



*University of Novi Sad
Technical faculty "Mihajlo Pupin"
Zrenjanin*



**PROCEEDINGS OF
INTERNATIONAL CONFERENCE
ON APPLIED INTERNET AND
INFORMATION TECHNOLOGIES**

Serbia, Zrenjanin, October 25, 2013



**UNIVERSITY OF NOVI SAD
TECHNICAL FACULTY "MIHAJLO PUPIN"
ZRENJANIN, REPUBLIC OF SERBIA**



International Conference

**International Conference on
Applied Internet and Information Technologies
ICAIIIT 2013**

P R O C E E D I N G S

**Zrenjanin
October 25, 2013**

Organizer:

University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin,
Republic of Serbia

Publisher:

University of Novi Sad, Technical Faculty "Mihajlo Pupin"
Djуре Djakovica bb, Zrenjanin, Republic of Serbia

For publisher:

Milan Pavlović, Ph. D, Full Professor, Dean of the Technical Faculty "Mihajlo Pupin"

Technical preparation and design:

Brтка Vladimir, Lacmanović Dejan, Zdravko Ivanković, Ljubica Kazi

Cover design:

Ognjenović Višnja

Printed by:

Printing office Dignet, Zrenjanin, Republic of Serbia

CIP - Каталогизација у публикацији
Библиотека Матице српске, Нови Сад

004(082)

INTERNATIONAL Conference on Applied Internet and Information Technologies (2 ; 2013 ; Zrenjanin)

Proceedings [Elektronski izvor] / [2nd] International Conference on Applied Internet and Information Technologies ICAИТ 2013, Zrenjanin, October 25, 2013 ; [organizer] Technical Faculty "Mihajlo Pupin", Zrenjanin. - Zrenjanin : Technical Faculty "Mihajlo Pupin", 2013. - 1 elektronski optički disk (DVD) : tekst, slika ; 12 cm

Tiraž 250. - Bibliografija uz svaki rad.

ISBN 978-86-7672-211-2

1. Technical Faculty "Mihajlo Pupin" (Zrenjanin). - I.
ICAИТ (2 ; 2013 ; Zrenjanin) v. International Conference on Applied Internet and Information Technologies (2 ; 2013 ; Zrenjanin)

a) Информационе технологије - Зборници

COBISS.SR-ID 281228551

Circulation: 250

By the resolution no. 114-451-3096/2012-03, Autonomous Province of Vojvodina Provincial Secretariat For Science and Technological Development donated financial means for printing this Conference Proceedings.

The Conference is supported by the Provincial Secretariat for Science and Technological Development, Autonomous Province of Vojvodina, Republic of Serbia; Regional Chamber of Commerce Zrenjanin; BIZ, Business Incubator Zrenjanin.

International Scientific Committee

Mirjana Pejić Bach, University of Zagreb, Croatia
Evgeny Cherkashin, Institute of System Dynamic and Control Theory SB RAS, Russia
Madhusudan Bhatt, R.D. National College, University of Mumbai, India
Amar Kansara, Parth Systems LTD, Navsari, Gujarat, India
Narendra Chotaliya, H. & H.B. Kotak Institute of Science, Rajkot, Gujarat, India
Christina Ofelia Stanciu, Tibiscus University, Faculty of Economics, Timisoara, Romania
Zeljko Jungić, ETF, University of Banja Luka, Bosnia and Hercegovina
Saso Tamazič, Univerisity of Ljubljana, Slovenia
Marijana Brtko, Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, São Paulo Brazil
Zoran Cosic, Statheros, Split, Croatia
Istvan Matijevics, Institute of Informatics, University of Szeged, Hungary
Slobodan Lubura, ETF, University of East Sarajevo, Bosnia and Hercegovina
Zlatanovski Mita, Ss. Cyril and Methodius University in Skopje, Republic of Macedonia
Josimovski Saša, Ss. Cyril and Methodius University in Skopje, Republic of Macedonia
Edit Boral, ASA College, New York, NY, USA
Dana Petcu, West University of Timisoara, Romania
Marius Marcu, "Politehnica" University of Timisoara, Romania
Zora Konjović, Faculty of technical sciences, Novi Sad, Serbia
Siniša Nešković, FON, University of Belgrade, Serbia
Nataša Gospić, Faculty of transport and traffic engineering, Belgrade, Serbia
Željko Trpovski, Faculty of technical Sciences, Novi Sad, Serbia
Branimir Đorđević, Megatrend University, Belgrade, Serbia
Slobodan Jovanović, Faculty of Information Technology, Belgrade, Serbia
Zlatko Čović, Subotica Tech / Department of Informatics, Subotica, Serbia
Diana Gligorijević, Telegroup, Serbia
Borislav Odadžić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Miodrag Ivković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Biljana Radulović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Ivana Berković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Vladimir Brtko, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Branko Markoski, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dalibor Dobrilović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Željko Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dejan Lacmanovic, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Zdravko Ivankovic, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Ljubica Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia

Organizing Committee

Ph.D Borislav Odadžić, president, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Ph.D dr Miodrag Ivković, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Ph.D Vladimir Brtka, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Ph.D Biljana Radulović, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Ph.D Ivana Berković, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Ph.D Branko Markoski, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Ph.D Željko Stojanov, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Ph.D Dalibor Dobrilović, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Mr Dejan Lacmanović, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Mr Ljubica Kazi, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

MSc Zdravko Ivanković, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Olivera Dobrosavljev, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

Vesna Keljački, Technical Faculty “Mihajlo Pupin”, University of Novi Sad, Republic of Serbia

INTRODUCTION

Information Technologies and Internet as a part of Computer science creates new approaches and perspectives, new models and numerous services, which opens up and makes use of the world of information and symbolized knowledge. Advances in Information technology, including the Internet, have dramatically changed the way we collect and use public, business and personal information.

