IDENTIFY NAVIGATIONAL PATTERNS OF WEB USERS

Shiva Asadianfam and Masoud Mohammadi

Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran

ABSTRACT

RapidMiner is a software for machine learning, data mining, predictive analytics, and business analytics. The server will record large web log files when user visits the website. Extracting knowledge from such huge data demands for new methods. In this paper, we propose a web usage mining method with RapidMiner. At first, the redundant files in log file are deleted by Matlab and then we mines web log which has been pretreated with RapidMiner, it obtains the custom of different user to visit the website by processing and analyzing log file, and mines unusual rules, and provides the reference for the policy decision and construction of website. Experimental result analysis show that, applying RapidMiner in web usage mining, will obtain frequent model which user visits the website, manage to optimize the website structure and recommends for users.

KEYWORDS

Web Usage Mining, RapidMiner, Association Rule Mining, Web Log Analysis

1.INTRODUCTION

A framework for web usage mining is that suggested by Srivastava.This process consists of four phases: the input phase, the preprocessing phase, the pattern discovery phase, and the pattern analysis phase.(a) Input phase. At the input phase, three types of raw web server log files are retrieved access, referrer and agent logs.(b) Preprocessing phase. The raw log files do not reach in a format conducive to useful data mining. Therefore, substantial data preprocessing must be applied.The most common preprocessing tasks are data cleaning and filtering, de-spidering, user identification, session identification, and path completion.(c) Pattern discovery phase. In this phase use data mining methods for the purpose of discovering models[1]. These methods include standard statistical analysis, clustering algorithms, association rules, classification algorithms, and sequential patterns.(d) Pattern analysis phase. Not all of the models uncovered in the pattern discovery phase would be considered beneficial. Hence, in the pattern analysis phase, human analysts check the output from the pattern discovery phase and glean the most interesting, beneficial, and actionable models [1].

Many experts have probed on the correlation theory, some papers on the web server log mining have had initial result [2]. Y.T Wang recommended a compact graph structure to record information about the navigation paths of website visitors [3]. Jozef Kapusta suggested to optimize web portal in level of portal adaptation on the basis user and access hour on portal [4]. Resul Das proposed to preprocess log files and use path analysis technique to probe the URL information concerning access to electronic sources [5]. Nicolas proposed to use machine learning techniques and Markov-chain models to build Self-adaptive utility-based web session management [6]. Kleinberg and Tomkins have given the technique for web structure mining [7].

Federico proposed a new method towards automatic personalized recommendation by algorithm Apriori [8]. Shiyong Zhang suggested a framework that use a hidden Markov chain [9].In this paper, we use web log file based on platform RapidMiner.We analysis EPA web log file and then applied for RapidMiner, will obtain frequent model which user visits the website, manage to optimize the website structure.The rest of this paper is organized as follows. We describe RapidMiner in Section 2. We explain the implementation of web log file by Matlab in Section 3. Section 4 contains our simulation on the platform RapidMiner and section 5 shows our simulation results. Section 6 concludes the paper.

2. RAPIDMINER PLATFORM

The RapidMiner software was started in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Artificial Intelligence Unit of the Dortmund University of Technology [11]. RapidMiner, is a software for machine learning, data mining, predictive analytics, and business analytics [10]. It is used for research, education, training, application development, and industrial applications [10]. In a poll by KDnuggets, a data-mining newspaper, RapidMiner ranked second in data mining/analytic tools used for real projects in 2009 and was first in 2010. It is distributed under the AGPL open source license and has been hosted by SourceForge since 2004 [10]. RapidMiner can define analytical steps and be used for analyzing data generated by high-throughput instruments such as those used in genotyping, proteomics, and mass spectrometry. It can be used for text mining, multimedia mining, feature engineering, data stream mining, development of ensemble methods, and distributed data mining. RapidMiner functionality can be extended with additional plugins [11].

