AN EFFICIENT APPROACH TO PERFORM PRE-PROCESSING

S. Prince Mary
Research Scholar,
Sathyabama University,
Chennai- 119
princemary26@gmail.com

E. Baburaj
Department of Computer Science & Engineering,
Sun Engineering College, Nagercoil,
Kanyakumari Dist- 629902
alanchybabu@gmail.com

Abstract

Nowadays, WWW (World Wide Web) becomes more popular and user friendly for transferring information. Therefore people are more interested in analyzing log files which can offer more useful insight into web site usage. Web usage mining is one of the data mining fields, which deals with the discovery and extract useful information from web logs. There are three phases in web usage mining, pre-processing, Pattern Discovery and Pattern Analysis. This paper describes the importance of pre-processing methods and steps involved in retrieving the required information effectively.

Keywords: Pre-processing; Web usage mining; Web Log, Session

1. Introduction

Data Mining is the study of information retrieval techniques to discover and model mystical patterns in huge volume of crude data. Web data mining is referred as the application of Data mining techniques to web data. Web data mining is divided into three areas:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Web content mining is the process to extract useful data from the content of a web page. Example: image, audio, video, metadata and hyperlinks. Web structure mining is the study of web page schema of a collection of hyperlinks. Web usage mining is the technique to extract knowledge from web usage data or web server logs. Web usage mining process divided into three phases:

- Data Collection and Pre-processing
- Pattern discovery
- Pattern analysis.

This paper gives the detailed description about the data collection and pre-processing steps in web usage mining.

2. Data Collection

The data source of web usage mining includes web data stores like:

- Web Server Logs
- Proxy Server Logs
- Browser Logs

2.1 Web Server Logs

History of web page requests is maintained as a log file. Web servers are the costly and the most common data source. They collect large volume of information in their log files. These logs contain name, IP, date, and time of the request, the request line exactly came from the client, etc. These data can be bound together as a single text file, or divided into different logs, like access log, referrer log, or error log. However, user specific information is not stored in server logs.

A sample Apache server log file: Common log file format

```
125.125.125.125 - uche [20/Jul/2008:12:30:45 +0700] "GET /index.html HTTP/1.1" 200 2345
```

Sample combined log file format

```
125.125.125.125 - uche [20/Jul/2008:12:30:45 +0700] "GET /index.html HTTP/1.1" 200
2345
"http://www.ibm.com/" "Mozilla/5.0 (X11; U; Linux x86_64; en-US; rv:1.9a8)
Gecko/2007100619
GranParadiso/3.0a8" "USERID=Zepheira;IMPID=01234"
```

2.2. Proxy Server Logs

It acts as an intervening level of catching lies between client browser and web servers. Proxy caching is used to decrease the loading time of a web page as well as the reduce network traffic at the server and client side. The actual HTTP request from multiple clients to multiple web servers are tracked by the proxy server. The proxy server log is used as a data source for browsing behavior characterization of a group of unauthorized users sharing a common proxy server [1].

2.3. Browser Logs

On client side using JavaScript or Java applets the browsing history is collected. To implement client side data collection, user cooperation is needed.

Here pre-processing discussed using Web Server Logs. Web server logs are used in the web page recommendation to improve the E-Commerce usability.

In this paper first phase that is data collection and pre-processing is discussed.

3. Pre-Processing

Any real time data mining project usually spends 80% of the time on the data pre-processing step. The ground work of web data mining is done at preprocessing phase. Data preprocessing use log data as input then process the log data and produce the reliable data. The goal of data preprocessing is to choose cardinal features then remove irrelevant information and finally transform raw data into sessions. To achieve its goal Data pre-processing is divided into Data Cleaning, user identification, and Session Identification and Path Completion steps. In Fig. 1 steps involved in data preprocessing are shown.

