An application for clickstream analysis

C. E. Dinucă

Abstract— In the Internet age there are stored enormous amounts of data daily. Nowadays, using data mining techniques to extract knowledge from web log files has became a necessity. The behavior of Internet users can be found in the log files stored on Internet servers. Web log analysis can improve business firms that are based on a Web site through learning user behavior and applying this knowledge to target them for example to pages that other users with similar behavior have visited. The extraction of useful information from these data has proved to be very useful for optimizing Web sites and promotional campaigns for marketing, etc. In this paper I will focus on finding associations as a data mining technique to extract potentially useful knowledge from web usage data. I implemented in Java programming language, using NetBeans IDE, a program for identification of pages' association from sessions. For exemplification, I used the log files from a commercial web site.

Keywords—Apriori algorithm, Association rules, Clickstream analysis, Sessions' identification, Web server logs, Web usage mining.

I. INTRODUCTION

WEB mining is an area that lately has gained a lot of interested. This is due essentially to the exponential growth of the World Wide Web and its anarchic architecture and also due to the increase of its importance over the people's life. A Web site is a lot of interconnected web pages that are developed and maintained by a person or organization. Web mining studies analyzes and reveals useful information from the Web [11]. Web mining deals with the data related to the Web, they may be the data actually present in Web pages or the data concerning the Web activities. The Web can be viewed as the largest unstructured data source available, although the data on the Web sites, which composed them, is structured. This presents a challenging task for effective design of and access to Web pages. Web mining is a term used for applying data mining techniques to Web access logs [12]. Data mining is a non-trivial process of extracting previously unknown and potentially useful knowledge from large databases [5].

Web mining is an area that lately has gained a lot of interested. This is due essentially to the exponential growth of the World Wide Web and its anarchic architecture and also due to the increase of its importance over the people's life. Scientists and engineers want to extract information from it, in order to better understand and to improve its features. They applied data mining techniques on the web. Therefore, Web mining can be defined as the application of Data Mining techniques to the web related data.

Web mining can be divided into three categories: Web content mining, Web structure mining and Web usage mining [10]. Web content mining is the process of extracting knowledge from documents and content description. Web structure mining is the process of obtaining knowledge from the organization of the Web and the links between Web pages.

Web usage mining analyzes information about website pages that were visited which are saved in the log files of Internet servers to discover the previously unknown and potentially interesting patterns useful in the future. Web usage mining is described as applying data mining techniques on Web access logs to optimize web site for users.

Click-stream means a sequence of Web pages viewed by a user; pages are displayed one by one on a row at a time. Analysis of clicks is the process of extracting knowledge from web logs. This analysis involves first the step of data preprocessing and then applying data mining techniques. Data preprocessing involves data extraction, cleaning and filtration followed by identification of their sessions.

Due to the immense volume of Internet usage and web browsing in recent years, log files generated by web servers contain enormous amounts of web usage data that is potentially valuable for understanding the behavior of website visitors.

This knowledge can be applied in various ways, such as enhancing the way that the web pages are interconnected or for increasing the sales of the commercial web sites.

II. DATA PREPROCESSING

Log files are created by web servers and filled with information about user requests on a particular Web site. They may contain information about: domains, sub domains and host names; resources requested by the user, time of request, protocol used, errors returned by the server, the page size for successful requests.

Because a successful analysis is based on accurate information and quality data, preprocessing plays an important role. Preparation of data requires between 60 and 90% of the time from data analysis and contributes to the success rate of 75-90% to the entire process of extracting knowledge [3].

For each IP or DNS we determine user sessions. The log files have entries like these:

95.175.194.33 - - [27/Jul/2011:07:23:04 -0500] "GET /css/preview_style.css HTTP/1.1" 200 2553 "http://www.nicelayouts.com/preview.php?p=34062" "Mozilla/5.0 (Windows;

Manuscript received November 28, 2011: Revised version received November 28, 2011.

C. E. Dinuca is PhD Student at the University of Craiova, Craiova, 200585, Romania (e-mail: clauely4u@yahoo.com).

U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 (.NET CLR 3.5.30729)" 95.175.194.33 - - [27/Jul/2011:07:23:04 -0500] "GET /css/tabright.gif HTTP/1.1" 200 2095 "http://www.nicelavouts.com/css/preview_style.css" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 (.NET CLR 3.5.30729)" 95.175.194.33 - - [27/Jul/2011:07:23:04 -0500] "GET /css/tableft.gif HTTP/1.1" 200 377 "http://www.nicelayouts.com/css/preview style.css" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 (.NET CLR 3.5.30729)" 95.175.194.33 - - [27/Jul/2011:07:23:05 -0500] "GET /secure/none.gif HTTP/1.1" 200 827 "https://www.nicelayouts.com/secure/custom css.css" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/20100401 Firefox/3.6.3 (.NET CLR 3.5.30729)"

As can be noticed above, each record in the file contain the IP, date and time, protocol, page views, error code, number of bytes transferred. The steps needed for data preprocessing were presented in detail in [1]. For sessions' identification in the first case was considered that a user can not be stationed on a page more than 30 minutes. This value is used in several previous studies, as can be seen in the work [2]. The current study intends to add an improvement in sessions' identification by determining an average time of page visiting the sites for the visit duration determined by analysis of web site visit duration, data which can be found in the log files of the site. Thus, for each visited page, is calculated the visit duration, which is determined by the difference between two consecutive timestamps for the same user, which is identified by IP. For records of pages with the highest timestamp among those visited by a user is assigned a predefined value of our choice to 20,000 seconds. I calculate the average visit time for a page by the average of all the times spent on that page. When calculating the average visiting time we don't take into consideration the pages with the time less than 2 seconds and largest than 20,000 seconds. Thus for our analysis I selected only those log records that contained a web page, eliminating the required load images and other files adjacent to it, this information being considered not important for analysis. I kept only pages that have status code of class 200, a successfully loaded page. Thus, we calculated how long a user stayed on a page as the difference between consecutive timestamps of visited pages for the same person, same IP. I calculated the average visiting time for a page as the media of time spent for different users on that page and used this mean to better identify sessions. I have removed pages of double sessions and I just kept for review sessions with more than 1 page views.

After preprocessing stage we obtained a file containing the user sessions. I implemeted in Java the Apriori algoritm in order to obtain the association rules between the pages from the sessions. I applied this algorithm on the sessions obtained.

III. ASSOCIATIONS MINING

Items that occur often together can be associated to each other. These together occuring items form a **frequent itemset**. Conclusions based on the frequent itemsets form **association** **rules.** For ex. {milk, cocoa powder} can bring a rule *cocoa* powder \rightarrow milk.

Consider we have database D consists of events $T_1, T_2, ..., T_m$, that is $D = \{T_1, T_2, ..., T_m\}$. Let there be an itemset X that is a subregion of event T_k , that is $X \subseteq T_k$.

The support can be defined as :

$$\sup(X) = \frac{\left| \left\{ T_k \in D \mid X \subseteq T_k \right\} \right|}{\left| D \right|}$$

this relation compares number of events containing itemset X to number of all events in database.

Any frequent item set (support is higher than the minimal support): I frequent, $\sup(I) \ge \sup_{\min}$.

Properties of the Support of an Item Set are:

• No superset of an infrequent item set can be frequent, the well known Apriori property.

• All subsets of a frequent item set are frequent.

