^t possible partial colorings of $v_1, \dots, v_t$. The search tree, denoted by T, has n levels, the last level containing the colorings of the graph. Let T_t denote the set of partial colorings on level t. If $t < n$ and $w \in T_t$, then w has k children in the search tree: those partial colorings of $v_1, \dots, v_{t+1}$ that assign to the first t vertices the same colors as w.

In each partial coloring w, the backtrack algorithm considers the children of w and visits only those that are valid. T depends only on n and k, not on the specific input graph. However, the algorithm visits only a subset of the nodes of T, depending on which vertices of G are actually connected. The number of actually visited nodes of T will be used to measure the complexity of the given problem instance.

3 The expected number of visited nodes of T

For each $w \in T$, we define the following random variable (the value of which depends on the choice of G):

$$Y_w = \begin{cases} 1 & \text{if } w \text{ is valid,} \\ 0 & \text{else.} \end{cases}$$

Let $p_w = Pr(Y_w = 1)$. Moreover, we define one more random variable (whose value also depends on the choice of G): $Y =$ the number of visited nodes of T.

Since the algorithm visits exactly the valid partial colorings, it follows that $Y = \sum_{w \in T} Y_w$, and thus $E(Y) = \sum_{w \in T} E(Y_w)$. Moreover, it is clear that $E(Y_w) = p_w$. It follows that the expected number of visited nodes in T is:

$$E(Y) = \sum_{w \in T} p_w.$$

Let $Q(w) := \{\{x, y\} \in V^2 : x \neq y, color(x) = color(y)\}$, where V^2 is the set of unordered pairs of elements of V. Let $q(w) := |Q(w)|$. Clearly, w is valid if and only if, for all $\{x, y\} \in Q(w)$, x and y are not adjacent. It follows that $p_w = (1-p)^{q(w)}$ and thus the expected number of visited nodes of T is:

$$E(Y) = \sum_{w \in T} (1-p)^{q(w)}.$$

Note that computing $E(Y)$ through this formula is not tractable since $|T|$ is exponentially large in n.

4 Efficient calculation using dynamic programming

Before presenting our algorithm, we need to introduce some further notions. Our first aim is to compute the maximum possible value of $q(w)$ within T_t. We denote by $s(w, i)$ (or simply s_i if it is clear which partial coloring is considered) the number of vertices of G that are assigned color i in the partial coloring w.

Proposition 1. *For all $w \in T_t$, $q(w) \leq \binom{t}{2}$.*

Proof.

$$\begin{aligned} q(w) &= \sum_{i=1}^{k} \binom{s_i}{2} = \frac{1}{2} \left(\sum_{i=1}^{k} s_i^2 - \sum_{i=1}^{k} s_i \right) \leq \\ &\leq \frac{1}{2} \left(\left(\sum_{i=1}^{k} s_i \right)^2 - \sum_{i=1}^{k} s_i \right) = \frac{1}{2} \left(t^2 - t \right) = \binom{t}{2}. \end{aligned}$$

□

It is also possible to derive a formula for the minimum of $q(w)$ [13], but it is not necessary for our purposes.

Let $R(q, t, k) := |\{w \in T_t : q(w) = q\}|$ denote the frequency of value q among the $q(w)$ values of nodes in T_t.

If we could determine all the $R(q, t, k)$ values explicitly, that would enable us to calculate the exact value of $E(Y)$:

$$E(Y) = \sum_{w \in T} (1-p)^{q(w)} = \sum_{t=0}^{n} \sum_{q=q_{min}(t)}^{q_{max}(t)} R(q, t, k)(1-p)^q.$$

Determining the $R(q, t, k)$ values is possible with the following recursion:

Proposition 2.

$$R(q, t, k) = \sum_{j=0}^{t} \binom{t}{j} R\left(q - \binom{j}{2}, t - j, k - 1\right).$$

Proof. Assume that color class 1 contains j vertices. There are $\binom{t}{j}$ possibilities to choose these j vertices. The remaining $t - j$ vertices must be colored with $k - 1$ colors. Moreover, the j vertices of color 1 already account for $\binom{j}{2}$ vertex pairs with identical colors. Hence, the remaining $t - j$ vertices must be colored in such a way that the number of vertex pairs with identical colors out of these $t - j$ vertices equals $q - \binom{j}{2}$. □

Based on this recursive formula, we can use dynamic programming to compute the $R(q, t, k)$ values and store them in a 3-dimensional table. We fill this table according to increasing values of k. For a given k, we must iterate through

Algorithm 1 Dynamic programming algorithm to compute $E(Y)$

```
for t=0 to n
{
    R((t choose 2), t, 1) = 1
}

for k=2 to number of colors
{
    for t=0 to n
    {
        for q=q_min to q_max
        {
            R(q,t,k) = 0
            for j=0 to t
            {
                if q - (j choose 2) >= q_min(t-j, k-1)
                {
                    R(q,t,k) = R(q,t,k) + (t choose j) R(q - (j choose 2), t-j, k-1)
                }
            }
        }
    }
}

k=number of colors
result=0
for t=0 to n
{
    for q=q_min to q_max
    {
        result=result+R(q,t,k)(1-p)^q
    }
}
E(Y)=result
```

the possible values of t from 0 to n, and for each such t, we must fill the table for all possible values of q from q_{min} to q_{max}. As a starting point, when $k = 1$, then for all values of t, $q_{min} = q_{max} = \binom{t}{2}$ and for this value of q we have $R(q, t, k) = 1$. As additional boundary conditions, we have $R(q, t, k) = 0$ in all cases when $t < 0$ or $q < q_{min}$. See Algorithm 1 for details.

Since $t = O(n)$, $j = O(n)$ and $q_{max} = O(n^2)$, the runtime of Algorithm 1 is $O(kn^4)$. This is polynomial in the size of the graph, though quite high. On the other hand, the calculation of the $R(q, t, k)$ values is the most time-consuming part of the algorithm and these values can be pre-computed and stored. Afterwards, we can compute $E(Y)$ more quickly for different values of n, p, k.

5 Numerical results

The presented method enables us to gain some insight as to how the complexity of graph coloring changes for different values of the parameters n, k, p. Fig. 1 shows an example: $E(Y)$ as a function of n and k, for fixed p. We can conclude from the figure that for small values of k, the problem is easy, even if n becomes large. This is consistent with previous results on the relatively low average-case complexity of graph coloring [19, 18]. However, as k increases, this increases the complexity of the problem dramatically (note the exponential scale on the vertical axis). It is still true that the complexity saturates, i.e. increasing n does not increase the complexity significantly after some threshold. However, this saturation takes place at a much higher value than in the case of small k.

A more detailed empirical analysis using the tool BCAT [14] will be part of a future extended version of this paper.

6 Conclusion and future work

We have investigated the complexity of a typical backtracking algorithm for coloring random graphs of the class $G_{n,p}$ with k colors. Using the expected size of the search tree as the measure of complexity, we devised a polynomial-time algorithm for predicting complexity.

In this paper, we only dealt with uncolorable problem instances. Our future work will focus on extending the presented results to colorable problem instances.

Acknowledgements

This work was partially supported by the Hungarian National Research Fund and the National Office for Research and Technology (Grant Nr. OTKA 67651).

References

[1] Dimitris Achlioptas and Assaf Naor. The two possible values of the chromatic number of a random graph. In *36th ACM Symposium on Theory of Computing (STOC '04)*, pages 587–593, 2004.

[2] Noga Alon and Michael Krivelevich. The concentration of the chromatic number of random graphs. *Combinatorica*, 17(3):303–313, 1997.

[3] Edward A. Bender and Herbert S. Wilf. A theoretical analysis of backtracking in the graph coloring problem. *Journal of Algorithms*, 6(2):275–282, 1985.

[4] Béla Bollobás. The chromatic number of random graphs. *Combinatorica*, 8(1):49–55, 1988.

[5] Preston Briggs, Keith D. Cooper, and Linda Torczon. Improvements to graph coloring register allocation. *ACM Transactions on Programming Languages and Systems*, 16(3):428–455, 1994.

[6] Pál Erdős and Alfréd Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5:17–61, 1960.

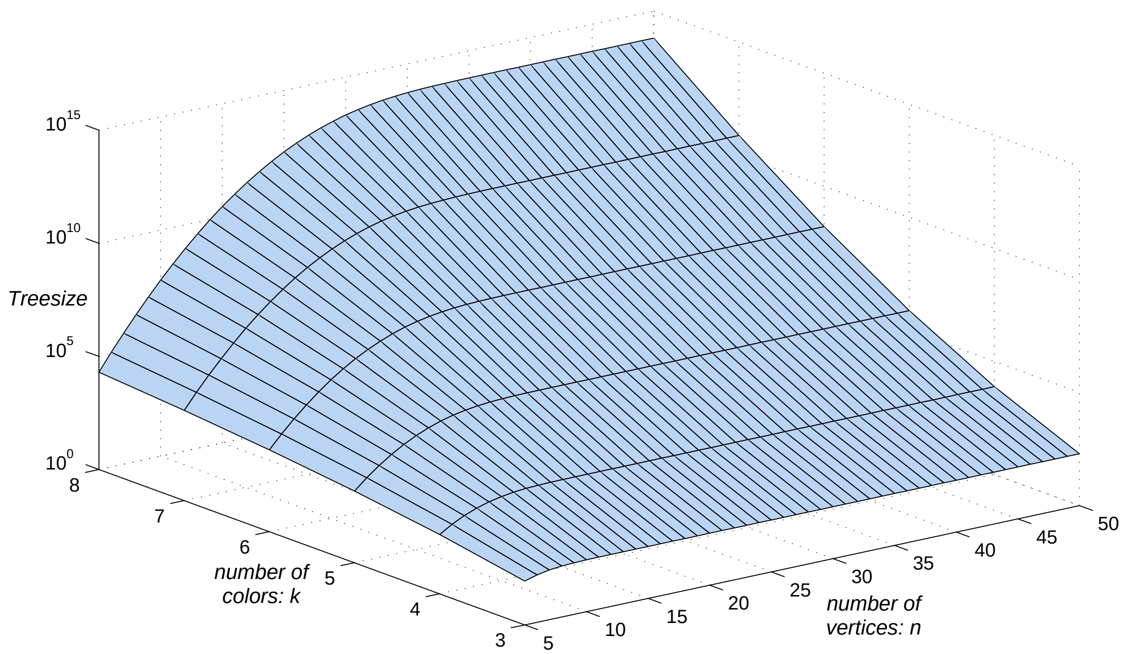


Figure 1: Expected size of the search tree for $p = 0.5$, as a function of n and k.