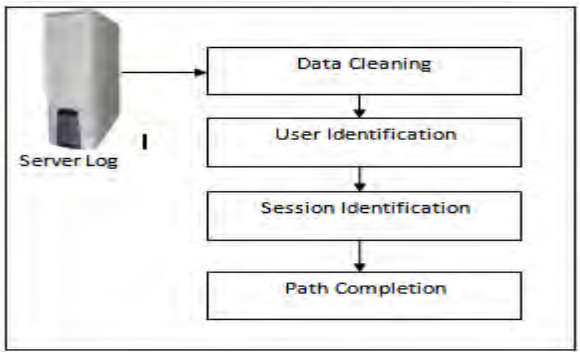

Fig.1 Data pre-processing Steps

3.1. Data cleaning

Data cleaning is to remove all the useless data used in data analysis and mining. Data cleaning is necessary for increasing the mining efficiency. In Fig. 2 the steps involved in data cleaning are given. It includes removal of local and global noise, elimination of graphic records, videos and the format efficiency, elimination of HTTP status code records, robots cleaning [2]. In Fig. 2 steps in data cleaning are shown below.

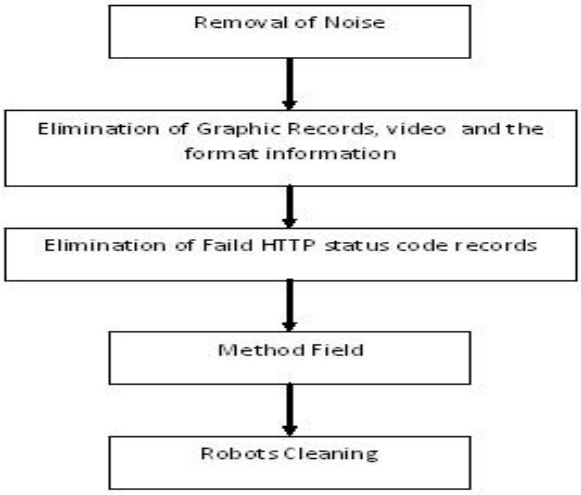

Fig. 2 Steps in Data Cleaning

3.1.1. Elimination of local and global Noise:

Local Noise: This is also called as inter- page noise, which includes irrelevant data in the web page. Local noise includes Decoration pictures, navigational guides, banner etc. It is better to remove local noise for efficient results.

Global Noise: Irrelevant objects with high granularities which are larger than the web page are belongs to global noise. This noise includes duplicated web pages, mirror web sites and previous version web pages.

3.1.2. The Records of- graphics, video and the format information:

JPEG, GIF, CSS filename extension is found in the every record on URI field, this can be eliminated from the log file. The files with these extensions are the documents embedded in the web page. So it is not necessary to include these files in identifying the user interested web pages [3]. This process support to identify user interested patterns.

3.1.3. Failed HTTP- status code:

This cleaning process will reduce the evaluation time for finding the user's interested patterns. In this process the status field of every record in the web access log is checked and the status code over 299 or below 200 are removed.

3.1.4. Method- field:

Records which contain methods like POST or HEAD are used to get more absolute referrer information.

3.1.5. Robots- Cleaning:

It is also called as spider or not, it is a software tool that scans a website periodically to extract the content. All the hyperlinks from a web page are automatically followed by WR. The uninterested session from the log file is removed automatically when WR is removed.

3.2. User identification

Each different user accessing the website is identified in the user identification process. The aim of this process is to retrieve every user's access characteristics, then make user clustering and provide recommendation service for the users. There are three conditions to identify the user:

- Some user has a unique IP address
- Some user has two or more IP addresses
- Some user may share one IP address due to the proxy server.

User identification Rules:

- Different IP addresses refer to different users.
- Same IP with different browsers or different operating systems considered as different users.
- A new user can be identified whether the requesting page can be reached by accessed pages before, according to the topology of the site, even the IP address, browser, operating system is all same.

4. Session Identification

It defines the number of times the user has accessed a web page. It takes all the page reference of a given user in a log and divides them into user sessions. These sessions can be used as an input data vector in classification, clustering, prediction and other tasks. Based on a uniform fixed timeout a traditional session identification algorithm is used. A new session is identified when the interval between two sequential requests exceeds the timeout [4].