Algoritmul Apriori

Apriori algorithm defined in 1994 by Agrawal and Srikant is the benchmark among unsupervised learning system based on association rules. Apriori algorithm is the first and most important efficient algorithm for discovering association rules.

The general scheme of the Apriori algorithm after Borgelt[8]:

• Determine the support of the one element item sets and discard the infrequent items.

• Form candidate item sets with two items (both items must be frequent), determine their support, and discard the infrequent item sets.

• Form candidate item sets with three items (all pairs must be frequent), determine their support, and discard the infrequent item sets.

• Continue by forming candidate item sets with four, five etc. items until no candidate item set is frequent.

It is based on two main steps: candidate generation and pruning. All frequent item set mining algorithms are based on these steps in some form.

Apriori uses a scroll in depth strategy to compute the support sets of elements and uses a function to generate candidates that uses circumscribed lower of support property.

If we consider the time for the selected elements, then we have a sequential association.

In the case of clickstream analysis we can appply both of them. So, we can determine association with pages, sequential association, associations rules and sequential rules in order to determine navigation paths models from the log files.

I implemented an application for applying data mining algorithms on log files in order to extract intersting knowleadges from data.

IV. CASE STUDY

To implement the algorithm presented earlier and the entire models extraction application I used the Java programming language, the code is written using NetBeans IDE. For implementation I created the NetBeans project ClickstreamAnalysis and it's componenets can be seen in the following image :

	<u>% %a ⊳ 85 · 03 ·</u>		Q. Search (Chi+1)
ChickstreemAnalysis Source Participation Source Participation Definition	Start Page 🗴 🧟 Restaurants Java 🗴 健 3QL Con	nord 2 × 20 trataretesumismo × 20 calendarismo	x 👔 Siga Command 6 x 🗟 moething Jawa 💷 💌
Control of the second sec	Learn & Discover Take a Tour Try & Sample Project What's New Community Cerner	Affantierans Berne & Tuleranie Meren & Tuleranie Meren & Antwich opperanie Korf & Antwich opperanie Meren M	Contractions Co
olar teor Analdes	batpat		

Fig. 1. The ClickstreamAnalysis project into NetBeans IDE

The application contains all the preprocessing steps needed to obtain the data in a form necessary for using it as input to algorithms. Using this developed application we can do the following operations. First we can read the data from web log file, clean the data and insert them in a database table in order to be able to perform the next preprocessing steps. Then we execute the following operations: the pages are codified, we calculate the average time for every page, we calculate the sessions with the method of in which we consider a user can't stay longer than 30 minutes on a page and then we also create the sessions using the method proposed by us with the average time for each sessions.

electare	Preprocesare	Explorarea datelor	Aplicare algoritmi	Despre
	Inserare Logu	ri		
	Codificare Pag			
	Caculare Timp			
		p Mediu Pagina		
	Creare Sesiun			
	Creare Sesiun	ii Timp Mediu 🔸		
	Afisare Date			

Fig. 2. The main window with the preprocess menu

After each preprocessing step the data can be seen in a window, as the one from figure 3, so the analyst is able to monitor and modify the data at each moment.

