Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise, Cookies and Web Robots

P. Nithya
Doctoral Student in Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India.

P. Sumathi, PhD.
Asst. Professor, Dept.of Computer Science Government Arts College, Coimbatore, India

ABSTRACT
Today internet has made the life of human dependent on it. Almost everything and anything can be searched on net. Web pages usually contain huge amount of information that may not interest the user, as it may not be the part of the main content of the web page. Web Usage Mining (WUM) is one of the main applications of data mining, artificial intelligence and so on to the web data and forecast the user's visiting behaviors and obtains their interests by investigating the samples. Since WUM directly involves in applications, such as, e-commerce, e-learning, Web analytics, information retrieval etc. Weblog data is one of the major sources which contain all the information regarding the users visited links, browsing patterns, time spent on a particular page or link and this information can be used in several applications like adaptive web sites, modified services, customer summary, pre-fetching, generate attractive web sites etc. There are varieties of problems related with the existing web usage mining approaches. Existing web usage mining algorithms suffer from difficulty of practical applicability. This paper continues the line of research on Web access log analysis is to analyze the patterns of web site usage and the features of users behavior. It is the fact that the normal Log data is very noisy and unclear and it is vital to preprocess the log data for efficient web usage mining process. Preprocessing is the process comprises of three phases which includes data cleaning, user identification, and pattern discovery and pattern analysis. Log data is characteristically noisy and unclear, so preprocessing is an essential process for effective mining process. In this paper, a novel pre-processing technique is proposed by removing local and global noise and web robots. Preprocessing is an important step since the Web architecture is very complex in nature and 80% of the mining process is done at this phase. Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web Dataset are used for evaluating the proposed preprocessing technique.

General Terms
Data mining, web usage mining, preprocessing.

Keywords
Preprocessing, Data Cleaning, Path Completion, Travel Path set, Content Path Set.

1. INTRODUCTION
Web has evolved into a network of data with no proper organizational structure. In addition, survival of plentiful Millions of electronic data are included on hundreds of millions data that are previously on-line today. With this significant increase of existing data on the Internet and because of its fast and disordered growth, the World Wide data in the network and the varying and heterogeneous nature of the web, web searching has become a tricky procedure for the majority of the users. This makes the users feel confused and at times lost in overloaded data that persist to enlarge. Moreover, e-business and web marketing are quickly developing and significance of anticipate the requirement of their customers is obvious particularly. As a result, guessing the users' interests for improving the usability of web or so called personalization has turn out to be very essential.

Web personalization can be depicted as some action that builds the web experience of a user personalized according to the user's interest.

Generally, three kinds of information have to be handled in a web site: content, structure and log data. Content data contains of anything is in a web page, structure data is nothing but the organization of the content and usage data is nothing but the usage patterns of web sites. The usage of the data mining process to these dissimilar data sets is based on the three different research directions in the area of web mining: web content mining, web structure mining and web usage mining [6, 8].

Web usage mining [16, 17] consists of three main steps. Weblog file is usually given as input.

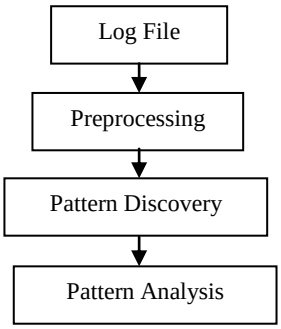

Figure 1: Steps in Web Usage Mining

Preprocessing is an important step because of the complex nature of the Web architecture which takes 80% in mining process. The raw data is pretreated to get reliable sessions for efficient mining. It includes the domain dependent tasks of data cleaning, user identification, session identification, and path completion and construction of transactions. Data cleaning is the task of removing irrelevant records that are not necessary for mining. Data cleaning includes

1. Elimination of Local and Global Noise,
2. Removal of records of graphics, videos and the format information
3. Removal of records with the failed HTTP status code
4. Robots cleaning
5. Removal of cookies