The 2nd **International Conference on Applied Internet and Information Technologies** is an international refereed conference dedicated to the advancement of the theory and practical implementation of both knowledge of Information Technologies and Internet and knowledge of the special area of their application.

The objectives of the **International conference on Applied Internet and Information Technologies** are aligned with the goal of regional economic development. The conference focus is to facilitate implementation of Internet and Information Technologies in all areas of human activities. The conference provides forum for discussion and exchange of experiences between people from government, state agencies, universities and research institutions, and practitioners from industry.

The key Conference topic covers a broad range of different related issues from a technical and methodological point of view, and deals with the analysis, the design and realization of information systems as well as their adjustment to the respective operating conditions. This includes software, its creation and applications, organizational structures and hardware, different system security aspects to protocol and application specific problems. The Conference Topics are:

1. Information systems
2. Communications and computer networks
3. Data and system security
4. Embedded systems and robotics
5. Reliability and maintenance
6. Process assessment and improvement
7. Software engineering and applications
8. Computer graphics
9. ICT Support for decision-making
10. Management in IT
11. E-commerce
12. Internet marketing
13. Customer Relationship Management
14. Business intelligence
15. ICT practice and experience

The Conference Organizing Committee would like to thank for the support and cooperation to the Regional Chamber of Commerce Zrenjanin, BIZ – Business Incubator Zrenjanin, University of Novi Sad and Provincial Department of Science and Technological Development.

Special thanks to the authors of papers, reviewers and participants in the Conference who have contributed to its successful realization.

**President of the Organizing Committee
Ph.D Borislav Odadžić**

Zrenjanin, October 2013

We are very grateful to

*Provincial Department of Science and Technological Development,
Autonomous Province of Vojvodina,
Republic of Serbia*

*Ministry of Education, Science and Technological Development,
Republic of Serbia*

*for financial support in preparing the Conference Proceedings and organizing the
Conference.*

ORGANIZATOR WITH PARTNERS:

**Technical Faculty "Mihajlo Pupin" Zrenjanin
University of Novi Sad
Zrenjanin, SERBIA
<http://www.tfzr.uns.ac.rs/>**

**Faculty of computer science
Irkutsk State Technical University
Irkutsk, RUSSIA
<http://www.istu.edu/structure/57/9518/1801/>**

**Faculty of Technical Sciences
University of St. Clement Ohridski
Bitola, MACEDONIA
<http://www.tfb.edu.mk/>**

**Faculty of Economics
Tibiscus University of Timisoara
Timisoara, ROMANIA
<http://www.fse.tibiscus.ro/>**

CONTENT:

"The Art of Modeling": How Can AIIT Be Modeled? Pece Mitrevski	1
Comparison of Approaches to Energy Efficient Wireless Networks Borislav Odadžić, Dragan Odadžić	7
Method for Construction of all Bent Functions based on concatenating Functions of n-1 Variables Dragan Lambić, Miroslav Lambić	13
Information Technology as a support of energy efficiency monitoring Saša Bošnjak, Zita Bošnjak, Olivera Grljević	17
Business intelligence as a support to marketing analysis and decision-making Ivana Berković, Dušanka Lečić, Milan Ceković	22
Building Ontologies in Protégé Zoltan Kazi, Biljana Radulović, Ljubica Kazi	26
Web Integration of REST Enabled Wireless Sensor Networks for Fire Detection Vladimir Vujović, Mirjana Maksimović, Dijana Kosmajac, Vladimir Milošević, Branko Perišić	30
IT Higher Education In India Naisargee Chotaliya, Ljubica Kazi, Narendra Chotaliya	36
Comparison of ICT usage and market trends in Romania and Serbia Mira Sisak, Dalibor Dobrilović, Robert Molnar	41
Gap Between Service Requestor and Service Builder Aleksandar Bulajić, Radoslav Stojić, Samuel Sambasivam	47
Analysis of Serbian Malware Statistics Petar Ćisar, Sanja Maravić-Ćisar, Branko Markoski, Miodrag Ivković, Dragica Radosav	53
Tools for WLAN IEEE 802.11 security assessment Stefan Jäger, Dalibor Dobrilović	56
The benefits of standardization for business intelligence tools Margarita Janeska, Dejan Zdraveski, Suzana Taleska	63
Multi-Objective Automatic Calibration of the Distributed Hydrological Model Milan Stojković, Nikola Milivojević, Vladimir Milivojević, Vukašin Ćirović	67
Machine Learning Approach for Performance Based Cloud Pricing Model Monika Simjanoska, Saško Ristov, Marjan Gusev	74
Learning approaches based on information and communication technologies Jovan Savičić	79
Data gathering from websites Zdravko Ivanković, Branko Markoski, Dejan Savičević	84
Data retrieval from database Zdravko Ivanković, Dragica Radosav, Dejan Lacmanović	88
Can Cloud Virtual Environment Achieve Better Performance? Saško Ristov, Velkoski Goran, Marjan Gusev	92
Triangulation of convex polygon: Parallel Programming approach Selver Pepić, Borislav Odadžić, Stanimir Čajetinac	98
The role of visualization in the Building Management System Vladimir Vujović, Ines Perišić, Mirjana Maksimović, Igor Kekeljević	102
Predicting the EUR/RSD exchange rate with wavelet and neural network Jovana Božić, Đorđe Babić	108

