ISSN:0975-9646

Web Usage mining for Predicting User Access Behaviour

Amit Dipchandji Kasliwal[#1], Dr. Girish S. Katkar[#2]

[#1]*Dept. of Computer Science, M.S.G. College, Malegaon, Nashik,Maharashtra, India*

[#2] *Dept. of Computer Science, Arts, Sci., Commerce College, Koradi ,Nagpur, Maharashtra, India*

Abstract— **In current trend of computer science, there are lots of software available for data mining, predictive analytics, and business analytics. In every sort of software where expectation comes for prediction then there is need for a web log files of user visit to particular website stored at server side. The interesting information for extracting knowledge from such huge data demands for new logic and new methods. In this paper work, we proposed a web usage mining method with a well-known mining tool software. At first, we are having a log file from KDD repository then we performed deletion on the redundant information from file and then we mining log file which has been submitted with mining tool, then going to get the customized form of access done by different user to visit the website by performing processing and analyzing phase on log file, and mine for so called unusual rules, and suggesting rules the as reference for the decision making and construction of website. At the end, we collected experimental result and analysis on result show that, applying web usage mining, will can look of frequent model which user visits the website, manage to optimize the website structure and recommends for users.**

Keywords- **Web Log File, Web Usage Mining, Knowledge Discovery, Association Rule Mining, Web Log.**

I. INTRODUCTION

A process of discovery of knowledge from a large set of database was suggested by Srivastava. That process of extracting interesting information from web log file is called as Web Usage Mining, it is process consists of four task from start to reading web log file till developing the virtualization on the web log file referred as the input phase, the preprocessing phase, the pattern discovery phase, and the pattern analysis phase.

In the input phase, few types of raw web server log files are retrieved and saprated by their records as access, referrer and agent logs. In Preprocessing phase, the raw log files that do not reach in a format as useful for data mining and specially for web usage mining. Therefore, necessary data preprocessing must be performed. IN preprocessing, the most common tasks are data cleaning, data filtering, session identification, user identification, and path completion. In Pattern discovery phase, use of data mining methods is done with the intension of extracting or discovering models[1]. In data mining methods and techniques, it includes algorithm for building association rules, classification algorithms, clustering algorithms, sequential patterns and standard statistical analysis. As per the researcher's involvement, not all of the models handled in the pattern discovery phase but fairly speaking it would be considered beneficial it any of them is suitable or there research work. In last phase Pattern analysis, individual analysts follow the output that they got from the pattern discovery phase and serve the interesting or more knowledgeable, beneficial and actionable models [1].

Many experts have convinced on the theoretical discussion, some papers on the web server log mining have had initial result [7]. An author named Y.T Wang recommended structure to record information about the navigation paths of website visitors using a compact graph [11]. Another named Jozef Kapusta provided a suggestion for optimize web portal in level of portal adaptation through user and access hour on portal [9]. 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 techniques to use machine learning and Markov-chain models to build Self-adaptive utility-based web session management [6]. Supported by the techniques proposed by Kleinberg and Tomkins for web structure mining [2]. Federico proposed a new method towards automatic personalized recommendation by algorithm Apriori [8]. Shiyong Zhang suggested a framework that uses a hidden Markov chain [3].

This paper presenting the work strongly working towards web usage mining for predicting access behavior. For this, We are using web log file based on developed platform named RapidMiner. We analysis EPA web log file and then provided it to RapidMiner, and will obtain frequent model which user visits the website, manage to optimize the website structure. The actual processing of this paper is presented as follows. In another section we describing RapidMiner software. The use of MatLab is for implementation of web log file in presented in another part. Next part shows our simulation on the platform and our simulated results is presented next. And at last, we have concluding about the paper.

II. WEB USAGE MINING: A PROCESS

Web mining can be defined as discovering and analysing of useful information from the World Wide Web in the

form of web access log data. Based on the different emphasis and different ways to obtain information of web users appear the form of web log files or access log files. For every user's request from a browser to a web server, a response is produce automatically in a file stored at server side of hosted website. So, Log file is the collection of that all requests to the user's browser from a web server. This text file may be comma delimited, space-delimited, or tab-delimited. Following is the way that is follow for getting the knowledgeable information form log file.

A. Input Stage

For our research purpose, we are using a web log data file available from the data repository at http://archive.org. Every line in log file represents a request made by a browser of user. Fig.1 shows the web log file and we are using it for mining process in my program.

