Improve Efficiency of Web Pattern Analysis Through Web Mining

Hemwati Kumawat
M.Tech Scholar
Govt. Engg College Ajmer

Vinesh Jain
Assistant Professor
Govt. Engg. College Ajmer

ABSTRACT
In commercial and academic areas of Web usage mining techniques, the interest has been increased due to the information explosion in World Wide Web. Through web user navigation patterns, proposed a approach which predict users' future request that provide valuable information for web designers which quickly respond to their individual needs for the efficient organization of the website by mining the weblog files using Web weblog Expert and Universal Extractor. The approach is based on the combined mining to discover knowledge from web server log files which derive user navigation profiles and the contents of the retrieved web pages. The proposed algorithm implemented, and the system achieves the bandwidth of nearly 80% and the efficiency of about 75% of extracted log files, which is about 20% higher than original log files by mining Web server logs. This approach has been carried out in order to validate and facilitate better web personalization and website organization.

General Terms
Weblog File, Association Rule Mining, Field Extractor Algorithm.

Keywords
web usage mining; web content mining; user navigation patterns; weblog files;

1. INTRODUCTION
The digital revolution, explosive growth of the Internet and the knowledge available on the World Wide Web ensured that huge volume of high dimensional multimedia data are available to all users which lacks an integrated structure. This information is often mixed with different data types such as text, image, audio, hypertext, graphics and video components interspersed with each other and stored in files, databases, and other repositories have abundant of information, which needs to be extracted and analysed [2]. However, it becomes much more difficult for users to access relevant information efficiently. The problem is to mine useful information or patterns from the huge datasets [1].

Web mining, which is the art of identifying, extracting and analysing useful information from the browsing, usage patterns, web documents and services has attracted much attention to the webmasters and then this information can be used by the web developers on one hand for predicting the navigation behaviour of current users, thus aiding in web personalization for designing and developing better web designs [3]. Web Log mining, also referred as Web usage Mining which is a category of web wining. It is the automatic discovery of user access patterns from data describing the usage of web resources. The access logs of HTTP servers are the best source of information. The knowledge gained from web usage mining gives guidelines on user behaviour, usage patterns for mass customization and personalization.

We propose an experimental system to investigate whether associating a content mining approach with regular web usage mining could result in a more accurate classification of user navigation patterns, and consequently lead to a more accurate prediction of users' future requests. From the user perspective, the classification of navigation patterns can enhance the quality of personalized web recommendations that estimate the best profile describing navigation behaviour of the current user, and find related unrequested web pages of great potential to be the next pages that the user wants to see. As the recommendation, the links of these pages will then be inserted into the currently requested page dynamically for display. This will help to user access their favourite information efficiently. Instead of being arranged purely according to one view of the web site content, a site will be adjusted in terms of the desires of users. For instance, necessary links will be added between the web pages, which seemingly do not share the same topic, but were visited one after another by plenty of users. Also, pages which drew lots of clicks will be highlighted from their categories of topics, while pages which were not visited for a period of time will be moved or discarded. Organizing websites by topics is both static and reactive. Since users' navigation patterns will be learned periodically, the change of their navigation interest can be captured regularly and then the site organization can be adjusted accordingly. This is a dynamic and proactive way of managing websites. As a result, the passing visitors will be enticed to become consumers or users of the site while current users are willing to remain loyal to the site.

The objectives of this paper are to improve the efficiency of log file by extract the original log file that helps webmaster and web developers to improve the design of a web page by analyzing the extracted weblog files. The rest paper is organized as follows: In Section 2, recent review research on related work, advances web usage, association rule and content mining. Section 3, the approaches and algorithms of system proposed for user navigation patterns and future requests. The results are presented in Section 4 and Section 5 concludes the work with future directions

2. RELATED WORK
All Data mining applications can employ a mixture of parameters to inspect the data. They comprise associations among different data objects, path where one data object leads to another data object, classification which identifies new patterns [6], association rule which group the similar data objects. Web usage mining, the art of analyzing user interactions with a web page, has been dealt by several researchers using different approaches. Association rule mining method ([4], [5]) generates rule describing the relationship between user access and a web page. Though very successful, the number of association rules generated is often very huge and are often difficult to process and analyze which are being used to mine web log files. Web usage mining [7] information provide recent clients, maintain existing clients, and track clients who are leaving the web site.