Information Dispersal for Big Data Storage Miloš Stević, Radoje Cvejić	113
TYPESCRIPT, a new OpenSource way to program JavaScript Miloš Stević, Radoje Cvejić	117
Tag-Based Collaborative Filtering in e-learning systems Aleksandar Kotevski, Cveta Martinovska-Bande, Radmila Kotevska	122
Game development in java Netbeans platform – Sudoku application Nemanja Bilinac, Miroslav Eremić, Radovan Adamov, Dalibor Dobrilović, Vladimir Brtko	126
Cognitive mapping in robotics using genetic algorithms Ramona Markoska, Mitko Kostov, Mile Petkovski, Aleksandar Markoski	131
Web service and mobile application for exam registration Petar Bjeljac, Dijana Kosmajac, Vladimir Vujović	135
Concordances based linguistic search algorithm applied on Serbian - Slavonic language Dejan Lacmanović, Branko Markoski, Izabela Lacmanović, Zdravko Ivanković, Predrag Pecev	138
Plume Boundaries Extraction by Multiresolution and Least Squares Approximation Mitko Kostov, Aleksandar Markoski, Mile Petkovski, Ramona Markoska	142
Fuzzy Screening in Cryptography Vladimir Brtko, Eleonora Brtko, Višnja Ognjenović	146
Development of an interactive educational game for mobile phones Zlatko Čović, Suzana Palfi, Andor Nagl, Andor Sipos	150
Product packaging design with Harmony Nada Jovanović, Višnja Ognjenović, Ivana Berković, Vesna Jevtić	154
Discretization influence on data reduction Višnja Ognjenović, Vladimir Brtko, Ivana Berković	158
Tracking Failures of Auxiliary Mechanization in an Open-Pit Mine Sonja Dimitrijević, Snežana Pantelić, Gradimir Ivanović, Dragana Bogojević, Radiša Đurić, Dragan Stević	162
Risk Assessment Concept in the New Approach Directives Ana Bašić, Igor Lavrnić, Dejan Viduka, Boban Panajotović	168
The Application of the Polynomials in Cryptography Marijana Brtko, Jelena Danikov, Biljana Goševski, Vladimir Brtko	174
Multi-Criteria Analysis of Data for Ranking in Construction of Regional Irrigation System in the Republic of Serbia Tihomir Zoranović, Svetlana Potkonjak, Ivana Berković	177
Review of the CFD Software Packages Milena Todorović, Dragan Pavlović	181
On the Performance of Scalable Video Coding for Use in P2P Live Video Streaming Zoran Kotevski, Pece Mitrevski	187
Improving Performance of e-Commerce Systems by Vertical Scaling Ilija Hristoski, Pece Mitrovski	191
Conceptual SWOT Analysis on eCommerce in Terms of Services Marketing Daniel Kysilka	197
Drools Rule Language – A new approach to building business layers Predrag Pecev, Dragana Glušac, Sanja Maravić-Čisar, Dejan Lacmanović, Nedžad Osmankač	201
Using Linear Regression for Estimating Useful Energy for Solar Collectors Based on Real Project Data and Data Available on Internet Kristijan Vujičin, Željko Stojanov	207
Predicting the outcome of disease in patients with hepatitis using machine learning algorithms Jasmina Novaković, Alempije Veljović	211
Implementation of Data Security Measures in Information Systems Emir Skejić, Osman Džindo, Suad Kasapović	216
Rendering 3D Graphics on Android Operating System using OpenGL ES Emir Skejić, Samer Abud	219
Reflections on Some Validity and Ethical Issues in Mixed Methods Research on Investigating English Language Usage at IT Departments in Serbia Tijana Dabić, Željko Stojanov	225

Automatic baum tests' classification Florentina Anica Pintea, Dan Lacrama, Corina Musuroi, Tiberiu Karnyanszky	230
Calculation of the Quality and (un)availability of the RR link Suad Kasapović, Emir Skejić, Amir Hadžimehmedović	234
The Role of Human Resource Information Systems in EU based on CRANET research Agneš Slavić, Nemanja Berber	238
Security Aspects Of The Social Network Facebook: Some Empirical Results Andreja Samčović, Svetlana Čičević	244
Intelligent Organizations Instead of Rigid Organization Forms Deniz Ahmetagić, Jelena Rodić, Boris Saulić	248
A document content logical layer induction on the base of ontologies and processing changes Evgeny Cherkashin, Polina Belykh, Danil Annenkov, Kristina Paskal	252
LiveGraphics3D Potential Applicability in Primary School Geometry Dinu Dragan, Dragan Ivetic, Natalia Dragan	258
IIS Based Remote Monitoring Of Distributed Technical Systems In Real Time Slobodan Janković, Dragan Kleut, Vladimir Šinik	264
An Approach to Developing Information Systems with Service Orientation using Form Types Marko Knežević, Salaheddin Elheshk, Vladimir Ivančević, Ivan Luković	270
Measuring the performance of eXtremeDB solutions in gesture recognition systems Veljko Petrović, Dragan Ivetic	275
Promoting Robotics Education and Curriculum Edit Boral, Ivana Berković	280
Refine Edge method – analysis of parameters for hair selection Marko Kresojević, Dragan Mijajlović, Višnja Ognjenović, Ivana Berković	284
Decision support system for management of the forest resources Evgeny Cherkashin, Alexander Larionov, Anastasia Popova, Igor Vladimirov	288
Identification and Evaluation of Pertinent Parameters used for Cost-Modeling of a Wide Area Network Basri Ahmed, Pece Mitrevski	294
IT jobs market in Serbia – a preliminary analysis Ljubica Kazi, Biljana Radulović, Miodrag Ivković, Madhusudan Bhatt, Ofelia Stanciu	300
Decision making on using Internet for WAN platform: the case of state-owned banks in countries in transition Asmir Handžić, Dragica Radosav	305
Flow indicator broadcasting time TV show - as a mandatory part of the digital television Bratislav Blagojević	310
Storage systems: Comparing different MySQL types Selver Pepić, Borislav Odadžić, Stanimir Čajetinac	313
Controlling computer games through web camera with motion detection Dimitrija Angelkov, Cveta Martinovska-Bande	317
Analyzing Web Server Access Log Files Using Data Mining Techniques Marjan Velkoski, Cveta Martinovska-Bande	321
Protecting Critical Information Infrastructures by Increasing its Resilience Goran Murić, Nataša Gospić, Milica Šelmić	327
Integrating RFID-Based Classroom Management System into Quality Assurance System Danijel Mijić, Ognjen Bjelica	331
Technical and regulatory aspects of vectoring deployment Sanja Vukčević-Vajs, Stefanović Aleksandra, Cvetković Tatjana	336
Android Application for Data Acquisition Jelena Tucakov, Srđan Popov, Jovana Simić	341
Semantic Web recommender system for e-learning materials Milica Ćirić, Aleksandar Stanimirović, Leonid Stoimenov	344
Evaluation of Mobile Touch-Screen Devices as Media for Reaction Time Measurement Svetlana Čičević, Milkica Nešić, Andreja Samčović, Aleksandar Trifunović	350

