Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Study of Preprocessing Methods in Web Server Logs

Dr. Sanjeev Dhawan
Asst. Professor, UIET
KUK, INDIA

Mamta Lathwal
M.Tech Scholar,UIET
KUK, INDIA

Abstract: Web log mining can be described as the discovery and analysis of access patterns of users through mining of log files. 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. But the data stored in these log files do not present an accurate picture of the users' accesses to the Web site. So the preprocessing of web log data is a pre-requisite phase before it can be used for mining tasks. The preprocessed web data then is suitable for web mining. This paper presents various steps involved in preprocessing of web log files.

Keywords: Web Server Logs, Data Preprocessing, Data cleaning, User Identification, Session Identification, Path Completion

I. Introduction

World Wide Web is expanding everyday in number of websites as well as number of users. The WWW is serving as a huge widely distributed global information service center for technical information, news, advertisement, e-commerce and other information service. But WWW is an open, dynamic and heterogeneous global distribution network. There is too huge information but Web pages lack a unified structure. So information retrieval is difficult. Web mining, which is the application of data mining technologies in Web. Web mining can extract useful and interesting pattern and potential knowledge from relevant record. Web mining consists of three active research areas: Web content mining, Web structure mining and Web usage mining. Web Content Mining is the process of extraction of useful information from the contents present in Web documents. The content data is a collection of facts, a Web page is designed to convey to users. It may consist of images, text, audio, video, or lists and tables. Web structure mining is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data can be discovered by using the web structure schema through database techniques for Web pages. This link connection allows a search engine to pull data relating to a search query directly to the linking Web page. This paper mainly studies Web usage mining. Web usage mining, also called web log mining, is the process of extraction of interesting patterns in web access logs[1]. Whenever a user requests for resources, the web server of a website stores the data about user interaction in the log file that serves as a valuable pool of information. Analyzing the web access logs of different websites can help understand the user behavior and web structure thus improving the website design. Web log mining consists of three steps as data preprocessing, pattern discovery, pattern analysis. These three phases are connected to each other to form a complete web log mining methodology.

A. Data Sources

The data sources used in web log mining may include web data repositories like:

1) Web Server Logs

These are logs which maintain a history of page requests. The WWWC maintains a standard format for server logs, but some other proprietary formats also exist. More recent entries are appended to the end of the file. The information about the request, which includes client IP address, HTTP code, page requested, request date/time, bytes served, user agent, and referrer are typically added. This data can be combined into a single text file, or separated into various logs, such as an access log, error log, or referrer log. However, server logs do not collect user-specific information [2].

2) Proxy Server Logs

A Web proxy is a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of Web pages as well as the network traffic load at the server and client side. Proxy server logs contain the HTTP requests from multiple clients to multiple Web servers. So a proxy server may serve as a data source to discover the usage pattern of a group of users, who share a common proxy server.

3) Browser Logs

Various browsers like Mozilla, Internet Explorer etc. can be modified or various JavaScript and Java applets can be used to collect client side data. This implementation of the client-side data collection requires user cooperation, either to enable the

functionality of the JavaScript and Java applets, or to use the modified browser. Client-side collection scores over server-side collection because it reduces both the bot and session identification problems.

B. Web Preprocessing:

The most important task of the web log mining process is data preparation. The success of the website is highly correlated to how well the data preparation task is executed. So this is important to ensure that every nuance of this task is taken care of. A Web server usually registers a Web log entry for every access of a Web page. There are many types of Web logs due to different server and different setting parameters. But all the Web files have the same basic information. Web log is usually saved as text (.txt) file. Due to large amount of irrelevant information in the Web log, the original log can't be directly used in the Web log mining procedure. By data cleaning, user identification, session identification and path completion the information in the Web log can be used as transaction database for mining procedure. The Web site's topological structure is also used in session identification and path completion.

Purpose of data preprocessing is to offer structural, reliable and integrated data source to pattern discovery[3].

C. Pattern Discovery

In this phase of log mining statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns. The objective of mining process is to discover sequential association rules. This knowledge will form the knowledge base which can be used in recommendation and personalization systems.[4]

D. Pattern Analysis

The patterns discovered are analyzed using knowledge query management mechanism, OLAP tools, and intelligent agent to filter out the uninteresting rules/patterns. The result of such analysis might include:

1. the frequency of visits per document,
2. most recent visit per document,
3. who is visiting which documents,
4. the frequency of use of each hyperlink, and
5. most recent use of hyperlinks.

II. PREPROCESSING TECHNIQUE

Data preprocessing includes data cleaning, user identification, session identification and path completion.