RapidMiner provides a GUI to design an analytical pipeline. The GUI produce an XML (eXtensible Markup Language) file that defines the analytical processes the user wishes to apply to the data [11]. Alternatively, the engine can be called from other programs or used as an API. Individual functions can be called from the command line. RapidMiner is open-source and is presented free of charge as a Community Edition released under the GNU AGPL. There is also an Enterprise Edition offered under a commercial license for integration into closed-source projects [11]. We use RapidMiner 5.0.000 beta. You can download it by the website http://www.rapidminer.com.

3. WEB LOG MINING

Information of web users appear the form of web server log files or log files. For every request from a browser of user to a web server, a reaction is produce automatically.So, Log file is the sum of all requests to the user's browser from a web server. This text file may be comma-delimited, space-delimited, or tab-delimited [1]**.**

3.1. Input Stage

We use EPA web log data available from the Internet Traffic Archive at http://ita.ee.lbl.gov/html/traces.html. Every line in EPA log file illustrates a special action requested by a browser of user [1]. In Fig.1 is shown a part of EPA web log file. We use EPA file in the first stage of our program.

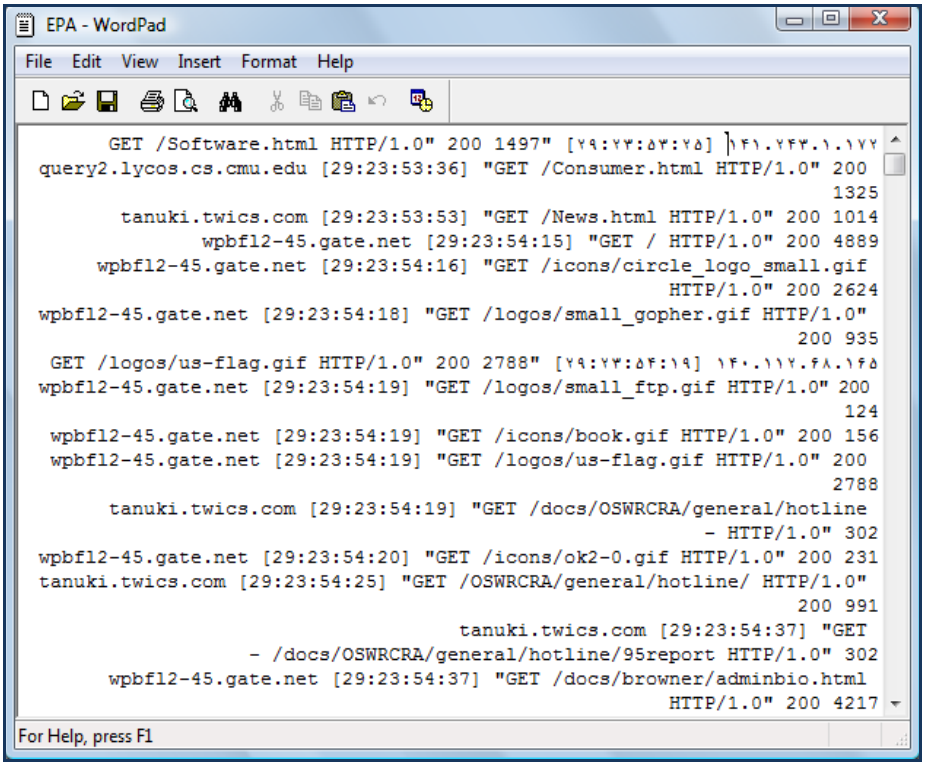


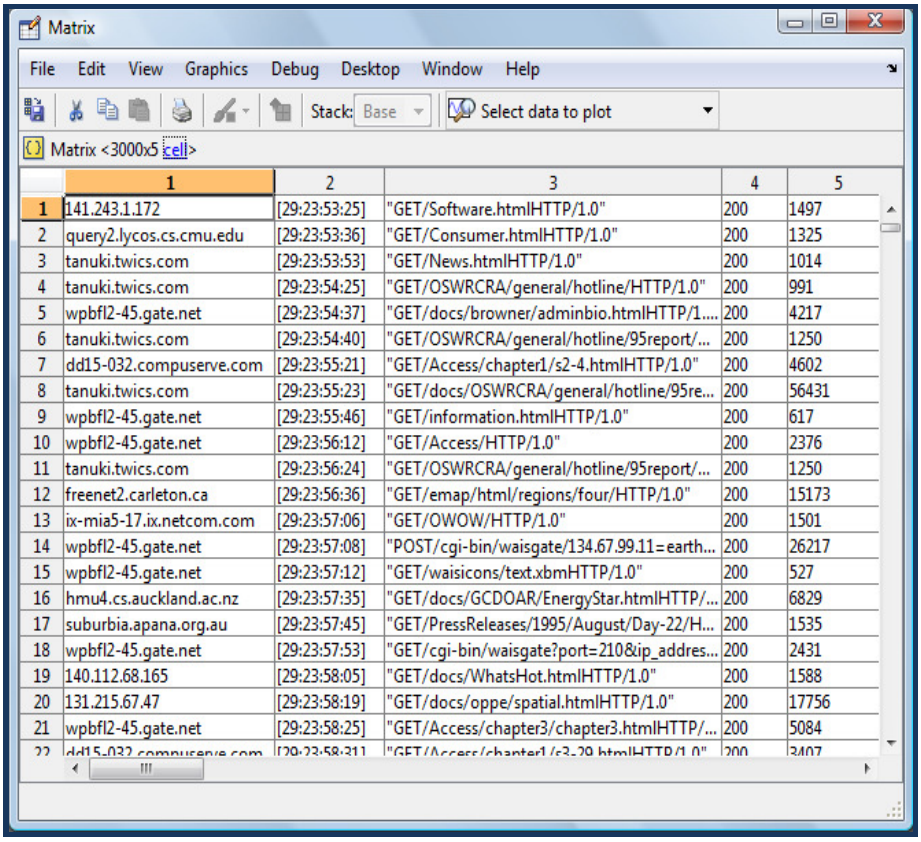
Figure 1. Part of EPA web log

3.2. Preprocessing Stage

The raw log files do not reach in a format conducive to useful data mining [1]. Hence, substantial data preprocessing must be applied. The most common preprocessing tasks are data cleaning and filtering, de-spidering, user identification, user session identification, and path completion. Preprocessing phase demands between 60 and 90% of the time essential for data analysis. Therefore, preprocessing phase help to the success rate of 75-90% to the all of process of knowledge discovery[12].

3.2.1. Data Cleaning And Filtering

In this paper, we use Matlab R2010A**.** First, the EPA web log with text format is loaded into Matlab,and convert to matrix format, and then is extracted IP Address, Date/Time Filed, HTTP Request, Protocol Version, Status Code and Transfer Volume Variables. Then the redundant files are deleted by Matlab rules, such as .gif entries, null requests and all of Status codes except 200 series. In Fig.2 is shown a part of EPA web log file after preprocessing stage.