Automated Reasoning System Based on Linguistic Variables Vladimir Brtka, Aleksandar Stojkov, Eleonora Brtka, Ivana Berković	356
Basic English Acronyms For Information Technology Students Erika Tobolka	361
PACS systems based on the Web Ivan Tasić, Dragana Glušac, Jelena Jankov, Dajana Tubić	365
Ontology driven decision support system for scoring clients in government credit funds Laszlo Ratgeber, Saša Arsovski, Petar Čisar, Zdravko Ivanković, Predrag Pecev	369
Brute Force attacks on web applications Branko Markoski, Predrag Pecev, Saša Arsovski, Miodrag Šešlija, Bojana Gligorović	374
QR Codes and its applications Miodrag Šešlija, Branko Markoski, Predrag Pecev	379
Software support to fashion design Niyazi Baltali, Ljubica Kazi	383
The potentials of corporate blogging Ljubinka Manovska, Antonio Stamatovski, Bojana Gligorović, Predrag Pecev, Dušanka Milanov	386
HMM Optimization Based On Genetic Algorithm In Speech Recognition: A review Ivan Filipović, Miljan Vučetić	390
Biological modeling of software development dynamics Valentina Paunović	395
The application of Customer Relationship Management in customer retention and relationship development Milan Vujašanin	401
Application of multi linked lists technique for the enhancement of traditional access to the data Đorđe Stojisavljević, Eleonora Brtka	403
Review of group buying websites in Serbia Jelena Rodić, Deniz Ahmetagić	407
Decision support system for mechanical engineering Nataša Glišović, Marija Milojević	413
NoSQL databases – example of use in a Lost and found website Petar Bjeljic, Igor Zečević, Ines Perišić	417
Analyzing the impact of administrative and demographic data on students' performance Snježana Milinković, Mirjana Maksimović	421
Benefits of establishing project management office in an IT company Srđan Grbavac	426
The Concepts of private cloud computing solutions in the public sector Jovan Ivković	432
Advanced programming techniques for data validation in Excel Đorđe Stojisavljević	438
Heron web data mining system Jasmin Pavlović, Rade Milović, Atila Vaštag, Katarina Zorić, Zdravko Ivanković	441
Persons with Disabilities Evacuation – Pathfinder Application Jovana Simić, Tanja Novaković, Nenad Duraković, Gordana Mijatov, Ljiljana Popović, Maja Sremački, Srđan Popov	446
E- commerce and the importance of electronic data interchange (EDI) Milica Stanković	450
The Implications of Adopting E-Commerce Technology for Rural Business in Serbia Boris Saulić, Deniz Ahmetagić	454
The role of Internet marketing in the creation of product and company image Jasmina Markov, Biljana Stankov	458
Wrappers methods and supervised learning algorithms on the example of diagnosis Parkinson's disease Jasmina Novaković	464
A methodological approach to software development process David Maravić, Nemanja Tešić, Eleonora Brtka	469

Fuzzy classification of knowledge of experts to assess the quality of machine tools Sophia Sosinskaya, Elena Kopylova	473
Visualization of 3D structural analysis data Aleksandar Borković	477
Importance of UML in Modeling as part of information systems' development Sofija Krneta	481
Terminal for Remote Sensing in Tax Administration Darko Marjanović	486

Analyzing Web Server Access Log Files Using Data Mining Techniques

Marjan Velkoski and Cveta Martinovska Bande

University Goce Delcev, Computer Science Faculty, Stip, Macedonia

marjanvel@gmail.com, cveta.martinovska@ugd.edu.mk

Abstract - Nowadays web is not only considered as a network for acquiring data, buying products and obtaining services but as a social environment for interaction and information sharing. As the number of web sites continues to grow it becomes more difficult for users to find and extract information. As a solution to that problem, during the last decade, web mining is used to evaluate the web sites, to personalize the information that is displayed to a user or set of users or to adapt the indexing structure of a web site to meet the needs of the users. In this work we describe a methodology for web usage mining that enables discovering user access patterns. Particularly we are interested whether the topology of the web site matches the desires of the users. Data collections that are used for analysis and interpretation of user viewing patterns are taken from the web server log files. Data mining techniques, such as classification, clustering and association rules are applied on preprocessed data. The intent of this research is to propose techniques for improvement of user perception and interaction with a web site.

