

# An Efficient Preprocessing Mechanism for Web usage Mining

Dipika Sahu<sup>1</sup> Yamini Chouhan<sup>2</sup>

<sup>1,2</sup>Department of Computer Science & Engineering

<sup>1,2</sup>Faculty of Engineering and Technology Shri Shankaracharya Technical Campus, Junwani, Bhilai, District-Durg, Chhattisgarh-490020, India

**Abstract**— Web mining is the application of data mining techniques to extract knowledge from Web data, including Web document, hyperlink between document, usages logs of web sites, etc. Data preprocessing has a fundamental role in Web Usage Mining applications. There are two stages [1] Data clean consist of removing all the data records in web logs that are useless for mining purpose e.g.: request for graphical page content (e.g., jpg and gif images); requests for any other file which may be include into a web page; or even navigation sessions performed by robots and web spiders. [2] Session Identification and Reconstruction consists set of pages visited by the same user within the duration of one particular visit to a web site. Data Formatting is the final step of preprocessing. Once the previous phases have been complete, data are properly formatted before applying mining techniques, stores data extracted from web logs into a relational database using a click fact scheme, so as to provide superior support to log querying finalized to frequent pattern mining.

**Key words:** Web Mining, Data Cleaning, Session Identification, Data Formatting and Log Querying

## I. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from Web data, i.e. Web Contents, Web Structures and Web Usage data. The attention paid to Web mining, in research, software industry, and Web-based organization, has led to accumulation of a lot of experiences. It is our attempt in this paper to capture them in a systematic manner, and recognize directions for future research.

The web is extremely enormous, diverse, flexible, and dynamic. The World Wide Web continues to develop both in huge volume of traffic and the size and complexity of Web sites. With the increasing growth of information accessible in net, it is difficult to identify the relevant information present in the web. Meanwhile much information is unstructured. It is essential to use automated tool for obtaining the necessary information from the huge collection of information.

Web Mining is used to extract information from the raw unstructured data. The emerging field of web mining objective at finding and extracting relevant information that is hidden in Web related data, in particular in text files published on the Web. Web mining is performed in three ways they are 1) web usage mining 2) web content mining 3) web structure mining. Web usage mining gives the support for the web site design, giving personalization server and the other business making decisions, etc. Web content mining is the process of extracting knowledge from the content of files or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent; based technology might also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and the link

between references and referents in the Web. Finally, web usage mining, also known as the Web Log Mining, is the process of extracting interesting patterns in web access logs. In order to superior serve for the user, web mining apply the data mining, the artificial intelligence and the chart technologies and so on to web data and traces user visiting characteristic, and then extracts the users' using pattern. It has rapidly become one of the most significant areas in Computer and Information Sciences because of its direct applications in the e-commerce, CRM, Web analytic, information retrieval and filtering, and Web information systems.

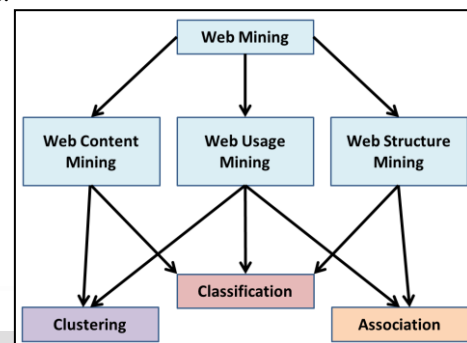


Fig. 1: Taxonomy of Web Mining

Data preprocessing has a fundamental role in Web Usage Mining applications. The preprocessing of web logs is generally complex and time demanding. It comprises four different tasks:

- The data cleaning.
- The identification and the reconstruction of users' session.
- The retrieving of information about page content and structure.
- The data formatting.

### A. Data Cleaning

This step consist of removing all the data tracked in web logs that are useless for mining purpose e.g.: request for graphical page content (e.g., jpg and gif images); requests for any other file which may be include into a web page; or even navigation sessions performed by robots and web spiders. While requests for graphical contents and documents are easy to eliminate, robots' and web spiders' navigation patterns must be explicitly identified. This is generally done for instance by referring to the remote hostname, by referring to the user agent, or by checking the access to the robot.txt files. However, some robots actually send a false user agent in HTTP request. In this case, a heuristic based on navigational behavior can be used to separates robot sessions from actual users' sessions.