[7] Michael R. Garey and David S. Johnson. The complexity of near-optimal graph coloring. *Journal of the ACM*, 23:43–49, 1976.

[8] Michael R. Garey, David S. Johnson, and L. J. Stockmeyer. Some simplified NP-complete graph problems. *Theoretical Computer Science*, 1:237–267, 1976.

[9] Carla P. Gomes and Bart Selman. Algorithm portfolios. *Artificial Intelligence*, 126(1-2):43–62, 2001.

[10] G. R. Grimmett and C. J. H. McDiarmid. On colouring random graphs. *Mathematical Proceedings of the Cambridge Philosophical Society*, 77(2):313–324, 1975.

[11] Tomasz Luczak. The chromatic number of random graphs. *Combinatorica*, 11(1):45–54, 1991.

[12] Tomasz Luczak. A note on the sharp concentration of the chromatic number of random graphs. *Combinatorica*, 11(3):295–297, 1991.

[13] Zoltán Á. Mann and Anikó Szajkó. Improved bounds on the complexity of graph coloring. In *12th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2010.

[14] Zoltán Á. Mann and Tamás Szép. BCAT: A framework for analyzing the complexity of algorithms. In *8th IEEE International Symposium on Intelligent Systems and Informatics*, 2010.

[15] Zoltán Ádám Mann and András Orbán. Optimization problems in system-level synthesis. In *3rd Hungarian-Japanese Symposium on Discrete Mathematics and Its Applications*, pages 222–231, 2003.

[16] Nirbhay K. Mehta. The application of a graph coloring method to an examination scheduling problem. *Interfaces*, 11(5):57–65, 1981.

[17] Eli Shamir and Joel Spencer. Sharp concentration of the chromatic number on random graphs $G_{n,p}$. *Combinatorica*, 7(1):121–129, 1987.

[18] Jonathan S. Turner. Almost all k-colorable graphs are easy to color. *Journal of Algorithms*, 9(1):63–82, 1988.

[19] Herbert S. Wilf. Backtrack: an O(1) expected time algorithm for the graph coloring problem. *Information Processing Letters*, 18:119–121, 1984.

Community detection and its use in Real Graphs

András Bóta
Department of Computer Science
University of Szeged, Hungary
bandras@inf.u-szeged.hu

László Csizmadia
Department of Computer Science
University of Szeged, Hungary
cheeseme@freemail.hu

András Pluhár
Department of Computer Science
University of Szeged, Hungary
pluhar@inf.u-szeged.hu

ABSTRACT

We survey and unify the methods developed for finding overlapping communities in Small World graphs in the recent years. The results have impact on graph mining; we give some demonstration of this.

1. INTRODUCTION

The discovery of Small World graphs has changed the direction of interest in graph theory profoundly. These graphs are different from those that were studied before, and also the questions that were asked about those. It is not easy to collect the information to build such a graph, or give models to generate it. The sheer size of the real problems prohibits most of the time consuming algorithms, so the researcher has to fall back on simpler heuristics, sometimes derived from physical intuition [2, 4, 19]. Following the usual notation, a graph G has vertex set $V(G)$, edge set $E(G)$. If the later one consists of ordered pairs, then G is directed, and an edge might be also weighted.

An intriguing question is the classification of vertices of a graph. One can consider the usual clusters and also overlapping sets, that we call communities. Here we concentrate on the possible definitions, search and use of communities. While for clustering both the top down and bottom up algorithms are used for defining and finding the classes, all known algorithms for communities are bottom up.

2. SOME ALGORITHMS

Here we consider only three algorithms. The selection is arbitrary, although has some justification. Maybe the first algorithm that was used for finding communities is the N^{++}. However, since we could get no permission to use the data set it was designed for, it has not been published yet in English. After that several similar algorithm were proposed; unfortunatelly the qualities of implementations differ so it is not easy to compare them. The k-Clique percolation method was the first widely known algorithm, which was also applied to real world problems. Edge clustering is the third algorithm we mention; it has mainly theoretical interest.

2.1 The N^{++} algorithm.

[22, 9] It is a generic algorithm, with arbitrary functions

$$f : 2^{V(G)} \times V(G) \to \mathbb{R}$$

and $c : \mathbb{N} \to \mathbb{R}$. Here $f(A, x)$ describes the strength of a community A with a vertex x. Then the algorithm joins x to A if $f(A, x) \geq c(|A|)$. The **Build** routine gets the first approximation of communities $\mathcal{K}$ in a bottom up way.

The pseudo-code of Build

begin(Build)

Input G, k, c (`max` k`-size` c`-communities`)

Let $\mathcal{K} := V(G)$ (`nodes are communities.`)

For $i = 1$ to k

$\forall A \in \mathcal{K}$, $x \in V(G)$ if $f(A, x) \geq c(|A|)$ then put $A \cup \{x\}$ into $\mathcal{K}$.

Remove all $A \in \mathcal{K}$, for which $A \subset B \in \mathcal{K}$, and $A \neq B$.

Print $\mathcal{K}$, "c-communities of G up to size k."

end(Build)

After running Build, we use **Merge** to glue communities that are almost identical. Let C be a graph, where $V(C) = \mathcal{K}$, and $(A, B) \in E(C)$ if $A \cap B$ is "big" then changes $\mathcal{K}$ to $(\mathcal{K} \setminus \{A, B\}) \cup \{A \cup B\}$. Then the components of C are declared to be the communities. The practice suggested the following set-ups. The big means the 60% of the smaller set. The function $f(A, x)$ depends on the number of paths with length one and two from x to A. That is to get the communities containing x, it is enough to search $N^{++}(x) := N(N(x))$.

Some similar methods are listed in [12].

2.2 k-Clique percolation.

[20] Here a $k \in \mathbb{N}$ is fixed. After finding all k-size clique in G, the graph Q_k is considered such that the vertices of Q_k are these cliques, and $(A, B) \in E(Q_k)$ iff $|A \cap B| = k - 1$. Finally a k-community is the unions of cliques of a connected components.

2.3 Edge clustering.

[21, 11] One chooses an arbitrary clustering on the set of edges. Then the communities are defined as the set of endpoints of the clusters.

These methods differ in output, i. e. in the type of communities, and in the computing costs. Although the edge clustering is easy to compute, it has serious drawbacks in use. (First of all is that the overlap among communities is maximum one vertex.) The N^{++} and Clique percolation are more promising; here the implementation issues are crucial. For small world graphs both can perform almost in linear time, which is a natural requirement if one wants to deal with real problems.[1]

2.4 A unified view

These algorithms, and those that were mentioned but not listed, has a common core. Their execution consists of two steps. In the first a hypergraph $\mathcal{F} = (V, \mathcal{H})$ is defined (and computed), where $V = V(G)$, the original point set of the graph G, and $\mathcal{H} \subset 2^V$. The elements of $\mathcal{H}$ can be considered as the building blocks of the communities. In the second step one endowes the set $\mathcal{H}$ with an appropriate d distance function and thereby establishes a metric space $\mathcal{M} = (\mathcal{H}, d)$. Then a chosen clustering algorithm is executed on $\mathcal{M}$, yielding a set of clusters $\mathcal{C}$. Finally, the arising clusters are associated to the subsets of V such that $K_i = \cap_{H \in C_i \in \mathcal{C}} H$, where K_i, the ith community corresponds to C_i, the ith cluster and K_i is just the union of those hyperedges that belong to C_i.

In the case of the mentioned algorithms $\mathcal{H}$ consist of vertex sets of the small dense subgraph, k-cliques and the edges, respectively. The distance functions are represented by an appropriate graph $\mathcal{D}$, take the value one if there is an edge, infinity otherwise. In the first case $(K_i, K_j) \in \mathcal{D}$ if $|K_i \cap K_j|$ is big enough, in the second if $|K_i \cap K_j| = k - 1$, while we left this as a parameter in the third case.

3. EVALUATION

Since more or less all community (or cluster) definitions are arbitrary [17], there are several ideas to measure their goodness. This is a crucial point and naturally the viewpoint of researches differ. There are direct and indirect methods to assess the usefulness of communities, the following list is far from being complete.

3.1 Appearance, parametrization

First of all, one has to run the algorithms, get the outputs and possible make mathematical predictions for certain graph classes. That is an important factor is the speed of these algorithms. However, it is not easy to compare the real speed of these algorithms since it depends strongly on the implementations and test graphs (being real or theoretical). Definitely all three algorithms, and perhaphs most algorithms in that family we described in subsection 2.4 are fast, and designed to solve huge problems. In subsection 3.3 we recur to this problem, and report some date on time and a goodness measure (modularity) of the solutions.

The clique percolation method is appealing from both theoretical and practical view. For Erdős-Rényi random graphs the clique percolation process is thorougly studied and well understood, [5]. It was reported to be useful also in practice, [1]. However, it sometimes gives too large communities and the parametrization is elusive, since one has to decide for which value k to be chosen?

The N^{++} algorithm looks arbitrary, and do not yield for theoretical investigations. Its main advantages are the speed, the small diameter of the communities and its robustness. The edge clustering methods are not well studied or tested in practice. Their inherent problem is that communites derived this way may have only one common element, what is too restrictive in real graphs.