I. INTRODUCTION

During the last decade web is not only considered as a network for acquiring data, buying products and obtaining services but as a social environment for interaction and sharing information [1]. Web-based data mining can be used for knowledge discovery in recommendation engines, to personalize the Web pages displayed to set of users, for understanding communities or modeling user search [2].

Many works describe different implementations of web mining techniques with the intent to improve user interaction with a web site. For example, Perkowitz and Etzioni propose automatic adaptation of the indexing structure of a web site [3], Spiliopoulou describe a tool Web Log Miner for evaluation whether the expected navigation patterns between pages are met by the majority of the visitors [4], Mobaster et al. [5] describe a tool Web Personalizer for creating usage profiles using association rules and clustering.

Several commercial and free web server log analyzers are available which produce statistical data, like the number of visitors accessing the site, the browsers they use, the length of their sessions, pages with maximal hits, errors that occur while accessing the site, etc. Goel and Jha [6] provide a comparative study of several log analyzer tools. These summary statistics of web site activity can serve as additional data for discovering patterns in web data.

In this work we propose a methodology for web usage mining with the intent to increase the web server efficiency. Data collections that are used for analysis and interpretations of user viewing patterns are taken from web server log files of the Secretariat for European Affairs (SEA) for the process of integration of the Republic of Macedonia in the European Union and are obtained during the user web-based sessions.

Particularly we are interested to discover user access patterns and whether the topology of the web site matches the desires of the users and based on the results we plan to adapt the link structure to better meet the needs of the users. Experimental work is accomplished using WEKA [7].

II. WEB MINING

The appearance of the www service caused a need for analysts to aim their attention towards extracting useful information and knowledge using the techniques of data mining. Web mining represents the use of data mining techniques to extract knowledge from web data including web documents, hyperlinks between the documents, the use of web site logs and similar.

Figure 1 depicts the steps in the process of web mining, starting from preprocessing to identification of

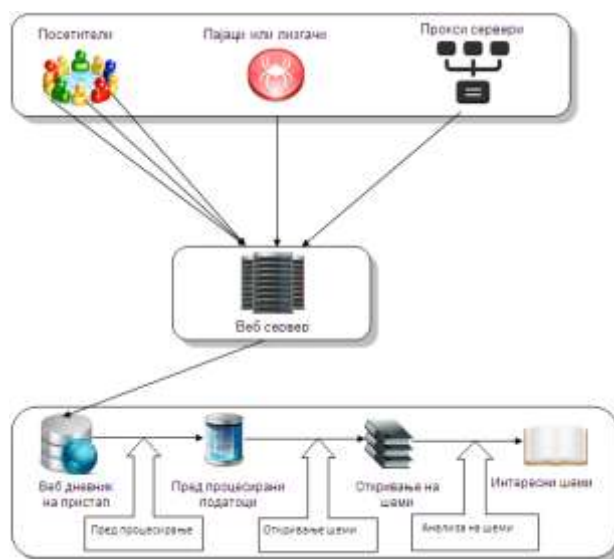


Figure 1. Steps of web mining process

useful patterns. The phase of preprocessing consists of

cleaning the data and user and session identification. The second phase, discovering patterns, involves algorithms and techniques of data mining. The last phase is analysis of the discovered patterns and evaluation of user interests.

Based on the primary type of data which are used in the process of data mining, the web mining can be categorized into three types: structure mining, content mining and web usage mining, depending on which part of the web is researched [8].

A. Structure mining

The goal of web structure mining is to discover useful knowledge from the hyperlinks which represent the structure of the web, i.e. a classification of the web pages can be made based on their organization. Structure mining can be used to categorize the web pages based on hyperlinks or document-structure. The content of a web page can also be organized in a tree-structured form, based on different HTML and XML tags on the page.

The structure of a web page can be represented on a typical web graph which consists of web pages as nodes and hyperlinks as edges which connect the connected pages.

B. Content mining

Content mining is a process of extraction of useful information from the content of web documents by application of the data mining algorithms. The contents of web documents are *true* data for which the web page is designed to transfer to the users. The pages can be of different types of data, so it results in existence of different categories of content mining. It is connected to data mining because many techniques of data mining can be used in content mining but at the same time it is different because web data are mainly semi-structured and/or unstructured while data mining mainly addresses structured data. It is also connected to text mining because a big part of the web content is text, but also different because the web is semi-structured while text mining is focused on unstructured texts.

C. Web usage mining

The third kind of web mining is user access mining. The user access logs enable monitoring the user's activity with the web site and improving the structure of the web site. If established by analysis that the visitors stay a long time it is a sufficient indicator of the need for restructuring of the web page in order to help the visitors to reach the wanted information quickly. By user mining based on information about user preferences, interesting content can be offered. To achieve this goal it is recommended to use adaptive web sites which use information about the access schemes of the user in order to improve their organization and presentation [5].

III. DATA PREPARATION

A. Filtering web access logs

As mentioned before, all records available as a result of user web access and browsing are stored in web

server log files generated by Microsoft IIS 6.0. Server log files provide information in Extended Log Format because web site of SEA is developed and hosted using Microsoft Windows 2003 platform. As Figure 2 shows the fields associated with the extended log format are date, time, request, host address, browser type, referring page, status and bytes. Data preprocessing is performed using Perl scripts from WUMprep tool [9] which is part of open source project HypKnowSys.

