Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Effective Web Log Mining: An Implementation View

Ankita Kusmakar
M. Tech Scholar Computer Science
Department L.N.C.T, Bhopal, India

Sadhna Mishra
Professor Computer Science
Department L.N.C.T, Bhopal, India

Vineet Richhariya
HOD Computer Science
Department L.N.C.T, Bhopal, India

Abstract— With the increase in popularity of web in the past few decades, web mining has attracted lots of attention. An important area in Web mining is Web usage mining, the discovery of patterns in the browsing and navigation data of Web users.* Web *log mining is application of data mining techniques to discover usage patterns from web data in order to better serve the needs of web based applications. The user access log files present very significant information about web server. This paper describes about analysis of Web Log Data to find information about a web site, top priority pages, navigation pattern of the users of particular website and other information which will help system administrator and Web designer to improve their system by determining the patterns and the usage of web pages. The results thus obtained from the study will be used in the further development of the web site in order to increase its effectiveness.

Keywords— web usage mining, web log, navigation pattern, data mining

I. INTRODUCTION

Web usage mining (WUM) is the process of applying data mining techniques to the discovery of usage patterns from web data [1]. It can be used for various purposes such as personalization, system improvement and site modification etc. But the main problem with the web mining and specifically with web usage mining is the nature of the data that they are dealing with. As the use of internet has increased rapidly during the past few decades, the data available on web has also increased tremendously and a lot of transactions are taking place by seconds. The rapidly growing knowledge available on World Wide Web has been lacking an integrated structure or schema, so it has become difficult for users to access relevant information efficiently. Also the substantial increase in the number of websites presents a challenging task for web administrators to organize the contents of websites to cater to the need of users. Thus a lot of pre-processing and parsing is needed before the actual extraction of the required information.

In this paper, we present a framework for extracting navigation pattern of users in the tabular form. This table contains clusters of pattern of pages visited by user**.** The rest of the paper is organized as follows: Section 2, briefly discusses about phases of web usage mining. Section 3 reviews some previous approaches related to our work. Section 4 discusses about various steps of proposed work. Section 4 explains the proposed work. Finally, section 5 summarizes the paper and introduces future work.

II. Web Usage Mining Process

There are three main phases of web log mining as shown in figure 1.These phases are briefly explained as follows:

a) Preprocessing: This is the first step of web mining process. It retrieves raw data from the web resources and transforms the data into a format that will be more easily and effectively processed.

b) Pattern Discovery: Mines effective, potentially useful and understandable information and knowledge using mining algorithm. Includes techniques such as clustering, classification, association rule etc.

c) Pattern Analysis: This is the last step in the overall web usage mining process. The goal is to eliminate the irrelevant patterns and to extract the interesting patterns from the output of pattern discovery process.

