University of Novi Sad
Technical faculty "Mihajlo Pupin"
Zrenjanin

AIIT

International Conference

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 ICAIIT 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”
Djure 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:
Brtka Vladimir, Lacmanović Dejan, Zdravko Ivankovic, Ljubica Kazi

Cover design:
Ognjenović Višnja

Printed by:
Printing office Diginet, 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 [Elektrosnki izvor] / [2nd] International Conference on Applied Internet and Information Technologies ICAIIT 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. ICAIIT (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 Pejic 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 Jungic, ETF, University of Banja Luka, Bosnia and Hercegovina
Saso Tamazic, Univerisity of Ljubljana, Slovenia
Marijana Brtka, 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
Željen 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 Gligorijevic, 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 Brtka, 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/

International Conference
on Applied Internet and Information Technologies
October 25, 2013

CONTENT:

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

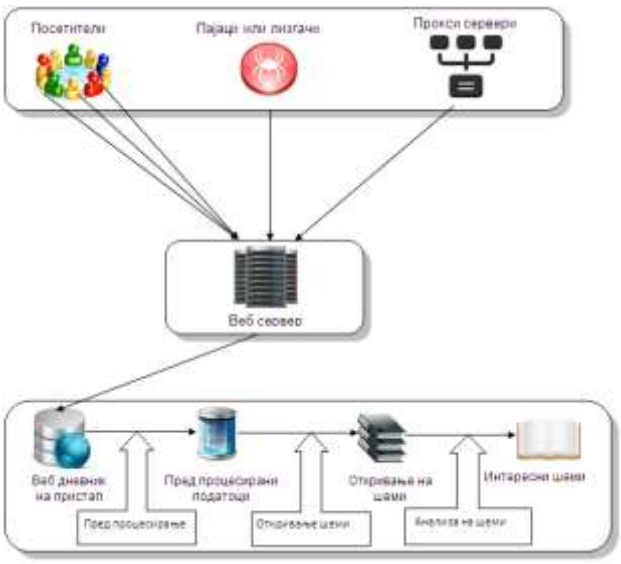

Figure 1. Steps of web mining process

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.

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).

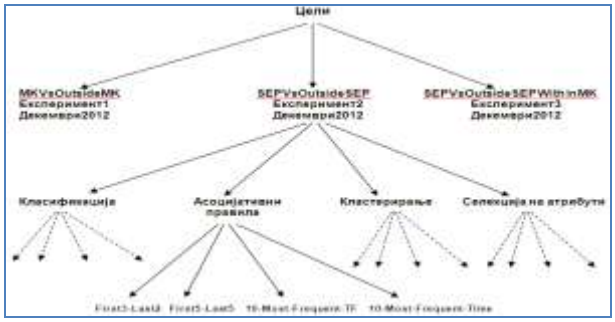

Figure 3. Organization of the performed experiments

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 (Waiakato 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-TF	NO
		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.

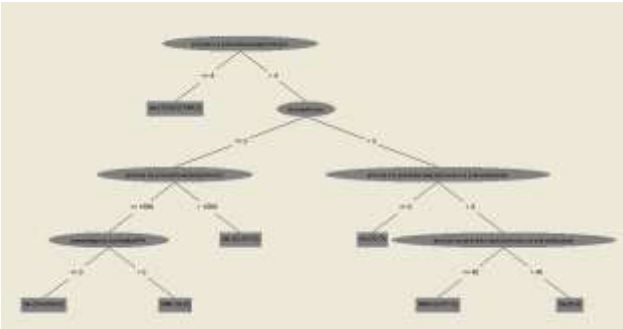

Figure 4. Decision tree obtained with J48 algorithm using 10-Most-Frequent-Time group of instances

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
                                Cluster
Attribute                        0       1
                              (0.21)  (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, Zaïane 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 ACMSIGKDD 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