An Efficient Preprocessing Methodology of Log File for Web Usage Mining

A. Deepa,
Research Scholar,
Ayya Nadar Janaki Ammal College,
Sivakasi – 626 124

P. Raajan, Ph.D
Assistant Professor of MCA,
Ayya Nadar Janaki Ammal College,
Sivakasi – 626 124

ABSTRACT

Now a day, WWW has become important and huge data storage. All users' activities will be stored in log file. The log file shows the interest on the particular website. With a wide usage of internet, the log file size is growing rapidly. Web mining is the process of extracting information from web data. The raw log file won't reveal the users' accessing pattern. Thus, preprocessing has become an important process in web mining. Web Usage Mining is the important domain area of web mining to extract and analyze the usage pattern of users from the server log file. The quality of the input decides the quality of the output. Preprocessing is the noteworthy process before mining the interesting information from data. In this paper we have implemented the preprocessing techniques to convert the log file into user sessions which are suitable for mining and reduce the size of session file by filtering the least requested pages.

Keywords

WWW, Web Usage mining, log file, preprocess, filter, session.

1. INTRODUCTION

World Wide Web has become important data warehouse and growing rapidly. Data Mining is the process of extracting the information from the huge amount of data. It discovers the patterns and relationships exist among the data using various techniques like classification, clustering and association rule mining. Data mining is primarily used today by companies with a strong consumer focus like retail, financial, communication, and marketing organizations.

Web mining is the application of data mining techniques based on web data. It can be broadly classified into three domains based on the mining object: web content mining, web structure mining and web usage mining. Web content mining also called as text mining which extracts knowledge from the web documents. Web structure mining involves in link analysis to extract the hidden knowledge from web page links.

Web usage mining, also known as Web Log Mining, is the process of mining interesting Patterns in access logs [1]. It intends to uncover the users' behavior who accessing the particular web site through the server log file. The outcome of Web usage mining can be used in the applications like personalization, system improvement, site modification. The process of log mining is divided into 3 phases: Preprocessing, pattern discovery and pattern analysis. The taxonomy of web usage mining is shown in the following figure 1.

In preprocessing phase, the raw log file is transform into user sessions. First, the log file is cleaned by removing irrelevant request and error code. In session identification steps, each user page references are divided into session. After preprocessing phase, the pattern, rules and statistics are extracted from the session file by applying data mining techniques in discovery phase. Finally the extracted patterns are verified and visualized in the pattern analysis phase.

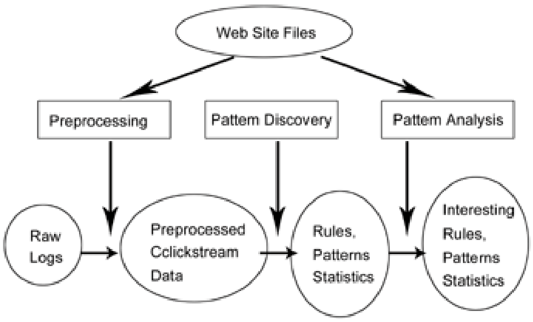
Web Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Cclickstream Data
Rules, Patterns Statistics
Interesting Rules, Patterns Statistics

Figure1. Web Usage Mining Taxonomy

In preprocessing phase, the raw log file is transform into user sessions. First, the log file is cleaned by removing irrelevant request and error code. In session identification steps, each user page references are divided into session. After preprocessing phase, the pattern, rules and statistics are extracted from the session file by applying data mining techniques in discovery phase. Finally the extracted patterns are verified and visualized in the pattern analysis phase.

Preprocessing is the significant and time consuming process. The work is deals with the conversion of raw server log file into formatted user session. This paper focuses on the following:

- Converts the raw log file into user session,
- Filters the minimum requested pages to minimize the session file size.

This paper is organized as: section 2 about the related works, section 3 discusses the steps in preprocessing, section 4 provides the experimental details and section 5 contains the conclusion.

2. RELATED WORKS

In preprocessing phase, the whole process deals with the conversion of raw Web server logs into a formatted user session file in order to perform effective pattern discovery and analysis phases. Variuos author had done research on preprocessing the log file.Raiyani et al, [1] proposed a technique called Distinct User Identification (DUI) to identify the users. The traditional approach is based on the site structure which will reduce the efficiency of process. To avoid the drawback, the author proposed a new technique. The new algorithm is based on the IP address, Agent, session time and referrer pages on selected session time. The new technique optimizes the performance of preprocessing technique.