Solicat Ota Ora Metoda Pegna Period Cod Erace N Ebit Timestamp Cod Pagna Id Sesume In Sesume Timp Medu Pag 242 0.3.09 ZUM4001107 CFT Jimas-Asendai HTTPI.1.1 200 22109.0 2227012 131 498 490 200 70 195 10.3.29 ZUM201107 CFT Jimas-Asendai HTTPI.1.0 200 22179 2227015 170 1239 1255 5000 172 2470 12.02 ZUM201107 CFT Jimas-Asendai HTTPI.1.1 200 21479 222705 509 101 221 496 157 5000 125 249 72.28 ZUM201107 CFT Jimas-Asendai HTTPI.1.1 200 21499 2227479 193 1116 1214 2000 196 74.50 21 Jimas Martines HTTPI.1.1 200 21499 227373 431 491 496 111 220 227373 431 449	Clickstream	Analysis										_ 6
249.72.20 201/2011.01	Help											
2/22.4.01 0/2 0/2 2/20.01 0/2	d Solicitant											
194111.230 1971 1279 1279 1279 1279 1299 1299 1391 1429 1391 1429 1391 1429 1391 1429 1391 1429												78
242.8.6.6 21/A021147 CT proof												
2449.72.20 271/2011.01.01.01.01.01.01.01.01.01.01.01.01.												
758.7.1.00 271.4/2011.07 (2T (Pish-Steepide HTP).1.1 200 2179 22749 340 137 137 2000 169 7.2.1.05.2.00 271.4/2011.07 (2T (Port-Marker) HTP).1.1 200 2179 227479 153 116 124 20000 169 7.2.1.05.2.00 271.4/2011.07 (2T (Port-Marker) HTP).1.1 200 21499 2277731 458 161 621 469 101 149.1.1.2.20 271.4/2011.07 (2T (Port-Marker) HTP).1.1 200 21499 227731 458 110 621 469 70 627 559.1.6 70 627 559.1.6 70 627 559.1.6 70 70 559.1.6 70 70 559.1.6 70 70 559.1.6 70 70 70 70 559.1.6 70												
21/2001107												
27,218,20 27/200/201.0107 Corr Proceedenation Profit.10 2000 2199 227331 442 441 449 910 2000 198 198.125.20 2773/201107 Corr Prof-rake-tend HTP11.0 2000 21979 2273731 463 1239 1237 483 535 198.125.20 2773/201107 Corr Prof-rake-tend HTP11.1 200 2146 2273731 66 1239 1237 483 1368 146 147 198.126.20 2773/201107 Corr Prof-rake-tend HTP11.1 200 2149 227408 200 307 30510 146 147 149 1237 486 111 119 200 207 207/100107 Corr Prof-rake-tend HTP11.1 200 2162 227/100107 Corr Prof-rake-tend HTP11.1 200 2162 227/100107 Corr Prof-rake-tend HTP11.1 200 2162 227/10107 Corr Prof-rake-tend HTP11.1 200 2163 227/201												
249.72.20 27/M(201.01												
199:11.2.29 277/A(00.110.7												
136.25.09 77/M0110107 CT (stoweb-tangles, HTTP) 1.1 200 216.30 227795 952 105 105 2000 107 240.21/2 77/M0110107 CT (stoweb-tangles, HTTP) 1.1 200 21977 227400 100 811 621 146.6 77 19511.329 77/M01107 CT (kto-store) NTTP) 1.0 200 21493 227464 443 1239 1157 339 229 19511.329 77/M01107 CT (kto-store) NTTP) 1.0 200 21491 227464 443 1239 1157 339 229 135 339 229 1352 339 23000 100 55 56.67 77/M01106 GT (kto-store) 111 100 21602 227512 448 1181 119 20000 57 56.67 77/M0100.00 67 50.66 77 79 50.66 77 79 50.66 77 79 50.66 77 79 50.66 77 79 50.66 77 79 50.66												
62.18.0.1 27.140(0.11.0												107
2927.0220 72/M201107 (ET //web-templase, HTTP/1.1 200 21997 2274020 100 611 621 1466 777 1951.13.29 72/M201107 (ET //kor-set-templa HTTP/1.0 200 21493 227495 448 1239 1257 1262 773 1951.13.29 72/M201107 (ET //kor-set-templa HTTP/1.1 200 21662 227518 177 775 756 750 2000 150 216.20.11 72/M201106 (ET //joonfa-templase HTTP/1.1 200 21662 227518 104 776 786 20000 55 2000 100 105 118 1199 20000 55 21620 227312 438 118 1199 2000 426 431 431 323 338 333 338 <td></td> <td>143</td>												143
199:11.329 27/J4/201107: (ET /kon-sets-templ HTTP/1.0 200 21493 227994 443 1239 1237 1242 77. 198:26.307 27/J4/201107: (ET /joonta-templase HTTP/1.1 200 21662 227312 433 1181 1199 20000 150. 216:20:107 27/J4/201108: (ET /joonta-templase HTTP/1.1 200 21662 227312 436 1181 1199 20000 157. 2407.223 27/J4/201108: (ET //jacAstemplase HTTP/1.1 200 21433 227566 402 611 621 177. 781 89. 31.332 394 134 710.108 C7/J4/201108: (ET //wodpress-the HTTP/1.1 200 21575 227601 226 975 991 20000 427 713.108 C7/J4/201108: (ET //wodpress-the HTTP/1.1 200 21575 227646 205 291 252 20000 33 73.55.1 C7/J4/201108: (ET //wodpress-the HTTP/1.1 200 21575 227646 309 659 650 20000	249.72.228				200	21937	2274200	100	811	821	1406	712
98.92.02 0 2/Jul/201109	195.113.239				200	21179	2274565	389	1239	1257	389	2939
212.02.01.1 27.04/2011.08 (ST joond-templete, HTTP/1.1 200 227822 436 1181 1199 20000 57. 240.70.2 27.04/2011.08 (ST jied-templete, HTTP/1.1 200 21483 227566 402 811 821 1172 931 240.70.2 27.04/2011.08 (ST jied-templete, HTTP/1.1 200 21483 227566 402 811 821 934 934 934 70.130.08 27.04/2011.08 (ST jied-templete, HTTP/1.1 200 2157 227501 226 975 994 2000 432 70.130.08 27.04/2011.08 (ST jied-templete, HTTP/1.0 200 21575 227640 205 291 232 2000 331 73.753.1 27.04/2011.08 (ST jied-templete, HTTP/1.1 200 21575 227646 205 291 232 2000 331 73.753.1 27.04/2011.08 (ST jied-templete, HTTP/1.1 200 21575 227646 309 659 2000 331 73.753.1 27.04/2011.08 (ST jied-templete	195.113.239	27/Jul/2011:07: GET	/icon-sets-templ	. HTTP/1.0	200	21943	2274954	443	1239	1257	1262	732
21/202011 21/20201108 (ET joon14-tengles, HTTP/1.1 200 21/62 227821 436 1181 1199 20000 55 262.070 21/20201108 (ET jited-4-tengles, HTTP/1.1 200 21/483 227864 402 611 621 1172 313 260.070 21/2021108 (ET jited-4-tengles, HTTP/1.1 200 21/87 227864 402 611 621 1172 313 70.150.08 27/1/2021108 (ET jited-4-tengles, HTTP/1.1 200 21/87 227861 266 975 594 2000 472 70.150.08 27/1/2021108 (ET jited-4-tengles, HTTP/1.0 200 21/76 227646 205 291 292 2000 333 73.052.1. 27/1/2021108 (ET jited-4-tengles, HTTP/1.1 200 21/55 227646 205 291 292 2000 333 73.052.1. 27/1/2021108 (ET jited-4-tengles, HTTP/1.1 200 21/55 227646 309 639 650 2000 33 73.052.1. 27/1/2021108 (ET </td <td>9.85.226.87</td> <td>27/Jul/2011:07: GET</td> <td>/flash-8-templat</td> <td>. HTTP/1.1</td> <td>200</td> <td>21682</td> <td>2275181</td> <td>77</td> <td>755</td> <td>765</td> <td>20000</td> <td>1003</td>	9.85.226.87	27/Jul/2011:07: GET	/flash-8-templat	. HTTP/1.1	200	21682	2275181	77	755	765	20000	1003
29/7.223 27/Ju/201108 (ET Predneseth (HTP).1.1 200 21483 227868 402 611 621 1172 313 594137 27/Ju/201108 (ET Prodpreseth (HTP).1.1 200 21875 227801 266 975 991 2000 442 70.10.010 27/Ju/201108 (ET Prodpreseth (HTP).1.1 200 21875 227801 266 975 991 2000 442 70.10.010 27/Ju/201108 (ET Prodpreseth (HTP).1.1 200 2175 227645 205 291 232 2000 333 51.0.47 27/Ju/201108 (ET /verodpreseth (HTP).1.1 200 21573 227645 309 659 2000 333 50.10.69 27/Ju/201108 (ET /verodpreseth (HTP).1.1 200 21573 227645 319 657 490 2000 333 50.10.69 27/Ju/201108 (ET /verodpreseth (HTP).1.1 200 21575 227649 317 312 316.32 2000 68 447.23 200 68 407.24 207.42 207.42 <td>.216.220.117</td> <td></td> <td>/joomla-template.</td> <td> HTTP/1.1</td> <td>200</td> <td>21602</td> <td>2275212</td> <td>436</td> <td>1181</td> <td>1199</td> <td>20000</td> <td>52</td>	.216.220.117		/joomla-template.	HTTP/1.1	200	21602	2275212	436	1181	1199	20000	52
54.94.17 27/J4/201108 (ST jwodgness-ba HTTP:1.1 200 21971 227976 28 331 332 3949 134 173.0108 27/J4/201108 (ST jwodgness-ba HTTP:1.1 200 21760 2278216 340 1239 1257 94497 2931 195.13.230 27/J4/201108 (ST jwobd-20-tengia HTTP:1.1 200 21760 2278216 340 1239 1257 94497 2931 15.174.09 27/J4/201108 (ST jwobd-resch HTTP:1.1 200 2155 2276307 72 446 488 20000 333 35.81.69 27/J4/201108 (ST jwodgness-ba HTTP:1.1 200 2156 2276497 317 327 338 20000 163 37.91.19 27/J4/201108 (ST jwodgness-ba HTTP:1.1 200 2156 2276497 317 327 338 20000 468 397.15.2. 27/J4/201108 (ST jwodgness-ba HTTP:1.1 200 2157 2276497 317 812 424 432 434 447 426 432 435 446 42	62.209.70				200	21961	2275343	104	776	786	20000	772
