An Effective Preprocessing Method for Web Usage Mining

K. Sudheer Reddy, G. Partha Saradhi Varma, and M. Kantha Reddy

***Abstract*—Web usage mining (WUM) is one of the categories of data mining technique that identifies usage patterns of the web data, so as to perceive and better serve the requirements of the web applications. The working of WUM involves three steps – preprocessing, pattern discovery and analysis. The first step in WUM - Preprocessing of data is an essential activity which will help to improve the quality of the data and successively the mining results. This research paper studies and presents several data preparation techniques of access stream even before the mining process can be started and these are used to improve the performance of the data preprocessing to identify the unique sessions and unique users. The methods proposed will help to discover meaningful pattern and relationships from the access stream of the user and these are proved to be valid and useful by various research tests. We have concluded this paper by proposing the future research directions.**

***Index Terms*—Web usage mining, data preprocessing, weblog, user session, path completion.**

I. Introduction

Data mining is the study of data-driven techniques to discover patterns in large volumes of raw data. Web mining can be referred as the transformation of the data mining techniques to web data. Web mining has three distinct phases involved – content, structure and usage mining of web data. Mining the content involves extracting the relevant information, structure mining studies the structure and prototype and usage mining is the analysis of the discovered patterns. Web usage mining (WUM) is all about identifying user browsing patterns over WWW, with the aid of knowledge acquired from web logs. The outcomes of the WUM can be used in web personalization, improving the performance of the system, modification of the site, business intelligence, usage characterization etc.

The working of WUM has three steps –preprocessing of the data, pattern discovery and analysis of the patterns. Results of the pattern discovery directly influenced the quality of the data processing. Good data sources not only discover quality patterns but also improve the WUM algorithm. Hence, data preprocessing is an important activity for the complete web usage mining processes and vital in deciding the quality of patterns. In data preprocessing, the collection of various types of data differs not only on type of data available but also the data source site, the data source size and the way it is being implemented.

The data preprocessing of WUM is focused research field nowadays. This research paper studies the preprocessing of data in Web usage mining. This research paper has organized in various sections, includes – related research in this area, brief descriptions about the preprocessing of the usage mining and the proposed algorithms, investigation study to verify the productivity and effectiveness of the algorithms suggested and finally, conclusions and future scope of study has been proposed.

Manuscript received December 28, 2013; revised February 17, 2014.

K. Sudheer Reddy is with the Dept. of Computer Science & Engineering of Acharya Nagarjuna University, Guntur, AP, India (e-mail: sudheercse@gmail.com).

G. P. Saradhi Varma is with the Dept. of Information Technology, SRKR Engineering College, Bhimavaram, AP, India.

Kantha Reddy is with Indo US Collaboration for Engineering Education (IUCEE), UML, Lowell, USA.

II. Research Overview

This section, we can bring together some related research in data preprocessing. Web usage mining research has been a focus area for researchers across various research organizations and academia. Researchers found lot of interesting facts in this area. Though, preprocessing of the data in web usage mining has got fewer contributions and more focus was put on pattern discovery. Robert Cooley et.al [1] have investigated methods for identification of the user, session, episode, page view and path completion and also proposed some informal methods to address the complexities during preprocessing of data. Li Chaofeng proposed various data preprocessing algorithms for improving the efficiency. However, not even a single method has serious issues. In other research [2], the authors are compared time-based heuristics for visit reconstruction.

III. Data Preprocessing

The data preprocessing is the initial step in the data preparation process, aims to reformat the original logs to identify user's sessions. This process is most time consuming and intensive step. A user session file is an input to the web usage mining process that gives information on who accessed the page of a web site, what pages accessed, the order in which the pages accessed and total time spent on each page. Web server writes information whenever a user requests a resource from the site. A web server generally stores all user based activities of the web site in the form of server logs.