For data mining relevant log file fields are those fields that enable determining the sequence of clickstreams followed by each user as they navigate through the web site. It is important to create a session file, which contains sets of page-views requested by a single user from a single web server. A single page-view consists of one or more page files and is marked with a unique URI.

In the preprocessing phase the first step is elimination of irrelevant entries. Filtering of unnecessary elements, such as graphics or sound can be accomplished by checking the suffixes of URL names. Records of images and videos, records of servers inter-mediators, records with failed requests (non-existing pages, server failures) except requests with code 2/x/x and double records are not appropriate for the experiments and they are removed using the script LogFilter.pl.

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2012-11-05 07:57:01
#Fields: date time cs-method cs-uri-stem cs-uri-query cs-username c-ip cs-
version cs(User-Agent) cs(Referer) sc-status sc-bytes
2012-12-21 20:53:32 GET
/Content/Publications/Documents/Dogovor+od+Lisabon(1).pdf - - 77.19.26.111
HTTP/1.1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/537.11+
2012-12-21 20:57:17 GET /default.aspx ContentID=47 - 173.195.114.119 HTTP/1.1
"Mozilla/5.0+(compatible;+AhrefsBot/4.0;++http://ahrefs.com/robot/)" - 200
27256
2012-12-21 20:58:09 GET
/Content/Publications/Documents/Dogovor+od+Lisabon(1).pdf - - 89.205.15.152
HTTP/1.1
Mozilla/5.0+(Windows+NT+6.1)+AppleWebKit/537.11+(KHTML,+like+Gecko)+Chrome/23
.0.1271.97+Safari/537.11
http://www.pfk.uklo.edu.mk/index1.php?page=recutatatisgrad= 200 191331
```

Figure 2. A sample of web log in extended log format

Similarly, the records made by crawlers, spiders, indexers and other robots have to be discarded from the web logs. Web robots access the file "robots.txt" for permissions created by administrators which helps in the process of their identification. The script RemoveRobots.pl is used for removal of web robot accesses.

B. Session Identification

Creating a session file is not a simple task. Some problems related to identification of sessions are discussed in [10, 11]. To extract the individual server session one has to identify each user having in mind that several users may be accessing the site from the same host. Host address has to be combined with the referring page to distinguish one user session from another.

Several authors describe ways to identify sessions, as for example, using reference length [12] or maximal forward references [13]. In this work is used method named “time window”. When the period between two accesses from the single user is greater than a certain threshold then these accesses are considered as different sessions. The time period of 30 minutes is considered as appropriate threshold to identify the sessions. The session file in this work is created using the script Sessionize.pl.

Sessions that have at least 5 visited pages are considered as useful for data mining in this work.

C. Mining the data

Once the session file is created different techniques can be applied such as association rules or clustering methods. Several recent works describe web log analysis using data mining techniques [14, 15, 16].

Association rules give the instances (pages) that appear together in a single session record. If the direct link do not exist between these pages the rule may warrant modifying the indexing structure of a web by placing direct links between the pages.

Unsupervised clustering can be used to form clusters of similar instances in file sessions.

IV. EXPERIMENTAL RESULTS

The aim of this work is to find certain interesting patterns for the visitors of the web site of the Secretariat for European Affairs for the process of integration of the Republic of Macedonia in the European Union (SEA).

Figure 3 gives a hierarchical overview of the performed tasks for discovery of models in these experiments. Moreover, this figure shows the organization of the experiments. All the experiments are conducted using user access logs from December 2012. In total 3 experiments were conducted:

- Experiment 1: MKVsOutsideMK (visitors from Macedonia and visitors outside of Macedonia),
- Experiment 2: SEPVsOutsideSEP (visitors from SEA and visitors outside SEA),
- Experiment 3: SEPVsOutsideSEPWithinMK (visitors from SEA and visitors outside SEA but from Macedonia).



Figure 3. Organization of the performed experiments

For each experiment 4 data mining techniques are used: classification, association rules, clustering and attribute selection. From the processed data, 4 different groups of instances were separated: First3-Last2, First5-Last5, 10-Most-Frequent-TF and 10-Most-Frequent-Time. The above mentioned techniques are applied to each of these groups to discover potentially interesting patterns.