Maideen et al, [2] presented a framework namely MS Log cleaner. It cleans the unnecessary entries from the log file. It combines numerous log file from different server and removes the needless noise data in log file based on the user requirement. It optimizes the performance and time complexity. Castellano, G et al, [3] developed a tool called Log Data Preprocessor (LODAP) to preprocess the server log file. After the data cleaning process, they find out the session and remove the unnecessary resources. Finally they produce the summary of the log file.

Sumathi, C.P et al, [4] discussed that how the log files are preprocessed. They did data fusion, data cleaning, user identification and session identification. They have used CTI log file for their research and finally they have shown the how the log files is converted into session file. Pamutha, T et al, [5] developed a method on preprocessing the log file from NASA center. They focused on the session identification process. Also they produced the statistical information like total unique IPs, total unique pages, total sessions, Session length and the frequency visited pages. Smith, K. A et al, [6] had implemented the SOM algorithm to cluster the pages to identify the usage pattern. They done preprocessing tasks and reduce the dimension by filtering the session using k-means algorithm before formatting the log file suitable for mining.

Ramya, c et al, [7] discussed the complete steps involve in preprocessing phase. After the session identification, they discover the access pattern using dynamic ART1 neural network algorithm. They cluster the pages accessed by the user. Eltahir, M et al, [8] analyzed the server log from www.interactivegt.com in IIS log format using deep log analyzer software. First the clean the data. Next the cleaned log file is transferred to the tool and produces the summary of log file. Bhawsar, S et al, [9] implemented morkov model to predict the users' future session after effectively found out the user session. Langhnoja, S.G et al, [10] identified the user and session of each user after the cleaning process. They cluster the session file after the session identification.

3. PREPROCESSING

Preprocessing aims to reduce the size of file and enhance the quality of data. It comprises a various steps data cleaning, user identification, session recognition, data filtering and data formatting. The preprocessing steps are shown in the following figure2.

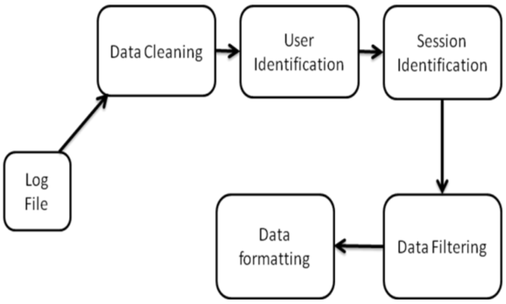

Figure2. Preprocessing steps

The server log file is the input for the usage mining. A log file is the text file that records all the actions of user while accessing the web site. It can be stored in web server, proxy server or client browser. Basically there are 3 standard formats to store log file such as: Common Log Format (CLF), Extended Common Log Format (ECLF) and IIS. All formats stores the information about the user like IP address of the client, date and time of access and the page url address. The details in IIS log file are, Client IP address, user name, date, time, service and instance, server name, server IP address, time taken, client bytes sent, server bytes sent, status code, windows status code and request type. The sample IIS log file format is shown in the following Figure 3.

```
2002-04-01 00:00:10 1cust62.tnt40.chi5.da.uu.net -
w3svc3 bach bach.cs.depaul.edu 80 get
/courses/syllabus.asp course=323-21-
603&q=3&y=2002&id=671 200 156 http/1.1
www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+
9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21)
http://www.cs.depaul.edu/courses/syllabilist.asp
```

Figure3. IIS log file format

3.1 Data Cleaning

Data cleaning is the first step of preprocessing phase. It removes the irrelevant request from the log file. It eliminates the graphic file such as (gif, jpg, jpeg, mov, png, mp3, and swf), failure status code, unwanted method and spider navigation.

Status code is a value that indicates the success or failure of user request. 100 series indicates that the request is in processing. 200 series indicates that the request is successfully processed. 400 series indicates that the client error and 500 errors indicate the server error. Remove the method except GET and Post.

Web robots (also known as Web crawlers or Web spiders) are programs that automatically search and download complete Web sites by following every hyperlink on every page within the site in order to update the index of search engine. Requests created by Web robots are not considered usage data and, consequently, have to be removed. All records containing the name "robots.txt" in the requested resource name (URL) are identified and straightly removed [3]. The algorithm is as follows:

```
ALGORITHM: Data Cleaning
Input: Server Log File.
Output: Cleaned log database
Step1: Read Log Record from Web Server Log File.
Step2: If status code is 200 and method is GET then
Step3:  Retrieve the page extension
Step4: if extension is not audio, video, css format, js
Format then
Step5: Retrieve the url page
Step6: If the url page is not robot.txt then
                    Save the record in cleaned table.
Step5: Go to step 1.
Step6: Stop, If EOF.
```

3.2 User Identification

The goal of user identification is to find out whom accessing the website and what type of pages. It creates logical clusters of pages for every user. The unique users can be identified by using IP address and user agent.