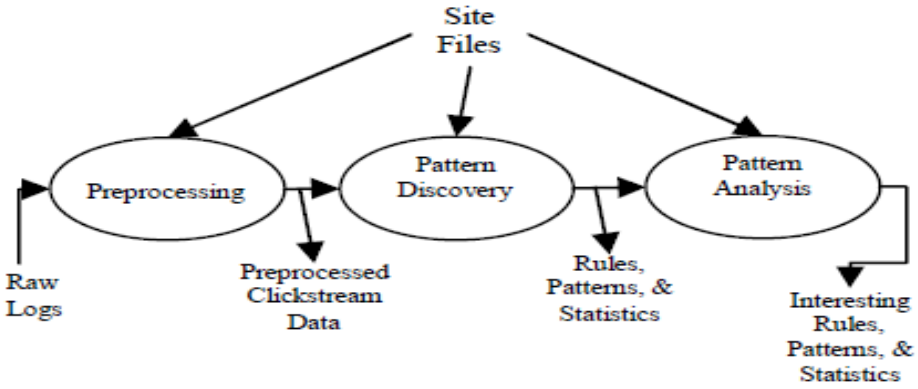


Fig.1. High Level Web Usage Mining

III. RELATED WORK

The focus of the literature survey is to study or collect information about web usage mining which is used to find out web navigation behavior of user. Data mining efforts associated with the Web, called Web mining, can be broadly categorized into three areas of interest based on which part of the Web to mine; Web Content mining, Web Structure mining, and Web Usage Mining (Kosala and Blockeel) [2]. Baraglia and Palmerini [3] proposed a WUM system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance. Yi Dong, Huiying Zhang and Linnan Jiao [4] proposed a new recommendation method to be applied to web log mining by integrating user clustering and association rule mining techniques to improve the effectiveness of electronic commerce. However, the resulting association patterns did not perform well in predicting future browsing patterns. Jalali et al. (2008a [5] and 2008b [6]) proposed a system for discovering user navigation patterns using a graph partitioning model. An undirected graph based on connectivity between each pair of Web pages was considered and weights were assigning to edges of the graph. Dixit and Gadge [7] presented another user navigation pattern mining system based on the graph partitioning. An undirected graph based on connectivity between Referrer and URI pages was presented along with a preprocessing method to process unprocessed web log file and a formula for assigning weights to edges of the undirected graph. But all the graphical approaches are complex in nature and also very cost expensive. An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner[8] by M.P. Yadav, P.Keserwani, S.Ghosh is an approach involving the support and confidence of sequential pattern of web pages and candidate set pruning to reduce the repetitive scanning of database containing the web usage information and thus reducing the time. However the algorithm produces the partial result with a limited set of information which turns out to be a great demerit of the algorithm. The algorithm only list the item sets and the IP addresses from where these item sets were accessed. This partial information may be helpful in some cases but not always; this raises an expectation for better and efficient algorithm. Bhushan, R and Nath R[9] propose a web recommendation approach which is based on learning from web logs and recommends user a list of pages which are relevant to him by comparing with user's historic pattern. Finally, search result list is optimized by re-ranking the result pages.

IV. PROPOSED WORK

We would like to propose a system which would discover interesting patterns from the weblogs. This section shall be discussing about the methodology and implementation used in this study. This section presents the steps involved in the server log process and analysis. Also discusses the preprocessing of the server logs and pattern mining. The methodology consists of the following phases.

A) Collection of Web Usage Data :

Log data is the primary source of data which is collected automatically in Web servers. Server log is a log file automatically created and maintained by a server of activity performed by it. It is a file to which the Web server writes information each time a user requests a resource from that particular site.

In this phase, raw log file were collected from Tutor.com. This portal focuses on education and provides more information related to education purposes such as tutorials, Student Information, Teaching material, and etc. for the analysis purposes, data dated on 24 November that consists of 82 683 records was retrieved from the server and needed to be preprocessed.

The raw log files consists of 19 attributes such as Date, Time, Client IP, AuthUser, ServerName, ServerIP, SetverPort, Request Method, URI-Stem, URI-Query, Protocol Status, Time Taken, Bytes Sent, Bytes Received, protocol Version, Host, User AGENT, Cookies, Referer. One of the main problems encountered when dealing with the log files is the amount of data needs to be preprocessed.

B) Preprocessing:

In this phase, the starting point and critical point for successful log mining is data preprocessing. The required tasks are data cleaning, user identification and session identification.

An entry of web server log contains the time stamp of a traversal from a source to a target page, the IP address of the originating host, the type of request (GET and POST) and other data, many entries that are considered uninteresting for mining were removed from the data files. The filtering is an application dependent. While in most cases accesses to embedded content such as image and scripts are filtered out. However, before applying data mining algorithm, data preprocessing must be performed to convert the raw data into data abstraction necessary for the further processing (see Table 1).

TABLE1 sample of preprocessed log file

S.No.	Datetime	Server IP	Method	URI Stem	Port	Client IP
1	2012-05-23 00:04:05	69.10.57.50	GET	/MAHindi.aspx	80	100.43.83.146
2	2012-05-23 00:16:13	69.10.57.50	GET	/DeptofElectronics.aspx	80	66.249.71.162
3	2012-05-23 00:43:03	69.10.57.50	GET	/robots.txt	80	207.46.13.114
4	2012-05-23 00:43:40	69.10.57.50	GET	/Labs.aspx	80	207.46.13.114
5	2012-05-23 00:43:42	69.10.57.50	GET	/Labs.aspx	80	207.46.199.33
6	2012-05-23 00:52:32	69.10.57.50	GET	/AboutUs.aspx	80	66.249.71.162
7	2012-05-23 00:54:42	69.10.57.50	GET	/DeptofMusic.aspx	80	180.76.5.177

There are many more attributes other then specified above. After preprocessing completed, the pattern mining was performed to mine the access pattern.