71.130.108 27.14/2011.08 (ST /wordpress-te HTTP/1.1 200 21875 227891 266 975 991 20000 427 15.13.298 27.14/2011.08 (ST /wordpress-te HTTP/1.1 200 2176 227642 205 291 322 20000 333 15.17.49 27.14/2011.08 (ST /wordpress-te HTTP/1.1 200 21575 227646 309 659 650 20000 333 35.10.47 27.14/2011.08 (ST /wordpress-te HTTP/1.1 200 21573 227645 309 659 650 20000 333 35.10.47 27.14/2011.08 (ST /wordpress-te HTTP/1.1 200 21573 227645 309 659 20000 333 35.18.09 27.14/2011.08 (ST /wordpress-te HTTP/1.1 200 21575 227675 317 612 621 733 135 27.14/2011.08 (ST /wordpress-te HTTP/1.1 200 21575 227767 317 612 612 733 135 27.14/2011.08 (ST /wordpress-te HTTP/1.1 200 21575 <td>.249.72.228</td> <td>27/Jul/2011:08: GET</td> <td>/flash-8-templat</td> <td>. HTTP/1.1</td> <td>200</td> <td>21483</td> <td>2275606</td> <td>402</td> <td>811</td> <td>821</td> <td>1172</td> <td>315</td>	.249.72.228	27/Jul/2011:08: GET	/flash-8-templat	. HTTP/1.1	200	21483	2275606	402	811	821	1172	315
1991.13.23 27/Ju/201108 (ET //web-20-englia (FTP) 00 21700 2278.216 340 1239 1257 4947 3937 1517-49 27/Ju/201108 (ET //cs-web-tengla (FTP) 00 2155 227646 05 291 329 2000 333 52.10.47 27/Ju/201108 (ET //cs-web-tengla (FTP) 00 2155 227647 317 327 328 20000 1333 308.106.09 27/Ju/201108 (ET //wodpres-the (FTP) 00 2156 227649 317 327 328 20000 1333 308.106.09 27/Ju/201108 (ET /wodpres-the (FTP) 00 2156 227649 317 327 328 20000 468 307.152.1 27/Ju/201108 (ET /wodpres-the (FTP) 00 2157 227649 317 612 621 434 446 426 2477.228 27/Ju/201108 (ET /wodpres-th (FTP) 00 21575 227749 266 1619 499 420 427 426 137 4216 426 426 <	54.84.137	27/Jul/2011:08: GET	/wordpress-the	HTTP/1.1	200	21371	2275776	28	331	332	3048	1344
15.17.49 27Jul/201108 (GT /csswebtem/a (TTP).1. 00 2176 227646 005 291 926 0000 933 73.282.10.47 27Jul/201108 (GT /sswebtem/a (TTP).1 00 21573 227646 309 639 650 20000 933 73.282.10 27Jul/201108 (GT /wordpress-b (TTP).1 00 21573 2276415 309 639 650 20000 933 83.06.06 27Jul/201108 (GT /wordpress-b (TTP).1 200 2157 227645 399 639 650 20000 633 81.06.06 27Jul/201108 (GT /wordpress-b (TTP).1 200 2157 227657 39 927 49 2000 646 249.72.28 27Jul/201108 (GT /wordpress-b (TTP).1 200 21575 227678 317 612 621 733 1158 249.72.28 27Jul/201108 (GT /wordpress-b (TTP).1 200 21575 227757 317 612 6149 9090 424 426 426 426 426 427 420	70.130.108	27/Jul/2011:08: GET	/wordpress-the	HTTP/1.1	200	21875	2275901	266	975	991	20000	427
52.110.47 27.104/201108 (ST jcss-web-tengio (HTP)1.1 200 2155 2276307 72 446 468 20000 317 73.251. 27.104/201108 (ST jwodpress-be (HTP)1.1 200 2157 2274457 317 327 328 20000 138 08.016.06 27.104/201108 (ST jwodpress-be (HTP)1.1 200 2157 2274457 317 327 328 20000 168 87.152.1 27.104/201108 (ST jwodpress-be (HTP)1.1 200 2157 227658 399 927 493 20000 466 87.152.2 27.104/201108 (ST jwodpress-be (HTP)1.1 200 21575 227758 39 927 493 2000 466 2497.72.28 27.104/201108 (ST jwodpress-be (HTP)1.1 200 21575 227749 266 6169 169 499 402 426 426 426 426 426 426 426 426 426 426 426 426 426 426 426 426 426 426 426 <t< td=""><td>.195.113.239</td><td>27/Jul/2011:08: GET</td><td>/web-2-0-templa.</td><td> HTTP/1.0</td><td>200</td><td>21780</td><td>2276216</td><td>340</td><td>1239</td><td>1257</td><td>4947</td><td>2919</td></t<>	.195.113.239	27/Jul/2011:08: GET	/web-2-0-templa.	HTTP/1.0	200	21780	2276216	340	1239	1257	4947	2919
72.852.1 27.Jul/201108 (ST /wodpress-ba HTTP1.0 200 21573 227415 399 639 639 630 30300 323 801.806.90 27.Jul/201108 (ST /wodpress-ba HTTP1.1 200 21576 227658 349 1217 1238 20000 989 87.15.1 27.Jul/201108 (ST /wodpress-ba HTTP1.1 200 21576 227658 349 1217 1238 20000 989 87.15.2 27.Jul/201108 (ST /wodpress-ba HTTP1.1 200 2157 227675 39 927 93 2000 989 249.72.28 27.Jul/201108 (ST /wodpress-ba HTTP1.1 200 2156 227678 317 612 621 73 15.8 249.72.28 27.Jul/201108 (ST /wodpress-ba HTTP1.1 200 2157 227757 317 612 641 69 2000 422 249.72.28 27.Jul/201108 (ST /wodpress-ba HTTP1.1 200 2157 227757 317 616 612 621 621 621 621 621 621 <td>15.174.49</td> <td>27/Jul/2011:08: GET</td> <td>/css-web-templa.</td> <td> HTTP/1.1</td> <td>200</td> <td>21736</td> <td>2276246</td> <td>205</td> <td>291</td> <td>292</td> <td>20000</td> <td>333</td>	15.174.49	27/Jul/2011:08: GET	/css-web-templa.	HTTP/1.1	200	21736	2276246	205	291	292	20000	333
388.108.08 271/4021.108	52.110.47	27/Jul/2011:08: GET	/css-web-templa.	HTTP/1.1	200	21535	2276307	72	486	488	20000	318
11/194/19 27/Ju/201108	73.252.1	27/Jul/2011:08: GET	/wordpress-the	HTTP/1.0	200	21573	2276416	309	639	650	20000	320
89.715.21 27/Jul/201108 (grt) //// 000 (grt) /////// 000 (grt) ///////////////////////////////////	0.88.108.69	27/Jul/2011:08: GET	/wordpress-the	HTTP/1.1	200	21360	2276487	317	327	328	20000	1353
2497.228 27.14/2011 08 (GT wordpress-th TTP1.1.1 200 21670 227778 317 812 821 733 135 6670.39 27.14/2011 08 (GT wordpress-th TTP1.1.2 200 21675 227752 266 812 83 469 1499 2000 422 6670.39 27.14/2011 08 (GT wordpress-th TTP1.1.2 200 21675 227752 266 1649 1499 2000 422 677.28 27.14/2011 08 (GT wordpress-th TTP1.1.2 200 21675 227752 266 1649 1499 2000 422 677.28 27.14/2011 08 (GT wordpress-th TTP1.1.2 200 21675 227752 266 1649 1499 2000 422 677.28 27.14/2011 08 (GT wordpress-th TTP1.1.2 200 21675 227752 366 1649 1499 2000 422 677.28 27.14/2011 08 (GT wordpress-th TTP1.1.2 200 21675 227752 366 1649 1499 2000 422 677.28 27.14/2011 08 (GT wordpress-th TTP1.1.2 200 21675 227752 366 1649 1499 2000 422 71.14.08 27.757 368 10.08	4.17.194.149	27/Jul/2011:08: GET	/wordpress-the	HTTP/1.1	200	21876	2276584	249	1217	1235	20000	890
29/72/2011/09 (CT //wodgress-bu (HTTP).1.1 200 21975 227741 266 912 921 449 422 267.03 27/Ju/2011/09 (CT //wodgress-bu (HTTP).1.1 200 21975 227743 266 1912 921 4499 422 267.03 27/Ju/2011/09 (CT //wodgress-bu (HTTP).1.1 200 21975 227743 986 1919 2000 422 27.4.102.01.07 (CT //wodgress-bu (HTTP).1.1 200 21975 92776 986 1919 200 422 3040401 7/Ju/2011/07 (CT //wodgress-bu (HTTP).1 200 21975 92776 986 1909 1100 90000 422 407 7/Ju/2011/07 (CT //wodgress-bu (HTTP).1 200 21975 92776 986 1909 1009 9000 422 4030 GT //wodgress-bu (HTTP).1 200 21976 //wodgress-bu (HTTP).1 900 100 100 100 100 100 100 100 100 100 100 100 100 100 1	8.97.15.21	27/Jul/2011:08: GET	/jquery-templates	s/ HTTP/1.1	200	21451	2276758	93	927	943	20000	463
68/70.9 27/Jul 2011.06 jCT //or 2019ressette TTP/1.1 200 21875 227752 266 1649 199 2000 425 79.126.232.59 79.126.232.59 79.126.232.59 79.126.232.59 79.126.232.59 79.126.232.59 79.127	.249.72.228	27/Jul/2011:08: GET	/wordpress-the	. HTTP/1.1	200	21360	2276778	317	812	821	703	1353
2 # 10 # 201 Uni2011 to fet Incomediations in 1001 to 1001 to 1001 to 1001 to 1000 t	.249.72.228	27/Jul/2011:08: GET	/wordpress-the	. HTTP/1.1	200	21875	2277481	266	812	821	469	427
904822.99 904822.99 904801.97.35.99 9019<	.68.70.39											427
ava 7/J/2011/07:35:99 ava 7/J/2011/07:35:99 bit 7/J/2011/07:35:99 bit 7/J/2011/07:35:90 bit			locowebstempla	HTTD/1 1	200	21228	2277530	386	1003	1100	20000	054
def def initial /cs.wab-tangkate/catagor/govt/ def times/catagor/govt/ times/catagor/govt/ def times/catagor/govt/ <	iolicitant:	79.126.232.59										
initial issueb-templates/catagory/sport/ itemplates/catagory/sport/ itemplates/catagory/sport/ ite	a Ora:	27/Jul/2011:07:35:59										
true ITTP/1.1 true 20 true 20.3 status 20.3 status 20.3 status 20.3 status 20.3 status 20.4	oda:	GET										
Erose Od Biros 203 Station 203799 Station 203 Ipages 26 Ipages 204	jina:	/css-web-templates/category	/sport/									
Res I233 stance I23759 stance IS stan	tocol:	HTTP/1.1										
Pagina 227759 IPagina 352 Isaure 106 Isaure 2000 Pagena 2000	Eroare:	200										
Vertifier Vertifier Pagino 522 Isolation 106 Isolation<	aci:	21261										
Pagina: 552 Desure: 106 Desure: 106 Desure: 2000 Media: Pagina: 107												
deskne 106 issund 206 p Pag: 2000 Medu Pag: 107												
estine: 106 P Pg: 2000 Medu Pgg: 107												
p Pag: 2000												
0 Među Pog: 107												
	-											
Row Force Force Same	prisona ragi									New	Delete D-6	ach Course
										New	Delece Kell	ENI 28V6