The server log files acts as a primary data sources in Web usage mining, which include - access logs of the web server and application server logs. The important task in the preprocessing phase is field extraction. The log files containing log entries which represents the single click stream. The log entry comprises of several fields which need to be isolated for further processing. The process of isolating

DOI: 10.7763/IJCTE.2014.V6.900

various fields from a single line of the log file is known as field extraction. We can use programming logic to separate various fields from the log files. All the log files collected from the data source are sorted and joined together in a single log file.

Due to different server setting parameters, there exists several types of web logs, but typically these log files (a log file is a simple text file), share the basic information, such as: user IP address, request time, requested Uniform Resource Locator, HTTP status code, referrer, and so on. Data sets, which has web log records for 2658 users were collected from SRKR Engineering College website. Web log consist of 17 attributes, each represents a data value in the form of records. The following is the fragment of the IIS server logs:

date time c-ip cs-username s-sitename s-computername sip s-port cs-method cs-uri-stem cs-uri-query sc-status timetaken cs-version cs-host cs(User-Agent) cs(Referer)

Generally, data cleaning, identification of user, session and path completion are the various steps involved in preprocessing (see Fig. 1).

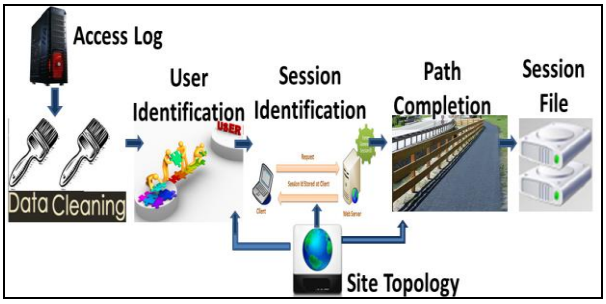

Fig. 1. Various phases of data preprocessing.

A. Data Cleaning

The aim of WUM is to acquire the traversal pattern. Thus, the data cleaning task is essential which involves removing the log entries which are irrelevant and redundant. This activity is usually a site-specific. There are two kinds of irrelevant data needed to clean: extraneous references to embedded objects and error requests.

Due to stateless or connectionless nature of the HTTP protocol, a user's request to browse a page results in several log entries since graphics and other scripts are downloaded along with HTML file content. Since, the main objective of web usage mining is to have a clear portrait of the web user behavior, hence elimination is required for files having the suffixes such as, jpeg. Jpg, gif, css, cgi, etc., which we can found in cs_uri_stem field.

Error codes are not relevant and not useful for mining. We can check the status and they can be removed, if found irrelevant. There are four different categories of status codes – Success (series starts from 200), Redirect (series starts from 300), Failure (series starts from 400), Server Error (series starts from 500). From these, we can eliminate all error codes say, 401 (failed authentication), 404 (file not found) which are not required for analysis process and they are cleaned from the logs. There is a need to eliminate at least some of the data fields using this cleaning process.

Once preprocessing done, data integration from multiple sources will be done and then transforming data to an acceptable form, which serves as an input to various mining processes.

B. User Identification

The aim of the user identification process is to find out the different users from the web access log file. Different users are being distinguished by using their Internet Protocol (IP) addresses. The method used for this process is a referrer-based method. User identification is complex due to the presence of resident caches, firewalls and proxy servers. To deal this problem, we can employ the WUM methods that rely on user cooperation. However, it's difficult because of high security and privacy. We have the following heuristics used in our testing methodology to identify the user:

1) Each IP address represents one user;
2) If the IP address is same for more logs, but the agent log displays a change in browser or operating system, the IP address represents a different user;
3) If there is a same IP address, browser and operating system, the referrer information can be considered. If a user requested page is not directly accessible by a link from any of these pages, hence with the same IP there is another user.