	1	2	3	4	5
1	141.243.1.172	[29:23:53:25]	"GET/Software.htmlHTTP/1.0"	200	1497
2	query2.lycos.cs.cmu.edu	[29:23:53:36]	"GET/Consumer.htmlHTTP/1.0"	200	1325
3	tanuki.twics.com	[29:23:53:53]	"GET/News.htmlHTTP/1.0"	200	1014
4	tanuki.twics.com	[29:23:54:25]	"GET/OSWRCRA/general/hotline/HTTP/1.0"	200	991
5	wpbfl2-45.gate.net	[29:23:54:37]	"GET/docs/browner/adminbio.htmlHTTP/1....	200	4217
6	tanuki.twics.com	[29:23:54:40]	"GET/OSWRCRA/general/hotline/95report/...	200	1250
7	dd15-032.compuserve.com	[29:23:55:21]	"GET/Access/chapter1/s2-4.htmlHTTP/1.0"	200	4602
8	tanuki.twics.com	[29:23:55:23]	"GET/docs/OSWRCRA/general/hotline/95re...	200	56431
9	wpbfl2-45.gate.net	[29:23:55:46]	"GET/information.htmlHTTP/1.0"	200	617
10	wpbfl2-45.gate.net	[29:23:56:12]	"GET/Access/HTTP/1.0"	200	2376
11	tanuki.twics.com	[29:23:56:24]	"GET/OSWRCRA/general/hotline/95report/...	200	1250
12	freenet2.carleton.ca	[29:23:56:36]	"GET/emap/html/regions/four/HTTP/1.0"	200	15173
13	ix-mia5-17.ix.netcom.com	[29:23:57:06]	"GET/OWOW/HTTP/1.0"	200	1501
14	wpbfl2-45.gate.net	[29:23:57:08]	"POST/cgi-bin/waisgate/134.67.99.11=earth...	200	26217
15	wpbfl2-45.gate.net	[29:23:57:12]	"GET/waisicons/text.xbmHTTP/1.0"	200	527
16	hmu4.cs.auckland.ac.nz	[29:23:57:35]	"GET/docs/GCDOAR/EnergyStar.htmlHTTP/...	200	6829
17	suburbia.apana.org.au	[29:23:57:45]	"GET/PressReleases/1995/August/Day-22/H...	200	1535
18	wpbfl2-45.gate.net	[29:23:57:53]	"GET/cgi-bin/waisgate?port=210&ip_addres...	200	2431
19	140.112.68.165	[29:23:58:05]	"GET/docs/WhatsHot.htmlHTTP/1.0"	200	1588
20	131.215.67.47	[29:23:58:19]	"GET/docs/oppe/spatial.htmlHTTP/1.0"	200	17756
21	wpbfl2-45.gate.net	[29:23:58:25]	"GET/Access/chapter3/chapter3.htmlHTTP/...	200	5084
22	dd15-032.compuserve.com	[29:23:58:31]	"GET/Access/chapter1/s3-29.htmlHTTP/1.0"	200	3407

Figure 2. Part of EPA table after preprocessing

3.2.2. User Identification

User identification phase will be processed after preprocessing. This step should be to identify unique users. If you use firewalls and proxy servers will be complex to record this information [13]. In EPA web log, each user has individual IP address. So, each IP address represents different user.

3.2.3. User Session Identification

The purpose of user session identification is to determine the division of access each user has a separate session.The simplest method is to use an expiration time, ie the time spent in a page passes a certain threshold, it is assumed that the user has started a new session.. The default time for user session identification is thirty minutes [12]. In this paper for user session identification is considered 30 min expiration time.This default value is used in various studies [13,14].The end of this section, we obtain a ARFF file for association rule mining in RapidMiner platform. In Fig.3 is shown a ARFF file which derived from EPA log file.