### B. Session Identification and Reconstruction

This step consists of (i) identifying the different users' sessions from the generally very poor information available in log files and (ii) reconstructing the users' navigation path

within the recognized sessions. Most of the problems encountered in this phase are caused by the caching performed either by proxy server either by browser. Proxy caching causes a single IP address (the one belonging to the proxy Server) to be associated with diverse users' sessions, so that it becomes impossible to use IP addresses as users identifies. This trouble can be partially solved by the use of cookies, by URL rewriting, or by requiring the user to log in when entering the web sites. Web browsers caching is a further difficult issue. Log from web servers cannot include any information about the utilized of the back button. This can generate inconsistent navigation paths in the users' sessions. However, by utilizing further information about the web site structure is still possible to reconstruct a consistent path by means of heuristic. Because the HTTP protocols are stateless, it is nearly impossible to determine when a user really leaves the web site in order to determine when a session should be considered finished. This problem is referred to as sessionization. Described and compared three heuristics for the recognition of sessions termination; two were based on the time between users' pages request; one was based on information about the referrer. Proposed an adaptive time out heuristic. Proposed a method to infer the timeout threshold for the specific web site. Other authors proposed different thresholds for time oriented heuristic base on empiric experiment.

### C. Content and Structure Retrieving

The vast majority of Web Usage Mining applications utilize the visited URLs as the main source of information for mining purposes. URL is however a poor source of information since; for example, they do not convey any information about the actual page content has been the initial to employ content based information to enrich the web log data. If an adequate cataloging is not known in advance, Web Structure Mining techniques can be employed to develop one. As in search engines, web page is classified according to their semantic regions by means of Web Content Mining techniques; this classification information can then be utilized to enrich information extracted from logs. For instance, proposes to use Semantic Web for Web Usage Mining; web page is mapped into ontologies to add meaning to the observed frequent paths. Introduces concept-based path as an alternative to the usual users navigation paths; concept-based path are a high level generalization of usual path in which ordinary concepts are extracted by means of intersection of raw user pats and similarity measure.

### D. Data Formatting

This is the final step of preprocessing. Once the previous phases have been complete, data are properly formatted before applying mining techniques, stores data extracted from web logs into a relational database using a click fact scheme, so as to provide superior support to log querying finalized to frequent pattern mining. Introduces a method based on signature tree to index log stored in databases for efficient pattern queries. A tree structure, WAP-tree, is also introduced in to register access sequence to web pages; this structure is optimized to exploit the sequence mining algorithm.

## II. RELATED WORK

Pooja Mehta et al. (2012) have worked on "Web Personalization Using Web Mining: Concept and Research Issue" and the web mining was the application of the data mining which was useful to extract the knowledge. Web mining has been explored to different techniques have been proposed for the variety of the application. Most research on Web mining has been from a 'data-centric' or information based point of view. Web usage mining, Web structure mining and Web content mining were the types of Web mining. Web usage mining was used to mining the data from the web server log files. Web Personalization was one of the areas of the Web usage mining that can be defined as delivery of content tailored to a particular user or as personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it.

Ajebaraj Ratnakumar (2007) has worked on "An Implementation of Web Personalization Using Web Mining Techniques" and the web mining was the application of data mining techniques to extract knowledge from Web. Web mining has been explored to a vast degree and different techniques have been proposed for a variety of applications that includes Web Search, Classification and Personalization etc. Most research on Web mining has been from a 'data-centric' point of view. In this paper, they highlight the significance of studying the evolving nature of the Web personalization. Web usage mining was used to discover interesting user navigation patterns and can be applied to many real-world problems, such as improving Web sites/pages, making additional topic or product recommendations,

Sheng-Tang Wu and Yuefeng Li (2013) have worked on "Pattern-Based Web Mining Using Data Mining Techniques" and the several data mining techniques have been proposed for fulfilling various knowledge discovery tasks in order to achieve the goal of retrieving useful information for users. Data mining techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximum pattern mining and closed pattern mining. However, how to effectively exploit the discovered patterns was still an open research issue, especially in the domain of Web mining.