C) Pattern Mining :

On completion of preprocessing, pattern mining can be performed in order to find the useful patterns that can be used to improve the sites. Pattern mining will return several findings such as:-

1. General statistics
   General statistics are the summary of the whole log file. Usually it provides the Total Hits, Page Views, and Total Visitor.
2. Access Statistics
   Access statistics provides information such as Most Popular Access Page and Most Downloaded Files.
3. Visitors Information
   Visitor's information will provide the information such as the most active country which accesses the website.
4. Referrer
   Referrer will provide information such as the most used search engines and phrases, and keyword used.
5. Error
   Error is important for the system administrator's website in order to improve the site as well as to reduce the error such as "404 file not found".

D) Generalized Association Rules:

In this phase, Generalized Association Rules is used to mine the data in order to obtain the support and confidence for each rule. Generalized association rule is one of the commonly used web usage mining technique. Once pattern mining analysis has been successfully executed, generalized association rules were applied to mine the useful patterns using support and confidence counting. From the server logs, hierarchy of the websites is determined. To perform this task, generalized association rules is applied until level 3. Comparing with the standard association rules, generalized association rules allow rules at different levels. Generalized association rules were also used to tackle the data diversity problems.

The two important measures of association rule mining are *support* and *confidence* which are briefly discussed below

1. SUPPORT: Support measure how often the rules occur in database. To determine the support for each rules produced, several arguments have been identified in calculating the support such as Total Transaction in database and number of occurrence for each rules. The formula for support is shown below.
   Input:-Total Transaction in Database
   No. of occurrences each item {x,y}
   $$\text{SUPPORT} = \frac{\text{support count of XUY}}{\text{Total transaction in Database}}$$
2. CONFIDENCE: It is a measure of strength of association rules. Confidence is used how confident we are about our patterns By finding confidence we get information regarding tendency to appear web pages after another one. The formula for confidence is shown below

$$\text{CONFIDENCE} = \frac{\text{support (XUY)}}{\text{Support(X)}}$$

E) RESULTS:

The results of web log processing are described in two ways. The general description about the access patterns, Access statistic, Visitors, Referrer, Error will be displayed. In addition, the support and confidences of different levels of server portal accessed will be illustrated. The system administrator could make a decision from the result illustrated in order to improve or enhance the content, link, site navigation and facilities.

Finally, the flowchart summarizes all the steps of the proposed work as shown in figure 3.

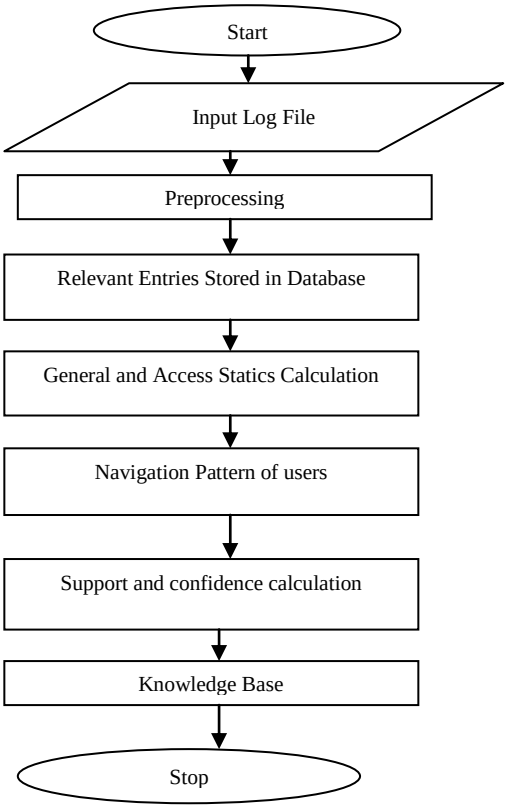


Fig.2. Flowchart of Proposed Work

V. Experimental Results and Evaluation Measures

A) Experimental Results

Implementation is done in dot net framework all programs are written in Microsoft Visual studio 2010 as front end and SQL server for database .We have taken input as web log file. A sample web log file is as follows:

#Software: Microsoft Internet Information Services 7.0
#Version: 1.0
#Date: 2012-05-23 00:04:05
#Fields: date time s-sitename s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-Agent) sc-status sc-substatus sc-win32-status sc-bytes cs-bytes time-taken 2012-05-23 00:04:05 W3SVC115 69.10.57.50 GET /MAHindi.aspx -80-100.43.83.146
Mozilla/5.0+(compatible;+YandexBot/3.0;++http://yandex.com/bots) 404 0 0 3739 279 631 2012-05-23 00:16:13 W3SVC115 69.10.57.50 GET /DeptofElectronics.aspx - 80 - 66.249.71.162 Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html) 404 0 0 3759 255 50

Selection of the log file, preprocessing step and database record storage has been shown in fig.3.