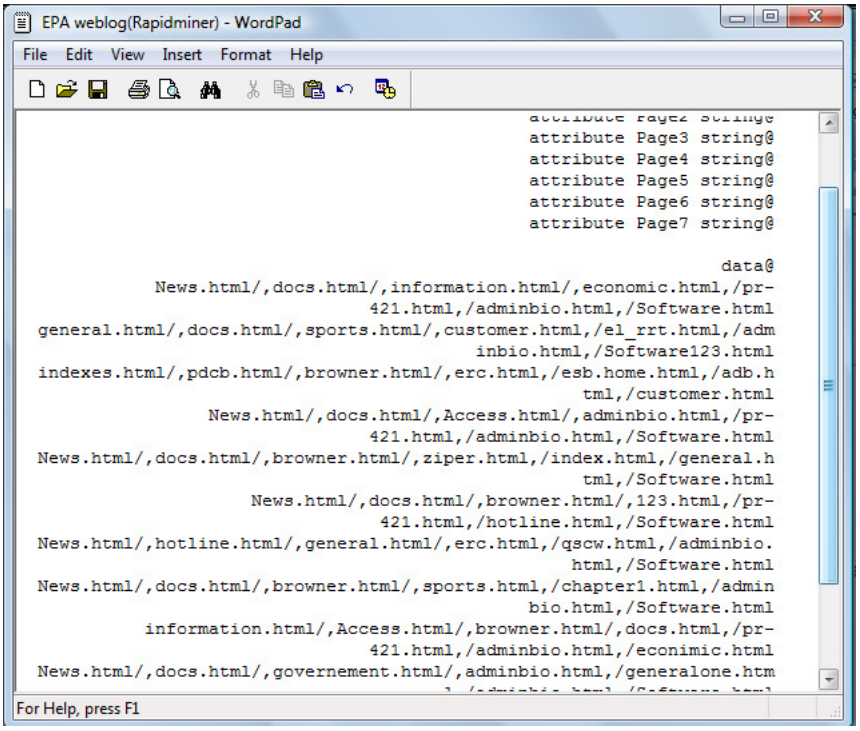


Figure 3. Part of ARFF file from EPA log file

4. ASSOCIATION RULE MINING IN RAPIDMINER

Association rule mining, one of the most important and well researched techniques of data mining [15]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [15]. Association rules scopes are web mining, risk management, telecommunication networks and etc [22]. Also, Association rules are widely used in web mining [16,17,18,19].

4.1. Pattern Discovery Stage

Pattern discovery is a key component. The phase algorithms and techniques from several research areas such as data mining, machine learning, statistical pattern recognition, it is convergent. The simplest method of log analysis as applied to the web usage mining process is taken into consideration .The goal of web usage mining, using statistical and data mining techniques to web log data preprocessing is useful for finding patterns .The most common and simplest method that can be used for such data to be analyzed .More advanced data mining techniques and algorithms appropriate for use in web domain consist of association rules, sequential pattern discovery, clustering and classification. The techniques adopted in the area, we used association rules by RapidMiner in this study**.**

The ARFF file obtained in section 3 is suitable for Rapidminer platform**.** Therefore, we are designed association rule mining in four block .First block of association rule mining process is " READ ARFF ". Thus, we read ARFF file and then convert to a matrix of discrete and binominal values. Next, we send the matrix for Fp_Growth block and then Association Rule Mining block execute. In Fig.4 is shown blocks of this process.

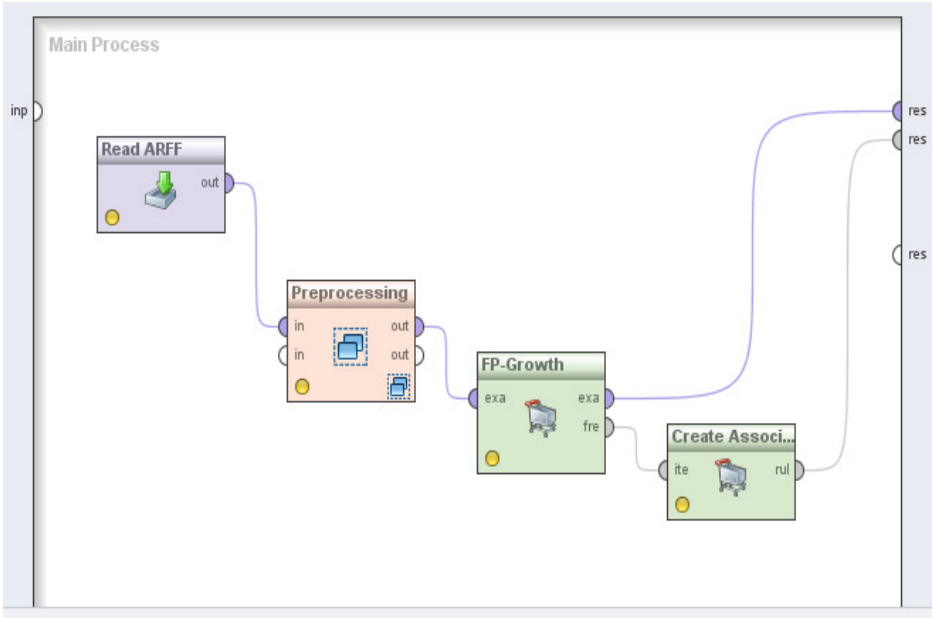


Figure 4. Flow of a association rule mining process

5. EXPERIMENTAL RESULT ANALYSIS

Association rules in web usage mining are used to find the relationship between user sessions that frequently appear together [20]. Each association rule is described by the ratio of Support and Confidence [21]. The Support of an item set is equal to the number of transactions that includes that item set. After setting mining parameter such as support, confidence, the web usage mining begins. The result of association rule mining is shown as figure 5.