Abdul Rahaman et al. (2012) have worked on "Web Mining –A Catalyst for E-Business" and in this world of information technology; everyone has the tendency to do business electronically. Today lot of businesses was happening on World Wide Web (WWW), it was very important for the website owner to provide a better platform to attract more customers for their site. Providing information in a better way was the solution to bring more customers or users. Customer was the end-user, who accessing the information in a way it yields some credit to the web site owners.

## III. PROBLEM IDENTIFICATION

Clearly improved data quality can improve the quality of any analysis on it. Data quality is a source of major difficulties for Web usage mining, and particularly for the sessionizing problem. The solution of a dedicated server recording all activities of each user individually was put

forward as a tested and successful - 8 - solution. In the absence of such a server, cookies and scripts can be used to distinguish among unique users. However, they are not always feasible or popular, due to privacy considerations.

A problem in the Web domain is the inherent conflict between the analysis needs of the analysts (who wants more detailed usage data collected), and the privacy needs of users (who want as little data collected as possible). In addition caching and proxy servers can make reconstructing reliable sessions difficult.

#### A. Log Files

In a web log file thousands of people access “log record” has been recorded by the server. Finding and reading useful data from them will be complex. Finding data from large record will be consumes much time. A web log file is a web server files automatically created and maintained by a web server to check the activity performed by it. This log files contain which pages are being accessed, by whom, and when as it maintains a history of page requests on its site.

#### B. Solution to the Problems

The solution of this problem is to clean and to remove the unnecessary entries in web log files. Typical web log cleaning methodologies mainly aim at removing image and picture files with extensions GIF and JPG (if analysis does not involves image examining). The analysis concerns the investigation of media/multimedia file but there are masses of other irrelevant files, which stay untouched during all the analysis process or even have a negative effect on the analysis. These can be internal web administrator actions, special purpose files, etc. The complexity of various web sites can influence the data cleaning process and impact the final results as well.

The Major issues in the process of web mining are:

- Web data sets can be very big; it takes ten to hundreds of terabytes to store on database.
- It can't mine on a single server so it wants large number of server.
- Proper organization of the hardware and the software to mine multi-terabyte data sets.
- Limited customization, restricted coverage, and limited query interface to the individual users.
- Automated data cleaning.
- Over fitting and under the fitting of data.
- Over sampling of the data.
- Scaling up for high dimensional data.
- Mining series and time series data.
- Difficulty in finding related information.
- Extracting new knowledge from the web.
- Data / Information Extraction concentrate on extraction of the structured data from web pages such as products and the search result.
- Web information integration and schema matching. The web contains huge amount of data, each web site accept similar information in a diverse way. Similar data discovery is an significant problem with lots of the realistic applications.
- Opinion extraction from the online sources i.e. customer makes sure of products, forums, blogs and chat room. Mining opinions are of large consequence for marketing intelligence and the product benchmarking.

- Automatically segmenting web pages and identifying noise is an interesting trouble in Web application. It could not have advertisements, navigation link and the copyrights notices. Hence, extracting the main contents of web page is important problem in web application.

### IV. METHODOLOGY

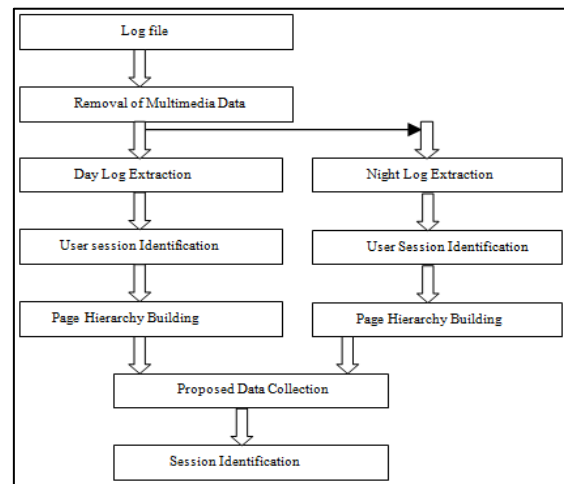


Fig. 2: Flowchart of the methodology

#### A. Step 1: Select the log files which are there in the browsing history

- In computing, the log files refers to the list of web pages a user has visited recently—and associated data such as page title and time of visit—which is recorded by web browser software as standard for a certain period of time.
- In this, a check list of the log files is to maintained and is selected in order to extract those and pre-process it for the further use.