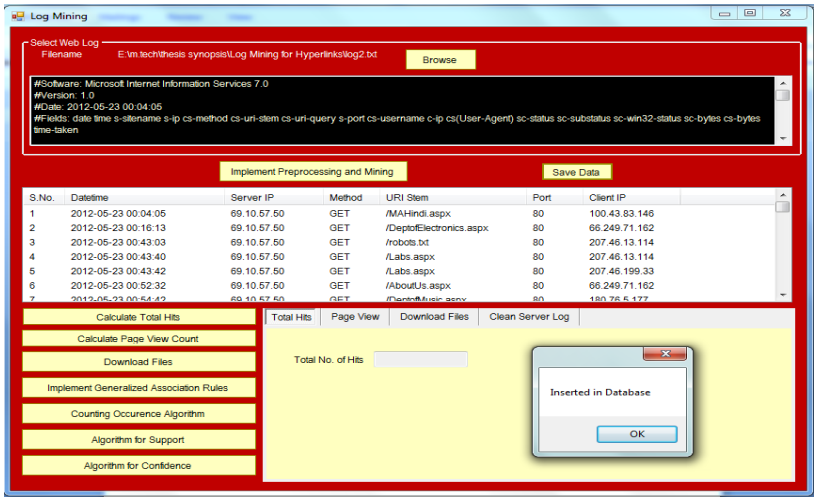
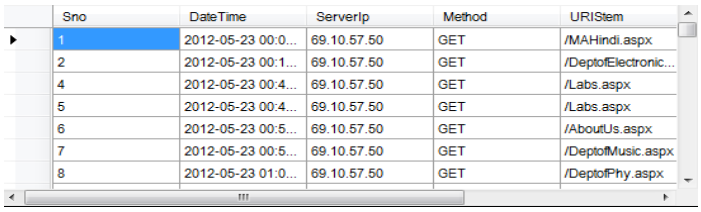


Fig.3.Selection of log file and database record storage

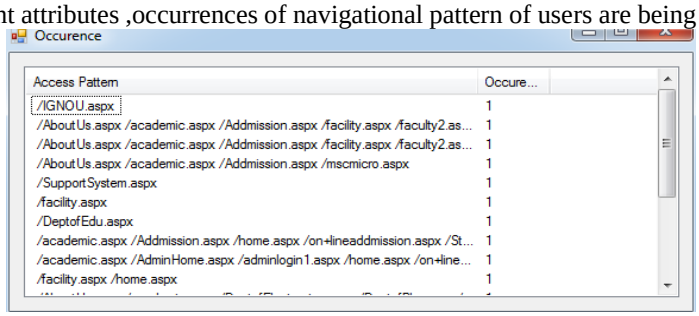
After removing the irrelevant entries database will look like this



Sno	DateTime	ServerIp	Method	URIStem
1	2012-05-23 00:0...	69.10.57.50	GET	/MAHindi.aspx
2	2012-05-23 00:1...	69.10.57.50	GET	/DeptofElectronic...
4	2012-05-23 00:4...	69.10.57.50	GET	/Labs.aspx
5	2012-05-23 00:4...	69.10.57.50	GET	/Labs.aspx
6	2012-05-23 00:5...	69.10.57.50	GET	/AboutUs.aspx
7	2012-05-23 00:5...	69.10.57.50	GET	/DeptofMusic.aspx
8	2012-05-23 01:0...	69.10.57.50	GET	/DeptofPhy.aspx

Fig.4.Relevant Attributes

After removal of irrelevant attributes ,occurrences of navigational pattern of users are being calculated as shown in fig.5.