The experiments are performed using WEKA (Waikato Environment for Knowledge Analysis) software that implements a collection of machine learning algorithms for data mining tasks.

Using WEKA the following results were obtained:

Experiment 1: Visitors from Macedonia most often visit the root page of the web site but they also visit pages directly while the visitors outside Macedonia most often visit specific pages directly. This is most likely because they use search engines.

Experiment 2: Results show that visitors from SEA most often visit the root page, but also they also visit pages directly and this is due to the fact that the SEA employees know the structure of web site well.

Experiment 3: Some of the discovered patterns are in line with those discovered in experiments 1 and 2. The discovered patterns show that visitors outside SEA usually spend less time on the web pages compared to visitors from SEA.

Table 1 shows the data mining techniques used in the first experiment and for which groups of instances significant patterns were discovered.

TABLE 1. EXPERIMENT1: MKVsOUTSIDEMK - SUMMARY OF RESULTS

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2012	First3-Last2	YES
		First5-Last5	YES
		10-Most-Frequent-TF	YES
		10-Most-Frequent-Time	YES
Association Rules	access2012	First3-Last2	YES
		First5-Last5	YES
		10-Most-Frequent-TF	NO
Clustering	access2012	First3-Last2	YES
		First5-Last5	YES
		10-Most-Frequent-Time	NO
Attribute Selection	access2012	First5-Last5	YES
		First3-Last2	YES
		10-Most-Frequent-TF	YES
		10-Most-Frequent-Time	YES

The classification results for the instances of different groups are obtained with OneR and J48 classifiers. These algorithms give sequences of visited pages. According to the obtained results visitors from Macedonia that access the root page as first page, also visit other pages which give information about reports and news from EU, pre-accession support, negotiation processes, the page that contain document register, the translation process, pages giving information about the structure, organization and work of the SEA, etc.

Visitors outside Macedonia which visited the root page also visited the Europe’s Bulletin and the page for pre-accession support.

The samples of 10-Most-Frequent-TF group are formed after selecting 10 most frequently visited pages.

The attribute for the page is T (true) if the page is visited in a particular session or F (false) otherwise.

The group 10-Most-Frequent-Time consists of the same instances as 10-Most-Frequent-TF except that the value of the attribute is time spent on a particular frequently visited page.

Figure 4 shows the decision tree obtained with J48 algorithm for 10-Most-Frequent-Time group.

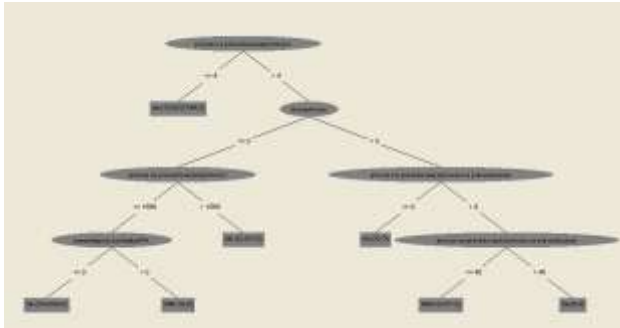


Figure 4. Decision tree obtained with J48 algorithm using 10-Most-Frequent-Time group of instances

Association rules are obtained using Apriori algorithm. As Figure 5 shows 7 rules are discovered for the group First 3-Last 2.

The interpretation of the first and fifth rule is as follows:

- if the first and next-to-last page is the root page then the visitor is from Macedonia.
- if the third visited page is the root page then the first visited page is also the root page.

Apriori
 =====
 Minimum support: 0.1 (54 instances)
 Minimum metric <confidence>: 0.9
 Number of cycles performed: 18

Generated sets of large itemsets:
 Size of set of large itemsets L(1): 4
 Size of set of large itemsets L(2): 5
 Size of set of large itemsets L(3): 2

- Best rules found:
1. F1=Home/home L2=Home/home 93 ==> Country=mk 88 conf:(0.95)
 2. L2=Home/home 107 ==> Country=mk 101 conf:(0.94)
 3. F3=Home/home Country=mk 106 ==> F1=Home/home 99 conf:(0.93)
 4. F1=Home/home F3=Home/home 106 ==> Country=mk 99 conf:(0.93)
 5. F3=Home/home 114 ==> F1=Home/home 106 conf:(0.93)
 6. F3=Home/home 114 ==> Country=mk 106 conf:(0.93)
 7. F1=Home/home 404 ==> Country=mk 367 conf:(0.91)

Figure 5. Partial output of the Apriori algorithm using First 3-Last 2 group of instances

Clustering is performed using EM algorithm. As expected two clusters are obtained: Cluster0 consisting of visitors outside Macedonia, that are 21% of the total number of visitors, and Cluster1 of visitors from Macedonia which are 79% of the total number of visitors.

Cluster1 which is formed of the visitors from Macedonia that start searching from the root page and also visit pages related to the organization and work of the SEA, news and procurements, advertisements and competitions in SEA. The users in this cluster have tendency to visit the following three pages: Home/home, Home/Novosti, Za nas/organizacija and NOK/Tenderi. Figure 6 shows partial output of EM algorithm for the group 10-Most-Frequent-TF.

