Preprocessing Techniques for Web Usage Mining

[1]Faizan I Khandwani, [2]Ashok P Kankale

[1]ME Scholar, [2]Professor
Department of Computer Science and Engineering,
RSCE, Buldana, India

***Abstract*—Web usage or log mining can be described as the discovery and analysis of access patterns of users through mining of log files. It is used to analysis Web log files and discovers user accessing patterns of Web pages; it can find user's access patterns automatically and quickly from the vast Web log data, such as frequent access paths, frequent access page groups and user clustering. For analyzing the customer's behavior, the data generated by the users visiting the website must be analyzed. The users' accesses to Web sites are stored in server log files. This will provide foundation for decision making of organizations. But the data stored in these log files do not present an exact picture of the users' accesses to the Web site. So preprocessing the web log data is a pre-requisite before this data can be used for mining tasks. It is a key technology in this mining activity. This preprocessed web data then will be suitable for web mining. Once the data preprocessing is done, the invalid data will be removed. Here we present various steps involved in preprocessing of web log files.**

***Index Terms*—Web Usage Mining, Web Server Logs, Data Preprocessing, Data cleaning, User Identification, Session Identification, Path Completion, Transactions Identification.**

I. Introduction

In today's generation, human life is dependent on internet. The World Wide Web is serving as a massive broadly spread global information service center for e-commerce, technical information, advertisement, news and other information service. However WWW is a continuously changing, open and heterogeneous global distribution network. There is lot of data available but Web pages are short of a proper organizational structure. Thus information retrieval is not easy. This problem can be solved efficiently by Web mining, which is the of data mining technology in Web. Web mining can dig out valuable and interesting pattern and potential knowledge from related record. It is divided further as: Web content, structure and usage mining. Web Content Mining is the method of mining of valuable information from the contents available in Web pages. It usually consists of images, text, audio, video, or lists and tables. Web structure mining is a means used to recognize the association among the Web pages connected by information or direct link connection. This structure data can be revealed by using the web structure schema through database techniques for Web pages. This link connection helps a search engine to extract data concerning to a search query directly to the linking Web page. Web usage mining, also called web log mining, is the method of mining of interesting patterns in web access logs. Web log mining has three steps as data preprocessing, pattern discovery, pattern analysis. [1][2]

II. Web Usage Mining

Web usage mining is the automatic discovery of user access patterns from Web servers. When a user requests for some resources, the web server of that website keeps the data about user interaction in the log file that is a valuable set of information. These log files are stored in various formats such as Common Log Format (CLF) or Extended Log Format (ELF). [3] Log record has much useful information as URL, IP address, time, etc. Studying the web access logs could help us to find more potential users of the web site and trace service quality of the site. [4] The preparation of a suitable data set is a pre-requisite and significant task for the mining of Web data. Hence, Web usage mining is divided into 4 major sub-tasks: [1]

A. Data Collection
B. Data Preprocessing
C. Pattern Discovery
D. Pattern Analysis

A. Data Collection:

The data collection can be either a server level collection or client level collection or a proxy level collection. [3] The data is collected from various sources. Some data sources used in web usage mining may include web data repositories like:

1) Web Server Logs

These are logs which retain a record of page requests. The WWWC keeps a standard format for server logs, although some other formats also exist. More new entries are added to the end of the file. The information about the request as client IP address, HTTP code, page requested, request date/time, bytes served, user agent, and referrer are added. This data can be brought into a single text file, or divided into different logs, such as an access log, error log, or referrer log. But, server logs do not gather user-specific information. [3][5]

2) Proxy Server Logs

A Web proxy is a caching system which is present between Web servers and client browsers. It helps in reducing the load time of Web pages as well as the network traffic load at the server and client side. Proxy server logs include the HTTP requests from several clients to several Web servers. So a proxy server might serve as a data source to determine the usage pattern of a group of users, who share same proxy server. [3][5]

3) Browser Logs

Different Java applets and JavaScript can be used or various browsers like Internet Explorer, Mozilla, Chrome, etc. can be tailored to gather client side data. This execution of the client-side data collection requires user support, either to use the tailored browser, or to enable the features of the JavaScript and Java applets. [3][5]