#### B. Step 2: Removal of Multimedia data entry

- Multimedia data refers to data that consist of various media types like text, audio, video, and animation. It refers to data that contain various media types such as text, graphics, animation, audio, and video.
- it should remove entries that have status of “error” or “failure”. It’s to - 3 - remove the noisy data from the data set. To accomplish it is quite easy.
- Secondly, some access records generated by automatic search engine agent should be identified and removed from the access log. Primarily, it should identify log entries created by so-called crawlers or spiders that are used widely in Web search engine tools. Such data offer nothing to the analyzing of user navigation behaviors.

#### C. Step 3: Extraction of the Files, which have been logged in at the day time

- Segregate the log files which have been browsed at the day time and the night time.
- In the first step, extract the log files which have been browsed at the day time.
- This procedure is done usually to avoid the robotic activities and their browsing because there are some possibilities of the robotic browsing at the night time, which is the reason the day logs and night logs are extracted separately.

**D. Step 4: Identify the time or session of the user, when he/she logged in and browsed**

- This task is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. The data recorded by a Web server are not sufficient for distinguishing among different users and for distinguishing among multiple visits of the same person.
- The time or the session of the user at which he/she was logged in and was browsing the files is taken into note.
- A Session ID is a piece of data that is used in network communications to identify a session, a series of related message exchanges.

**E. Step 5: Extraction of the Files, which have been logged in at the Night time**

- Segregate the log files which have been browsed at the night time.
- This procedure is done usually to differentiate the robotic activities and their browsing and the human activities, because there are some possibilities of the robotic browsing at this time.

**F. Step 6: Identify the time or session of the user, when he/she logged in and browsed at the night time**

- The time or the session of the user at which he/she was logged in and was browsing the files is taken into note.
- The night time is divided into the respective sessions and accordingly the session is identified, that in which session user was logged in.
- For Example – Mark was having a look in yahoo.com at 01.44 hours, and then the time is noted and is taken into the session from 00.00 to 03.00 hours session.

**G. Step 7: Build the Hierarchy of the web pages which were browsed at the respective sessions in the day time and night time separately**

- The User logs into the pages, where he/she can find so many links to browse one after another, so a hierarchy of the log files is made accordingly.
- The hierarchy of the web pages are prepared so that it could be known that which web page is logged in first and the sequence of the browsing of web pages is known.

**H. Step 8: Collect the data which has been pre-processed**

- We have pre-processed the data in the previous steps, now we have to accumulate the data for the Pre Processing.
- After collection of data, the data needs to be processed.

**V. RESULT**

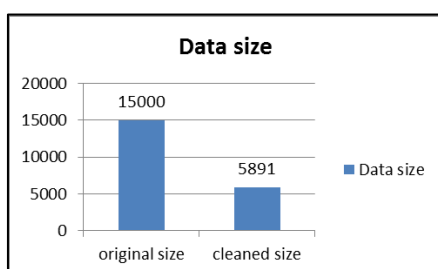


Fig. 3: Comparison of before and after data size cleaning.