3.2 Graphical.

Another way is to compare the communities with some visualized form of the graph; this was the most common approach in the early publications. Indeed, the clustering methods provide classes that conform the eye. Assessing communities (permitting overlapping) are harder, since visualization is not an obvious task anymore. Some ideas, like showing the intersection graph of communities can help. However, this approach has certain limits; it works only for small graphs and it is always subjective.[2]

3.3 Modularity.

The Newman modularity [19] is the following function of a graph G and its partition:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

where $m = |E(G)|$, A_{ij} is the adjacency matrix of G, k_i is the degree, c_i is the cluster of the ith vertex, and $\delta(c_i, c_j)$ is the Kronecker symbol. The clustering algorithms may be based on some mathematical/physical heuristics like edge-betweenness (EB), eigenvectors (EV), label propagation (LP), spin glass (SG), walk trap (WT), or try to maximize the modularity function itself on the set of all partitions with a greedy algorithm (Gr). The formula can be generalized to communities [18]. One write s_{ij} instead of $\delta(c_i, c_j)$, where s_{ij} is an arbitrary similarity measure between vertices i and j. (In [18] the u_i is a probability distribution of i over the communities, and $s_{ij} = \langle u_i, u_j \rangle$, but it could be $\|u_i - u_j\|$ form any norm.) On the other hand, it is possible to get communities by maximizing the modularity function. The findings of [15] show the cluster and community structure cannot be measured on the same scale, some additional weighting must be introduced to solve this. The algorithms were tested for some graphs, we illustrate the results on the well-known Zachary graph. The clusterings are followed the Clique Percolation (CPM) with clique sizes $k = 3$ and $k = 4$, and N^{++} with its default parameters. The running time is in seconds, $\#C$ stands for the number of clusters or communities, whatever it applies.

Method	Modularity	Running time	#C
EB	0.4013	0.0100	5
EV	0.3727	0.0000	3
Gr	0.3807	0.0000	3
LP	0.4020	0.0000	3
SP	0.4063	1.1500	6
WT	0.4198	0.0000	4
CPM 3 CPM 4	0.2438 0.2557	0.012	3 3
N^{++}	0.1947	0.6690	12

One can evaluate cluster/community algorithms in indirect ways. That is by taking a problem in which the communities might have predictive value, and check the usefulness of these. We have observed dependencies among functions in some social graphs (telecommunication, friendship, Erasmus contracts etc.), and practically all methods provided useful hints. However, here the use of communities greatly outperforms the methods which use only clusters.

[1]This means millions of vertices. The N^{++} available in the Sixtep software, while the Clique percolation in the CFinder.

[2]For graph visualization the so-called *force directed* algorithms performed best. However, these usually take $O(n^2)$ time that prohibit the use when n is several thousand or million.

3.4 Refinements, time and orders.

One can conduct similar studies like the graphical method if have some functions that are defined on the vertices or the edges. Again, we have seen some highly subjective but still robust phenomena that might deserve to be mentioned.

First of all, the clusters are usually much bigger than the communities, and their number is less.

The number of communities might follow power law, although even to test this is impossible.

The communities are usually within the clusters, and give a fine structure of those larger classes. However, the reverse direction is also detected, the clusters might give information on communities. To be more precise, the most interesting communities are those ones in which elements belong to several clusters.

In social graphs we confirmed the role of the weak links described in [13], and also tested the different algorithms. The communities given by N^{++} are containing strong edges almost exclusively, while most of the weak edges are among communities. On the other type of small world graphs, the so-called technical graphs[3] there are no such effects. We used data from [16]. (The CPM does not give good results with any k, perhaps its performance is too sensitive to the measurement errors, missing data.)

The social graphs might have natural vertex attribute, the time when a vertex has been joined to the net. This order may not be manifested in the clusters if one considers the whole graph, but shows remarkable coincidence when restricting the graph to the neighborhood of a fixed vertex. In that case the clusters usually can be interpreted with some interval of time or spatial restrain. Note, that communities may cross the borders of clusters.

3.5 Weights.

Dealing with weighted graphs is difficult. It turns out that for the indirect methods the numerical results are more reliable. While all these methods can be extended to weighted graphs, the performance of them is little known [6].

In the rest we outline a model which is an example for indirect evaluation. The infection models are central in applications of real graphs [3], but to build appropriate ones is far from being trivial. The main points are (i) which model to choose, (ii) what are the significant variables and (iii) how to decide the values of the parameters. Our investigations concentrated on two problems in corporate banking, default (failing in paying debt) [7] and delay (in paying debt) [8]. We have to stress, although the two problems look similar, there are subtle differences.

3.6 Independent Cascade Model (IC)

This model is due to Domingos and Richardson [10], but an equivalent is in [14]. Here an edge weighted graph G is given, where to the edge (v, w) a probability $p_{v,w}$ is associated. The process of infection goes as follows. In the 1st step the set of infected vertices F_1 considered active, that is $F_1 = A_1$. In general for a vertex $w \in V(G) \subset F_{i-1}$ gets infected with probability $p = \prod_{v \in A_{i-1}} p_{v,w}$, and in that case $w \in F_i$. Note that the infected vertices may transmit the disease only in the very next step, that is $A_i = F_i \setminus F_{i-1}$. If for an i $F_i = F_{i-1}$, then the process halts.

[3]In social graphs the presence of edges (x, y) and (x, z) increases the conditional probability of the the edge (y, z), while in the technical graphs this probability is decreased in that case.

3.7 Results.

The two most significant variables of IC model are the proportional flow one an edge and the number of communities containing of that edge. Based on this, we found an expected 3-4 to even 10-12 times lift in the different segments [7]. The fact that a vertex x is in a same community with an infected increases the chance of x's infection by a factor three [8]. The computations were carried out by the use of Sixtep software.

4. REFERENCES

[1] B. Adamcsek, G. Palla, I. J. Farkas, I. Dernyi, T. Vicsek CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics* **22**, 1021–1023 (2006).

[2] R. Albert, A. L. Barabási Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74** 2002.

[3] M. Boguñá, R. Pastor-Satorras, A. Vespignani, Absence of epidemic threshold in scale-free networks with connectivity correlations. Preprint cond-mat/0208163 (2002).

[4] B. Bollobás, *Modern Graph Theory,* Springer, New York (1998).

[5] B. Bollobás and O. Riordan, Clique percolation. *Random Structures Algorithms* **35** (2009), no. 3, 294–322.

[6] A. Bóta, Applications of Overlapping Community Detection, $(CS)^2$ - Conference of PhD Students in Computer Science, Szeged, 2010.

[7] A. Csernenszky, Gy. Kovács, M. Krész, A. Pluhár and T. Tóth, The use of infection models in accounting and crediting. Challenges for Analysis of the Economy, the Businesses, and Social Progress, Szeged, 2009.

[8] A. Csernenszky, Gy. Kovács, M. Krész, A. Pluhár and T. Tóth, Parameter Optimization of Infection Models. $(CS)^2$ - Conference of PhD Students in Computer Science, Szeged, 2010.

[9] L. Csizmadia, Recognizing communities in social graphs, MSc thesis, University of Szeged, 2003.

[10] P. Domingos, M. Richardson, Mining the Network Value of Costumers. *7th Intl. Conf. on Knowledge Discovery and Data Mining,* 2001.

[11] T. S. Evans and R. Lambiote, Edge Partitions and Overlapping Communities in Complex Networks. arXiv:0912.4389v1, 2009.

[12] S. Fortunato, Community Detection in graphs. arXiv:0906.0612

[13] M. Granovetter, The Strength of Weak Ties. *American Journal of Sociology* **78(6)** 1360–1380, 1973.

[14] M. Granovetter, Threshold models of collective behavior. *American Journal of Sociology* **83(6)** 1420–1443, 1978.

[15] E. Griechisch, Clustering and community finding methods in graphs, MSc thesis, University of Szeged, 2010.

[16] C. A. Hidalgo, B. Klinger, A. L. Barabási and R. Hausmann, The Product Space Conditions the Development of Nations. *Science* (2007) 317: 482–487.

[17] J. Kleinberg, An Impossibility Theorem for Clustering. *Advances in Neural Information Processing Systems (NIPS)* 15, 2002.

[18] T. Népusz, A. Petróczi, L. Négyessy and F. Bazsó, Fuzzy communities and the concept of bridgeness in complex networks. arXiv:0707.1646v3, 2007.

[19] M.E.J. Newman, The structure and function of complex networks. Preprint cond-mat/0303516 (2003).

[20] G. Palla, I. Derényi, I. Farkas and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814 (2005).

[21] A. Pluhár On the clusters of social graphs based on telephone log-files. (in Hungarian) Research Report 2001.

[22] A. Pluhár Determination of communities in social graphs by fast algorithms. (in Hungarian) Research Report 2002.

GREEDY HEURISTICS FOR DRIVER SCHEDULING AND ROSTERING

Viktor Árgilán, Attila Tóth
Institute of Applied Sciences
Juhász Gyula Faculty of Education
University of Szeged
Boldogasszony sgt. 6, 6725 Szeged, Hungary
e-mail: {gilan, attila}@jgypk.u-szeged.hu

Balázs Dávid, Csaba Kemény, Gábor Pongrácz
Institue of Informatics
Faculty of Science and Informatics
University of Szeged
Árpád tér 2., 6720 Szeged, Hungary
e-mail: {David.Balazs, Kemeny.Csaba, Pongracz.Gabor}@stud.u-szeged.hu

ABSTRACT

In this paper, we introduce greedy methods for both the driver scheduling and the driver rostering problems. The methods were created to be efficient parts of a decision support system in public transportation, and give a good result for real life instances in acceptable time.

1 INTRODUCTION

Driver scheduling and rostering are important problems arising in public transportation. Based on a feasible vehicle schedule, driver scheduling determines the shifts for each day of the planning period. This is followed by driver rostering, where the shifts are assigned to the drivers. All these steps are limited by different rules and constraints, usually defined by several authorities (government, EU, etc.).

Several different approaches have been introduced in literature for the problems above [3]. Solving the problems by formulizing exact mathematical models [1], using constraint programming [2,4,7], or evolutionary algorithms [5,6] give good theoretical solutions, but not all of them can be applied in practice. Both the driver scheduling and the driver rostering problems are NP-hard, and solving real life instances (in our case: Szeged, Hungary: 170.000 inhabitants, ~2700 trips on a usual workday) efficiently is not always possible using these approaches. The planning period is usually 1-2 months, which makes the size of the problem even larger.