B. Data Preprocessing: The most significant job of the web usage mining process is data preparation. The success of any website is highly associated to how well the data preparation job is implemented. Thus it is essential to ensure that every shade of this job is taken care of. A Web server typically enters a Web log entry for every access of a Web page. There are several types of Web logs due to different server and different setting parameters, each having the same basic information. Web logs are usually saved as text files, but due to huge amount of unnecessary information in the Web log, the original log can't be directly used in the Web log mining methods. By different preprocessing steps, the information in the Web log can be used as transaction database for mining method. The Web site's topological structure is also used in session identification and path completion. With the help of data preprocessing we can get structural, reliable and integrated data source for pattern discovery. [2][5]

C. Pattern Discovery: In this part of usage mining, statistical methods as well as data mining methods such as path analysis, Sequential patterns, Association rule, and cluster and classification rules are applied in order to discover interesting patterns. The aim of the mining process is to discover sequential association rules. This knowledge will build the knowledge base which can be used in personalization and recommendation systems. [6]

D. Pattern Analysis: The patterns identified are analyzed using knowledge query management mechanism, OLAP tools, and intelligent agent to remove the uninteresting patterns. The result of such analysis may include:

1. Most recent visit per document.
2. Who is visiting which document?
3. The frequency of visits per document.
4. The frequency of use of each hyperlink.
5. Most recent use of hyperlinks. [5]

II. PREPROCESSING TECHNIQUE

Data preprocessing consists of data cleaning, user identification, session identification, path completion and transactions identification. [7] Following diagram shows all the steps in data preprocessing:

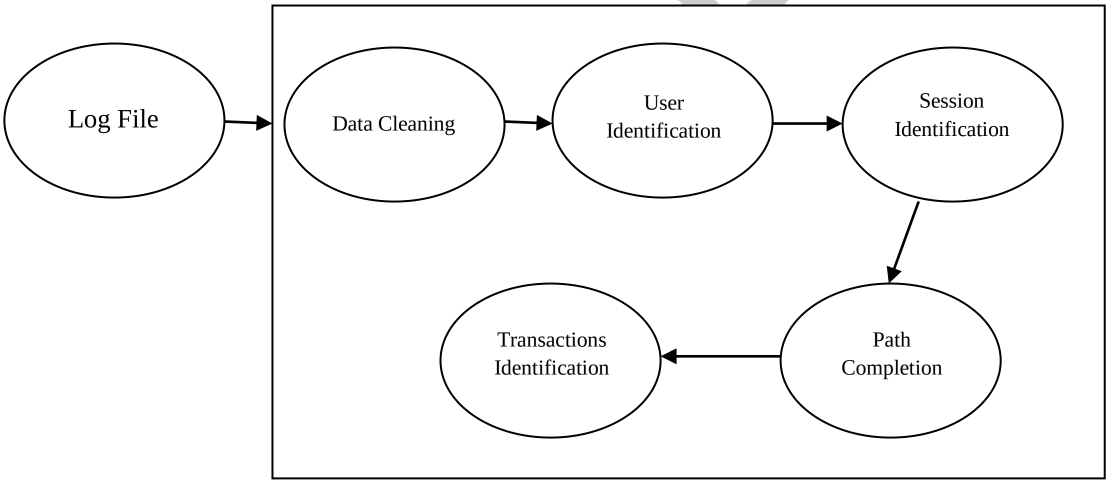

Figure 1: Steps in data preprocessing

A. Data Cleaning:

It is the method of removing the data which is irrelevant to the mining process. It reduces the size of log file. It is essential for improving the effectiveness of the mining process. It consists of removal of local and global Noise, removal of records of graphics, videos and the format information; removal of records with the failed HTTP status code, method field and robots cleaning. [1] Following diagram shows the complete process of cleaning:

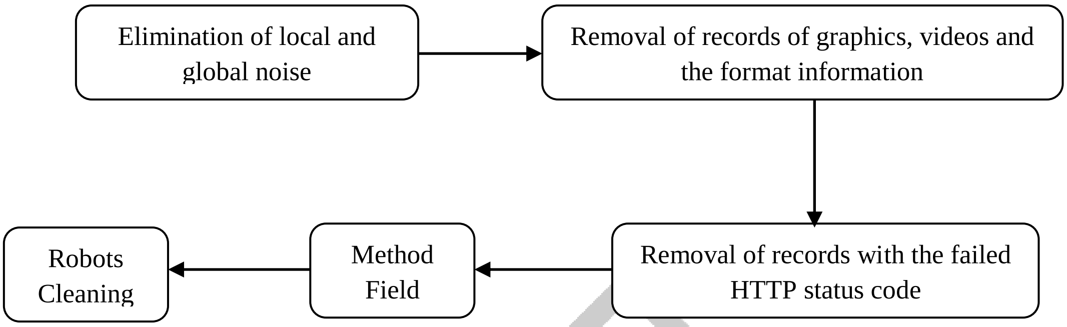
Elimination of local and global noise
Removal of records of graphics, videos and the format information
Removal of records with the failed HTTP status code
Method Field
Robots Cleaning

Figure 2: Steps in Data Cleaning

1) Elimination of Local and Global Noise:

Web noise is of two types according on its granularities as Global and Local Noise.

Global Noise: It refers to the unnecessary objects with higher granularities than individual pages. It consists of mirror sites, duplicated Web pages, previous versioned Web pages, etc.

Local Noise: It is also known as intra-page noise, it consists of unnecessary items within a Web page. It is generally illogical as compare to the major content of the page. It may be decoration pictures, navigational guides, banner ads, etc. It is mandatory to eliminate such noise for good results. It also deals with the user background knowledge, and can be identified from user's local information collections, such as emails, stored documents, browsed web pages, etc.

2) The records of graphics, videos and the format information:

From the log files, records with filename extension JPEG, GIF, CSS, etc. can be removed, which are generally available in the Uniform Resource Identifier (URI) field of the every record. Such files are not the user interested web page; rather it is just the documents embedded in the web page. Thus, it is not essential to add these files in finding the user interested web pages. With the help of data cleaning, this unnecessary analysis is removed which helps in quick discovery of user interested patterns.

3) The records with the failed HTTP status code:

In this step, the status field of each record in the web access log is evaluated and the records with status codes below 200 or above 299 are removed. It will further reduce the evaluation time in identifying the user's interested patterns.

4) Method field:

Records with value of POST or HEAD in method field are kept for getting more proper information regarding the referrer.

5) Robots cleaning:

A Web Robot is a software tool which regularly checks a website to mine its content. It is also known as spider. It automatically follows each and every one of the hyperlinks from present web page. Search engines like Google, regularly uses web robots to collect each and every page from a website for updating their search indexes. The number of requests from one web robot is most probably equals to the number of web site's URIs. If we eliminate WR-generated log entries, it will not only simplify the mining task that will follow, but will also remove the uninteresting sessions from the log file. Generally, a WR generates many requests on a web site and all these requests of a WR are out of the scope of analysis. There are two ways to identify WR's requests:

- In this method, all records having the name “robots.txt” in the requested resource name (URL) are recognized and directly removed.
- The next method 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. Thus, for each IP address, the browsing speed is calculated and every request with this value more than a threshold is considered as a request made by robots and is thus removed. The value of the threshold is set by analyzing the browser behavior occurring due to log files under study.

This helps in exact discovery of user interested patterns by providing only the appropriate web logs. Only the patterns that are a great deal interest to the user will be available in the last phase of identification if this cleaning process is completed before identifying the user interested patterns. [1]

B. User Identification:

User identification is the method of identifying each separate user accessing a particular web site. Here the main purpose is to extract access characteristic of each user, form user clustering and offer personal service for every user. All users have their unique IP address and every IP address corresponds to one user. However there are three conditions: i) some users have unique IP address; ii) some user has two or more IP addresses; iii) Due to proxy server, some user may share one IP address. Rules for user identification are:

- Different IP addresses correspond to different users.
- The same IP with different operating systems or different browsers must be considered as different users.
- While the IP address, operating system and browsers are all the same, new user can be determined whether the requesting page can be reached by accessed pages before, according to the topology of the site. [5]