5. Path Completion

Determining the missing important web page access due to the proxy server and the browser is essential for mining information. This is accomplished by path identification process. If the requested page is not linked to the previous accessed page by the unique user, then from which page request came is identified using the referrer log file. If the page is available in the user's history, then it is assumed that the user pressed back button. Hence each and every session reflects the complete path, including the web pages that have been backtracked. At the end of path completion the user session file gives the paths consisting of a group of page references including repeated page accesses made by a user.

A set of users in a user session file is created by data pre-processing Phase. Collection of referring pages made by a user during a one time visit to a website is known as user session. In some cases, user sessions are further divided into meaningful groups of web page references referred as a transaction. These transaction again can be identified using any one of the approaches such as maximal forward reference, reference length and time leading to the creation of a user transaction file [5].

6. Preprocessing Steps With Sample Web Server Log File

6.1. Algorithm for Data Cleaning

Read Record in Database.

For each Record in Database

Read fields (URI – stem) //URI- stem indicates

The target URL//

If fields = {*.gif,*.jpg,*.css} then

Remove Records

Else

Save Records

End if

Next record

The data like *.gif,*. Jpg, *.css are not needed for the purpose of pattern discovery are removed using the above algorithm.

In Table 1. The sample log file is given after cleaning it.

IP	TIME	URL	REFF	AGENT
192.167.100.101	0.04	A	-	IE5;Win2k
192.167.100.101	0.12	B	A	IE5;Win2k
192.167.100.102	0.13	A	-	IE6;XP
192.167.100.102	0.16	B	A	IE6;XP
192.167.100.102	0.21	C	B	IE6;XP
192.167.100.102	0.24	D	C	IE6;XP
192.167.100.101	0.29	C	B	IE5;Win2k
192.167.100.101	0.34	D	C	IE5;Win2k
192.167.100.102	0.36	D	C	IE6;XP

6.2. User Identification:

Web usage data analysis need knowledge about user and how they identified among different users. Server log records multiple sessions when a user visit more than once. In [6] used a phrase "user activity record" to refer to the sequence of logging activities of the same user. Here IP and Agent are used to identify the user.

Table.2. User 1

IP	TIME	URL	REFF	AGENT
192.167.100.102	0.13	A	-	IE6;XP
192.167.100.102	0.16	B	A	IE6;XP
192.167.100.102	0.21	C	B	IE6;XP
192.167.100.102	0.24	D	C	IE6;XP
192.167.100.102	0.36	D	C	IE6;XP

Table.3. User 2

IP	TIME	URL	REFF	AGENT
192.167.100.101	0.04	A	-	IE5;Win2k
192.167.100.101	0.12	B	A	IE5;Win2k
192.167.100.101	0.29	C	B	IE5;Win2k
192.167.100.101	0.34	D	C	IE5;Win2k

In Table.2,3 shows 192.167.100.101 , 192.167.100.102 IP visits more than one time, hence to identify a user here IP address and User Agent are important parameters. USER1 AND USER 2 is identified using IP and User Agent.

7. Session Identification

A sequence of pages viewed by a user during one visit is known as the Session. The session is recorded in the log file. In pre-processing it is necessary to find session of each user. There are two ways to capture session: Time Oriented and Structure Oriented [7].In this paper Time Oriented Session is used.