Real life applications demand a good running time, as the problems are usually considered as parts of decision-support systems, where the emphasiss is on long time planning. The user should be able to create several different solutions to the problem, compare them, or even re-calculate some results of the system, meaning that the algorithms will be executed several times. Since real life application of these problems in a decision support system requires running time measured in minutes, the running time of the methods is an important aspect, besides the quality of their given solutions.

This leads to the introduction of heuristic approaches, which can result in an effective solution in adequate time. Our goal was to introduce fast heuristic algorithms for both problems, which work effectively on real life data, giving an "quasi-optimal" solution. Greedy heuristics are classical methods of algorithm design, and are known to have a quick running time. Though they usually fail to achieve global optimum, we will show through our test results that they can be applied efficiently on the driver scheduling and rostering problems, and give a good solution for real-life problems.

The outline of this paper is the following: first, we deal with the driver scheduling problem, followed by the driver rostering problem. In both cases, we first present the traditional IP formulation, and show why this type of approach is unusable on a real life instance. Choosing other models would have been possible, but the problems arising from the size of the IP model apply to the other models as well. After this, we propose greedy methods for the problems. Finally, we present the results of the methods on a real life instance, and draw the conclusions.

The results of these algorithms are shown on instances provided by the Szeged Bus Company, Tisza Volán ZRt.

2 DRIVER SCHEDULING

The input of our driver scheduling method is a feasible vehicle schedule, which contains the properties of the assigned buses besides the the basic properties (starting time, ending time, starting place, ending place) of the tasks. The planning period is always one day so the optimization works day-by-day and considers only those rules which concern only the daily duty.

The most important rules are the rules of maximal working time, the minimal resting time (between two shifts), the length of the breaks, and their place and frequency. Apart from these, other special rules may apply, depending on the country, city, or the company.

The objective function minimizes the number and cost of the shifts.

Literature usually discusses two different solution methods. The first is Generate and Selection Approach (GaS), which initially generates sets of the candidate shifts. The size of these sets may vary from a couple of 100.000 to even 1.000.000. The shifts, giving the best solution, are selected from these sets. This problem can be reduced to the set covering or set partitioning problems, and formulized as the following:

Minimze

$$w_1 \sum_{j=1}^{n} c_j x_j + w_2 \sum_{j=1}^{n} x_j$$

subject to

$$\sum_{j=1}^{n} a_{ij} x_j \geq 1 \quad , i = 1,2,\ldots,m$$

xj = 0 or 1, j = 1,2,…,n

where

n = the number of candidate shifts

m = the number of tasks

x_j = shift variable, x_j = 1 if shift j is selected and 0 otherwise

c_j = cost of shift j

a_{ij} = 1 if taks i is covered by shift j, 0 otherwise

w_1 and w_2 are weights

There are several efficient methods to solve this. However, the main problem for a real-life instance (like Szeged) may have several million different shifts for a day, and the applied rules also increase the complexity further. Finding a feasible solution is really hard, as the number of candidate shifts is big, and generating the problem may take a very long time because of the numerous rules. Solving the IP problem above with such a size also requires significant running time. Real-life applications of such a decision support system demands that it should give results in an adequate time, so that the results could be modified, or several different results tested and compared. Another solution method is more suitable for this, which is called Constructive Approach. Our method uses this approach as well.

Another aspect is that the solution algorithm should be adjustable according to the needs of the user, with modifyable methods of solutions using different settings. The basis of our method based on these observations is a controlled greedy approach, which can give an adequate, feasible solution in a couple of seconds, according to the needs of the user.

The main idea is to divide the driver scheduling into sequential substeps, where different optimization methods can be used and combined. This way a big flexibility can be reached, since it is quite easy to adjust the process to the expectations, or it is even possible to improve the output at a step before the next one by hand, as sometimes a professional human mind can see what an optimization algorithm could not find.

The process goes through the following steps:

1. Creating work-pieces. This makes work-pieces by pulling the tasks together by the possiblity of the driver reliefing.
2. Creating pre-shifts. This generates initial solutions, called pre-shifts, ommiting the difficult rules.
3. Cutting. This step slices the pre-shifts into smaller feasible parts considering all the rules.
4. Joining. In this step the shift parts are joined together producing longer feasible shifts.
5. Refining. It improves the solution by changing work-pieces between the shifts.
6. Finishing. Finally the finishing step inserts all required duties into the shift and define an idle activity for all idle time.

The aim of (1) is to reduce the size of the problem. (2) generates the initial shifts. These shifts cover all the events without overlapping, but they do not necessarily satisfy all the given rules. The task of (3) is to create feasible pieces considering the rules. At this point, several different strategies can be applied, which are based on the length, structure, the cost, or on the load of the different periods of the day. (4) joins the resulting parts together, applying also different strategies. The joining of two parts can be executed with two different greedy methods, or also according to the load of the different periods of the day. (5) does not change the number of the shifts, only exchanges different tasks between shifts, if the cost can be reduced this way. (6) is a deterministic method, which inserts other events defined by the rules into the shifts.

The flexibility of the method is mainly provided by (2), (3) and (4), where the user can combine different methods, and manually alter the result of the different methods after every step. Since the running time for a day is around a few seconds for the whole planning period not more then a couple of minutes.

3 DRIVER ROSTERING

The driver rostering problem uses the output given by the driver scheduling algorithm, and also needs certain information about the drivers (contract type, etc.), and the driver constraints. These constraints come from rules specified by one or more authorities, usually the local government and the EU. The most important rules are the following:

1. The maximum weekly working time of a driver (48 hoursin our case).
2. Resting time must be given to a driver between 2 shifts (12 hours in our case).
3. The maximum number of consecutive working days of a driver (6 in our case).
4. The minimum number of monthly resting day of a driver (4 in our case).
5.The minimum number of monthly Sundays given as resting day to a driver (1 in our case).

The shifts given as the result of the driver scheduling have to be assigned to the drivers for each day of the planning period, with respect to all driver rules.

We present two main approaches for the driver rostering problem: first, we give a classical IP model for the problem, and show the limits of this approach. After that, we present a fast greedy solution method, which gives good results for the problem.

3.1 IP approach

The problem can be represented with an IP model, where we want to assign each shift to exactly one driver, without violating the rules for the drivers. The driver rules give us the problem constraints, and our objective function minimizes the cost of the assignment of the shifts.

Let

$F_t = \{f_{tj}\}_{j-1}^{n_t}$ be shifts of day t,

$F_t = \{f_{tj}\}_{j-1}^{n_w}$ be shifts of week w,

$F = \{F_t\}_{t-1}^{\tau}$ shifts of τ days, where $\tau \in \square$.

Let $S = \{s_i\}_{i-1}^{m}$ be the set of drivers, and let

ϕ (f_{tj}) = (beginning of day f_{tj}, end of day f_{tj}).

Using the rules mentioned before, the constraints for the IP formulation of the driver rostering problem are the following:

$$\forall i \forall w \sum_{j \in F_w} l_j x_{ij} <= 2880$$

$$\forall i \forall j \sum_{\substack{k \in F \\ if\, \varphi(f_{tj})[1] \le \varphi(f_{tk})[2]+720}} x_{ij} <=1$$

$$\forall i \forall j \sum_{\substack{k \in F \\ if\, \varphi(f_{tj})[1] \le \varphi(f_{tk})[2]+540}} x_{ij} <=1$$

$$\forall i \forall j \sum_{\substack{F_t \in F \\ t=x..x+7}} \sum_{\substack{k \in F_t \\ j \in F_x}} x_{ik} <=6$$

$$\forall i \forall j \sum_{F_w \in F} \sum_{j \in F_w} x_{ij} >=4, \text{ where w =weeks of the month}$$

$$\forall i \forall j \sum_{F_t \in F} \sum_{j \in F_t} x_{ij} >=1, \text{ where t=Sundays of the month}$$

Using these constraints, the objective function of our problem is:

$$\forall i \forall j \sum c_{ij} * x_{ij} \rightarrow \min.$$

The resulting IP is solved using a MILP solver.

The main drawback of this approach comes from the size of the problem. A model of general instance from our input cases has approximately 2.320.000 variables and 2.400.000 inequalities. A problem with this size can not be solved in adequate time using a MILP solver .

Since the running time of the solution of the IP model is too long to be used in a real life situation, a faster method had to be introduced, which gave us a good solution regarding the cost, while having an acceptable running time.

3.2 Greedy algorithm

Greedy algorithms are usually known to provide solution faster, than other optimization methods. Though it comes with the cost that the greedy choice guarantees only local optimum, a good pre-processing and proper selection of the initial driver set can improve the global solution also.

As a first step, the set of drivers needed for the planning period has to be specified. In our case, an estimate for the initial set of drivers is given by a special algorithm, which takes into consideration the the different types of driver contracts, and the number of available driver of each contract as an upper bound.

The algorithm iterates through every day of the planning period. Each day, the daily shifts are assigned to feasible drivers, who are chosen according to a certain strategy (fitness, random).

Let S be the set of shift, and D the set of drivers. Our greedy algorithm for the driver scheduling problem is:

```
GREEDY_DR(S,D)
    for days i=1..n
        S'=unassigned shifts of day(i)
        while S' ≠ ∅
            sj=a shift from S'
            D'=drivers of D, who can execute sj
            dk=a driver from D', chosen by a strategy
            assing si to dk
            S'=S'\{sj}
            S= S\{sj}
```

The method enables the user to choose between several selection strategies of the driver executing the chosen shift. The drivers can be selected randomly, with a first fit approach, or by using a fitness function to determine the best driver for the shift.
Our experience shows that a neighborhood of driver number combinations makes a plane of more or less strictly decreasing values in the search space, and the direction of the decrease is easily recognisable. The estimate values given by our special method is close to this plane. Executing the greedy algorithm above sequentially, using a neighbouring driver combination with an expected lower cost can lead us to a solution closer to the optimum.