Fig. 3. The window for monitoring all the preprocessing steps.

The analyst can delete records of data that are considered irrelevant or outliers from the data. The form has the ability of refreshing the data.

So, in the interface I gave the opportunity for the user to choose the file with sessions. The content of the file must be in the following format :

🖡 sesiuni_claudia1.txt - Notepad	
File Edit Format View Help	
295 402 85 230 294 351 418 419 199 224 291 300 350 279 390 42 97 398 410 13 7 115 128 307 420 111 113 130 212 264 443 12 56 151 211 322	×
202 354 32 90 117 156 162 283 386 438 56 139 225 289 421 51 247 278 310 341 402 26 45 93 176 226 280 282 297 332 348 353 384 402 412 449 26 45 93 176 226 280 282 297 332 348 353 384 402 412 449 25 67 99 108 149 163 180 287 295 296 355 377 390 413 446 450 454 15 160 178 199 202 231 336 345 4 54 69 127 136 262 369 204 284	11
52 399 88 100 164 201 392 397 12 22 41 94 106 187 204 330 334 335 348 374 417 431 446 146 178 306 412 236 414 60 162 257 86 145 175 198 204 256 298 406 452 145 288 406 407 452 86 193 311 79 140 256 28 143 205 28 62 148 274 285 401 415 424	
26 62 148 274 285 401 415 424 196 241 248 295 361 412 444 249 274 360 397 401 430	×

Fig.4.

As, it can be seen in the figure before, in the preprocessing stage, we codified the pages from the log files.

The user can also choose the minimum support threshold. After all these being set, we can run the algorithm.

After applying the Apriori algorithm we obtained some important associations between pages. For example, we obtained the result shown in Fig. 5.

🕹 Apriori		
Open File	Add Min. Sup.	Run
enerating K≕7 large itemsets enerating K⇒8 large itemsets enerating K⇒9 large itemsets enerating K⇒10 large itemsets enerating K=11 large itemsets enerating K=12 large itemsets enerating K=13 large itemsets enerating K=14 large itemsets		
REQUENT (LARGE) ITEM SETS (W	ith support counts)	
$ \begin{array}{l} (7) = 3 \\ (10) = 3 \\ (10) = 3 \\ (10) = 3 \\ (10) = 3 \\ (10) = 3 \\ (10) = 3 \\ (20) = 3 \\ (7, 10, 0) = 3 \\ (7, 10, 20) = 3 \\ (7, 10, 20) = 3 \\ (7, 10, 20) = 3 \\ (7, 10, 20) = 3 \\ (10, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 20) = 3 \\ (11) (7, 10) = 3 \\ (11)$		

When we decrease the support, we obtain more association between pages.

Another way of obtaining association rules on pages from a web-site is by transforming this file with sessions that we obtained before in a matrix containing 0 and 1 and create an .arff file from all these pages, having as attributes the pages that can take values 0 or 1 and the relation defined between being the sessions, the data from the .arff are the values of the sessions.

The .arff file can be defined in a sparse or dense manner. After obtaining the .arff file, it can be applied to the Apriori algorithm from Weka, or any open source data mining tool that accepts this format.

Generating sets of pages frequently visited together is determined by going to the main window and choosing the application algorithms option AplicareAlgoritmi \rightarrow FPGrowth. There is an option to generate steps of frequent pages accesed together with coded pages and with the exact name of the pages. FPGrowth algorithm is used to generate associations of pages frequently visited together [15].