C. Session Identification

Normally, Web log mining covers a long time periods, thus users might access the site more than once. Session identification is in order to divide the access records into several accessing sequences, in which the pages are requested at the same time. [9] It finds the number of times the user has accessed a web page and collects each page reference for a particular user in a log and breaks them up into user sessions. These sessions can be used as data vectors in classification, prediction, clustering and other mining functionalities. If URL in the referrer URL field in present record is not accessed before or if referrer URL field is empty then it is treated as a new user session. Reconstruction of exact user sessions from server access logs is a challenge task. Conventional session identification algorithm is based on a uniform and fixed timeout. While the interval between two sequential requests exceeds the timeout, new session is determined. According to some related researches, the value of timeout can be set as 25.5 minutes. [5][9]

D. Path Completion:

It is essential to find out the presence of significant accesses that are not available in the access log. Path completion is the addition of significant page accesses that are missing in the access log due to browser and proxy server caching. Similar to user identification, the heuristic assumes that if a page that is requested by the user is not directly associated to the previous page accessed by the same user, the referrer log can be referred to notice from which page the request came. If the page is in the user’s recent click history, it is understood that the user browsed back with the “back” button, using cached sessions of the pages. So each session reflects a complete path, including the pages that have been backtracked. [8]

E. Transactions Identification:

For some particular mining algorithm, user session still has a large scale, which needed to be divided for a smaller scale. Transaction identification means finding out important accessing sequences, which also called accessing transaction. Transaction identification is a method of grouping user session i.e. the goal of transactions identification is to create meaningful clusters of references for each user. Transaction identification is carried out by merges or divides approaches. To identify the user’s travel pattern and user’s interests, two kinds of transactions are defined. i.e., travel path transactions and content only transactions. The travel path is a blend of auxiliary and content pages accessed by a user. The content only transactions are only content pages which are used in mining to discover user’s interest and cluster users visiting the same web site. There are three methods present to identify transactions; they are identification by Reference Length, identification by Maximal Forward Reference and identification by Time Window.[7][8][9]

III. Conclusion:

Web log data is a set of huge information. Many interesting patterns are present in the web log data. But it is very difficult to mine the interesting patterns without preprocessing stage. Preprocessing stage helps to clean the records and determine the interesting user patterns and session creation. Data preprocessing is an important job of Web usage mining application. So, data must be processed before applying data mining methods to determine user access patterns from web log. The data preparation process is often the most time consuming as it includes different phases as data cleaning, user identification, session identification, path completion and transactions identification. The preprocessed data is then ready for further pattern discovery and analysis.

References

[1] P.Nithya and Dr.P.Sumathi "*Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots*", 2012 National Conference on Computing and Communication Systems (NCCCS), IEEE, 2012

[2] Fang Yuan, Li-Juan Wang, Ge Yu, "*Study On Data Preprocessing Algorithm In Web Log Mining*", Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an, 2-5 November 2003

[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan*," Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*", SIGKDD Explorations, Volume 1, Issue 2, January 2000.

[4] Liu Kewen, "*Analysis of Preprocessing Methods for Web Usage Data*", International Conference on Measurement, Information and Control (MIC), IEEE, 2012.

[5] Dr. Sanjeev Dhawan, Mamta Lathwal, "*Study of Preprocessing Methods in Web Server Logs*", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013.

[6] Paweł Weichbroth , Mieczysław Owoc and Michał Pleszkun "*Web User Navigation Patterns Discovery from WWW Server Log Files*", Proceedings of the Federated Conference on Computer Science and Information Systems, IEEE, 2012.

[7] J. Vellingiri and S. Chenthur Pandian **"***A Novel Technique for Web Log mining with Better Data Cleaning and Transaction Identification*", Journal of Computer Science, 7 (5), 2011.

[8] V.Chitraa and Dr. Antony Selvadoss Davamani, "*An Efficient Path Completion Technique for web log mining*", International Conference on Computational Intelligence and Computing Research, IEEE, 2010.

[9] He Xinhua and Wang Qiong "*Dynamic Timeout-Based A Session Identification Algorithm*", IEEE, 2011.