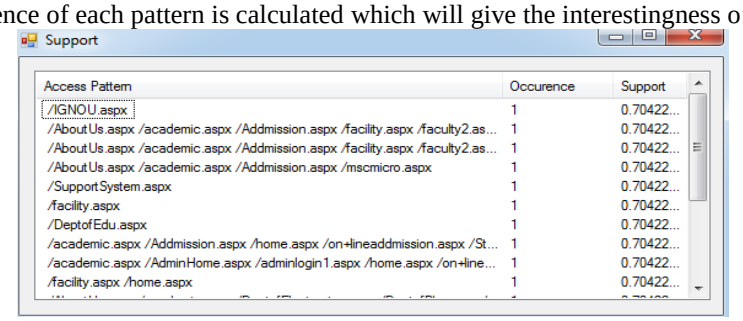


Occurence

Access Pattern	Occure...
/IGNOU.aspx	1
/AboutUs.aspx /academic.aspx /Addmission.aspx /facility.aspx /faculty2.as...	1
/AboutUs.aspx /academic.aspx /Addmission.aspx /facility.aspx /faculty2.as...	1
/AboutUs.aspx /academic.aspx /Addmission.aspx /mscmicro.aspx	1
/SupportSystem.aspx	1
/facility.aspx	1
/DeptofEdu.aspx	1
/academic.aspx /Addmission.aspx /home.aspx /on+lineaddmission.aspx /St...	1
/academic.aspx /AdminHome.aspx /adminlogin1.aspx /home.aspx /on+line...	1
/facility.aspx /home.aspx	1

Fig.5.Navigational Patterns of user

Now support and confidence of each pattern is calculated which will give the interestingness of pattern

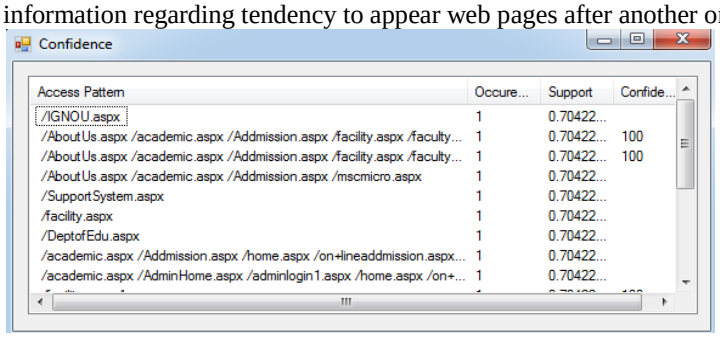


Support

Access Pattern	Occurence	Support
/IGNOU.aspx	1	0.70422...
/AboutUs.aspx /academic.aspx /Addmission.aspx /facility.aspx /faculty2.as...	1	0.70422...
/AboutUs.aspx /academic.aspx /Addmission.aspx /facility.aspx /faculty2.as...	1	0.70422...
/AboutUs.aspx /academic.aspx /Addmission.aspx /mscmicro.aspx	1	0.70422...
/SupportSystem.aspx	1	0.70422...
/facility.aspx	1	0.70422...
/DeptofEdu.aspx	1	0.70422...
/academic.aspx /Addmission.aspx /home.aspx /on+lineaddmission.aspx /St...	1	0.70422...
/academic.aspx /AdminHome.aspx /adminlogin1.aspx /home.aspx /on+line...	1	0.70422...
/facility.aspx /home.aspx	1	0.70422...

Fig.6.Support Calculation

Confidence is a measure of strength of association rules. Confidence is used how confident we are about our patterns By finding confidence we get information regarding tendency to appear web pages after another one



Confidence

Access Pattern	Occure...	Support	Confide...
/IGNOU.aspx	1	0.70422...	
/AboutUs.aspx /academic.aspx /Addmission.aspx /facility.aspx /faculty...	1	0.70422...	100
/AboutUs.aspx /academic.aspx /Addmission.aspx /facility.aspx /faculty...	1	0.70422...	100
/AboutUs.aspx /academic.aspx /Addmission.aspx /mscmicro.aspx	1	0.70422...	
/SupportSystem.aspx	1	0.70422...	
/facility.aspx	1	0.70422...	
/DeptofEdu.aspx	1	0.70422...	
/academic.aspx /Addmission.aspx /home.aspx /on+lineaddmission.aspx...	1	0.70422...	
/academic.aspx /AdminHome.aspx /adminlogin1.aspx /home.aspx /on+...	1	0.70422...	

Fig.7.Confidence Calculation

B) Evaluation Measures:

In order to evaluate the performance of our proposed work, we have used two important evaluation measures: coverage and precision. If we take RS as the recommended set of pages and take US as a set of pages that visitors accessed, then we can provide the definition for Coverage and Precision:

$$Coverage = \frac{|US \cap RS|}{|US|}$$

$$Precision = \frac{|US \cap RS|}{RS}$$

Meanwhile in order to get the best performance we need a measure to evaluate. Mobasher etc. provided a new evaluation measure M. M is called matching rate

$$M = \frac{2 \times coverage \times precision}{coverage + precision}$$

In order to evaluate the performance of Proposed work , we compared the results that come from Predictor 1.5 and Predictor 1.2 with our proposed work.