The procedure of information retrieval from the secondary data such as web server access logs, registration data, user sessions or transactions is defined as web usage mining.

The web pages [8] are assigned into a Weblog Expert based on likeness or other relationship measures in the web page. Data mining techniques are applied to the web logs for retrieving the user access patterns. The user web logs are extracted i.e., future behavior of user is forecasted. In web usage mining, the association rules identify the sets of pages that are accessed together which are not directly correlated to one another through hyperlinks.

The Association rule mining [9] is used to determine the frequently visited web pages collectively in a Session and interesting associations between huge set of data items. The association rules discovered through sessions and each session is elucidated as transaction. Association rule [10] can identify the series in which the web pages were viewed and these series of pages are called as paths. The association rule mining algorithms are applied to web mining for identifying associations between web pages and exciting access patterns. Association rules are measured to be exciting if both a minimum support and a minimum confidence is satisfied.

Field extraction algorithm is useful to handle huge data set due to their lower time and space complexity also it use the threshold value to decide if an data object must be positioned into an existing log file or a new log file has to be produced. The argument behind using Field extraction algorithm was that it is both memory efficient and has simple implementation procedure. The other experimental results of systems found to be very slow at convergence. To solve the convergence problem [11] and to increase the efficiency proposed a algorithm for mining user's navigation pattern using Weblog Expert and used Universal Extraction estimates of parameters in probabilistic, where the it depends on unobserved latent variables.

During experimentation, it was found that the performance of original and Extracted log files, the efficiency of Extracted logfiles could be increased by applying Field Extraction algorithm to classify user requests and combine the results to provide a more accurate web usage mining prediction system. This research work is focused on producing such a system.

3. PROPOSED METHODOLOGY

The main aim of the proposed methodology is to mine user navigation patterns by using weblog files. User navigation pattern is defined as common browsing characteristics among group of users. Different users have common browsing practices and navigation patterns needs to capture these common interests to identify user needs. In this paper, Field Extraction Algorithm is used to group users with similar browsing characteristics and to associate navigation behavior with these groups of users. The selected algorithm is memory efficient and easy to implement, with a profound probabilistic background.

3.1 Selecting proper schema for design

This research paper include methodology which is used to reduce the size of web-log files to increase the quality of log files with the help of data cleaning and merging the number of web-log files and also access the contents of user identification contents to check the records which is accessed by different and unique users.

3.2 Web data mining architecture

a) Integration and merging of sample web log files.
b) Pre-processing technique executed in which data cleaning, user identification and session identification occurs.
c) Pattern discovery
d) Pattern Analysis
e) Analysis reports

So, this step is the proposed methodology which we used to make our web-log files more useful and small.

3.3 Data Preprocessing

The main objective of this step is to reformat the raw log data into a format that identifies all web access sessions. Not every access made contain useful information. All those entries that are irrelevant should be removed. Examples of irrelevant entries include button image access, multimedia file access, non-human accesses etc. Redundant clicks and failed transactions should also be removed. The delimited set of pages of one particular visit within duration, visited by the same user to a Web site is a user session. Session identification is carried out as if a predefined period of time between two accesses is exceeded than a new session starts at that point. Web usage data is prepared for applying navigation patterns mining algorithms by doing these pre treatment tasks and are done manually in the present work.

3.4 Data Cleaning

Data cleaning technique is used to remove the useless or unwanted records those are not as useful as other data like user identification, user session, number of hits, number of visits, etc. So we propose a comparative algorithm for data cleaning which is very useful and also which reduces the size of web log file and improve the quality of contents in the log file.