User identification is the process of associating page references with same IP address with different users. Session identification is breaking of a user's page references into user

sessions. Path completion is used to fill missing page references in a session. Classifications of transactions are used to know the users interest ad navigational behavior. The second step in web usage mining [18] is knowledge extraction in which data mining algorithms like association rule mining techniques, clustering, classification etc. are applied in preprocessed data. The third step is pattern analysis in which tools are provided to facilitate the transformation of information into knowledge. Knowledge query mechanism such as SQL is the most common method of pattern analysis. This paper focuses on path completion process which is used to append lost pages and construction of transactions in preprocessing stage. In this study a referrer-based method is proposed to efficiently construct the reliable transactions in data preprocessing.

The remainder of this paper is organized as follows. The next section presents related work of log analysis. Section 3 provides the proposed methodology. Experimental results are illustrated in Section 4. Finally, the conclusions are drawn in Section 5.

2. RELATED WORKS

Various commercial available web server log analysis tools are not designed for high traffic web servers and provide less relationship analysis of data relationships among accessed files which is essential to fully utilize the data gathered in the server logs [25]. Web server log file is a simple plain text file which record information about each user. Log file contain information about user name, IP address, date, time, bytes transferred, access request. A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. When user submit request to a web server that activity are recorded in web log file. Log file range 1KB to 100MB. Log file gives significant information to web server. Web server logs contain more information about visitor's information in the access logs, usually in W3C format. There are also the error logs for each server that contains information on errors and problems that the server practiced. The statistical analysis introduces a set of parameters to describe user's access behaviors. With those parameters it becomes easy for administrators to define concrete goals for organizing their web sites and improve the sites according to the goals. But the drawback in this analysis is that the results are independent from page to page. Since user's behavior is expected to be different dependent on length of browsing time, the calculation of accurate browsing time is more important. The discovery of the users' navigational patterns using SOM is proposed by Etminani *et al.,* [1]. Jianxi *et al.,* [2] presented a Web usage mining technique based on fuzzy clustering in Identifying Target Group. Nina *et al.,* [3] suggests a complete idea for the pattern discovery of Web usage mining. Wu *et al.,* [4] given a Web Usage Mining technique based on the sequences of clicking patterns in a grid computing environment. The author discovers the usage of MSCP in a distributed grid computing surroundings and expresses its effectiveness by empirical cases. Aghabozorgi *et al.,* [5] proposed the usage of incremental fuzzy clustering to Web Usage Mining. Rough set based feature selection for web usage mining is proposed by Inbarani *et al.,* [7]. Jalali *et al.,* [8] put forth a web usage mining technique based on LCS algorithm for online predicting recommendation systems. For providing the online prediction effectively, Shinde *et al.,* [9] provides a architecture for online recommendation for predicting in Web Usage Mining System.

Zhang *et al.,* [10] given an intelligent algorithm of data pre-processing in Web usage mining. Nasraoui *et al.,* [11] provides a whole framework and findings in mining Web usage navigation from Web log files of a genuine Web site which has every challenging characteristics of real-life Web usage mining, together with evolving user profiles and external data describing an ontology of the Web content. Hogo *et al.,* [12] proposed the temporal Web usage mining of Web users on single educational Web site with the help of the adapted DeMin Dong, [13] SOM based on rough set properties. A development of data preprocessing technique for Web usage mining and the informations of algorithm for path completion are provided by Yan *et al.,* [14].

Baraglia *et al.,* [15] proposed a Web usage mining (WUM) system, called SUGGEST, which continuously creates the suggested connections to Web pages of probable importance for a user. Lee *et al.,* [19] put forth a Web Usage Mining technique based on clustering of browsing characteristics.

Kushmerick [22] has proposed a feature based method which identifies internet advertisements in a web page. It is mainly used for removing advertisements and does not remove other non-content blocks. Bar-Yossef and Rajagopalan [23] have proposed a method to identify frequent templates of Web pages and pagelets. Page level template detection is done by D.Chakraborti et al.[24] They examine the page's features and these features are used to score the DOM tree nodes. Page level templates are generated by doing isotonic smoothing on classifier scores.