C. Session Identification

The aim of the user session identification is to find out the different user sessions from the web access log file. A set of user clicks usually referred to as a click stream, across Web servers is defined as a user session. The user session identification involves - dividing the page accesses of every user into separate sessions. At present, we have the methods which will identify user session mainly include timeout mechanism [3] and maximal forward reference [4]. The following rules deployed to identify user session in our research:

1) If there is a new user, and hence, there is a new session;
2) If the refer page is null in one user session, there is a new session;
3) It is presumed that, the user is starting a new session, If the time frame between page requests exceeds a limit (usually 25.5 or 20 minutes)

D. Path Completion

There are many important user accesses that are not being recorded in the access log due to the existence of proxy server and local cache. The aim of the path completion is to acquire complete user access path by filling up the missing page references. The incomplete access path is recognized based on user session identification [5]. We can employ the same methods which used for user identification. For example, a user requested for a page, that is unlinked to the last page. We can use the referrer log to check what page the request came from? If the page is available in the users recent history, it is anticipated that the user has backtracked using the back button, bringing up the cached versions of the pages till a new page requested. The site topology can be used to if the referrer log in unclear to this effect. If in a start of the user session, Referrer as well URI has a data value, delete value of the referrer by adding a delimiter '-'. Web log preprocessing

helps in removing unwanted click-streams from the log file and also reduces the original file size by 50-55%.

IV. Data Preprocessing - the Process and Results

This section focuses on performance and results of the proposed model. To validate the efficiency of our methodology mentioned above, we have conducted research trial with the SRKR Engineering College web server log. The data source size is 42MB and the research trial conducted from November 6, 2011 to December 18, 2011. Our experiments were performed on a 2.8GHz core2duo processor, 2 GB of primary memory, Windows 2003 server operating system, SQL Server 2000 and JDK 1.6.

The following matrix depicts the entries of raw web logs, entries after cleaning, number of users accessed and sessions recorded.

TABLE I: The Process and Results

#Raw weblog entries	# entries after data cleaning	#Users accessed	#Sessions recorded
45692	5613	2658	3046

The results based on the above entries in the form of pictorial representations will be presented.

In the Fig. 2, we have shown the overall process items and results.

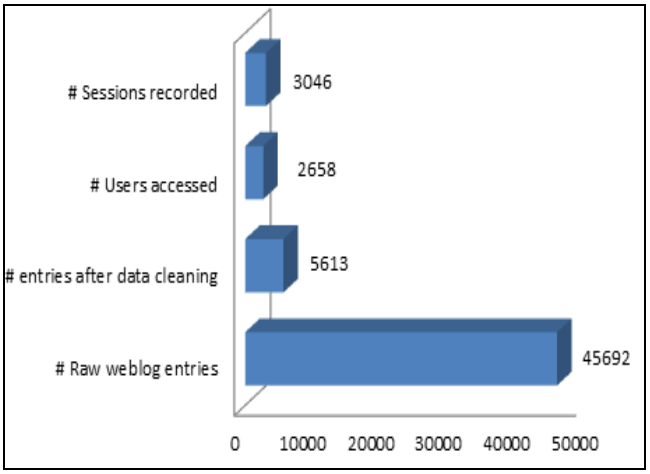

Fig. 2. Data cleaning process.

Only 5613 entries produced from over 45692 log entries, after the cleaning process. This shows that, only 12.2% data is being found as relevant and the remaining have been removed by the method suggested. The following table shows the data cleaning process at each level, i.e., after removal of GIF, CSS, JPEG, and other files.

TABLE II: Data Cleaning – by File Type and Results

#Raw weblog entries	#after removal of .GIF	#after removal of JPEG	#after removal of CSS	#after removal of error codes	#entries after data cleaning
45692	34362	22681	10864	8267	5613

In Fig. 3 shown below is an illustrative, which shows all relevant entries right from the total raw log entries, after removing the .GIF, .JPEG, .CSS and error codes and final figure after the completing cleaning process.

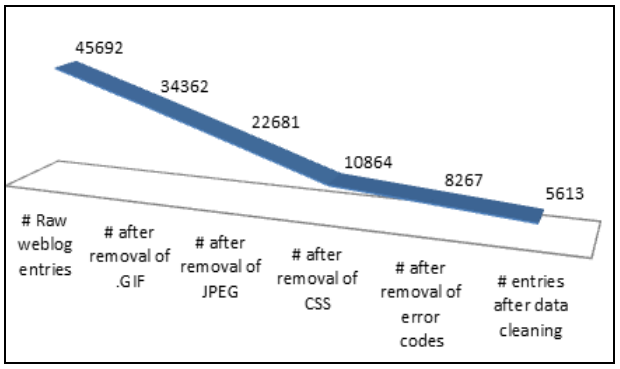

Fig. 3. Data cleaning process – after removal of GIF, JPEG, CSS.