🛓 Clickstream Analysis
Selectare Preprocesare Explorarea datelor Aplicare algoritmi Despre Selectare FREQUENT ITEMSETS L0 L1 patern 0. /con-sets-templates/dag/Staff support: 0.07 (8/112) patern 2. /lul-isite-templates/dag/Staff support: 0.08 (9/12) patern 2. /lul-isite-templates/dag/Staff support: 0.08 (9/12) patern 2. /lul-isite-templates/category/business/support: 0.08 (9/12) patern 4. /bpbb3-templates/category/business/support: 0.08 (9/12) patern 4. /bpbb3-templates/category/business/ifehack-blog/ support: 0.09 (10/12) patern 5. /wordpress-themes/category/business/ifehack-blog/ support: 0.09 (10/12) patern 6. /wordpress-themes/category/business/ifehack-blog/ support: 0.09 (10/12) patern 6. /wordpress-themes/category/business/ifehack-blog/ support: 0.09 (10/12) patern 6. /wordpress-themes/category/business/ifehack-blog/ support: 0.09 (10/12) patern 7. //con-sets-templates/category/business/ifehack-blog/ support: 0.09 (10/12) patern 7. //con-sets-templates/category/support: 0.011 (2/12) patern 7. //con-sets-templates/category/spord/ support: 0.01 (10/12) patern 11. //cs=web-templates/category/spord/ support: 0.14 (16/12) L2 patern 12. //joomia-templates/category/spord/ support: 0.04 (9/12) patern 13. //conporate-design///cs=-web-templates/category/spord/ support: 0.07 (8/12) patern 13. //conporate-design///cs=-web-templates/category/spord/ support: 0.07 (8/12) L2

Fig. 6. Frequent itemsets presented in the main window of the application.

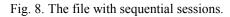
Explorarea Datelo	r Fisierului Log		
ID_SESIUNE	NrActiuniVizitare	DurataSesiunii	TimpPePagina
264	7	850	141.6666666666
278	4	211	70.3333333333
280	3	681	340.5000000000
293	2	173	173.0000000000
307	2	20	20.0000000000
312	2	4	4.0000000000
31.4	4	352	117.33333333333
316	2	173	173.0000000000
322	3	140	70.000000000
324	3	1207	603.5000000000
327	2	69	69.000000000
344	2	5149	5149.0000000000
346	3	1	0.500000000
358	3	612	306.0000000000
364	3	61	30.5000000000
374	3	1754	877.0000000000
378	2	107	107.0000000000
381	2	428	428.0000000000
383	3	1576	788.0000000000
384	35	6689	196.7352941176
385	13	7033	586.0833333333
200	2	0	4 000000000

Fig 7. Exploratory data analysis

The menu **Data Exploration** contains Sessions' Information submenu which provides information about resulted sessions such as session ID, number of visits in that session, session length, the average time per page in a session

as it is shown in the figure 4.

🖻 sesiuni_tmcc_secvential.txt - WordPad	- 2	×
Ele Edit View Insert Format Help		
		_
309 266 -1 -2		^
75 312 -1 -2		
309 433 -1 -2		
356 469 106 -1 -2		
438 480 -1 -2		
337 351 321 329 21 480 72 342 392 432 311 480 -1 -2 317 177 -1 -2		
31 (1/ -1 -2 93 433 -1 -2		
440 475 -1 358 181 -1 477 262 229 -1 -2		
403 131 -1 -2		
409 479 -1 93 146 -1 -2		
433 309 -1 -2		
433 465 433 -1 -2		
55 471 -1 -2 232 433 -1 -2		
433 160 -1 -2		
38 284 386 -1 -2		
160 274 -1 -2		
173 68 389 443 340 366 144 60 -1 119 96 -1 371 116 111 -1 145 255 438 37 430 -1 295 227 420 30 177 34 42 -1 380 230 90 394 -1 275 388 -1 437 59 134	226 191	
433 30 433 30 -1 -2		
28 309 -1 -2	Tracky Trenth I	
433 309 268 332 221 160 275 479 93 156 38 50 386 261 337 21 232 30 239 130 342 351 158 321 37 403 276 62 179 251 206 45 252 14 333 311 260 131 172	465 444	2
433 309 -1 -2 471 309 -1 -2		
32 433 280 165 161 160 477 401 210 275 107 60 117 473 415 440 95 395 422 345 221 -1 -2		
268 433 479 160 93 332 275 221 309 221 309 251 342 156 284 432 481 308 432 62 276 72 206 90 172 21 480 205 480 311 313 14 232 152 14 131 260 433 30	9 332 22	1
266 309 -1 -2		
266 271 266 -1 -2		
28 101 -1 -2		
433 309 -1 -2		
433 309 -1 -2		
266 429 -1 -2 170 8 -1 -2		
104 113 309 104 291 286 -1 -2		
433 309 -1 -2		
311 433 -1 -2		
160 397 -1 -2		
433 386 156 38 50 386 261 337 21 232 30 -1 239 433 239 130 342 351 158 321 37 403 276 -1 -2		
	104 145	
166 131 -1 7 57 357 434 383 89 307 222 225 148 482 109 287 478 394 461 103 96 70 187 416 389 452 1 380 405 48 341 91 166 288 111 242 85 46 375 254 439 309 -1 -2	404 147	
469 433 -1 -2		
317 368 -1 -2		Ξ
433 309 197 377 444 276 280 87 333 414 88 252 42 465 50 45 260 95 401 432 391 311 395 392 321 107 93 409 138 220 113 221 190 218 238 479 24 386 1 1	137 331 4	ε
433 320 160 215 221 275 192 93 479 309 -1 -2		
309 55 -1 -2		
356 309 -1 -2		
317 309 -1 -2 433 45 -1 -2		
133 13 -1 -2 46 839 -1 -2		
260 433 -1 -2		
266 429 368 104 -1 -2		
309 296 -1 -2		
		~
	>	2
For Help, press F1	P	JUM



Edit View Insert	nePagini.txt - WordPad
🖆 🖬 🍯 🖾	🙀 🖇 陆 🛍 🕫 🥵
FREQUE	ENT ITEMSETS LO
L1	
pattern 0: /	'icon-sets-templates/tag/House support : 0.06 (7/112)
pattern 1: /	css-web-templates/category/software/ support : 0.06 (7/112)
pattern 2: /	'flash-8-templates/category/society-culture/ support : 0.06 (7/112)
	/css-web-templates/category/art/ support : 0.06 (7/112)
	php-nuke-templates/tag/Makeup support : 0.06 (7/112)
	'dynamic-flash-templates/tag/ support : 0.06 (7/112)
	'css-web-templates/category/entertainment/ support : 0.06 (7/112)
	'joomla-templates/category/brewery/ support : 0.06 (7/112)
	'templates/category/exclusive/ support : 0.06 (7/112)
	flash-8-templates/category/art/ support : 0.06 (7/112)
	/wordpress-themes/category/fashion/ support : 0.06 (7/112)
	/templates/type/silverlight-templates/ support : 0.06 (7/112)
	/php-nuke-templates/tag/Forum support : 0.06 (7/112)
	/flash-templates/category/flash-intro-header/ support : 0.06 (7/112)
	/phpbb3-templates/tag/Cat support : 0.06 (7/112)
	/icon-sets-templates/tag/Staff support : 0.07 (8/112)
	/phpbb-templates/tag/Women support : 0.08 (9/112)
	/full-site-templates/category/business/ support : 0.08 (9/112)
	/joomla-templates/category/security/ support : 0.08 (9/112)
	/phpbb3-templates/tag/Bird support : 0.08 (9/112)
	/joomla-templates/category/halloween/ support : 0.08 (9/112)
	/wordpress-themes/category/business/lifehack-blog/ support : 0.09 (10/112) /icon-sets-templates/tad/liss support : 0.09 (10/112)
	/lcon-sets-templates/tag/Ltps Support : 0.09 (10/12) /last-added-website-templates/tag/Ltp support : 0.09 (10/112)
	/last-addea-website-templates/tag/Aft Support : 0.19 (10/12) /loomla-templates/category/internet/ support : 0.11 (12/12)
	/journate-englaces/category/internet/ support: 0.11 (12/12) /corborate-design/ support: 0.13 (15/112)
	/cs/polactacesign/seport. 0.13 (12/12) /css-web-templates/category/sport/support: 0.14 (16/112)
L2	/css-web-cemplates/categoly/sport/ support . 0.14 (10/12)
	/css-web-templates/category/software/ /css-web-templates/category/sport/ support : 0.06 (7/112)
	/css=web-templates/categoty/soluwate//css=web-templates/categoty/solut/ support : 0.06 (/112) /full=site=templates/categoty/business/ /ioomla-templates/categoty/internet/ support : 0.06 (/112)
	/lat-added-website-templates/tad/strics/ / journa-templates/tategory/interine/ support . 0.00 (//12) /lat-added-website-templates/tad/str /css-web-templates/tategory/sport/ support : 0.06 (//12)
	/last-added-website-templates/tady.nt /corporate-design/ support : 0.06 (//12)
	/last-addet website-templates/category/internet//css-web-templates/category/sport/ : 0.08 (9/112)
	/journal completes/category/moting/category/categor
	· · · · · · · · · · · · · · · · · · ·
	sociation rules
rule 0: /css	s-web-templates/category/software/ ==> /css-web-templates/category/sport/ support : 0.0625 (7/112) confidence : 1.0
rule 1: /ful	ll-site-templates/category/business/ ==> /joomla-templates/category/internet/ support : 0.0625 (7/112) confidence : 0.777777777
rule 2: /las	st-added-website-templates/tag/&rt ==> /css-web-templates/category/sport/ support : 0.0625 (7/112) confidence : 0.7
rule 3: /las	st-added-website-templates/tag/Art ==> /corporate-design/ support : 0.0625 (7/112) confidence : 0.7
rule 4: /joc	mla-templates/category/internet/ ==> /css-web-templates/category/sport/ support : 0.08035714285714286 (9/112) confidence : 0.75