3. METHODOLOGY

Web log data preprocessing is a complex process and takes 80% of total mining process. Log data is pretreated to get reliable data. The aim of data preprocessing is to select essential features clean data by removing irrelevant records and finally transform raw data into sessions.

3.1 Data cleaning

The process of data cleaning is removal of outliers or irrelevant data. Data Cleaning enables to filter out useless data which reduce the log file size to use less storage space and to facilitate upcoming tasks. Analyzing the huge amounts of records in server logs is a cumbersome activity. So initial cleaning is necessary. If a user requests a specific page from server entries like gif, JPEG, etc., are also downloaded which are not useful for further analysis are eliminated. The records with failed status code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be removed from log files. Thus removal process in the experiment includes

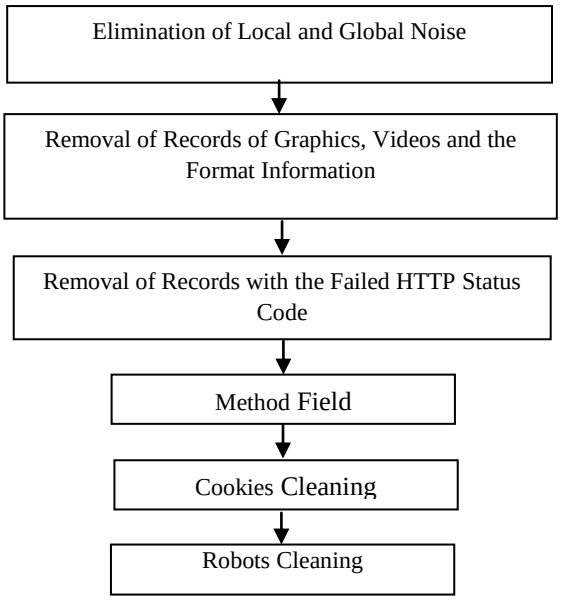

Figure 2: Steps in Data Cleaning

3.1.1 Elimination of Local and Global Noise

Web noise can be normally categorized into two groups depending on their granularities:

Global Noise: It corresponds to the unnecessary objects with huge granularities, which are no smaller than individual pages. This noise includes mirror sites, duplicated Web pages and previous versioned Web pages, etc. the noise from web pages in the front work, most pages still have some noise, such as "ad-serving", "contact", "company profiles", "copyright ","©", "all rights reserved" and other noise words.

Local (Intra-Page) Noise: It corresponds to the irrelevant items inside a Web page. Local noise is typically incoherent with the major content of the page. This noise includes banner ads, navigational guides, decoration pictures, etc. These noises have to remove for better results. The local noise also deals with the user background knowledge can be discovered from user local information collections, such as a user's stored documents, browsed web pages, and composed/received emails.

3.1.2 The records of graphics, videos and the format information

The records have filename extension of GIF, JPEG, CSS, and so on, which can be found in the URI field of the every record, can be removed. This extension files are not actually the user interested web page, rather it is just the documents embedded in the web page. So it is not necessary to include in identifying the user interested web pages. This cleaning process helps in discarding unnecessary evaluation and also helps in fast identification of user interested patterns.

3.1.3 The records with the failed HTTP status code

The HTTP status code is then considered in the next process for cleaning. By examining the status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. This cleaning process will further reduce the evaluation time for determining the used interested patterns.

3.1.4 Method field

It should be pointed out that different from most other researches, records having value of POST or HEAD in Method field are reserved in present study for acquiring more accurate referrer information.

3.1.5 Removal of Cookies

Cookies are usually randomly assigned IDs that a Web server gives to a Web browser the first time that the browser connects to a Web site. On subsequent visits, the Web browser sends the same ID back to the Web server, effectively telling the Web site that a specific user has returned. Cookies are independent of IP addresses, and work well on sites with a substantial number of visitors from ISPs. Authenticated usernames even more accurately identify individuals, but they require each user to enter a unique username and password, something that most Web sites are unwilling to mandate.