Conditions used in Time Oriented Session Identification:

- The difference between first and last request is less than or equal to 30 minutes
- The difference between first and next request is less than or equal to 10

Table5: Log File

IP	TIME	URL	REFERER URL	AGENT
192.167.100.103	0.15	A	-	IE5;Win2k
192.167.100.103	0.12	B	A	IE5;Win2k
192.167.100.103	0.16	C	B	IE5;Win2k
192.167.100.103	0.18	D	C	IE5;Win2k
192.167.100.103	0.21	D	C	IE5;Win2k
192.167.100.103	0.37	E	D	IE5;Win2k
192.167.100.103	0.45	F	C	IE5;Win2k
192.167.100.103	0.48	G	F	IE5;Win2k

Based on the above condition in the Table: 5 generates two sessions as follows:
Session 1

IP	TIME	URL	REFERER URL	AGENT
192.167.100.103	0.15	A	-	IE5;Win2k
192.167.100.103	0.12	B	A	IE5;Win2k
192.167.100.103	0.16	C	B	IE5;Win2k
192.167.100.103	0.18	D	C	IE5;Win2k
192.167.100.103	0.21	D	C	IE5;Win2k

Session 2

IP	TIME	URL	REFERER URL	AGENT
192.167.100.103	0.37	E	D	IE5;Win2k
192.167.100.103	0.45	F	C	IE5;Win2k
192.167.100.103	0.48	G	F	IE5;Win2k

8. Path Completion:

It depends on the URL and Referrer URL fields in the server log file. Using a graph model, in graph model node represents the web page and the edges shows the links between web pages [8]. In path completion the missing reference is not stored in the server log. It is cached on the client side.

Example:

URL	REFERER URL
A	-
B	A
A	-
B	A
C	B
C	C
D	C

For given example URL and Referrer URL Graph Model shown below Fig.5.

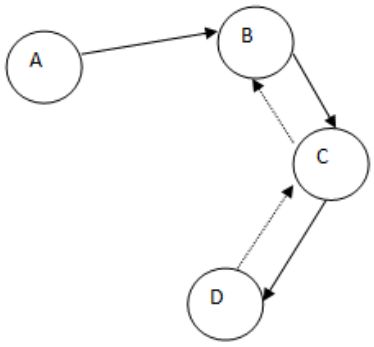

Fig.5 Web Site Visiting Structure

The Fig.5 represents the web site visiting structure the log file. The doted edge shows back button click. This is known as missing pages which is stored on the client side.

9. Conclusion

Web log mining is one of the recent areas of research in Data mining. To use the web usage mining efficiently, it is important to use the pre-processing steps. Steps of pre-processing are analyzed and tested successfully with sample web server log files. This paper delivers the steps of pre-processing including data cleaning, user identification, session identification and path completion. Once pre-processing is performed on web server log, then patterns are discovered using data mining techniques such as statistical analysis, association, clustering and pattern matching on pre-processed data, then the discovered patterns are analyzed for various applications such as web personalization, site improvement, site modification, business intelligence, etc.

References

[1] Vijayashri Losarwar, Dr. Madhuri Joshi, “Data Preprocessing in web usage Mining”, International Conference on Artificial Intelligence and Embedded Systems (ICAIES’2012) July-2012, Singapore.

[2] 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.

[3] 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
[4] He Xinhua and Wang Qiong "Dynamic Timeout-Based A Session Identification Algorithm" , IEEE 2011
[5] Mohd Helmy, Abd Wahab, Nik Shahidah. Development of Web usage Mining Tools to Analyze the Web Server Logs using Artificial Intelligence Techniques. The 2nd National Intelligence Systems and Information Technology Symposium (ISITS 207), Oct 30-31, 2007, ITMA-UPM, Malaysia.
[6] R.Cooley, Bamshad Mobasherand Jaideep Srivastava, "DataPreparation for Mining World Wide Web Browsing Patterns." Knowledge and Information Systems,1(1),1999,5-32.
[7] R.Cooley, B. Mobasher and J. Srivatsava, "Web mining: Information and pattern discovery on the World Wide Web." 9th IEEE Inernational Conference on Tools with Artificial Intelligence. CA, 1997, 558-567
[8] J.Srivatsava, R.Cooley, M.Deshpande and P.N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." ACM SIGKDD Explorat. Newsletter,2000,12-23.
[9] Pang-Ning Tan and Vipin Kumar. Modeling of Web Robot Navigational Patterns. In WEBKDD 2000 – web Marketing for E-Commerce-challenges and Opportunities, Second International Workshop August 2000.