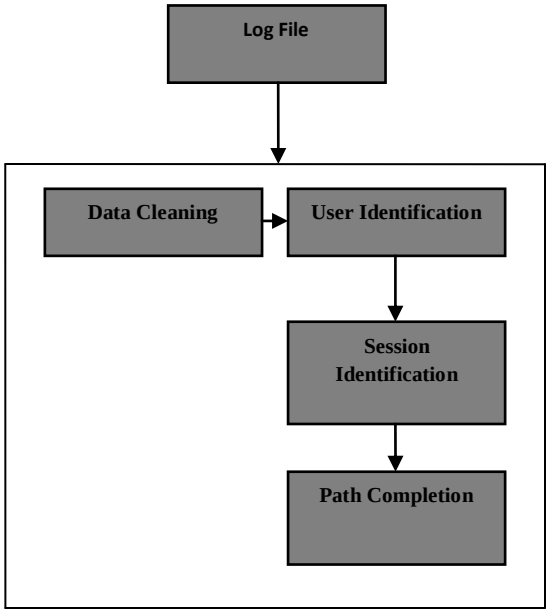

Figure 1: Web data preprocessing

A. Data Cleaning

Data cleaning refers to the process of deleting the data irrelevant to mining log algorithms in web server logs. This is necessary for improving the mining efficiency. Data cleaning includes 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, robots cleaning [5].

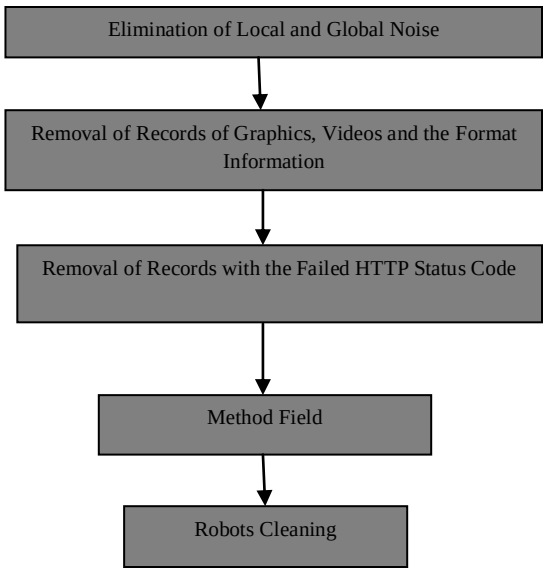

Figure 1: Steps in Data Cleaning

Thus the removal process includes:

1) Elimination of Local and Global Noise: Web noise is divided into two categories depending on its granularities:

Global Noise: Unnecessary objects with high granularities which are larger than individual pages correspond to global noise. This noise includes mirror sites, duplicated Web pages and previous versioned Web pages.

Local Noise: Local noise, also called inter-page noise, includes irrelevant items inside a Web page. This noise includes navigational guides, decoration pictures, banner ads etc. It is necessary to remove this type of noise for better results. The local noise also deals with the user background knowledge can be discovered from user's local information collections, such as a stored documents, browsed web pages, and emails.

2) The records of graphics, videos and the format information: The records with filename extension of JPEG, GIF, CSS and so on, which can be found in the URI field of the every record, can be removed from the log file. The files with these extensions 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 these files in identifying the user interested web pages[8]. The unnecessary evaluation is eliminated and this process helps in fast identification of user interested patterns.

3) The records with the failed HTTP status code: The HTTP status code is then considered in the next process for cleaning. In this step the status field of every record in the web access log is examined and the records with status codes over 299 or below 200 are removed. This cleaning process will reduce the evaluation time for determining the users interested patterns.

4) Method field: Records having value of POST or HEAD in Method field are used in present study for acquiring more accurate referrer information.

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

B. User Identification

User identification is the process of identifying each different user accessing Web site. Goal of user identification is to mine every user's access characteristic, and then make user clustering and provide personal service for the users. Each user has unique IP address and each IP address represents one user. But in fact there are three conditions **: (l)** Some users has unique IP address. **(2**) Some user has two or more IP addresses. **(3)** Due to proxy server, some user may share one IP address. Rules for user identification are:

- Different IP addresses refer to different users.
- The same IP with different operating systems or different browsers should 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.

C. Session Identification

Session identification defines the number of times the user has accessed a web page. Session identification takes all of the page references for a given 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 tasks. Traditional 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[6]. According to some related researches, the value of timeout can be set as 25.5 minutes.

D. Path Completion

It is necessary to determine the existence of important accesses that are not recorded in the access log. Path completion refers to the inclusion of important 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 linked to the previous page accessed by the same user, the referrer log can be referred to see from which page the request came. If the page is in the user's recent click history, it is assumed that the user browsed back with the "back" button, using cached sessions of the pages. So each session reflects a full path, including the pages that have been backtracked[7].

III. Conclusion

Web log data is a collection of huge information. Many interesting patterns are 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. Data preprocessing is an important task of Web log mining application. Therefore, data must be processed before applying data mining techniques to discover 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, and path completion. The preprocessed data is then available for further pattern discovery and pattern analysis.

References

[1] 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,November 2003

[2] 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- Pages 12-23, 2000.

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

[4] Paweł Weichbroth , Mieczysław Owoc and Michał Pleszkun "*Web User Navigation Patterns Discovery from WWW Server Log Files*", Proceedings Of The Fedcsis. Wrocław, 2012.

[5] 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

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

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

[8] 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): 683-689, 2011