If the IP address is different, requests are from different users. If it is same, check the user agent. If agent is different, the request is from different user. Otherwise both request from same user.

3.3 Session Identification

After the user identification, the next step is session identification. The session identification is the process of splitting the user into the group of pages according to the time interval. That is for every user find out the session. There are 2 heuristics method to find out the session. Such as: time oriented and structure oriented.

In this paper, we have applied time oriented heuristics. When the duration is exceeds the threshold value then starts the new session. The default threshold value is 30 minutes.

3.4 Data Filtering

The aim of the data filtering is to select the most accessed web page and to remove the least visited web page. After the session identification phase, compute the number of session NS_i for each page p_i. Remove the pages with the low frequency. This phase reduce the number of pages and the number of session in the session file. The algorithm is as follows:

```
ALGORITHM: Data Filtering
Input: Session file
Output: Filtered data
Step1: Read session file.
Step2: for all page pi do
Step3:  compute NSi
Step5: if (NSi < threshold)
Remove the page from database
Step5: end if.
Step6: end for
Step7: Stop, if EOF
```

3.5 Data Formatting

After the data filtering process, the next step is data formatting. Data formatting is the process of converting the session file into the suitable format for mining.

4. Result and Discussion

In our research, we have used CTI web log dataset. It is available in http://www/cs.depaul.edu. It is in IIS log file format. It consists of 2 weeks log data. We have used java for our experiment.

From our experiment we have observed that the cleaning algorithm optimize the log file size well. It reduces the size of file as 72 %. It reduces the size of file (KB) from 4344 to 1208 and number of records from 9893 to 7711. The result of data cleaning is shown in figure 4 and the statistical report is shown in table 1.

Table1. Statistical report of data cleaning

File size before cleaning	4344 KB
File size After cleaning	1208 KB
Number of records before cleaning	9893
Number of records after cleaning	7711
Number of pages	170
Number of users	984
Number of sessions	1003

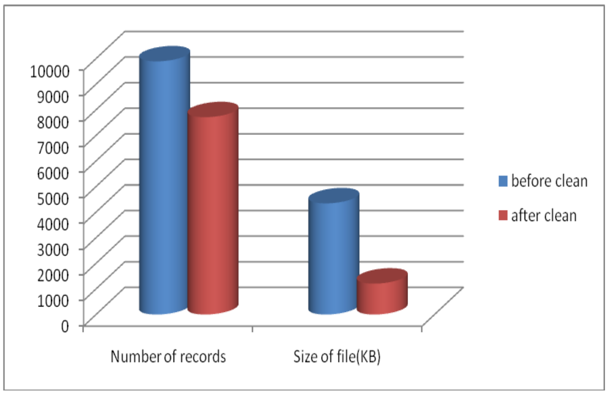

Figure 4 Result of Data Cleaning

After the cleaning process, the user is identified through IP address and user agent. After that session is identified for each user using time oriented method. Table 2 shows the session file.

Table2. Session File

sessionId	pageId	pageView
11	21	syllabisearch.asp
11	3	courses.asp
110	16	schedule.asp
110	20	syllabilist.asp
110	22	syllabus.asp
110	3	courses.asp
110	3	courses.asp
111	5	default.asp
112	11	login.asp
112	14	newsearch.asp
112	19	studentprofile.asp

Next to the session identification process, the data filtering step will be carried out. It eliminates the least requested resources and reduces the size of session file. The number of pages reduced from 170 to 25 and number of session from 1003 to 945. The result of data filtering is shown in figure 5.

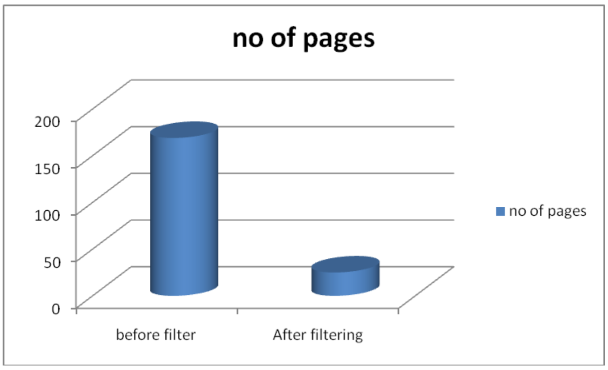

Figure5. Result of data filtering

The access frequency of each page indicates the importance of the page x axis shows the pages and y axis shows the frequency. The page access frequency is shown in the following figure6.