Web mining is mining of data, related to World Wide Web and categorized into three active research areas according to what part of web data are mined [12]: In Content mining knowledge extracted from content of websites; Structure mining, to obtain the topology of the interconnections between web objects, links and references are used within web pages; Usage mining, studies user access information to extract interesting usage patterns from logged server data and also known as web-log mining.

Web mining processes are applied to the server logs to mine the hidden navigation models of users. The user's requests from his current active session are recorded. By comparing these requests, we first combine usage and content mining to process web logs for building user navigation profiles, and then use these profiles to classify the site users. Next, we simulate the requests from active sessions to construct current navigation profiles of users. By matching the current profile with the navigation profiles, we are able to predict users' future requests.

3.5 Web Log File

Proxy Server Log file is collected from Proxy server and log file are composed and analyzed in this work. The data of the month Feb 2013 to June 2013 log file are merged to find source log file. This Log File is the input for the pre-processing point. Almost 2212 visitors visited the web site during this week. The size of the log file increases day by day. The total number of records found is 1,148,872 from this log file.

4. EXTRACTION ALGORITHM

A server log file consists of various data fields that should be separated before applying cleaning procedure. This process of

separating out different data fields from single server log entry is identified as data field extraction algorithm. A server uses different characters such as a comma or a space character which works as separators. The algorithm proposed below for data field extraction uses the space character as a separator to separate the fields of the log file. This algorithm is suitable for identifying the frequent web pages viewed by each of the grouped clusters. This method avoids the unnecessary usage of web.

4.1 Field Extraction Algorithm

Step 1: Begin
Step 2: Repeat to read multiple Log files from Proxy server web-logs
Step 3: Merge all Log files in to single sample web-log file
Step 4: Start data cleaning (If such type of extension like <jpeg, mpeg, js, gif, css>, < different error like http 404, refuse connection from proxy server> found then remove from single sample web-log records.
Step 5: Analyze the result log file from previous log file.
Step 6: Repeat step 4 and 5 until end of sample log file.
Step 7: End

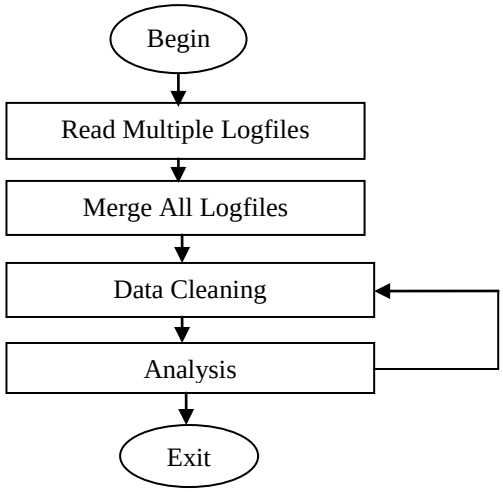

Fig 1: Representation of Field Extraction Algorithm

We have analyzed the visits history according to number of single site with different users and date.

5. RESULT AND DISCUSSION

The experiments are conducted in the proposed technique by using the log obtained from the reputed college web site for about a week in 2013. The obtained record consists of 60000 records in the log file. Then the data cleaning process is carries out. Initially, after removing records with graphics and videos format such gif, JPEG, 2520 records are obtained. Then by checking the status code, the total of 2450 records is resulted. Finally, 2360 records are resulted. In the proposed method the records accessed, agents are also cleaned.

The time required for determining user interested pattern after different data cleaning techniques. The total 60000 records are obtained initially. Then after removing the gif status, 2520 records are resulted. 2450 records are obtained, after gif status removal and finally 2360 records are obtained after cleaning process. As the number of irrelevant records is discarded, this helps in determining the user interested pattern more accurately in less time.