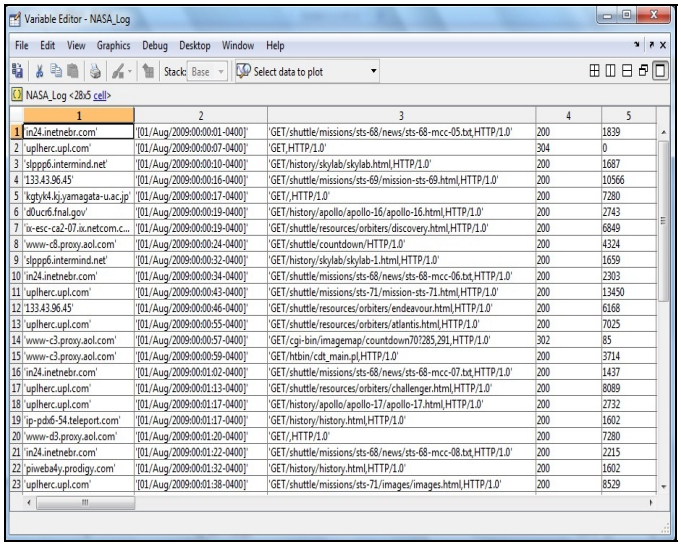
```
in24.inetnebr.com [01/Aug/2009:00:00:01-0400]
"GET/shuttle/missions/sts-68/news/sts-68-mcc-05.txt
HTTP/1.0" 200 1839
uplherc.upl.com [01/Aug/2009:00:00:07-0400] "GET
HTTP/1.0" 304 0
slppp6.intermind.net [01/Aug/2009:00:00:10-0400]
"GET/history/skylab/skylab.html HTTP/1.0" 200 1687
133.43.96.45 [01/Aug/2009:00:00:16-0400]
"GET/shuttle/missions/sts-69/mission-sts-69.html
HTTP/1.0" 200 10566
kgtyk4.kj.yamagata-u.ac.jp [01/Aug/2009:00:00:17-0400]
"GET/ HTTP/1.0" 200 7280
d0ucr6.fnal.gov [01/Aug/2009:00:00:19-0400]
"GET/history/apollo/apollo-16/apollo-16.html HTTP/1.0"
200 2743
ix-esc-ca2-07.ix.netcom.com [01/Aug/2009:00:00:19-
0400] "GET/shuttle/resources/orbiters/discovery.html
HTTP/1.0" 200 6849
www-c8.proxy.aol.com [01/Aug/2009:00:00:24-0400]
"GET/shuttle/countdown/HTTP/1.0" 200 4324
slppp6.intermind.net [01/Aug/2009:00:00:32-0400]
"GET/history/skylab/skylab-1.html HTTP/1.0" 200 1659
in24.inetnebr.com [01/Aug/2009:00:00:34-0400]
"GET/shuttle/missions/sts-68/news/sts-68-mcc-06.txt
HTTP/1.0" 200 2303
```

Fig. 1 Sample of web log file

B. Data Preprocessing

The raw log files do not reach in a format that can be adjusted to conduct useful data mining and therefore, essential data preprocessing must be applied. The most common preprocessing tasks are data cleaning, data filtering, user identification, session identification and path extraction. Data preprocessing is very essential for data analysis. Therefore, data preprocessing help to achieve the complete success in getting knowledgeable data tor the process of knowledge discovery.

1) Data Cleaning and Data Filtering: For our data preprocessing phase, We are using MatLab 7.0(R2011a). First of all, the web log which is in text format is loaded into MatLab and converted to matrix format using MatLab's tool for Matrix Formation then it's extracted IP Address, TimeStamp(Date/Time Field), Request (in HTTP form), Protocol Version, Status Code and Data Transfer Volume Variables. Then the redundant record from the web log files are deleted by MatLab rules, such as .gif entries, null requests and all the record with status code except 200 and its standard code series. Fig.2 is shows web log file after preprocessing stage.