TABLE II. Results from Predictor 1.5

RS	PRECISION	COVERAGE	M
6	0.865	0.553	0.675
9	0.827	0.638	0.72
12	0.8803	0.695	0.745
15	0.791	0.771	0.781
Average	0.822	0.664	0.73

TABLE III. Results from Predictor 1.2

RS	PRECISION	COVERAGE	M
6	0.762	0.451	0.567
9	0.693	0.537	0.605
12	0.585	0.573	0.579
15	0.476	0.626	0.541
Average	0.629	0.547	0.573

TABLE IV. Results from Proposed work

RS	PRECISION	COVERAGE	M
6	0.762	0.451	0.567
9	0.693	0.537	0.605
12	0.585	0.573	0.579
15	0.476	0.626	0.541
Average	0.629	0.547	0.573

Graph showing the performance of all three methods, clearly the proposed work shows high precision, coverage and matching rate as compared to other two methods .Hence the new methodology proves to have higher accuracy and power of navigational pattern mining.

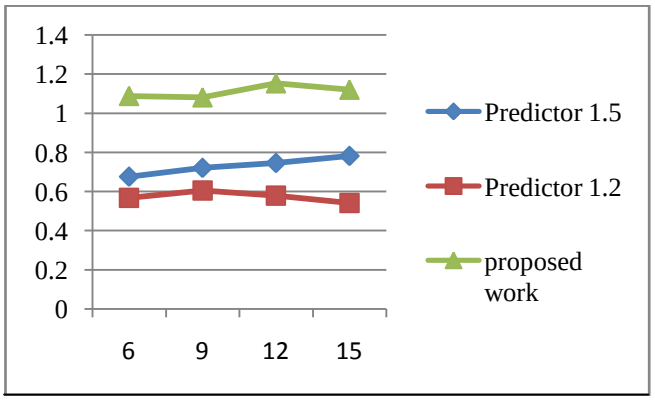


Figure.8. Graph showing performance of three methods

VI. Conclusion

In this paper we have tried to give a clear view of the web server logs, interesting patterns extracted from the web logs, calculation of support and confidence of each navigational pattern .Also we have used evaluation measures to evaluate Performance of our approach .The experimental results shows that our approach is better than the other two methods as the recommended set of pages found based on this algorithm are highly accurate. Future work on this study comprises of more refined techniques for data preprocessing and recognition of access sessions, in order to assuage common problems of Web Usage Mining. Other algorithms for pattern detection shall also be incorporated in the system, so as to generate substitute methods such as apriori algorithm and adaptive clustering technique which can be investigated for further enhanced analysis.

REFERENCES

[1] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explor. Newsl. 1(2),12–23 (2000).

[2] R. Kosala and H. Blockeel. Web mining research: a survey In ACM SIGKDD, pages 1–15, july 2000.

[3] R. Baraglia and P. Palmerini, "Suggest: A web usage mining system," in Proceedings of IEEE International Conference on Information Technology: Coding and Computing, April 2002.

[4] Yi Dong, Huiying Zhang, Linnan Jiao, "Research on Application of User Navigation Pattern Mining Recommendation", Intelligent Control and Automation, 2006. WCICA 2006, the Sixth World Congress, Volume 2.

[5] Jalali, M., Mustapha, N., Sulaiman, M. N. B. And Mamat, A. (2008a) OPWUMP: An Architecture for Online Predicting in WUM-Based personalization System, Communications in Computer and Information.

[6] Jalali, M., Mustapha, N., Sulaiman, N. B. and Mamat, A. (2008b) A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems,"12th International on Information Visualisation, IV'08, London, UK, Pp. 302-307.

[7] Ms. Dipa Dixit, Mr. Jayant Gadge, Automatic Recommendation for Online Users Using Web Usage Mining on International Journal of Managing Information Technology (IJMIT) Vol. 2, No. 3, August 2010.

[8] M. P. Yadav, P. Keserwani, S. Ghosh,"An Efficient Web Mining Algorithm for Web Log Analysis: E-Webminer" 978-1-4577-0697-4/12 IEEE 2012.

[9] Bhushan, R., Nath, R., "Recommendation of optimized web pages to users using Web Log mining techniques" 978-1-4244-4507-3/09/ IEEE 2013.