4 TEST RESULTS

The presented methods have been implemented and tested in Matlab, implementations under different programming environments might give different results.
The running time of the vehicle scheduling method is a few seconds for a daily schedule, which adds up to 2-3 minutes to a planning period of 2 months. The driver rostering method has a running time of around 2 minutes, which gives us a total running time around of 4-5 minutes (depending on the problem) for a planning period of 2 months for both algorithms together.
We have tested our methods on instances provided by Tisza Volán Zrt. The present practice of the company uses 264 drivers with different contract types. We used the vehicle schedule given by the company as the input for the driver scheduling and rostering algorithm. Our results (the number of drivers, and the differences between their daily working hours and their contract) can be seen in Table 1.

	Drivers	min. difference (%)	max. difference (%)	average difference (%)
Test1	222	-0,026	0	-0,004
Test2	232	-0,01119	0,022933	0,006725

Table 1: *Driver numbers and working hour differences.*

It can be clearly seen, that the methods resulted in a significant decrease in the number of drivers, while the average daily working hours of a driver always stay close to the contract type of the driver.
In the Test1 result, all the drivers had the same type of contracts (average daily 8 working hours), while the Test2 result contains drivers of 3 different contract types (average daily 8, 6 or 4 working hours). This test case is closer to the real-life practice of the company, as they also employ drivers of these 3 contract types.
The high number of the drivers used by the company comes from the fact that their roster also contains people sent on holiday, which we do not consider in our model.

5 CONCLUSION AND FUTURE WORK

This paper shows our results for solving the problems of both driver scheduling and driver rostering in public transportation. The algorithms we presented for the problems are both based on the greedy strategy which guarantees a fast running time, while the algorithms themselves give promising results on real-life instances.
Further improvement of the results is possible, using well chosen heuristic techniques on the output of the vehicle rostering algorithm. Presently a local-search based heuristic is being perfected and tested. This heuristic introduces a cost-function to every crucial rule of the driver rostering, with certain weights depending on the importance of the rule, and our goal is to minimize these costs.The heuristic examines each day of the planning period, and is able to swap the shifts of any two drivers on the given day. The shifts are swapped, if it decreases the overall cost of the rostering, and the algorithm has several termination criteria.
The introduced greedy methods are part of a research and development project, currently in the research phase. The Matlab implementation is mainly for research purposes, the methods will be implemented in the final project using a C++ enviroment.

Acknowledgment
This paper was partially supported by Szeged City Bus Company (Tisza Volán, Urban Transport Division).

References

[1] Cappanera, P. and Gallo, G.: A Multicommodity Flow Approach to the Crew Rostering Problem, Operations Research 52. pp. 583-596. 2004.
[2] A. Caprara, F. Focacci, E. Lamma, P. Mello, M. Milano,P. Toth, D. Vigo, Integrating constraint logic programming and operations research techniques for the crew rostering problem, Software Practice and Experience 28. pp. 49–76. 1998.
[3] A.T. Ernst, H. Jiang, M. Krishnamoorthy and D. Sier, Staff scheduling and rostering: A review of applications, methods and models, European Journal of Operational Research 153. pp. 3–27. 2004.
[4] N., Guerinik, M.V.m Caneghem: Solving Crew Scheduling Problems by Constraint Programming. In Proceedings of the 1st Conference of Principles and Practice of Constraint Programming pp. 481–498. 1995.
[5] M. Moz, M. Vaz Pato: A genetic algorithm approach to a nurse rerostering problem, Computers & Operations Research 34. pp. 667–691. 2007.
[6] Moz, M., Respcio, A., and Vaz Pato, M: Bi-objective Evolutionary Heuristics for Bus Drivers Rostering, Working paper 1-2007, Centro de Investigaao Operacional, Universidade de Lisboa, 2007.
[7] R.S.K. Kwan, A.S.K. Kwan, and A. Wren, Evolutionary driver scheduling with relief chains. Evolutionary Computation 9. pp. 445–460. 2001.

Generating pattern avoiding permutations by ECO - Extended abstract -

Phan-Thuan DO
School of Information and Communication Technology
Hanoi University of Technology
e-mail: thuandp@it-hut.edu.vn
and
Vincent VAJNOVSZKI
LE2I UMR-CNRS 5158, Université de Bourgogne
B.P. 47 870, 21078 DIJON-Cedex France
e-mail: vvajnov@u-bourgogne.fr

September 22, 2010

Abstract: ECO-method and its corresponding succession rules allow to recursively define and construct combinatorial objects. We introduce here *succession functions* which refine succession rules and apply them for the efficient exhaustive generation of classes of pattern avoiding permutations.

Keywords: ECO method, succession rule and function, pattern avoiding permutations, exhaustive generation.

1 Introduction and definition

Let $\mathfrak{S}_n$ be the set of permutations on $[n] = \{1, 2, \ldots, n\}$. We represent a permutation $\pi \in \mathfrak{S}_n$ in online notation: i.e. $\pi = \pi_1\pi_2 \ldots \pi_n$. A permutation $\pi \in \mathfrak{S}_n$ contains the pattern $\tau \in \mathfrak{S}_k$ iff a sequence of indices $1 \le i_1 < i_2 < \ldots < i_k \le n$ exists such that $\pi_{i_1}\pi_{i_2}\ldots\pi_{i_k}$ is ordered as τ. We denote by $\mathfrak{S}_n(\tau)$ the set of permutations of $\mathfrak{S}_n$ avoiding the pattern τ and for a set of patterns T, $\mathfrak{S}_n(T)$ denotes the set of permutations of $\mathfrak{S}_n$ avoiding each pattern in T.

A *barred* pattern $\bar{\tau}$ is a permutation in $\mathfrak{S}_k$ having a bar over one of its elements. Let τ be a permutation on $[k]$ identical to $\bar{\tau}$ but unbarred and $\hat{\tau}$ be the permutation on $[k-1]$ made up of the $(k-1)$ unbarred elements of $\bar{\tau}$, rewritten as a permutation on $[k-1]$. Then $\pi \in \mathfrak{S}_n$ avoids $\bar{\tau}$ if any patter $\hat{\tau}$ in π can be extended to a pattern τ. For example, if $\bar{\tau} = 4\bar{1}32$ then $\pi = 58132674 \in \mathfrak{S}_8(4\bar{1}32)$.

Succession rule and generating tree were introduced in [14] and extensively and systematically studied since then, see for instance [4, 6, 3, 10]. In this paper, we introduce *succession functions*, which refine succession rules and allow to construct general generating algorithms for the classes under consideration. Sometimes this view helps to find new classes easier than traditional methods. Here we bring together some basic results from [8], give many other succession and generalized succession functions for to known and new classes, and give general CAT exhaustive generating algorithms for these casses. Our main original results are given without proof in this short version, by the seven theorems in the last part of this note.

2 Regular patterns

We begin by explaining the generating tree technique in the context of pattern avoidance. The *sites* of $\pi \in \mathfrak{S}_n$ are the positions between two consecutive entries, before the first and after the last entry; and they are numbered, from right to left, from 1 to $n+1$. For a permutation $\pi \in \mathfrak{S}_n(T)$, where T a set of forbidden patterns, i is an *active site* if the permutation obtained from π by inserting $n+1$ into its ith site is a permutation in $\mathfrak{S}_{n+1}(T)$; we call such a permutation a *son* of π. The active sites of a permutation $\pi \in \mathfrak{S}_n(T)$ are *right justified* if the sites to the right of any active site are also active; and we denote by $\chi_T(i, \pi)$ the number of active sites of the permutation obtained from π by inserting $n+1$ into its ith active site. Obviously, for any $n > 1$ and $\pi \in \mathfrak{S}_n(T)$, by erasing n in π one obtains a permutation in $\mathfrak{S}_{n-1}(T)$; or equivalently, any permutation in $\mathfrak{S}_n(T)$ is obtained from a permutation in $\mathfrak{S}_{n-1}(T)$ by inserting n in one of its active sites.

A set of patterns T is called *regular* if:

- the length one permutation $1 \in \mathfrak{S}_1(T)$ has two sons (this condition is used only to start the recursively);
- all active sites are right justified;

- for any $n \geq 1$ and $\pi \in \mathfrak{S}_n(T)$, $\chi_T(i,\pi)$ does not depend on π but solely on i and on the number k of active sites of π. In this case we denote $\chi_T(i,\pi)$ by $\chi_T(i,k)$ and we call it *succession function.*

Let T be a regular set of patterns characterized by its succession function χ. The set of productions

$$(k) \rightsquigarrow (\chi_T(1,k))(\chi_T(2,k))\ldots(\chi_T(k,k)),$$

or more concisely $\{(k) \rightsquigarrow (\chi_T(i,k))\}_{i=1}^{k}$ for $k \geq 1$, is called the *succession rule* corresponding to the set of patterns T.

For any succession function (and generalized succession function later), one can derive a succession rule but the conversely is not true. So, we can say that the succession function gives more information than the succession rule for the class of combinatorial objects under consideration.

Exemple 1. The Catalan set of patterns $T = \{312\}$ and $T = \{321\}$ have the same succession rule $(k) \rightsquigarrow (2)(3)\ldots(k+1)$ whereas they have different succession functions. See (2.1.4) below.

Let σ and σ' be the permutations in $\mathfrak{S}_{n+1}(T)$ obtained from $\pi \in \mathfrak{S}_n(T)$ by the insertion of $n+1$ into its ith and $(i+1)$th active site. Then σ and σ' differ by a transposition, more precisely $\sigma' = \sigma \cdot (n-i+2, n-i+1)$. Generally, the insertion of an element in a list is not an efficient operation but the transposition of two element is an efficient elementary operation. Those are crucial in the CAT-ness of our generating algorithms.

Below we give several classes of regular patterns together with the $\chi(i,k)$ function. Notice that i is always less or equal to k.