	1	2	3	4	5
1	'in24.inetnebr.com'	'[01/Aug/2009:00:00:01-0400]'	'GET/shuttle/missions/sts-68/news/sts-68-mcc-05.txt,HTTP/1.0'	200	1839
2	'uplherc.upl.com'	'[01/Aug/2009:00:00:07-0400]'	'GET,HTTP/1.0'	304	0
3	'slppp6.intermind.net'	'[01/Aug/2009:00:00:10-0400]'	'GET/history/skylab/skylab.html,HTTP/1.0'	200	1687
4	'133.43.96.45'	'[01/Aug/2009:00:00:16-0400]'	'GET/shuttle/missions/sts-69/mission-sts-69.html,HTTP/1.0'	200	10566
5	'kgtyk4.kj.yamagata-u.ac.jp'	'[01/Aug/2009:00:00:17-0400]'	'GET/,HTTP/1.0'	200	7280
6	'd0ucr6.fnal.gov'	'[01/Aug/2009:00:00:19-0400]'	'GET/history/apollo/apollo-16/apollo-16.html,HTTP/1.0'	200	2743
7	'ix-esc-ca2-07.ix.netcom.c...	'[01/Aug/2009:00:00:19-0400]'	'GET/shuttle/resources/orbiters/discovery.html,HTTP/1.0'	200	6849
8	'www-c8.proxy.aol.com'	'[01/Aug/2009:00:00:24-0400]'	'GET/shuttle/countdown/HTTP/1.0'	200	4324
9	'slppp6.intermind.net'	'[01/Aug/2009:00:00:32-0400]'	'GET/history/skylab/skylab-1.html,HTTP/1.0'	200	1659
10	'in24.inetnebr.com'	'[01/Aug/2009:00:00:34-0400]'	'GET/shuttle/missions/sts-68/news/sts-68-mcc-06.txt,HTTP/1.0'	200	2303
11	'uplherc.upl.com'	'[01/Aug/2009:00:00:43-0400]'	'GET/shuttle/missions/sts-71/mission-sts-71.html,HTTP/1.0'	200	13450
12	'133.43.96.45'	'[01/Aug/2009:00:00:46-0400]'	'GET/shuttle/resources/orbiters/endeavour.html,HTTP/1.0'	200	6168
13	'uplherc.upl.com'	'[01/Aug/2009:00:00:55-0400]'	'GET/shuttle/resources/orbiters/atlantis.html,HTTP/1.0'	200	7025
14	'www-c3.proxy.aol.com'	'[01/Aug/2009:00:00:57-0400]'	'GET/cgi-bin/imagemap/countdown707285,291,HTTP/1.0'	302	85
15	'www-c3.proxy.aol.com'	'[01/Aug/2009:00:00:59-0400]'	'GET/htbin/cdt_main.pl,HTTP/1.0'	200	3714
16	'in24.inetnebr.com'	'[01/Aug/2009:00:01:02-0400]'	'GET/shuttle/missions/sts-68/news/sts-68-mcc-07.txt,HTTP/1.0'	200	1437
17	'uplherc.upl.com'	'[01/Aug/2009:00:01:13-0400]'	'GET/shuttle/resources/orbiters/challenger.html,HTTP/1.0'	200	8089
18	'uplherc.upl.com'	'[01/Aug/2009:00:01:17-0400]'	'GET/history/apollo/apollo-17/apollo-17.html,HTTP/1.0'	200	2732
19	'ip-pdx6-54.teleport.com'	'[01/Aug/2009:00:01:17-0400]'	'GET/history/history.html,HTTP/1.0'	200	1602
20	'www-d3.proxy.aol.com'	'[01/Aug/2009:00:01:20-0400]'	'GET/,HTTP/1.0'	200	7280
21	'in24.inetnebr.com'	'[01/Aug/2009:00:01:22-0400]'	'GET/shuttle/missions/sts-68/news/sts-68-mcc-08.txt,HTTP/1.0'	200	2215
22	'piweba4y.prodigy.com'	'[01/Aug/2009:00:01:32-0400]'	'GET/history/history.html,HTTP/1.0'	200	1602
23	'uplherc.upl.com'	'[01/Aug/2009:00:01:38-0400]'	'GET/shuttle/missions/sts-71/images/images.html,HTTP/1.0'	200	8529

Fig.2. Web Log table after preprocessing

2) User Identification: To identify users especially unique users User identification task processed after data preprocessing. It is somewhat tedious work to record this information if user is browsing with system on which firewalls and proxy servers are installed. In web log, each user has individual IP address representing different user.