Finally, the user details are shown in the table given.

TABLE III: User Details and Results

#Users	# users with unique IP	# users with same IP
2658	678	247

Fig. 4, given below is for User identification. Bar # 1 indicates number of users identified local proxy, Bar# 2 indicates, users with unique IP and agent and Bar# 3 indicates with the users identified only by IP address.

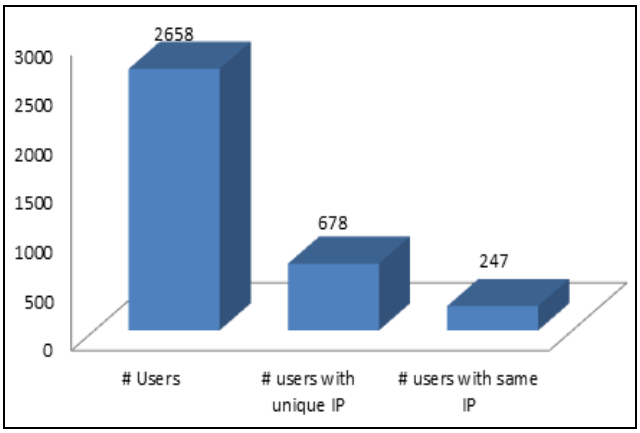

Fig. 4. Process of user identification.

Based on the user identification's results, 3046 sessions have been identified on a threshold of 25 minutes and path completion.

V. Conclusions

An important task in data mining application is the creation of a suitable data set to which mining and algorithms can be applied. This is an important activity in WUM due to the various characteristic features of the click stream data. The data preparation process is the most time consuming and intensive step in mining web usage data. This paper has presented various details about data preprocessing activities that are necessary to perform Web Usage Mining. In every phase of the data preprocessing, we give some rules to design and implement them easily and efficiently. Our experiments have estimate data preprocessing importance and our methodology's effectiveness. It is not only to reduce the size of the log file but also increases the quality of the data available. However, still there are problems remain such as

data collection, the accuracy metric of the user identification and the session identification and applying the results of the preprocessing to discover patterns.

REFERENCES

[1] S. Jaideep, R. Cooley, M. Deshpande, and P. Tan, "Web usage mining: discovery and applications of usage patterns from web data," *SIGKDD Explorations*, vol. 1, no. 2, pp. 1-12, 2000.

[2] H.-H. Dai and B. Mobasher, "Integrating semantic knowledge with web usage mining for personalization," *Web Mining: Applications and Techniques*, 2007.

[3] M. S. Chen, J. S. Park, and P. S. Yu, "Data mining for path traversal patterns in a web environment", in *Proc. 16th International Conference on Distributed Computing Systems*, 1996, pp. 385-392.

[4] G. T. Rajul and P. S. Satyanarayana, "knowledge discovery from web usage data: Complete preprocessing methodology," *International Journal of Computer Science and Network Security*, vol. 8 no. 1, January 2008.

[5] D. Tanasa and B. Trousse, "Advanced data preprocessing for intersites in Web usage mining," *IEEE Intelligent Systems*, vol. 19, pp. 59–65, 2004.

K Sudheer Reddy completed his doctoral degree from Acharya Nagarjuna University, AP, India and master's degree in computer science & engineering. He is a member of several international journals and conferences across the globe. Currently, he is guiding three Ph. D scholars in the area of data mining. He has published over 15 papers in International journals and conferences in the area of Computer Science & Engineering and Education space.

M. Kantha Reddy is working as a director in operations, India for Indo US Collaboration for Engineering Education (IUCEE), UML, Lowell, USA. He has presented several papers in national and international conferences and journals. He is also actively participating in designing academic curriculum with several reputed institutions.