After data cleaning, users are identified according to IP addresses, browsers and operating systems. Then the path completion technique is applied in order to determine the path accessed by the user. The path completed for a user by using log after cleaning. It can be observed that the irrelevant pages found are eliminated. Finally, it provides path completed for a user by using log. The most relevant web pages interested by the user is obtained, whereas, some of the irrelevant wed pages are considered for predicting the user interested patterns.

The result shows that the size of the log file reduces by removing unnecessary and irrelevant entries from the file due to proposed methodology. Earlier there were 60000 entries in the log file, but after cleaning only 2360 entries have been left. The original size of the log file before cleaning was 779.51 MB and after cleaning the size reduces to 14.26 MB as shown in Table 1.

Table 1. Comparison in Size Before and After Cleaning

	Size (MB)	No. of Records
Before cleaning	779.51	60000
After cleaning	14.26	2360

The change in size and number of records for the log file is graphically represented by means of a bar chart in Figure 2.

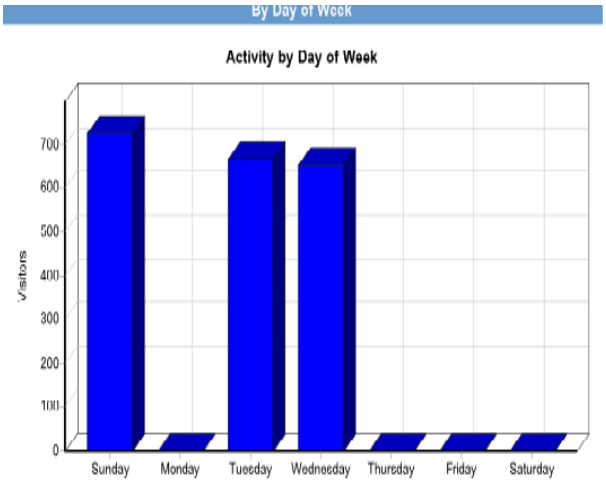

Fig 2: Illustrates the amount of data packets been sent and received

The graph makes it clear that there is a severe change in the size and number of records after data cleaning. Therefore, from this observation if calculate the percentage amount of decrease in the size of log file, it gives a reduction of 74% (see Table 2) which is quite a major value.

Table 2. Results of Data Cleaning

Web Server Log File	Result
Original Size	779.51 MB
Reduced Size	14.26 MB
Percentage in Reduction	74.00

The bandwidth utilization of each user (ip address) is identified from the web server logs. Fig. 3, illustrate the amount of data packets been sent and received by the experimental ip address. The experimental result shows that the amount of bandwidth received is comparatively higher than the sent packets. Using this result, the web administrator of educational institution, allocate the bandwidth to individual users based on the content viewed by those users.

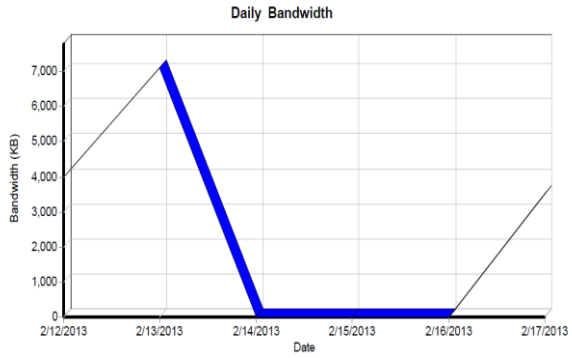

Fig 3: Bandwidth Utilization Of Each Users

The problem in web log mining is solved in this study. Initially, the logs are collected then the preprocessing steps are carried out here for removal of records of graphics, videos and the format information, the records with the failed HTTP status code, Method field. This will help in reduction of quantity of data to be passed to further processing and the users are identified by user identification phase. From this the sessions are identified. Next, the path completion step is carried out to identify missing pages due to cache and 'Back'. Path Set is the incomplete accessed pages in a user session. It is extracted from every user session set. Then, the user transactions are identified. Finally, from the obtained data, content path set are identified which will help in better web prediction. Then the experiment is conducted using the log obtained from the reputed college web site to evaluate the proposed technique. The experimental result shows the improvement in the web log mining.