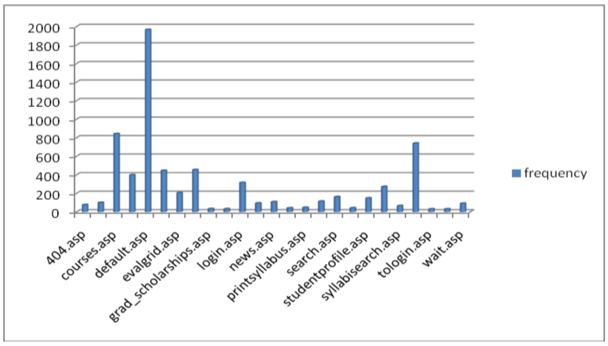

Figure6. Page Access frequency

The number of pages visited by user can be used to find out the user interest on the web site. The frequency of pages in each session is shown in figure 7.

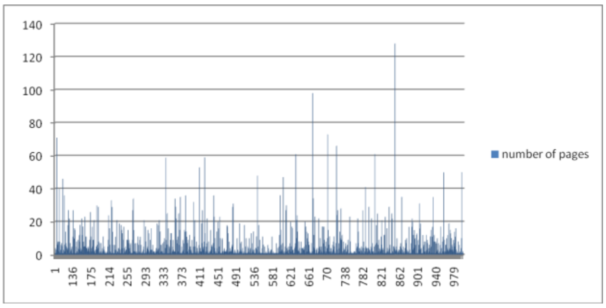

Figure7. frquency of pages in session

5. CONCLUSION

Data Preprocessing is one of the important tasks before applying mining algorithms. It converts the raw log file into user session. In this work, we have briefly introduced log file preprocessing and implemented it on CTI log file. Also we produce the summary of user session file. We have used filtering technique to remove least requested resources.

The preprocessed data will be used to discover the access pattern efficiently. In future the preprocessed log data will be applied to various data mining techniques to discover the usage pattern. The access pattern of user will resolve the user behavior.

6. REFERENCES

[1] Raiyani, ashwin g., and sheetal s. Pandya. "Discovering User Identification Mining Technique for Preprocessed Web Log Data."

[2] Maideen, C. M., & Palanivel, M. K. MS Log Cleaner: A framework to discover efficient use of web service.

[3] Castellano, G., Fanelli, A. M., and Torsello, M. A. 2007. Log data preparation for mining web usage patterns. In IADIS International Conference Applied Computing (pp. 371-378).

[4] Sumathi, C. P., et al., 2011. An Overview of Preprocessing Of Web Log Files For Web Usage Mining. Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645.

[5] Pamutha, T., Chimphlee, S., Kimpan, C., and Sanguansat, P. 2012. Data Preprocessing on Web Server Log Files for Mining Users Access Patterns.International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol, 2.

[6] Smith, K. A., & Ng, A. (2003). Web page clustering using a self-organizing map of user navigation patterns. Decision Support Systems, 35(2), 245-256.

[7] Ramya, C., and Kavitha, G. 2011. An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network. In Computer Networks and Intelligent Computing (pp. 198-204). Springer Berlin Heidelberg.

[8] Eltahir, M., & Dafa-Alla, A. F. (2013, August). Extracting knowledge from web server logs using web usage mining. In Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on (pp. 413-417). IEEE.

[9] Bhawsar, S., Pathak, K., Mariya, S., & Parihar, S. Extraction of Business Rules from Web logs to Improve Web Usage Mining.

[10] Langhnoja, S. G., Barot, M. P., & Mehta, D. B. (2013). Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery.International Journal.

[11] https://www.microsoft.com/technet/prodtechnol/WindowsServer2003/Library/IIS/be22e074-72f8-46da-bb7e-e27877c85bca.mspx?mfr=true

[12] Common Log Format – Wikipedia: http://en.wikipedia.org/wiki/Common_Log_Format

[13] A Log File Types Supported by Clickstream Intelligence, Oracle9iAS Clickstream Intelligence Administrator's Guide Release 2 (9.0.2) Part Number A90500-02 http://docs.oracle.com/cd/A97329_03/bi.902/a90500/admin-05.htm

[14] Markov, Z., & Larose, D. T. (2007). Data mining the Web: uncovering patterns in Web content, structure, and usage. John Wiley & Sons.