3) Session Identification: To determine the division of access each user has a separate session, session identification task is perform. This can be done using simplest method as to use an expiration time. The default time for user session identification is thirty minutes [4] so the expiration time means the time spent in a page passes a certain threshold, it is assumed that the user has started a new session. In this paper, I have considered 30 min expiration time for user session identification. And, at last of this phase I prepared the a suitable file format in ARFF file for association rule and that can be used for mining in RapidMiner platform. Fig.3 shows a ARFF file which prepared from web log file.

Fig.3 ARFF file prepared from Web log file

III. WEB MINING PLATFORM

In data mining, Association rule mining, one of the most important and well researched techniques in the current days. The aims behind it is the extracting or finding or building association rule to extract interesting relations or corelation, frequent items or patterns, associated or casual structures among sets of items in the transaction databases

or other data repositories. Association rules scopes are web mining, risk management, telecommunication networks and etc. Also, Association rules are widely used in web mining.

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 during preprocessing is suitable for Rapidminer platform. Therefore, we are designed association rule mining using some supported tools of Rapidmine. IN first tool for building association rule is 'READ ARFF'. We read ARFF file and then converted it to a matrix of numeric, nonnumeric, discrete and binominal values. We processed it forward to MatLab for creating matrix so that it can be accepted by 'Fp_Growth' and then Association Rule Mining block for final execution. In Fig.4 is shows this process.

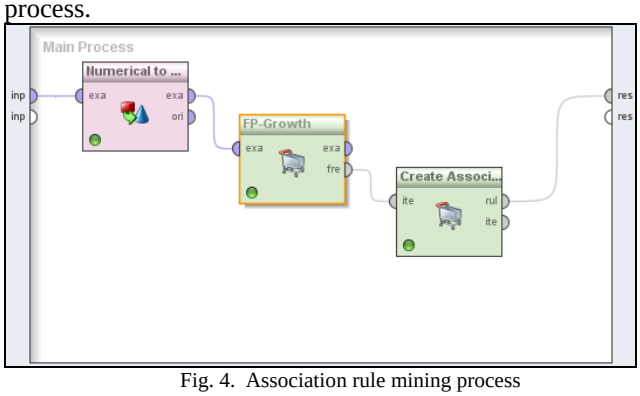


Fig. 4. Association rule mining process

The software RapidMiner was developed and implemented in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Artificial Intelligence Unit of the Dortmund University of Technology. The RapidMiner, is a well suited software for machine learning, data mining, predictive analytics, and business analytics and also recommended by many analysts also. It is used for future research, analytical education and training, application development, and industrial applications as per the industrial requirement. The wel known forum on Data Mining and Artificial Intelligence is KDnuggets had a poll about recent trends in data mining, a data mining newspaper, RapidMiner ranked second in data analytic tools used for real projects in 2009 and was first in 2010. It is distributed and made available under the AGPL open source license and has been hosted by SourceForge from last 10 years since 2004. RapidMiner can define analytical steps and be used for analyzing data generated for web mining, text mining, multimedia mining, feature engineering, data stream mining, development of ensemble methods, and distributed data mining. It can be used for analyzing the data generated by high throughput embedded devices. RapidMiner functionality can be extended with additional plugins.

RapidMiner is open-source and is available freely as freeware as a Community Edition released under the GNU AGPL. RapidMiner provides a wel suited user interface to design an analytical process through pipeline. Alternatively, the engine can be called from other programs or used as an API. Individual functions can be called from the command line. The GUI produce an XML file that defines the analytical processes the researcher or user wishes to apply to the data. There is also an Enterprise Edition offered under a commercial license for integration into closed-source projects. We use RapidMiner 5beta version. It is available for download at the website http://www.rapidminer.com.

IV. Result Analysis

Association rules in web usage mining are used to find the relationship between user sessions that appear together frequently [20]. In each association rule that described here is with respect to the relation of Support count and Confidence count. 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 in figure 5.

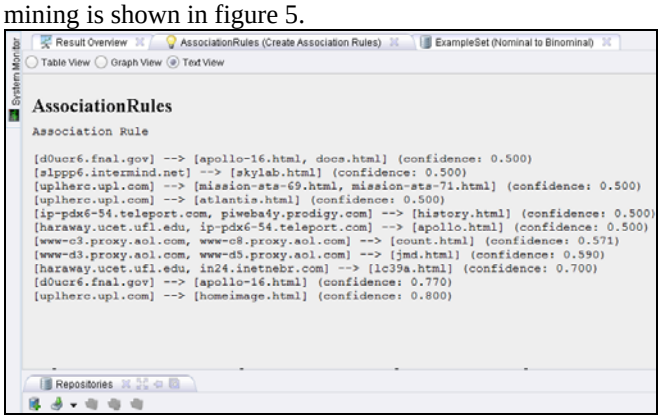


Fig 5. Association Rule

As we can see in the above figure, that after accessing page1 of address '*uplherc.upl.com*' the user also accessing the page3 '*mission-sts-71.html*' and so on. The table 1 given bellow is illustrating the use of various page accesses with respect to their support count and confidence count. These results are supposed to be for analysis.

