EXPLORATIONS ON ENGINEERING LETTERS (EEL)
VOLUME 1, ISSUE 1 (2016):PP.193-199
SANA ACADEMIC PRESS

Data Pre-processing of Web usage Mining-A Vital Stage in Information Retrieval

G. Rajakumari[1] and N. Venkatesan[2]
[1] M.Phil. Research Scholar, St. Peter's University, Chennai.
[2]Assistant Prof., Dept of Computer Applications, St. Peter's University, Chennai.
grajiapk@gmail.com

Abstract – *The World Wide Web has an ocean of information helping every feature of our lives in modern era. Having enormous data, mining the right information becomes a challenging task. Web mining uses many data mining techniques but it is not an application of usual data mining due to heterogeneity and shapeless nature of the data on web. Web mining tasks can be characterized into three types: Web Content Mining, Web Structure Mining and Web Usage Mining. The goal of Web Usage Mining is to capture, model and analyse the behavioural patterns and profiles of users interacting with a Website. Web Usage Mining consists of many stages but our study focuses on first stage i.e., Data Pre-processing and its techniques for an efficient information retrieval mechanism. Data Pre-processing consists of data cleaning, page view identification, sessionization, data integration and data transformation.*

Keywords: *Pre-processing, Web Content mining, Web Structure Mining, Web Usage Mining*

I. INTRODUCTION

World Wide Web is become a huge storeroom of web pages and links. The internet users can search out huge quantity of data i.e., information. Information Retrieval is the science of searching for documents, information within documents, metadata about documents, relational databases from the World Wide Web. Web Mining is the use of Data Mining techniques to automatically discover and extract information from Web Documents and Services. Web Mining is divided into (i) Web Content Mining, is the scanning and mining of text, pictures and graphs of a Web Page. This mining is used to gather, categorize, organize and provide the best possible solution in the internet to the user's request. (ii) Web Structure Mining, is a tool to identify the relationship between the web pages linked by information. (iii) Web Usage Mining, allows collecting the web access information for the Web Page. When a user access a web page an entry is created in web server's log file. So the log entries are increasing. There are three inter dependent stages (i) Data Preprocessing (ii) Pattern Discovery and (iii) Pattern Analysis.

Data Preprocessing: Log file consists of lot of irrelevant entries which is to be removed. To enhance the efficiency of mining noise is to be removed before mining. A series of process like data cleaning, page view identification, sessionization, data integration and data transformation are handled.
Pattern Discovery: Application of various data mining techniques to processed data like statistical analysis, association, clustering, pattern matching and so on.

Pattern Analysis: Uninteresting rules are ruled out and analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

Preprocessing is the first step in Web Usage Mining, which is essential activity which will help to improve the quality of the data and successively the mining results.

This paper has organized as follows, in section II, related research works of data preprocessing in web usage mining is discussed. In section III, we will discuss about the preprocessing of the usage mining and web log file structure. In section IV, we will discuss about investigation study to verify the productivity and efficiency. In section V, we will discuss on the conclusion and future research works.

II. RELATED RESEARCH WORKS

Preprocessing of the raw data in the server log file is an important step in web usage mining to ensure the quality of information used in mining.

Wasavand et al., 2014, identified user's navigation pattern by data cleaning on web log file. They also used classification algorithms to identify user's interested website.

Manisha.V, 2014, conducted data cleaning and distinct user identification technique which enhances the preprocessing steps of web log usage data. Using user identification they found out the distinct user based on their attended session time. This will help in personalizing the websites.

Mitali, S. et al., 2015, proposed an algorithm for the extraction of data sorted based on the analysis of time duration. Also, an algorithm is proposed in data cleaning to remove almost all irrelevant files, irrelevant HTTP methods and wrong HTTP status codes. After experiment it is analyzed that raw log data reduces to almost 80% which shows the importance of initial phases of data preprocessing.

Shahu, M.S et al., 2015, achieved data cleaning by removing Global and Local Noise, Multimedia Images, HTTP Status code, Request from web robots.
Cooley et al., 1997, presented methods for user identification, session identification, page view identification, path completion, and episode identification. They proposed some heuristics to deal with the difficulties during data preprocessing.