Fig. 9. Frequent itemsets and association rules obtained from web pages

Issue 1, Volume 6, 2012

In order to determine frequent web pages accessed together I use FP-Growth algorithm, and then the algorithm for generation of association rules created by Agraval and Srikant, 1994 [9] from these frequent pages.

In order to apply data mining methods and algorithms for data mining the sessions are saved in a file like the one from figure 1, the sessions containing the coded pages. For algorithms which determine frequent sequential pages we use for input a text file which contains the user sessions in a temporal order as in Fig. 8. Thus, the value -1 is used to separate sessions from the same IP and the value -2 is used to separate sessions from different IPs.

Sequence mining is the task of finding temporal patterns over a database of sequences, in this case a data base of click streams. Sequence mining is considered to be an extension of associations mining that only finds nontemporal patterns.

This technique can have a very important role in knowledge discovery in web log data, due to the (temporally) ordered nature of click-streams.

The type of patterns that results from the application of this technique, can have an example like this:

"If user visits page X, and then page Y, it will visit page Z with c% of chance". The algorithms for sequence mining inherited much from the association mining algorithms, and many of them are extensions of the firsts, where the main difference is that in sequence mining intersequence patterns are searched, where in the association mining the patterns searched are intra-sequence patterns.

For the determination of successive sets of frequently accessed together pages using PrefixSpan algorithm developed by Pei and others, 2004. Running the algorithm for generating frequent sets of sequential pages with minimum support 0.09 we obtain the results that can be seen below.

Data are implemented in the developed program. We run the algorithm for generating sequential rule by using the following path AplicareAlgoritmi \rightarrow

GenerareReguliSecventialeRuleGen.

EguliSecventialeRuleGen	
Suport minim	Confidenta minima
5	0.9
Denumire fisier pentru salvare rezultat	ReguliSecventiale
Apelare A	Igoritm

Fig. 10. Window used for algorithm call for generating sequential rules

After setting the minimum support, minimum confidence and the file name where you want to save the result file with the obtained rules we click on the button "Apelare Algoritm". For convenience these rules may be obtained with coded pages. Having a minimum support threshold and a minimum confidence, we determine sequential rules by using RuleGen algorithm [14]. First, this algorithm applies another algorithm to determine frequent sequential pages, in this case we use PrefixSpan, and then frequent pairs of models are combined to determine sequential rules from pages. In the following images it can be seen the sequential rules obtained from web pages with the coded pages in Fig. 11. and exact name of the pages in Fig. 12.

seq_rules_by_ruleGen - Notepad	
Eişier Editare Format Vizualizare Ajutor	
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	
< III	E. ▲

Fig. 11. Sequential rules obtained with the coded pages.

INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS

🖡 ReguliSecventialeNumePagini.txt - Notepad
Ele Edit Format Yew Help
<pre>(/joomla-templates/category/music/) ==> (/joomla-templates/category/beauty/ /joomla-templates/category/music/) sup= 8 conf= 0.666666666666666666666666666666666666</pre>
<pre>(/corporate-identity-templates/tag/Staff) ==> (/full-site-templates/category/entertainment/ /corporate-identity-templates/tag/Staff /joomla-templates/category/beauty/) sup= 9 conf= 0.5 (/joomla-templates/category/beauty/) ==> (/corporate-identity-templates/tag/Staff /joomla-templates/category/beauty/) sup= 9 conf=</pre>
<pre>0.5294117647058824 (/corporate-identity-templates/tag/Staff /joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment// /corporate-identity-templates/tag/Staff /joomla-templates/category/beauty/) sup= 9 conf= 1.0 (/live-chat) ==> (/full-site-templates/category/entertainment//live-chat) sup= 12 conf= 0.8 (/live-chat) ==> (/full-site-templates/category/entertainment//live-chat /joomla-templates/category/beauty/) sup= 8 conf= 0.533333333333 (/joomla-templates/category/beauty/) ==> (/live-chat /joomla-templates/category/beauty/) sup= 9 conf= 0.5294117647058824 (/live-chat / joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment//live-chat /joomla-templates/category/beauty/) sup= 9 conf= 0.5294117647058824 (/live-chat / joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment//live-chat /joomla-templates/category/beauty/) sup= 9 conf= 0.5294117647058824</pre>
(/joomla-templates/category/beauty/) ==> (/joomla-templates/category/beauty//joomla-templates/category/music/) sup= 8 conf= 0.47058823529411764
<pre>(/joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment/ /corporate-identity-templates/tag/Staff /joomla-templates/category/beauty/) sup= 9 conf= 0.5294117647058824 (/joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment/ /joomla-templates/category/beauty/) sup= 10 conf= 0.5882352941176471</pre>
(/joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment/ /live-chat /joomla-templates/category/beauty/) sup= 8 conf= 0.47058823529411764
(/templates/category/exclusive/) ==> {/full-site-templates/category/entertainment/ /templates/category/exclusive/ } sup= 19 conf= 0.37254901960784315
(/full-site-templates/category/entertainment/) ==> (/full-site-templates/category/entertainment/ /templates/category/exclusive/) sup= 19 conf= 0.41304347826086957
<pre>(/full-site-templates/category/entertainment/) ==> (/full-site-templates/category/entertainment/ /flash-intro-header/tag/Puppy) sup= 8 conf= 0.17391304347826086</pre>
(/full-site-templates/category/entertainment/) ==> (/full-site-templates/category/entertainment//corporate-identity-templates/tag/Staff) sup= 11 conf= 0.2391304347826087
(/full-site-templates/category/entertainment/) ==> (/full-site-templates/category/entertainment/ /corporate-identity-templates/tag/Staff /joomla-templates/category/beauty/) sup= 9 conf= 0.1956521739130435
<pre>(/full-site-templates/category/entertainment/) ==> (/full-site-templates/category/entertainment/ /joomla-templates/category/beauty/) sup= 10 conf= 0.21739130434782608</pre>
<pre>(/full-site-templates/category/entertainment/) ==> (/full-site-templates/category/entertainment/ /live-chat) sup= 12 conf= 0.2608695652173913 (/full-site-templates/category/entertainment/) ==> (/full-site-templates/category/entertainment/ /live-chat) joomla-templates/category/beauty) sup= 8 conf= 0.17391304347826086</pre>
(/flash-intro-header/tag/Puppy) ==> (/full-site-templates/category/entertainment/ /flash-intro-header/tag/Puppy) sup= 8 conf= 0.666666666666666
<pre>(/full-site-templates/category/entertainment/ /corporate-identity-templates/tag/Staff) ==> (/full-site-templates/category/entertainment/ /corporate-identity-templates/category/entertainment/ /joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment/ /corporate-identity-templates/tag/Staff /joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment/ /corporate-identity-templates/tag/Staff /joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment/ /joomla-templates/category/beauty/) ==> (/full-site-templates/category/entertainment/ /live-chat /joomla-templates/category/entertainment/ /live-chat) ==> (/full-site-templates/category/entertainment/ /live-chat</pre>
<pre>(/full-site-templates/category/entertainment/ /live-chat) ==> (/full-site-templates/category/entertainment/ /live-chat /joomla-templates/category/beauty/) sup= 8 conf= 0.6666666666666666</pre>