2.1 Classes given by counting sequences

2.1.1 2^{n-1} [6]

- $T = \{321, 312\}$: $\chi_T(i,k) = 2$.
- $T = \{321, 231\}$; $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \\ 1 & \text{otherwise.} \end{cases}$

2.1.2 Pell numbers [6, 3]

- $T = \{321, 3412, 4123\}$: $\chi_T(i,k) = \begin{cases} 3 & \text{if } i=1 \\ 2 & \text{otherwise.} \end{cases}$
- $T = \{312, 4321, 3421\}$: $\chi_T(i,k) = \begin{cases} 3 & \text{if } i=2 \\ 2 & \text{otherwise.} \end{cases}$

2.1.3 even index Fibonacci numbers [6, 3]

- $T = \{321, 3412\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \\ 2 & \text{otherwise.} \end{cases}$
- $T = \{321, 4123\}$: $\chi_T(i,k) = \begin{cases} 3 & \text{if } i=1 \\ i & \text{otherwise.} \end{cases}$
- $T = \{312, 4321\}$: $\chi_T(i,k) = \begin{cases} 3 & \text{if } k=3 \text{ and } i=3 \\ i+1 & \text{otherwise.} \end{cases}$

2.1.4 Catalan numbers [13]

- $T = \{312\}$: $\chi_T(i,k) = i+1$.
- $T = \{321\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \\ i & \text{otherwise.} \end{cases}$

2.1.5 Schröder numbers [3]

- $T = \{1234, 2134\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \text{ or } i=2 \\ i & \text{otherwise.} \end{cases}$
- $T = \{1324, 2314\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \text{ or } i=k \\ i+1 & \text{otherwise.} \end{cases}$
- $T = \{4123, 4213\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=k-1 \text{ or } i=k \\ i+2 & \text{otherwise.} \end{cases}$

2.1.6 Central binomial coefficients $\binom{2n-2}{n-1}$ [3]

- $T = \{1234, 1324, 2134, 2314\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \\ 3 & \text{if } i=2 \\ i & \text{otherwise.} \end{cases}$
- $T = \{1324, 2314, 3124, 3214\}$: $\chi_T(i,k) = \begin{cases} 3 & \text{if } i=1 \\ i+1 & \text{otherwise.} \end{cases}$

2.1.7 Motzkin numbers [3]

- $T = \{321, 3\bar{1}42\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \\ i-1 & \text{otherwise.} \end{cases}$

2.1.8 The numbers of the left factors in Motzkin words [4]

- $T = \{321, 4\bar{1}523\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \\ 2 & \text{if } i=2 \\ i-1 & \text{otherwise.} \end{cases}$

2.1.9 Fibonacci numbers [7]

- $T = \{321, 312, 231\}$: $\chi_T(i,k) = \begin{cases} 1 & \text{if } i=2 \\ 2 & \text{otherwise.} \end{cases}$

2.2 Variable length patterns

2.2.1 A pattern of length 3 and a generalized pattern [6, 9]

(a) $T = \{321, (p+1)12\ldots p\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \text{ and } k<p \\ p & \text{if } i=1 \text{ and } k=p \\ i & \text{otherwise.} \end{cases}$

(b) $T = \{321, p(p+1)12\ldots(p-1)\}$: $\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \\ i & \text{if } 1<i<p-1 \\ p-1 & \text{otherwise.} \end{cases}$

(c) $T = \{312, (p+1)p\ldots 21\}$: $\chi_T(i,k) = \begin{cases} p & \text{if } k=p \text{ and } i=p \\ i+1 & \text{otherwise.} \end{cases}$

Notice that $\mathfrak{S}_n(321, (p+1)12\ldots p) = \mathfrak{S}_n(321, p(p+1)12\ldots(p-1)) = \mathfrak{S}_n(312, (p+1)p\ldots 21)$.

2.2.2 A pattern of length 3, a pattern of length 4 and a generalized pattern [6]

$T = \{321, 3412, (p+1)12\dots p\}$:

$$\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \text{ and } k<p \\ p & \text{if } i=1 \text{ and } k=p \\ 2 & \text{otherwise.} \end{cases}$$

2.2.3 p-Generalized Fibonacci numbers [6]

$T = \{321, 231, (p+1)12\dots p\}$:

$$\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \text{ and } k<p \\ k & \text{if } i=1 \text{ and } k=p \\ 1 & \text{otherwise.} \end{cases}$$

2.2.4 A pattern of length 3 and two generalized patterns

(a) $T = \{321, p(p+1)12\dots(p-1), (p+1)12\dots p\}$:

$$\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \text{ and } k<p \\ p & \text{if } i=1 \text{ and } k=p \\ p-1 & \text{if } i=p \text{ and } k=p \\ i & \text{otherwise.} \end{cases}$$

(b) $T = \{321, p(p+1)12\dots(p-1), (m+1)12\dots m\}$:

Theorem 1. *The succession function $\chi_T(i,k)$ corresponding to $T = \{321, p(p+1)12\dots(p-1), (m+1)12\dots m\}$ is*

$$\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \text{ and } k<m \\ m & \text{if } i=1 \text{ and } k=m \\ i & \text{if } 1<i<p \\ p-1 & \text{otherwise.} \end{cases}$$

2.2.5 Generalized Schröder numbers [5, 11, 12]

(a) $T = \cup_{\tau\in\mathfrak{S}_{p-1}}\{(p+1)\tau p\}$:

$$\chi_T(i,k) = \begin{cases} k+1 & \text{if } k\le p \text{ or } i\ge k-p+1 \\ i+p-1 & \text{otherwise.} \end{cases}$$

(b) $T = T_1 \cup T_2$, where $T_1 = \{\sigma \in \mathfrak{S}_{p+1} | \sigma_1 = (p+1), \sigma_{p+1} = p\}$ and $T_2 = \{\sigma \in \mathfrak{S}_{p+1} | \sigma_1 = (p+1), \sigma_p = p\}$:

Theorem 2. *The succession function $\chi_T(i,k)$ corresponding to $T = T_1 \cup T_2$ is*

$$\chi_T(i,k) = \begin{cases} k+1 & \text{if } k<p-1 \text{ or } k-p+4\le i\le k \\ p-1 & \text{if } k\ge p-1 \text{ and } 1\le i\le 2 \\ i+p-3 & \text{otherwise.} \end{cases}$$

2.2.6 p-Generalized Motzkin numbers [4]

$T = \{321, (p+2)\bar{1}(p+3)2\dots(p+1)\}$:

$$\chi_T(i,k) = \begin{cases} k+1 & \text{if } i=1 \\ i & \text{if } 2\le i\le p \\ i-1 & \text{otherwise.} \end{cases}$$

Notice that for each class of variable length pattern in this section, for some particular values of p or m, we retrieve some cases discussed before. For example, the case (2.2.4.b), when $m = p$, becomes the case (2.2.4.a) above; when $m < p$, we deduce the case (2.2.1.a) above; if $p = 2$, then we retrieve the case (2.1.3) above; if $p = 3$, the case (2.1.2.a); and if $m = \infty$, the case (2.1.1.b).

Here we generalize the previous results to more general classes called *colored-regular*. In this 'abstract' version we restrict ourselves with the statement of the five following main results, and a general generating algorithm for these classes.

Theorem 3-7

During the presentation and in an extended version we will provide five theorems giving the succession functions for the classes of permutations avoiding the following patterns: $T = \{312, 4321, 2431\}$ (enumerated by the binomial transform of Padovan sequence), $T = \{312, 2431\}$ (enumerated by the even index Fibonacci numbers), $T = \{312, 2(p+1)p\dots 431\}$ for $p > 3$, $T = \{312, 4321, 34\dots(p+1)21\}$ and $T = \{312, 2431, (p+1)p\dots 21\}$.

```
procedure Gen_Avoid(size, k)
local i
if size = n then
  Print(π)
else
  size := size + 1
  π := [π, size]
  gen_Avoid(size, χ(1, k))
  for i := 2 to k do
    π := π · (size − i + 2, size − i + 1)
    gen_Avoid(size, χ(i, k))
  end for
  for i := k downto 2 do
    π := π · (size − i + 2, size − i + 1)
  end for
end if
end procedure
```

Figure 1: Algorithm 1 [8]: Pseudo code for generating permutations avoiding a set T of regular patterns characterized by the succession function $\chi(i,k)$. After the initialization of π by the length 1 permutation, the call of Gen_Avoid(1, 2) produces $\mathfrak{S}_n(T)$.

References

[1] E. Barcucci, A. Del Lungo, E. Pergola and R. Pinzani. *Permutations avoiding an increasing number of length-increasing forbidden subsequences.* Discrete Mathematics and Theoretical Computer Science 4(1), 31–44 (2000).

[2] S. Brlek, E. Duchi, E. Pergola and S. Rinaldi, On the equivalence problem for succession rules, *Proc., FPSAC'02*, Melbourne (Australie), 8–12 July, 2002.

[3] O. Guibert, Combinatoire des permutations à motifs exclus en liaison avec mots, cartes planaires et tableaux de Young, *PHD Thesis*, Université Bordeaux 1, 1995.

[4] E. Barcucci, A. Del Lungo, E. Pergola and R. Pinzani, From Motzkin to Catalan permutations, *Discrete Mathematics*, 217, Issue 1–3, 33–49, 2000.

```
procedure Gen_Avoid(size, k, c)
local i, u, v
if size = n then
   Print(π)
else
   size := size + 1
   π := [π, size]
   (μ, ν) := χ(1, k, c)
   gen_Avoid(size, μ, ν)
   for i := 2 to k do
      π := π · (size − i + 2, size − i + 1)
      (u, v) := χ(i, k, c)
      gen_Avoid(size, μ, ν)
   end for
   for i := k downto 2 do
      π := π · (size − i + 2, size − i + 1)
   end for
end if
end procedure
```

Figure 2: Algorithm 2: Pseudo code for generating permutations avoiding a set T of c-regular patterns characterized by the 3 variable succession function $\chi(i, k, c)$. After the initialization of π by the length 1 permutation, the call of Gen_Avoid$(1, 2, 0)$ produces $\mathfrak{S}_n(T)$. In particular, when only color $c = 1$ is allowed we obtain Algorithm 1.

[5] E. Barcucci, A. Del Lungo, E. Pergola and R. Pinzani, Permutations avoiding an increasing number of length-increasing forbidden subsequences, *Disc. Math. Theor. Comp. Sci.*, 4, 31–44, 2000.

[6] E. Barcucci, A. Bernini and M. Poneti, From Fibonacci to Catalan permutations, *Pu.M.A.*, 17(1–2), 1–17, 2006.