Addanki Ramya et al., 2012, used multi-layered network architecture with a back propagation learning mechanism to discover and analyze useful information from the available web log data. The discovered data is used to predict the user's behavior E-Commerce applications.

Bagilon et al., 2003, aimed at extracting models of the navigational behavior of website users. After the data cleaning process they carried out two experiments: the first one tries to predict the sex of a user based on the visited web pages, and the second one tries to predict whether a user might be interested in visiting a section of the site.

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 website, 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 website 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 various fields from a single line of the log file is known as field extraction. The programming logic was employed 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 nearby 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-hosst cs(User-Agent) cs(Referer)

Generally, data cleaning, identification of user, session identification and path completion are the various steps involved in preprocessing.

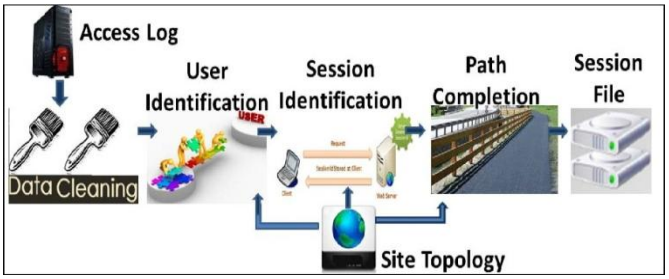

Fig. 1.Various phases of data preprocessing.

A.Data Cleaning

The aim of web usage mining is to attain 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: irrelevant 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 goal of web usage mining is to have a clear illustration of the web user behavior; hence, Elimination is required for files having the suffixes such as, jpeg,jpg, gif, css, cgi, etc., that were found in cs_uri_stem field.

Error codes are not relevant and not useful for mining. The status can be checked and they can be removed, if it found as 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, all error codes can be eliminated 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, the web usage mining methods were employed that rely on user cooperation. However, it's difficult because of high security and privacy. The following heuristics were used in testing the proposed methodology to identify the user:

Each different IP address represents different user;

- Even if the IP address is same, if the agent log displays a change in browser or operating system. So, each different agent type for an IP address represents a different user.

- If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user assumes that there is another user with the same IP address.

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, the methods that are currently available will identify user session mainly include timeout mechanism and maximal forward reference. The following rules deployed to identify user session in the proposed research:

If there is a new user, and hence, there is a new session;

- If the refer page is null in one user session, there is a new session;
- 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 significant user accesses that are not being recorded in the access log due to the continuation of proxy server and local cache. The aim of the path completion is to obtain complete user access path by filling up the missing page references. The incomplete access path is acknowledged based on user session identification. The same methods were employed that used for user classification. For example, a user requested for a page, that is unlinked to the last page. The referrer log used to check what page the request came from? If the page is available in the users recent history, it is expected 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 result. 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 the proposed methodology, several research trials were conducted with the nearby engineering college web server log. The data source size is 42MB and the research trial conducted from January 31, 2016 to March 10, 2016. 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 are presented below, based on the above table entries in the form of pictorial representations. In the figure 2, the overall process items and results are shown.

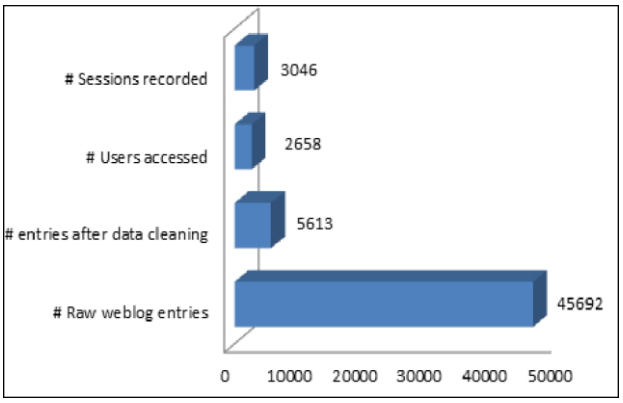

Fig. 2. Data Cleaning Process.

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 the figure 3, 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.