The cookie does not contain the mailing or credit card information; that information typically was collected when the visitor entered it into a form on the Web site. The cookie merely confirms that the same computer is back during the next site visit.

If a Web site uses cookies, information will appear in the cookie field of the log file, and can be used by a Web traffic analysis software to do a better job of tracking repeat visitors.

Unfortunately, cookies remain a misunderstood and controversial topic. A cookie is not an executable program, so it can't format your hard drive or steal private information. The cleaning process using the preprocessing process helps to clean the history of cookies, if the browsers do not turn on the cookie processing on. It improves the safety of the users.

3.1.6 Robots Cleaning

Web robot (WR) (also called spider or bot) is a software tool that periodically scans a web site to extract its content. Web robots automatically follow all the hyperlinks from a web page. Search engines, such as Google, periodically use WRs to gather all the pages from a web site in order to update their search indexes. The number of requests from one WR may be equal to the number of the web site's URIs. If the web site does not attract many visitors, the number of requests coming from all the WRs that have visited the site might exceed that of human-generated requests.

Eliminating WR-generated log entries not only simplifies the mining task that will follow, but it also removes uninteresting sessions from the log file. Usually, a WR has a breadth (or depth) first search strategy and follows all the links from a web page. Therefore, a WR will generate a huge number of requests on a web site. Moreover, the requests of a WR are out of the analysis scope, as the analyst is interested in discovering knowledge about users' behavior.

Most of the Web robots identify themselves by using the user agent field from the log file. Several databases referencing the known robots are maintained [Kos, ABC]. However, these databases are not exhaustive as each day new WRs appear or are being renamed, making the WR identification task more difficult.

To identify web robots' requests, the data cleaning module implements two different techniques.

- In the first technique, all records containing the name "robots.txt" in the requested resource name (URL) are identified and straightly removed.
- The next technique is based on the fact that the crawlers retrieve pages in an automatic and exhaustive manner, so they are distinguished by a very high browsing speed.

Therefore, for each different IP address, the browsing speed is calculated and all requests with this value more than a threshold are regarded as made by robots and are consequently removed. The value of the threshold is set up by analyzing the browser behavior arising from the considered log files.

This helps in accurate detection of user interested patterns by providing only the relevant web logs. Only the patterns that are much interested by the user will be resulted in the final phase of identification if this cleaning process is performed before start identifying the user interested patterns.

4. EXPERIMENTAL RESULTS

In order to evaluate the proposed preprocessing phase with robots cleaning, experiments were carried out using UCI Machine Learning Repository (University of California, Irvine). This repository contains 211 datasets. For the purpose of evaluating the proposed robot cleaning preprocessing phase, it is evaluated against,

- Initial log file and
- Preprocessed log file without removing robots.

Four standard datasets from the UCI Machine Learning Repository datasets and a real dataset is collected from reputed college were selected for the evaluation purpose. Following are the data sets used for evaluating the proposed preprocessing phase with robots cleaning.

- Anonymous Microsoft Web Dataset [20],
- MSNBC.com Anonymous Web Dataset [21]

4.1 Anonymous Microsoft Web Dataset

This dataset consists of 37711 records in the log file. Then the data cleaning process is carried out. Initially, after removing records with local and global noise, graphics and videos format such gif, JPEG, etc., 29862 records are obtained. Then by checking the status code and method field, the total of 26854 records is resulted.18452 records are resulted after applying cookies cleaning process and finally, 12659 records are resulted after robot cleaning it is shown in table 4.1.

Table 4.1

Number of Records Resulted After Three Data Cleaning Phases in Anonymous Microsoft Web Dataset

Data Cleaning Phase	Number of Records
Initial Log	37711
After removing local and global noise, graphics and videos format records	29862
After checking status code and method field	26854
After Cookies cleaning process	18452
After Robot cleaning	12659

Figure 4.1: Time Taken for User Interested Pattern Prediction in Anonymous Microsoft Web Dataset

Figure 4.1 shows that the time required for the prediction of user interested pattern using initial log is 165 seconds, whereas, 126 seconds and 71 seconds after cleaning by gif status removal and Finally, it takes only 58 seconds.