Table 1

Rule No	Support Count	Confidence Count	Association Rule
1	1.000	0.500	[d0ucr6.fnal.gov] --> [apollo-16.html, docs.html]
2	1.000	0.500	[uplherc.upl.com] --> [mission-sts-69.html, mission-sts-71.html]
3	0.800	0.571	[www-c3.proxy.aol.com, www-c8.proxy.aol.com] --> [count.html]
4	0.750	0.590	[www-d3.proxy.aol.com, www-d5.proxy.aol.com] --> [jmd.html]
5	0.750	0.700	[haraway.ucet.ufl.edu, in24.inetnebr.com] --> [lc39a.html]
6	0.600	0.800	[uplherc.upl.com] --> [homeimage.html]

In web usage mining process, web log data is very important for extracting the hidden pattern and discovering the hidden rule in between access. As an example, the table 1 prepared for analyzing purpose shows the some discovery of association rule that might help the website administrator to improve the website structure so that the web pages could be available for user on first attempt. The rule set we got here according to that the rule like 1 has the one with high reliability, as it indicate that most of users visit the government website they also visits the Apollo-16.html webpage from another website. This result has analyzed only for study purpose. In another rule like rule 5, if users are accessing the two different websites then they are also accessing the page lc39a.html means this shows that the a web page with high confidence count can also have the considerable support count and depending on that website admin can make the link to that page from this multiple stes so that it can help the user to access. I rule 2, if there is the maximum support count for the rule then website administrator can make the changes or admin can consider the changes in website structure to make direct link between the uplherc.upl.com website to the pages 'mission-sts-69.html' and to the page 'mission-sts-71.html' and hence rule [uplherc.upl.com] --> [homeimage.html] indicate that the admin can put a direct link between the two pages that may help users to find their interest. The association rule like 3, 4 shows that almost 60% users visiting 'count.html' and 'jmd.html' pages together. So, we recommend these two pages linked overseas together.

V. CONCLUSIONS

In this research study, in web usage mining association rules mining techniques are used. For this purpose, a web usage mining platform consisting some tools for input phase, preprocessing and pattern discovery. In this work, we performed pattern analysis also. The framework named RapidMiner is used as a tool for discovering association rules which is used for pattern analysis. All research work is performed on the NASALog file of user behavior and characteristics of these data sets discovered. With the results shown in analysis prepared using the RapidMiner in web usage mining, it help to model the frequent user visits the website had accessed during the specific period. Also, the rules for managing and optimizing the website structure and users are discussed can be suggested to be used.

REFERENCES

[1]M. Zdravco, D. T. Larose (2007) Data Mining the Web – UncoveringPatterns in Web Content, Structure and Usage. Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

[2] 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.

[3] 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.

[4] C. E. Dinuca, D. Ciobanu,(2011)" improving the session identification using the mean time", international journal of mathematical models and methods in applied sciences.

[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] Xiu-yu Zhong, (2011)"The Research And Application of Web Log Mining Based On The platform Weka", Procedia engineering 15,pp 4073 – 4078.

[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] 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.

[10]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.

[11]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.

[12]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.

[13]Claudia Elena Dinuca, Dumitru Ciobanu,(2011) "On an Algorithm for Identifying Sessions from Web Logs", Acta Universitatis Danubius, Vol. 7, No. 4.

[14]Jiawei Han, Micheline Kamber,(2001)Data Mining: Concepts and Techniques, Morgan Kaufmann.

[15]X. Fu, J. Budzik, K.J. Hammond , (2000) " Mining navigation history for recommendation ", Intelligent User Interfaces, Vol. 10, No. 1.

[16] 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.

[17] www.archive.org
[18] www.rapidminer.com
[19] www.kdnuggets.com
[20] www.csis.pace.edu
[21] www.di.unito.it
[22] dictionary.sensagent.com
[23] msdn.microsoft.com