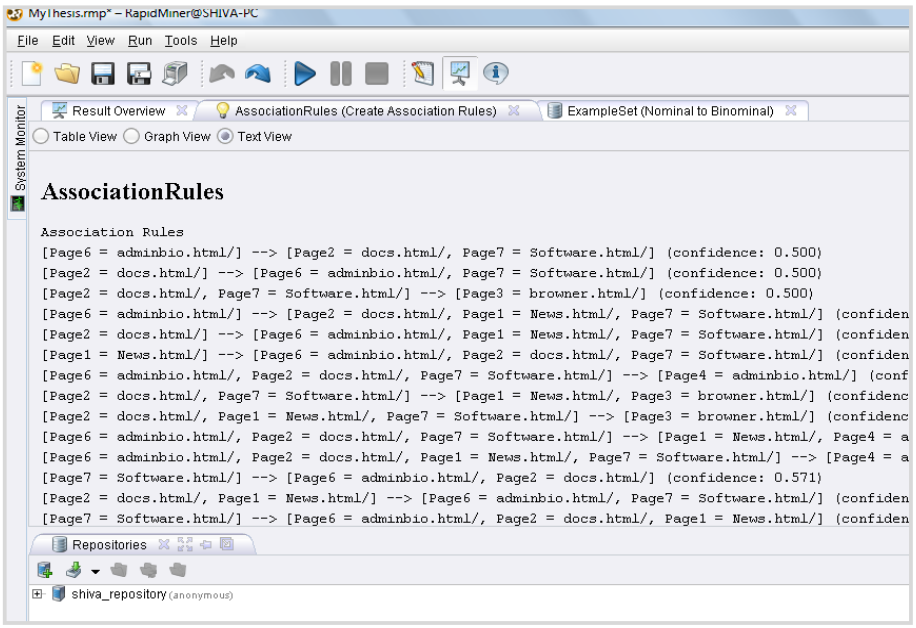


Figure 5. Some result of association rule mining

Some results in figure 5 are selected for analysis as table 1.

Table 1. Some association rule

RowNo.	Support	Confidence	Association Rule
1	0.5	1.000	[Page7 = Software.html/, Page3 = comp.html/] --> [Page2 = docs.html/]
2	0.667	1.000	[Page5 = politics.html/] --> [Page3 = religion.html/]
3	0.7	1.000	[Page3 = sports.html/] --> [Page6 = News.html/, Page1 = publicnews.html/]

5.1. Pattern Analysis Stage

The discovery of association rules in transactional web data have many advantages. For Example according to table1, A rule like the one with high reliability, This may indicate that most of users visit the online catalogs of software after seeing the software package. For example, if a website is no direct link between the Software.html/ and docs.html/ pages, discovery rule {Software.html/}=>{docs.html/} indicate that put a direct link between the two pages may help users find their favorite.
The second association rule shows that almost all users see politics and religion pages together. So, we recommend politics and religion pages together.

The third association rule shows that customers are more interested in the sports pages, reports in this field have also called for a visit.

6. CONCLUSIONS

In this study, association rules mining techniques are used in web usage mining. For this purpose, a web usage mining framework that consists Phases input, preprocessing, pattern discovery and pattern analysis is applied. Also, the RapidMiner tools for discovering association rules were used. The simulation was performed on the EPA log file of user behavior and characteristics of these data sets discovered. Analysis results show that using the RapidMiner in web usage mining, we can model the frequent user visits the website has achieved. Also, the rules for managing and optimizing the website structure and users are advised to be used.