4.2 MSNBC.com Anonymous Web Dataset

This dataset consists of 989818 records in the log file. Then the data cleaning process is carried out. Initially, after removing records with local and global noise, graphics and videos format such gif, JPEG, etc., 865412 records are obtained. Then by checking the status code and method field, the total of 824509 records is resulted. 631489 records are resulted after applying robot cleaning process and finally, 600181 records are obtained after applying cookies cleaning process it is shown in table 4.2.

Table 4.2

Number of Records Resulted After Four Data Cleaning Phases in Msnbc.Com Anonymous Web Dataset

Data Cleaning Phase	Number of Records
Initial Log	989818
After removing local and global noise, graphics and videos format records	865412
After checking status code and method field	824509
After cookies cleaning process	631489
After robot cleaning	600181

Figure 4.2 shows that the time required for the prediction of user interested pattern using initial log is 287 seconds, whereas, 189 seconds and 46 seconds after cleaning by gif status removal and finally, it takes only 40 seconds.

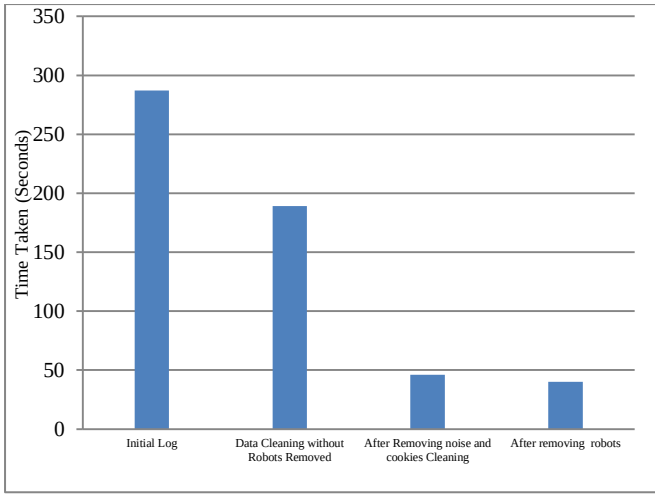

Figure 4.2: Time Taken for User Interested Pattern Prediction in MSNBC.com Anonymous Web Dataset

5. CONCLUSION

Web log data is a collection of huge information. Many interesting patterns available in the web log data. But it is very complicated to extract the interesting patterns without preprocessing phase. Preprocessing phase helps to clean the records and discover the interesting user patterns and session construction. But understanding user's interest and their relationship in navigation is more important. For this along with statistical analysis data mining techniques is to be applied in web log data. Data preprocessing treatment system for web usage mining has been analyzed and implemented for log data. Data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the failed HTTP status code and finally robots cleaning. Different from other implementations records are cleaned effectively by removing local and global noise and robot entries. This preprocessing step is used to give a reliable input for data mining tasks. Accurate input can be found if the byte rate of each and every record is found. The data cleaning phase implemented in this paper will helps in determining only the relevant logs that the user is interested in. Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web Dataset are used for evaluating the proposed preprocessing technique and it reveals that number of records.

6. REFERENCES

[1] Etminani, K., Delui, A.R., Yanehsari, N.R. and Rouhani, M., "Web Usage Mining: Discovery of the Users' Navigational Patterns Using SOM", First International Conference on Networked Digital Technologies, Pp.224-249, 2009.

[2] Jianxi Zhang,Peiying Zhao, Lin Shang and Lunsheng Wang, "Web Usage Mining Based On Fuzzy Clustering in Identifying Target Group", International Colloquium on Computing, Communication, Control, and Management, Vol. 4, Pp. 209-212, 2009.

[3] Nina, S.P., Rahman, M., Bhuiyan, K.I. and Ahmed, K., "Pattern Discovery of Web Usage Mining", International Conference on Computer Technology and Development, Vol. 1, Pp.499-503, 2009.