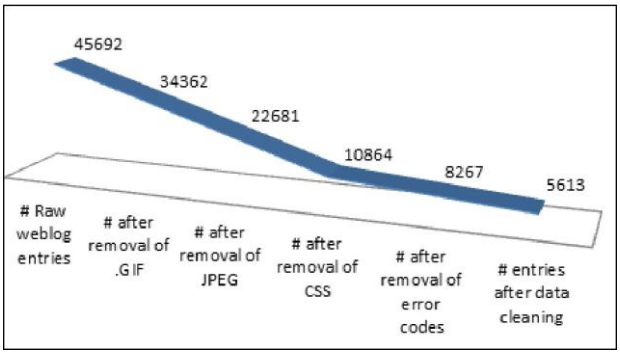

Fig.3. Data Cleaning process–After removal of .gif, .jpeg, .css.

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

TABLE III. USER DETAILS AND RESULTS

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

Figure 4, refers to the user identification details. Bar #1 indicates number of users identified by the local proxy, Bar #2 indicates, number of users with unique IP and agent and Bar #3 indicates with the users identified only by IP address.

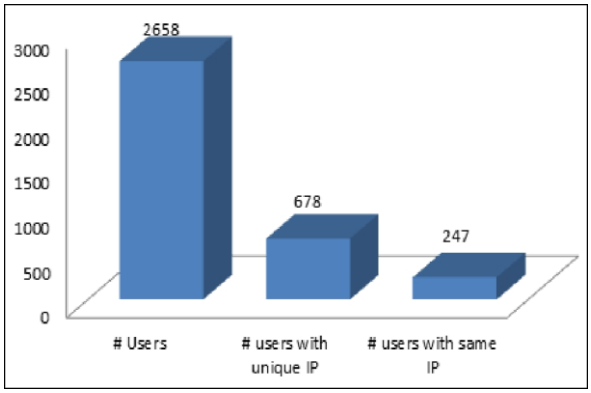

Fig. 4. Process of user identification

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

V. CONCLUSIONS

A vital task in data mining application is the creation of an appropriate data set to which mining and algorithms can be applied. This is a significant activity in web usage mining due to the various characteristic features of the click stream data. This process is the most time consuming and

intensive step in web usage mining. This paper dealt about various details of data preprocessing activities that are necessary to perform Web Usage Mining. In every stage of the data preprocessing, some regulations given to design and implement them without problems and resourcefully. Our experiments have estimate data preprocessing significance and our methodology's effectiveness. It reduced the size of the log file and also increases the quality of the data available. However, still there are problems remain such as data collection, the exactness metric of the user identification and the session identification and applying the results of the preprocessing to discover patterns.

REFERENCES

[1]. Bagilon, M. & Ferrara, U. & Romei, A. & Ruggieri, S. & Turini, F. "Preprocessing and Mining Web Log Data for Web Personalisation". *Advances in Artificial Intelligence Lecture Notes in Computer Science, Volume 2829, 2003, pp 237-249.*

[2]. Cooley R, Mobasher B, Srivastava J.,"Web Mining: Information and Pattern Discovery on the World Wide Web". *In International Conference on Tools with Artificial Intelligence, pages 558-567, IEEE, 1997.*

[3]. Manisha V. "A Step up in Data Cleaning and User identification of Preprocessing on Web Usage data". *International Journal of Advanced Research in Computer Engineering and Technology IJARCET, 2014*

[4]. Mitali, S. & Rakhi, G. & Mishra, P. K. "Analysis of Data Extraction and Data Cleaning in Web Usage Mining", Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015).

[5]. Ramya, A. & Sreenu, K. & Ratna, K. "Preprocessing and Unsupervised Approach for Web Usage Mining*", International Journal of Social Networking and Virtual Communities, Vol. 1, Issue 2, 2012.*

[6]. Shahu, M.S & Leena "A Survey on Frequent Web Page Mining with Improving Data Quality of Log Cleaner", *International Journal of Advanced Research in Computer Engineering & Technology, 2015.*

[7]. Srinaganya.G. "A Technical Study on Information Retrieval using Web Mining Techniques". *IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication systems (ICIIECS) 2015.*

[8]. Wasavand, C. & Devale, P.R & Ravindra, M. "Data Preprocessing Method of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern". *IJISET-International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, 2014.*