REFERENCES

[1] M. Zdravco, D. T. Larose (2007) Data Minig the Web – Uncovering Patterns in Web Content, Structure and Usage. Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
[2] Xiu-yu Zhong, (2011)"The Research And Application of Web Log Mining Based On The platform Weka", Procedia engineering 15,pp 4073 – 4078.
[3] Yao-Te Wang, Anthony J.T. Lee,(2011)" Mining Web navigation patterns with a path traversal graph[J]", Expert Systems with Applications, Vol. 38, No. 6, pp 7112-7122.
[4] Michal Munk, Jozef Kapusta,(2010)" Data preprocessing evaluation for web log mining: reconstruction of activities of a webvisitor[J]", Procedia Computer Science, Vol. 1, No.1, pp 2273-2280.

[5] Resul Das, Ibrahim Turkoglu,(2009)" Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method[J]", Expert Systems with Applications, Vol. 36, No. 3, pp 6635-6644.
[6] Nicolas Poggi, Toni Moren,(2009)"Self-adaptive utility-based web session management[J]", Computer Networks, Vol. 53, No. 10,pp 1712-1721.
[7] J.M.Kleinberg and A.Tomkins,(1999)"Application of linear algebra in information retrieval and hypertext analysis",In Proc.18 th,ACM Symp. Principles of Database Systems (PODS), Philadelphia, PA, (5), pp 185-193.
[8] Enrique Lazcorreta, Federico Botella, (2008)"Towards personalized recommendation by two-step modified Apriori data mining Algorithm Expert Systems with Applications", Vol. 35, No. 3,pp 1422-1429.
[9] Shiyong Zhang ,Jianping Zeng, (2008)" A framework for WWW user activity analysis based on user interes"t . Knowledge-Based Systems, Vol. 21, No. 8,pp 905-910.
[10] http://dictionary.sensagent.com/RapidMiner/en-en/
[11] http://www.gabormelli.com/RKB/RapidMiner_System
[12] C. E. Dinuca, D. Ciobanu,(2011)" improving the session identification using the mean time", international journal of mathematical models and methods in applied sciences.
[13] Chu-Hui Lee, Yu-lung Lo, Yu-Hsiang Fu,(2011)"A novel prediction model based on hierarchical characteristic of web site", Expert Systems with Applications, Vol. 38, No. 4, pp 3422-3430.
[14] Claudia Elena Dinucă, Dumitru Ciobanu,(2011)"On an Algorithm for Identifying Sessions from Web Logs", Acta Universitatis Danubius, Vol. 7, No. 4.
[15] Sonali Manoj Raut, Dhananjay Dakhane,(2012)"Comparative Study of Clustering and Association Method for Large Database in Time Domain", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 12, pp. 41-45.
[16] X. Fu, J. Budzik, K.J. Hammond , (2000) " Mining navigation history for recommendation ", Intelligent User Interfaces, Vol. 10, No. 1.
[17] J. Li, O. R. Zaiane,(2004)"Combining usage, content and structure data to improve web site recommendation", 5th International Conference on Electronic Commerce and Web.
[18] M. Eirinaki, C. Lampos, S. Paulakis, M. Vazirgiannis, (2004) "Web personalization integrating content semantics and navigational patterns", Proceedings of the sixth ACM workshop on Web Information and Data Management WIDM.
[19] A.M. Wasfi, (1993) "Collecting user access patterns for building user profiles and collaborative filtering", Proceedings of the 1999 International Conference on Intelligent User Interfaces.
[20] Jiawei Han, Micheline Kamber,(2001)Data Mining: Concepts and Techniques, Morgan Kaufmann.
[21] http://www.di.unito.it/~meo/Pubblicazioni/capitolo-libro-IGI-rare-patterns.pdf
[22] http://www.csis.pace.edu/~ctappert/dps/d861-13/session2-p1.pdf

Authors

Shiva Asadianfam was born in 1986 in Iran. Ms. Asadianfam received her B.Engr. degree from University of Urmia, and her M.S. degree from Islamic Azad University, Zanjan branch (Zanjan, Iran) in computer engineering. Her research interests include data mining, web mining and recommendations.

Masoud Mohammadi was born in Iran. He is studying his Phd degree from Islamic Azad University, Tehran branch (Tehran, Iran) in computer engineering.