[7] J.L. Baril and P.T. Do, ECO-generation for p-generalized Fibonacci and Lucas permutations, *Pu.M.A.*, 17(1–2), 19–37, 2006.

[8] W.M.B. Dukes, M.F. Flanagan, T. Mansour and V. Vajnovszki, Combinatorial Gray codes for classes of pattern avoiding permutations, *eprint arXiv:0704.2048*, http://arxiv.org/abs/0704.2048.

[9] T. Chow and J. West, Forbidden sequences and Chebyshev polynomials, *Disc. Math.*, 204, 1999, 119–128.

[10] L. Ferrari, E. Pergola, R. Pinzani and S. Rinaldi, Jumping succession rules and their generating functions, *Discrete Math*, 271, 2003, 29–50.

[11] D. Kremer, Permutations with forbidden subsequences and a generalized Schröder number., *Disc. Math.*, 218, 2000, 121–130.

[12] D. Kremer, Postscript: “Permutations with forbidden subsequences and a generalized Schröder number”, *Disc. Math.*, 270, 2003, 332–333.

[13] J. West, *Generating trees and the Catalan and Schröder numbers*, *Disc. Math.*, 146, 1994, 247–262.

[14] F. R. K. Chung, R. L. Graham, V. E. Hoggatt and M. Kleiman, The number of Baxter permutations, *J. Comb. Th.*, (A) 24, 1978, 382–394.

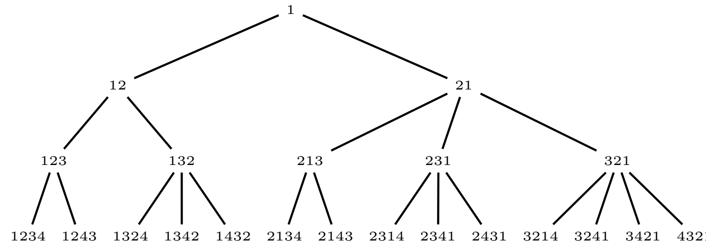


Figure 3: The first four levels of the generating tree corresponding to $\chi(i, k) = i + 1$ producing $\mathfrak{S}(312)$ in 2.1.4. The generating tree coincides with the computation tree of the generating procedure.

A NOTE ON CONTEXT-FREE GRAMMARS WITH REWRITING RESTRICTIONS*

Zsolt Gazdag
Department of Algorithms and their Applications
Faculty of Informatics
Eötvös Loránd University
gazdagzs@inf.elte.hu

ABSTRACT

In this work two special types of random context grammars are investigated. In one of these grammars one can apply a context-free rule to a nonterminal A only if the current sentential form does not contain a particular nonterminal associated to A. The other grammar works in the opposite direction, it requires the presence of the associated nonterminal to apply the rule. Here we show that these systems can generate non-context-free languages.

1 INTRODUCTION

Context-free grammars are among the most investigated topics in the formal language theory. They were used first in the study of human languages, but later it turned out that the specification of programming languages can also be modelled by these grammars. On the other hand, it is well known that context-free grammars can not cover all aspects of natural and programming languages. Therefore several types of grammars that are based on context-free rules but have larger generative power were introduced. This larger generative power is often achieved by adding a regulated rewriting mechanism to context-free grammars. Representatives of these grammar systems are, for example, random context grammars (see e.g. [3]). In these systems two sets of nonterminals, a permitting and a forbidding one, are associated to every context-free rule. Moreover, a rule is applicable, if it is applicable in the context-free sense and

(1) nonterminals from the associated forbidding set do not occur, while

(2) every nonterminal from the permitting set does occur

in the current sentential form.

This simple regulation makes these systems computationally complete if ε-rule is allowed. In [5] an even simpler regulation on the application of context-free rules was made and the resulting grammars were investigated. In this paper we call these grammars context-free grammars with rewriting restrictions.

These systems are in fact special random context grammars, but here the permitting and forbidding sets are associated to the nonterminals rather than to the rules of the system. Moreover, one of these sets is always a singleton set and the other one is empty. In more detail, a context-free grammar with rewriting restrictions is essentially a context-free grammar G enriched with a function f which associates to every nonterminal of G a nonterminal and a sign symbol ($+$ or $-$). We say that a rule rewriting the nonterminal A is applicable if it is applicable in the context-free sense and, additionally, the following holds. If f associates to A the nonterminal B and the sign symbol $-$ (resp. the sign symbol $+$), then B does not occur (resp. does occur) in the current sentential form.

It is shown in [5] that the above described systems are still Turing-equivalent if the context-free sub-system is not ε-free. On the other hand, it is noted in [5], that these systems do not generate all recursively enumerable languages if the system is such that f associates the sign $-$ to every nonterminal. The same applies when f associates the sign $+$ to every nonterminal. Nevertheless, these special types were not further investigated in [5]. In this work we show that these special types can also generate non-context-free languages.[1]

2 PRELIMINARIES

In this paper we assume that the reader is familiar with the basic concepts of the formal language theory, such as alphabets, words, languages, etc. For comprehensive guides to these topics the reader is referred to [4] and [6]. For an alphabet Σ, Σ^* denotes the *set of all words over* Σ. The *empty word* is denoted by ε.

A *context-free grammar* (CFG in short) is a 4-tuple $G = (V, \Sigma, R, S)$ where V and Σ are alphabets of *nonterminal* and

*This research was supported by the project TÁMOP-4.2.1/B-09/1/KMR-2010-003 of Eötvös Loránd University.

[1] Here we note that, independently from us, an even stronger result was proved in [2], but only in the permitting case, i.e., when f associates the sign $+$ to every nonterminal. Namely, it was shown that in this case context-free grammars with rewriting restrictions are equivalent to permitting random context grammars, a subclass of random context grammars, where the forbidding sets are all empty. A poster with this result was presented at DLT 2010 conference [1] after the submission of our work.

terminal symbols, respectively (it is assumed that $V \cap \Sigma = \emptyset$), $S \in V$ is the *start symbol*, and R is a finite set of *production rules* of the form $A \to \alpha$, where $A \in V$ and $\alpha \in (V \cup \Sigma)^*$. The *derivation relation* of G is defined as follows. For every words $u_1, u_2, v \in (V \cup \Sigma)^*$ and $A \in V$, we say that $u_1Au_2 \Rightarrow u_1vu_2$ if and only if there is a rule $A \to v$ in R. The *reflexive, transitive closure of* $\Rightarrow$ is denoted by $\Rightarrow^*$. The *language generated by* G is the language $L(G) = \{u \in \Sigma^* \mid S \Rightarrow^* u\}$. We say that a language is *context-free* (CF) if it is generated by a CFG. The family of all CF languages is denoted by $\mathcal{L}(\text{CF})$.

A *context-free grammar with rewriting restrictions* (CFGrr in short) is a tuple $G = (V, \Sigma, R, S, f)$, where V, Σ, R, S are the same as in a CFG and $f : V \to \{+, -\} \times V$ is a function. If, for every $A \in V$, $f(A) \in \{-\} \times V$, then G is a *context-free grammar with forbidding rewriting restrictions* (CFGfrr, for short). Likewise, if, for every $A \in V$, $f(A) \in \{+\} \times V$, then G is a *context-free grammar with permitting rewriting restrictions* (CFGprr). Since in a CFGfrr (resp. CFGprr), for every nonterminal A, the first component of $f(A)$ is the symbol $-$ (resp. $+$), we will define in these grammars f as a function $f : V \to V$.

The *derivation relation* of G is defined similarly as in the case of CFGs with the following additional restriction. For every $u_1, u_2 \in (V \cup \Sigma)^*$ and rule $A \to v \in R$, $u_1Au_2 \Rightarrow u_1vu_2$ provided that either

(1) $f(A) = (+, B)$ and B appears in u_1Au_2, or

(2) $f(A) = (-, B)$ and B does not appear in u_1Au_2.

The *language generated by* G is defined in the same way as in the case of CFGs.

Finally, we denote by $\mathcal{L}(\text{CFGrr})$, $\mathcal{L}(\text{CFGprr})$, and $\mathcal{L}(\text{CFGfrr})$ the families of languages generated by the corresponding context-free grammars with rewriting restrictions.

3 THE MAIN RESULT

In this section we show that both types of grammars introduced above are strictly more powerful than CFGs.

Theorem 1. $\mathcal{L}(\text{CF}) \subset \mathcal{L}(\text{CFGfrr})$.

Proof. It can be easily seen that $\mathcal{L}(\text{CF}) \subseteq \mathcal{L}(\text{CFGfrr})$. Indeed, let $G = (V, \Sigma, R, S)$ be a CFG. We can construct an equivalent CFGfrr $G' = (V', \Sigma, R, S, f)$ as follows. Let $V' = V \cup \{*\}$ where $*$ is a new nonterminal symbol not occurring in V and, for every $A \in V$, let $f(A) = *$. Clearly $*$ is a nonterminal that does not occur in the sentential forms of the derivations of G'. Thus $*$ never blocks a derivation of G', which means that $L(G) \subseteq L(G')$. On the other hand, $L(G') \subseteq L(G)$ also holds, thus we have $L(G) = L(G')$.

Next we are going to give a CFGfrr G_f and show that $L(G_f) \notin \mathcal{L}(\text{CF})$. The idea behind the construction is the following. We construct G_f such that it generates words over the alphabet $\{a, b, c, d, \#, \$\}$. In fact, we want G_f to generate, among others, all words in $\{a^i\#c^j\$b^kd^l \mid i > k \geq 0, j > l \geq 0\}$ but no words in $\{a^i\#c^j\$b^kd^l \mid k \geq i \geq 0, l \geq j \geq 0\}$. To achieve this, we do the following. The only rule that can introduce the symbol $\#$ is constructed such that after its application the number of $b's$ can not be increased after the position of $\#$ in the sentential form. We construct a similar rule with the symbol \$ on the right-hand side.