[4] Chih-Hung Wu, Yen-Liang Wu, Yuan-Ming Chang and Ming-Hung Hung, "Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment", International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 6, Pp. 2909-2914, 2010.

[5] Aghabozorgi, S.R. and Wah, T.Y., "Using Incremental Fuzzy Clustering to Web Usage Mining", International Conference of Soft Computing and Pattern Recognition, Pp. 653-658, 2009.

[6] Maratea, A. and Petrosino, A., "An Heuristic Approach to Page Recommendation in Web Usage Mining", Ninth International Conference on Intelligent Systems Design and Applications, Pp. 1043-1048, 2009.

[7] Inbarani, H.H., Thangavel, K. and Pethalakshmi, A., "Rough Set Based Feature Selection for Web Usage Mining", International Conference on Conference on Computational Intelligence and Multimedia Applications, Vol. 1, Pp. 33-38, 2007.

[8] Jalali, M., Mustapha, N., Sulaiman, N.B. and Mamat, A., "A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems", 12th International Conference Information Visualisation, Pp. 302-307, 2008.

[9] Shinde, S.K. and Kulkarni, U.V., "A New Approach for on Line Recommender System in Web Usage Mining", International Conference on Advanced Computer Theory and Engineering, Pp. 973- 977, 2008.

[10] Zhang Huiying and Liang Wei, "An intelligent algorithm of data pre-processing in Web usage mining", Fifth World Congress on Intelligent Control and Automation, Vol. 4, 3119- 3123, 2004.

[11] Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R., "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 2, Pp. 202-215, 2008.

[12] Hogo, M., Snorek, M. and Lingras, P., "Temporal Web usage mining", International Conference on Web Intelligence, Pp. 450-453, 2003.

[13] DeMin Dong, "Exploration on Web Usage Mining and its Application", International Workshop on Intelligent Systems and Applications, Pp. 1-4, 2009.

[14] Yan Li, Boqin Feng and Qinjiao Mao, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, Vol. 1, Pp. 554-559, 2008.

[15] Baraglia, R. and Palmerini, P., "SUGGEST: a Web usage mining system", International Conference on Information Technology: Coding and Computing, Pp. 282-287, 2002.

[16] Jian Chen, Jian Yin, Tung, A.K.H. and Bin Liu, "Discovering Web usage patterns by mining cross-transaction association rules", International Conference on Machine Learning and Cybernetics, Vol. 5, Pp. 2655-2660, 2004.

[17] Wu, K.L., Yu, P. S. and Ballman, A., "SpeedTracer: A Web usage mining and analysis tool", IBM Systems Journal, Vol. 37, No. 1, Pp. 89-105, 1998.

[18] Labroche, N., Lesot, M.J. and Yaffi, L., "A New Web Usage Mining and Visualization Tool", 19th IEEE

International Conference on Tools with Artificial Intelligence, Vol. 1, Pp. 321-328, 2007.

[19] Chu-Hui Lee and Yu-Hsiang Fu, "Web Usage Mining Based on Clustering of Browsing Features", Eighth International Conference on Intelligent Systems Design and Applications, Vol. 1, Pp. 281-286, 2008.

[20] http://archive.ics.uci.edu/ml/datasets/Anonymous+Microsoft+Web+Data

[21] http://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data

[22] N. Kushmerick , "Learning to remove internet advertisements , " In third annual Conf. on Autonomous Agents , ACM press, NY 1999.

[23] Z. Bar-Yossef and S. Rajagopalan . Template detection via data mining and its applications. In the Eleventh International World Wide Web Conference (WWW 2002). ACM press,New York, 7-11May 2002.

[24] D.Chakraborti,R.Kumar,K.Punera, "Page level template detectiojn via isotonic smoothing", in WWW'07, 2007.

[25] Cooley, R., Mobasher, B., and Srivastava, J. (1999). "Data preparation for mining World Wide Web browsing patterns", Knowledge and Information Systems, 1999.