```
EM
==
Number of clusters: 2
```

Attribute	Cluster	
	0 (0.21)	1 (0.79)
F1		
Home/home	24.1236	381.8764
F2		
za nas/organizacija	2.2247	49.7753
Home/Novosti	1.0803	43.9197
F3		
Home/home	6.5413	109.4587
za nas/organizacija	8.7708	32.2292
L2		
Home/home	6.6491	102.3509
NOK/tenderi	3.1326	21.8674
Home/Novosti	4.0535	20.9465
L1		
Home/home	2.4074	52.5926
Home/Novosti	2.7317	35.2683

```

Clustered Instances
0  113 ( 21%)
1  422 ( 79%)
Log likelihood: -14.16034
Class attribute: Country
Classes to Clusters:
  0  1 <-- assigned to cluster
 15 37 | NMK
 98 385 | mk
Cluster 0 <-- NMK
Cluster 1 <-- mk
Incorrectly clustered instances :      135.0    25.2336 %
```

Figure 6. Partial output of the EM algorithm for the group 10-Most-Frequent-TF

The results show that visitors from Macedonia most often visit the SEA web site to access information about news, procurements, competitions and advertisements in SEA.

The results of Attribute Selection for the group First 3-Last 2 are obtained using “cfssetEval” attribute evaluator with “BestFirst” search method. According to these results attributes F1, F3 and L2 are pointed as the most discriminative and relevant attributes. F1 corresponds to the first visited page in the session, F3 to the third page and L2 corresponds to the next-to-last page in the session.

V. APPLIED WEB LOG ANALYSIS TOOLS

A. Deep Log Analyzer

In order to obtain statistical information about the activity on the web site we used Deep Log Analyzer. The reports generated with this tool contain information about all

accessed resources on the web site, the activity of the visitors and their navigation, web sites through which visitors have come to the analyzed web site, robots which accessed the web site, used search engines and operating systems by the visitors, errors on the web servers, etc.

Figure 7 shows one of the reports obtained by this web log analyzer for December 2012. The reports show the following information: number of total visits: 24428, average number of visits per day 788, average visit duration 22:28 min, top referring website <http://www.google.com> 3804, top search engine Google, Spider requests 21475, most popular browser IE 9.0, most popular OS Win7, most popular entry page /default.aspx, most popular exit page /default.aspx, most popular download /final_europa_A_DO_S.pdf, number of unique visitors 8568, repeat visitors 2320, visitors who visited once 6248, average visits per visitor 2,85 and etc.



Figure 7. Overview report for SEA web site obtained with Deep Log Analyzer

B. Aqunetix

Without adequate safety protection and efficient security management, web sites can be abused for attack of the data integrity of the information systems and network connections. Using the data mining techniques, attacks as well as attack profiles can be discovered.

There are several typical security attacks which show the weaknesses that are abused to perform the attacks, like Denial-of-Service, SQL injection, Cross-Site Scripting and HTTP GET attack.

The following weaknesses on the web site were discovered using the Acunetix WVS (Fig. 8):

(1) high level risk- weaknesses for two SQL Injection attacks and one for a possible attack through which one can access important files and folders which are usually not visible,

(2) medium level risk- a few error messages were found regarding the code of the application,

(3) low level risk – weaknesses were found out of which one regards the HTTP method Options through which the hackers can prepare and advanced attack, four regard the

possibility to discover sensitive folders and two more regarding session cookies, and finally

(4) information level signal – one weakness is found regarding searching of a page which does not exist i.e. the page with error returns information about the server version and the list of available modules.



Figure 8. Results of the scanning with Aqunetix

VI. CONCLUSIONS

Using the data mining techniques interesting access patterns were discovered as well as certain differences in the access patterns of the users from Macedonia, outside Macedonia and the employees of SEA.

Analyzing the web site with Deep Log Analyzer a complete statistics is produced regarding the use of the web site and information of the accessed resources, the activity of the visitors, sliders searching the web site, search engines and operational systems the visitor use, web server errors, etc.

Web log analysis tools do not provide comprehensive analysis and access patterns which can be obtained using data mining techniques.

Future steps in our project related to analyzing web log access files will be:

- analysis of the sessions with less than 5 visited pages, because in reality such visits exist and many interesting access patterns can be discovered,
- applying techniques of exact identification of locations, visitors and sessions with the goal to get meaningful information in the defining of patterns,
- optimization of the web site design based on the statistical data and data obtained using data mining techniques,
- improvement of the security and integrity of the web site, based on the discovered weaknesses.

VII. REFERENCES

- [1] Spiliopoulou, M., B. Mobasher, O. Nasraoui, and O. Zaiane. 2012. Guest editorial: special issue on a decade of mining the Web. Data Mining and Knowledge Discovery, pp. 1_5.
- [2] Nasraoui O, Spiliopoulou M, Zaiane O, Srivastava J, Mobasher B (eds) (2008) 10th international workshop on knowledge discovery

- on the web, WEBKDD'08: 10 years of knowledge discovery on the web. ACM, Las Vegas. In conjunction with the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2008)
- [3] Perkowitz, M., and Etzioni, O. (2000) Adaptive Web Sites. Communications of the ACM, 43, 8, pp.152-158
- [4] Spiliopoulou, M. (2000) Web Usage Mining for Web Site Evaluation. Communications of the ACM, 43, 8, pp.127-134
- [5] Mobaster, B., Cooley, R., and Srivastava, J. (2000) Automatic Personalization Based on Web Usage Mining. Communications of the ACM, 43, 8, pp.142-151
- [6] Goel, N. and Jha, C. H. (2013) Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool. International Journal of Computer Applications. Vol. 62– No.2, pp.29-33.
- [7] <http://www.cs.waikato.ac.nz/ml/weka/index.html>
- [8] R. Kosala and H. Blockeel. Web mining research: A survey. ACM SIGKDD, 2(1):1–15, 2000.
- [9] WUMprep, (Web mining pre-processing), <http://sourceforge.net/projects/hypknowsys/>
- [10] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, 1(2):12-23,2000.
- [11] H. A. Edelstein. Pan for Gold in the Clickstream. Information Week. March 12. 2001
- [12] R. Cooley, B. Mobasher, and J. Srivastava. Grouping web page references into transactions for mining world wide web browsing patterns. IEEE Knowledge and Data Engineering Exchange Workshop (KDEX '97), 1997.
- [13] M.S. Chen, J.S. Park, and P.S. Yu. Data mining for path traversal patterns in a web environment. pages 385-392, 1996.
- [14] Akshay Upadhyay, Balram Purswani. Web Usage Mining has Pattern Discovery. International Journal of Scientific and Research Publications, Volume 3, Issue 2, February 2013
- [15] Ankit R Kharwar, Viral Kapadia, A Complete PreProcessing Method For Web Usage Mining. Ganpat university journal of engineering & technology, volume 1, issue 1, March 2011
- [16] K. R. Suneetha, Dr. K. R. Kreishnamoorthy, "Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, pp. 327-332, April 2009