Let $G_f = (V, \Sigma, R, S, f)$ be a CFGfrr where

- $V = \{S, A, A_1, B, B_1, C, C_1, D, D_1\}$,
- $\Sigma = \{a, b, c, d, \#, \$\}$,
- $f(S) = A_1$, $f(A) = B_1$, $f(B) = A$, $f(A_1) = B$, $f(B_1) = A_1$, $f(C) = D_1$, $f(D) = C$, $f(C_1) = D$, $f(D_1) = C_1$.

Moreover, let R be the following set of rules. R is devided into three parts: $R = R_1 \cup R_2 \cup R_3$ where

- $R_1 = \{S \to ACBD\}$,
- $R_2 = \{A \to aA_1 \mid \varepsilon, \quad B \to bB_1 \mid \varepsilon, \quad A_1 \to A \mid \#A_1AB, \quad B_1 \to B\}$,
- $R_3 = \{C \to cC_1 \mid \varepsilon, \quad D \to dD_1 \mid \varepsilon, \quad C_1 \to C \mid \$C_1CD, \quad D_1 \to D\}$.

We observe the following things. First of all, in every derivation of G_f, the rule $S \to ACBD$ is applied exactly once, namely at the first derivation step. After that step the nonterminal S does not occur again in the derivations. Thus the forbidding nonterminal $f(S) = A_1$ never blocks a derivation step that involves S (this also means that we could have chosen any nonterminal in V as $f(S)$).

Secondly, the rules in R_2 and R_3 work independently from each other in the following sense. The nonterminals $f(A)$, $f(A_1)$, $f(B)$, $f(B_1)$ do not occur in the right-hand sides of the rules of R_3, and similarly, $f(C), f(C_1), f(D), f(D_1)$ do not occur in the rules of R_2. Thus applying a rule from R_2 can not introduce a nonterminal that would block the application of rules in R_3 and vice versa.

Finally, we note that the structure of the rules in R_2 is the same as that of the rules in R_3.

By the above observations, to see what words can be derived from S, it is enough to examine those sub-derivations of G_f that use rules from R_2 only. Those sub-derivations that use rules from R_3 only look similarly. We will not describe the language $L(G_f)$ precisely as we are interested only in its intersection with the regular language $\{a^i\#c^j\$b^kd^l \mid i, j, k, l \geq 0\}$.

Clearly the first derivation step in every derivation of G_f is $S \Rightarrow ACBD$. In the word $ACBD$ the B can not be rewritten because of the presence of A. Thus we can apply only $A \to aA_1$. Now A_1 can not be rewritten because of B, so we apply $B \to bB_1$. Thus we have the sub-derivation

$$ACBD \Rightarrow^* aA_1CbB_1D.$$

Now we can apply only a rule that rewrites A_1. If we apply the rule $A_1 \to \#A_1AB$, then we get the sentential form $a\#A_1ABCbB_1D$. In this case no rule from R_2 can be applied (because of the common presence of nonterminals A, B, A_1, B_1). Therefore no word from Σ^* can be derived in this way. Thus we chose the rule $A_1 \to A$, then the rule $B_1 \to B$ and get the following sub-derivation:

$$aA_1CbB_1D \Rightarrow^* aACbBD.$$

Repeating the above applied rules, G_f can increase the number of $a's$ and $b's$ simultaneously.

Continuing our derivation, we apply the rule $A \to aA_1$ (note that $A \to \varepsilon$ could be applied also) and then the rule $B \to \varepsilon$. In this way B_1 did not appear in the sentential form, thus we can apply now the rule $A_1 \to \#A_1AB$ getting the following derivation:

$$S \Rightarrow^* aACbBD \Rightarrow^* aaA_1CbD \Rightarrow aa\#A_1ABCbD.$$

Now we have the situation when the number of $a's$ before the symbol $\#$ still can be increased, but the number of $b's$ in between the nonterminals C and D is fixed.

Next we continue by applying the rules $A \to \varepsilon$, $B \to \varepsilon$, $A_1 \to A$, $A \to \varepsilon$ that yields the following derivation:

$$S \Rightarrow^* aACbBD \Rightarrow^* aaA_1CbD \Rightarrow aa\#A_1ABCbD \Rightarrow^* aa\#CbD.$$

Finally, applying rules to $aa\#CbD$ from R_3 similarly, we can get, for example, the following derivation:

$$S \Rightarrow^* aa\#CbD \Rightarrow^* aa\#ccc\$bdd.$$

Next we show that $L(G_f) \notin \mathcal{L}(\text{CF})$. Assume on the contrary that $L(G_f) \in \mathcal{L}(\text{CF})$. Note that there are words in $L(G_f)$ which do not contain the symbol $\#$ or \$, and words which contain more than one occurrence of $\#$ or \$. To eliminate these words from $L(G_f)$ we define the regular language $L = \{a^i\#c^j\$b^kd^l \mid i, j, k, l \geq 0\}$. Clearly,

$$L' := L(G_f) \cap L = \{a^i\#c^j\$b^kd^l \mid i > k, j > l\}.$$

It is known that context-free languages are closed under the intersection with regular languages (cf. e.g. Theorem 7.27 of [4]), thus $L' \in \mathcal{L}(\text{CF})$.

Next we define a homomorphism that erases the symbols $\#$ and \$ from the words of L'. Let $h : \Sigma^* \to \{a, b, c, d\}^*$ be the homomorphism such that $h(x) = x$ for every $x \in \{a, b, c, d\}$, and $h(\#) = h(\$) = \varepsilon$. It is easy to see that

$$L'' := h(L') = \{a^ic^jb^kd^l \mid i > k, j > l\}.$$

As context-free languages are also closed under homomorphisms (cf. e.g. Theorem 7.24 of [4]), we get that L'' is in $\mathcal{L}(\text{CF})$. On the other hand, it is well known that L'' is not a context-free language. This is a contradiction that finishes the proof of the theorem. □

Next we show a similar result concerning the language class $\mathcal{L}$(CFGprr).

Theorem 2. $\mathcal{L}(\text{CF}) \subset \mathcal{L}(\text{CFGprr})$.

Proof. It is trivial that $\mathcal{L}(\text{CF}) \subseteq \mathcal{L}(\text{CFGprr})$, thus it is enough to construct a CFGprr G_p such that $L(G_p) \notin \mathcal{L}(\text{CF})$.

Let $G_p = (V, \Sigma, R, S, f)$ be a CFGprr where

- $V = \{S, A, A_1, B, B_1, C, C_1, D, D_1\}$,
- $\Sigma = \{a, b, c, d\}$,
- $f(S) = S$, $f(A) = A$, $f(B) = A_1$, $f(A_1) = B_1$, $f(B_1) = A$, $f(C) = C$, $f(D) = C_1$, $f(C_1) = D_1$, $f(D_1) = C$.

Again, R is divided into three parts: $R = R_1 \cup R_2 \cup R_3$ where

- $R_1 = \{S \to ACBD\}$,
- $R_2 = \{A \to aA_1 \mid \varepsilon,\ B \to bB_1,\ A_1 \to A,\ B_1 \to B \mid \varepsilon\}$,
- $R_3 = \{C \to cC_1 \mid \varepsilon,\ D \to dD_1,\ C_1 \to C,\ D_1 \to D \mid \varepsilon\}$.

First we note that here, similarly as in the case of G_f, the application of the rules in R_2 is independent from the application of the rules in R_3. The idea behind the construction of G_p is the following. The rules with A as the left-hand side can be applied always when A is in the sentential form. However, if both A and B are in the sentential form, then applying $A \to \varepsilon$ stops the derivation as this way A_1 can not be introduced and so B can not be rewritten anymore. The same applies to the nonterminals C and D. Thus G_p produces a's and b's simultaneously, and eventually applies the rule $B_1 \to \varepsilon$. After this several a's still can be produced before the application of the rule $A \to \varepsilon$. The derivation goes in the same way concerning the nonterminals C and D.

Now, it can be seen that the language defined by G_p is the language

$$\{a^ic^jb^kd^l \mid i \geq k \geq 1, j \geq l \geq 1\}.$$

Clearly this language is not in $\mathcal{L}$(CF). □

4 CONCLUSION

We have considered two simple types of context-free grammars with rewriting restrictions. These systems involve a very simple regulation above the application of the rules. Despite of this, they can still generate languages outside of the $\mathcal{L}$(CF) class. This also means that the sets of derivation trees of these systems are not recognizable by finite tree automata. It follows that even those powerful tree transducers (macro and pebble tree transducers, for example) that induce tree transformations with recognizable domains can not check whether the input tree is a valid derivation tree of a CFGfrr or a CFGprr.

Concerning the future work, the characterization of the $\mathcal{L}$(CFGfrr) class is still missing.

5 ACKNOWLEDGEMENTS

The author is grateful for the useful comments and suggestions of the referees, which led to the improvement of this paper.

REFERENCES

[1] J. Dassow, T. Masopust, On Restricted Context-Free Grammars, in: Y. Gao, H. Lu, S. Seki, S. Yu (Eds.), *Proceedings of DLT 2010*, pp. 434–435, LNCS 6224, 2010.

[2] J. Dassow, T. Masopust, On restricted context-free grammars, *unpublished manuscript*.

[3] J. Dassow, Gh. Paun, *Regulated Rewriting in Formal Language Theory*, Springer-Verlag, New York, 1990.

[4] J.E. Hopcroft, R. Motwani, J.D. Ullman, *Introduction to automata theory, languages, and computation*, 2nd edition, Addison Wesley, 2000.

[5] T. Masopust, Simple restriction in context-free rewriting, *Journal of Computer and System Sciences* 76, pp. 837–846, 2010.

[6] M. Sipser, *Introduction to the Theory of Computation*, 2nd edition, Thomson, Boston, 2006.

Indeks avtorjev / Author index

Uredniki:

prof. dr. Marko Bohanec
prof. dr. Matjaž Gams

prof. dr. Vladislav Rajkovič
prof. dr. Tanja Urbančič
Mojca Bernik

dr. Dunja Mladenić
Marko Grobelnik

prof. dr. Marjan Heričko

dr. Urban Kordeš
prof. dr. Olga Markič

prof. dr. Jadran Lenarčič
dr. Leon Žlajpah
dr. Andrej Gams

dr. Andrej Brodnik