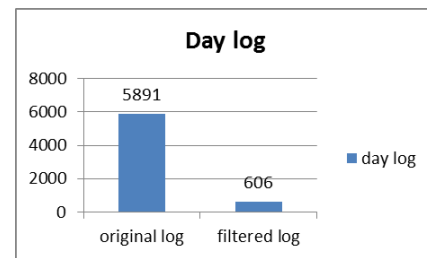


Fig. 4: Cleaning of original and filtered Day log file size

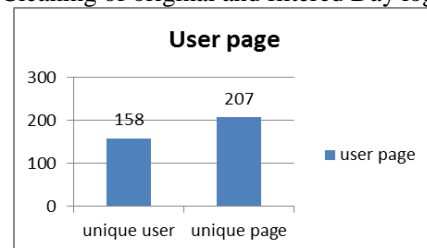


Fig. 5: Comparison of Unique users and unique page

**VI. CONCLUSION**

In this paper we did the research of the mining, focusing on the Preprocessing task. Typical web log cleaning methodologies mainly aim at removing image and picture files with extensions GIF and JPG. In this study, an effective web log mechanism is designed that makes preprocessing of the log data easier and efficient and cleans the log entries that are accessed by the user. The Web mechanism removes the unnecessary logs files, and cleans it by further reducing the size of record. 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. The work easily removes the garbage files, which occupies large amount of space, which is responsible for the perfect data cleaning. It also segregates the log files of day and night separately to identify the files of the user and robotic access. The analysis concerns the investigation of media/multimedia file but there are masses of other irrelevant files, which stay untouched during all the analysis process or even have a negative effect on the analysis. In future scope, the remaining part of the web mining can be taken into the consideration as only the preprocessing is emphasized in the paper.

**REFERENCES**

- [1] Ms. Dipa Dixit, Fr.CRIT, Vashi, M Kiruthika," PREPROCESSING OF WEB LOGS", (IJCS) International Journal on Computer Science And Engineering, Vol. 02, No. 07, 2010, 2447-2452.
- [2] Dr. Sohail Asghar, Dr. Nayyer Masood," Web Usage Mining: A Survey On Preprocessing Of Web Log File Tasawar Hussain", 978-1-4244-8003-6/10@2010.
- [3] Theint Theint Aye "Web Log Cleaning for Mining of Web Usage Patterns".
- [4] S. K. Pani, et.al L "Web Usage Mining: A Survey On Pattern Extraction From Web Logs", International Journal Of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.
- [5] Chidansh Amitkumar Bhatt • Mohan S. Kankanhalli, "Multimedia Data Mining: State Of The Art And

- Challenges” Published Online: 16 November 2010© Springer Science+Business Media, LLC 2010.
- [6] Margaret H. Dunham, Yongqiao Xiao Le Gruenwald, Zahid Hossain,” A SURVEY OF ASSOCIATION RULES Web Usage Mining”.
- [7] Brijendra Singh<sup>1</sup>, Hemant Kumar Singh<sup>2</sup>,”WEB DATA MINING RESEARCH: A SURVEY”, 978-1-4244-5967-4/10/\$26.00 ©2010 IEEE.
- [8] Rajni Pamnani, Pramila Chawan 1 Qingtian Han, Xiaoyan Gao, “Web Usage Mining: A Research Area in Web Mining”.
- [9] Wenguo Wu, “Study On Web Mining Algorithm Based On Usage Mining”, Computer- Aided Industrial Design And Conceptual Design, 2008. CAID/CD 2008. 9th International Conference On 22-25 Nov.2008.
- [10] R. Kosala, H. Blockeel. “Web Mining Research: A Survey,” In SIGKDD Explorations, ACM Press, 2(1): 2000, Pp.1-15.
- [11] <http://www.kdnuggets.com>
- [12] <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.1602>
- [13] J Vellingiri, S.Chenthur Pandian, “A Survey on Web Usage Mining”, Global Journal of Computer Science and Technology .Volume 11 Issue 4 Version 1.0 March 2011.
- [14] Chen L, Mao X,Wei P, Xue Y, Ishizuka M (2012) Mandarin emotion recognition combining acoustic and emotional point information. Appl Intell 37(4):602–612.
- [15] Shang F, Jiao LC, Shi J, Wang F, Gong M (2012) Fast affinity propagation clustering: a multilevel approach. Pattern Recognition 45(1):474–486.
- [16] J. Shao, X. He, C. Bohm, Q. Yang, C. Plant, “Synchronization-Inspired Partitioning and Hierarchical Clustering,” IEEE Transactions on Knowledge and Data Engineering, 2012.
- [17] Ta, sdemir K (2012) Vector quantization based approximate spectral clustering of large datasets. Pattern Recognition 45(8):3034–3044.
- [18] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi,”Overview of Web Content Mining Tools”, The International Journal of Engineering and Science (IJES), Volume 2, Issue 6, 2013.