6. CONCLUSION

In this paper, the usage of Weblog file are the best source to predict user's behavior for studying user's navigation pattern that was enhanced to include user request pattern. Along with the useful information the raw log files also contains entries for unnecessary details like image access, failed entries etc. which are of no use from the perspective of the Web Usage Mining. Therefore, it becomes necessary to get rid of this irrelevant information. It has undergone various steps as data cleaning, path completion, users, session, and transaction identifications. Different from other implementations records are cleaned effectively by removing. For data cleaning, proposed algorithm was successfully tested on the log files. The results showed a positive response to the inclusion of pattern. Such results are suitable for website owners, for example, to put their advertisements on websites or to change the structure of the web site according to user navigational behavior. The web usage of individual users is identified for proper allocation of bandwidth and efficiency to that user in future. The information revealed from this methodology provides efficient utilization of web in educational institutions. This work can be extended by providing privacy for user web logs. The system was developed using Universal Extractor and the log files used were simulated. The discussed approach showed a quite salient reduction in the number of records and in the log files size and hence increases the quality of the available data.

7. REFERENCES

[1] Roughan, M. and Zhang, Y., "Secure distributed data-mining and its application to large-scale network measurements", ACM SIGCOMM Computer Communication Review, Vol.36, Issue 1, Pp.7- 14, 2011.

[2] Washio, T., Suzuki, E., Ting, K.M. and Inokuchi, A. "Advances in Knowledge Discovery and Data Mining", Proceedings of12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, Lecture Notes in Computer Science, Pp. 1-1102. SBN: 978-3-540-68124, 2012.

[3] Heydari, M., Helal, R.A. and Ghauth, K.I., "A graph-based web usage mining method considering client side data", International Conference on Electrical Engineering and Informatics, ICEEI, Vol. 01, Pp.147-153, 2011.

[4] Kazienko, P. "Mining Indirect Association Rules for Web Recommendation", International Journal of Applied Mathematics and Computer Science, Vol. 19, Issue 1, Pp. 165-186, 2012.

[5] Raju, V.V.R., Rao V.M. and Kumari, V. "Understanding User Behavior using Web Usage Mining", International Journal of Computer Applications, Published By Foundation of Computer Science, Vol.1, No. 7, Pp. 55–64, 2010.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A review. ACM Computing Survey (CSUR)", 31(3): 264-323, 2011.

[7] Kobra Etminani, Mohammad-R, Akbarzadeh-T, Noorali Raeeji Yanehsari., "Web Usage Mining: users navigational patterns extraction from weblogs using Ant-based Clustering Method", Volume 4, Pp. 55-64, IEEE 2012.

[8] Mrs. Kiruthika M and Mrs. Dipa Dixit**.,** **"**Mining Access Patterns Using Clustering", International Journal of Computer Applications (0975 –8887) Volume 4– No.11, August 2012.

[9] Resul DAŞ and İbrahim Türkoğlu, "Extraction of Interesting Patterns Through Association Rule Mining For Improvement Of Website Usability", Journal Of Electrical & Electronics Engineering, vol 9, No 2(2012).

[10] Bamshad Mobasher and Robert Cooley, Jaideep Srivastava, "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs", 2012.

[11] Mustapha, N., Jalali, M. and Jalali, M., "Expectation Maximization Clustering Algorithm for User Modeling in Web Usage Mining Systems", European Journal of Scientific Research, vol. 32, No. 4, Pp.467-467, 2011.

[12] Alam, S., G. Dobbie, et al, Particle Swarm, "Optimization Based Clustering of Web Usage Data", 2008 IEEE/WIC/ACM Internation Conference On Web Intelligence and Intelligence Agent Technology 978-0-7695- 3496-I/08 DOI 10.1109/WIIAT.2008.292 IEEE/WIC/ACM International Conference On Web 2012.