Fig. 12. Sequential rules from web pages with the exact name of pages.

V.CONCLUSIONS

Nowadays, the web is an important part of human life. The web is a very good place to do businesses. Today, large companies rethink their business using the Internet to improve business. Business carried on the Web offers the opportunity to potential customers or partners where our products and specific company can be found. To differentiate through the Internet economy, winning companies have realized that e-commerce transactions is more than just buying / selling, so the appropriate strategies are the key to improve competitive power. One effective technique used for this purpose is data mining. Data mining is the process of extracting interesting knowledge from data. Web mining is the use of data mining techniques to extract information from web data.

Web mining can be divided as was stated above in three categories: Web content minig, Web structure mining and Web usage mining. Data mining as applied to e-commerce is a breakthrough technology that can gather information in an automated fashion and build models used to predict customer purchasing decisions and navigation models with remarkable accuracy.

At the beginning I present the data preprocessing steps which has been performed on the log files from this commercial web site. Data preparation phase starts with data collection. Usually, the analyst does not participate in the process of data collection, so his goal is to select from existing data those that best fit the analyse it wishes to perform. Variables and records used depend on what it is desired to obtain. The primary source used for web usage mining are logs files of the server. The data used to analyze web usage mining may come from two sources: the period of testing and web logs. The log files from the testing periods are rarely used because of the large time required and high cost. Web log files consist of information which track web users work in their interaction with web servers. Logs files can be located as follows: on the web server, a proxy server or client computers. Logs available on Web servers are most often used because they contain accurate and complete data on site usage.

Here I presented the method that I proposed for session identification by adding the medium time that a user can spend on a specific page as a threshold for session identification. Having the data preprocessing step done, we can then go to another important step in web mining, the one of effectively extracting useful information from all this data. Mining the associations from web site pages is an important task as it helps web site designers to improve the design of the site. It gives better satisfaction for the final user. By mining associations of web pages from web logs the web site designer can discover the bad web page association and can change the design.

This article presents different ways of solving this problem. I apply different algorithms for discovering navigation patterns from data log files. The novelty brought by this work is represented by the Java application with a friendly graphical user interface, use the mean time to identify sessions and application of different data mining algorithms on Web logs for navigation patterns extraction.

Analyzes aim is improving the site design and so leading to customer satisfaction and increasing the number of visits, visitors and therefore sales.

By it's architecture, the application provides a highly flexible environment and can be easily modified by its content, distributed and improved.

For the future I consider adding new modules to the applications developed in order to execute various data mining analysis.

REFERENCES

- C. E. Dinucă, "The process of data preprocessing for Web Usage Data Mining through a complete example", *Annals of the "Ovidius" University, Economic Sciences Series* Volume XI, Issue 1, 2011
- [2] Z. Markov, D. T. Larose, DATA MINING THE WEB, Uncovering Patterns in Web Content, Structure and Usage, USA: John Wiley & Sons, 2007.
- [3] Y. Nong, The handbook of Data Mining, Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey, 2003.
- [4] R. Cooley, B. Mobasher and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. A survey paper. In *Proc. ICTAI-97*.
- [5] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer Berlin Heidelberg New York, 2006.
- [6] L. Clark, I. Ting, C. Kimble, P. Wrigth, D. Kudenko, Combining Ethnographic and Clickstream Data to Identify Strategies Information Research 11(2), 2006.
- [7] R. Kohavi, R. Parekh, Ten supplementary analysis to improve ecommerce web sites, Proceedings of the Fifth WEBKDD workshop, 2003.
- [8] C. Borgelt, *Frequent Pattern Mining*, Intelligent Data Analysis and Graphical Models Research Unit European Center for Soft Computing, 33600 Mieres, Spain, 2004.
- [9] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules, IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120,1994.
- [10] R. Cooley, B. Mobasher, J. Srivastava, Web mining: Information and Pattern Discovery on the World Wide Web. A survey paper. In: Proc. ICTAI-97, 1997.
- [11] O. Zaiane, Conference Tutorial Notes: Web Mining: Concepts, Practices and Research. In: Proc. SDBD-2000, 2000, pp. 410-474.
- [12] G. Piatetsky-Shapiro, U. Fayyad, P. Smith, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining., AAAI/MIT Press, 1996.
- [13] A. M. Yahya, MD. B. S. Nasir, M. Norwati, I. U. Nur, M. Zaiton, ARS: Web Page Recommendation System for Anonymous Users Based On Web Usage Mining, Proceedings of the WSEAS European conference of systems, and European conference of circuits technology and devices, and European conference of communications, and European conference on Computer science, pp. 115-120, 2010.
- [14] M. J. Zaki, SPADE: An Efficient Algorithm for Mining Frequent Sequences, Machine Learning, vol. 42, no.1-2, 2001, pp. 31-60.
- [15] J. Han, J. Pei, Y. Yin, R. Mao, Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach, Data Mining and Knowledge Discovery, Volume 8, Issue 1, 53–87, 2004.
- [16] G. Castellano, A. M. Fanelli, M. A. Torsello, Understanding Visitor Behaviors from Web Log Data, WSEAS Transactions on Computer Research, Vol. 2, No. 2, pp. 277-284, 2007.
- [17] G. Castellano , A. M. Fanelli , M. A. Torsello, LODAP: a log data preprocessor for mining web browsing patterns, Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, pp.12-17, 2007.
- [18] G. Castellano, A. M. Fanelli, M. A. Torsello, Mining usage profiles from access data using fuzzy clustering, Proceedings of the 6th WSEAS

International Conference on SIMULATION, MODELLING AND OPTIMIZATION (SMO '06), pp. 157-160, 2006.