An Overview of Pre-processing Techniques in Web usage Mining

J. Soonu Aravindan[#1], Dr. K. Vivekanandan[*2]

[#]*Ph.D. Research Scholar, School of Computer Science, Bharathiar University, Coimbatore, India*
[*]*Professor, BSMED, Bharathiar University, Coimbatore, India*

Abstract *— WWW (World Wide Web) is a huge repository of web pages and links. Enormous amount of web data is being generated every day. User accessed websites are recorded as a web log file, which may contain noisy & ambiguous data which may affect the results of the mining process. So there is a necessity to pre-process the web data before extracting knowledge from web log files. Web Usage Mining is the area of data mining dealing with discovery and analysis of usage patterns from web data in order to improve web based applications. This paper mainly focuses on the Major steps followed in Data Pre-Processing Stage in Web usage mining.*

Keywords - *Web Usage Mining, Data Pre-processing, Web logs.*

I. INTRODUCTION

World Wide Web is expanding every day in number of websites and users. It's a huge repository of web pages and links which are widely distributed for news, advertisements, and other information services. Since the information organized in web pages are unstructured, information retrieval becomes a difficult process. Web mining is the application of data mining techniques to find interesting patterns and potentially useful knowledge from web data. For mining purpose web log data, hyperlink structure of the web is being used. In general web mining is broadly classified into three categories [1] i.e. Web content mining, Web structure mining, and Web usage mining. Web content mining deals with discovering useful knowledge from the web page content. It mainly focuses on the webpage content rather than links and structure. Web structure mining deals with discovering the link structure of the web. It helps in finding the similarities between the web sites and discovering web communities. Web usage mining deals with understanding of user behaviour with his interaction towards web sites. This paper mainly focuses on topics covered under web usage mining and various techniques followed in data pre-processing stage in web usage mining. The remaining sections of the paper is organized as follows. Section 2 deals with web log, Section 3 deals with data pre-processing techniques, and Section 4 gives the conclusion.

II. WEB LOG

The primary data sources used in Web Usage Mining are the server log files, which include Web server access logs and application server logs. Web server log file is a simple plain text file which record information about each user when submitting request to a web server [2]. The log file contains information about the time and date of request, HTTP method used, the user agent (browser and operating system type and version), the referring web resource, client side cookies.

It shows that the IP address 192.168.111.145 is accessing the papers.html file from the server mit.cs.dept.edu. Next, the agent field captures the browser type and version along with operating system information of the client machine. The referrer field indicates that the user came to this page from the source page

'webminingresource.blogspot.com'

Web log file is located in three different locations (i) Web server logs (ii) Web proxy server (iii) Client Browser

Server logs consist of four types (i) Access log file – records all requests that are processed by the server (ii) Error log file – records information whenever an error occurs on the page requested by client to server (iii) Agent log file – records information about a user's browser (iv) Referrer log file – records information about links and redirects visitor to site.

Log file is stored as three different formats (i) W3C Extended log file format (ii) NCSA common log file format (iii) IIS log file format.

III. DATA PRE-PROCESSING

The primary data sources used in Web Usage Mining are the server log files, which include Web server access logs and application server logs. Web server log file is a simple plain text file which record information about each user when submitting request to a web server [2]. The log file contains information about the time and date of request, HTTP method used, the user agent (browser and operating system type and version), the referring web resource, client side cookies.

A. Data Fusion

Data Fusion refers to merging of log files from several Web and application servers [3]. A user surfs

many things from the web, it is not necessary that all the information he/she needed may have located in a single web server. So there is a need for global synchronization across these servers. It can be done with the help of 'referrer' field in server logs along with sessionization and user identification methods. Once when the log files from different resource are merged together, user sessions can be easily identified by IP address and User Agent. Once when the user sessions are identified all the non-requested user logs has to be removed which is done in the cleaning phase. If suppose, when a user browse a particular information from more than one location, log files will be created in every location. After merging, all the repeated log files of the user has to be removed.

B. Data Cleaning

Data cleaning is the process of removing irrelevant or useless references from the web log files for performing better analysis [4]. It includes elimination of noise, removal of records of graphics, videos, failed HHTP status code. When a user browses a webpage, it may contain pre-loaded advertisements in several formats (audio, video, images, animations), which makes the mining process quite complex. Since we are interested only in the user requested data and not any system generated data, we need to make sure that only user requested data is present in the server logs. Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod [8] has divided the process of data cleaning into three different stages. In first stage access of JPEG, GIF files, Java Scripts are removed from the web logs since they are executed not on user's request. In second stage all the error codes are discarded from the log files. In third stage the entries occurred from the crawlers or spiders are eliminated. Theint Theint Aye [9] has presented an algorithm for data cleaning which not only clean noisy data but also reduce incomplete, inconsistent and irrelevant request according to status code and IP link.

C. Pageview Identification

Pageview is nothing but a collection of web resources linked together in a particular page representing a user event. When a user clicks a web site, all the resources and content regarding to that website are grouped together and displayed in a single web page. Identification of pageviews is heavily dependent on the intra-page structure of the site, as well as on the page contents and the underlying site domain knowledge [7]. Each pageview has a pageview id, static pageview type and other content attributes (e.g., keywords).

D. User Identification

Once when the web log data has been cleaned, the next process is to identify different users from that. It is not a simple process, since one user may use several IP address, or several users may use the same IP address using proxy servers or even a single user may use several browsers which may result in session differences or even user may visit a site more than once. Server logs record multiple sessions for each user referred as user activity record. Users are identified using IP address and User Agent in log files. Sudheer Reddy K, Kantha Reddy M, Sitaramulu V [10] has used some heuristics to identify the users such as, each IP address represents one user, If the IP address is same for more logs, but the agent log displays a change in browser or operating system, the IP address represents a different user, If there is a same IP address, browser and operating system, the referrer information can be considered. Liu Kewen [12] has given some notations for user identification. Users i={User ID, User IP ,User Ur1,User Time, User RefferPage, User_Agent},O<i<n where i is the number of total users;User ID is user's ID have been identified.User IP is the user's IP address. User UrI is the Web page accessed. User Time is the time at which user accessed. User RefferPage is the last page that the user requested. Use Agent is the agent user used.

TABLE I

User Identification by IP and User Agent

Time	IP	URL	Ref	Agent
0.01	192.168.1.145	A	-	IE9;WinXP
0.07	192.168.1.145	B	A	IE9;WinXP
0.18	192.168.1.142	C	-	IE7;Win7
0.28	192.168.1.142	D	C	IE7;Win7
1.01	192.168.1.145	C	A	IE9;Winxp
1.45	192.168.1.150	F	-	IE6;WinXP;SP1
1.50	192.168.1.150	G	F	IE6;WinXP;SP1

In table [1] there are three users identified with the help of IP address and user Agent.

TABLE II

User Separation by IP and User Agent

User 1	0.01	192.168.1.145	A	-	IE9;WinXP
	0.07	192.168.1.145	B	A	IE9;WinXP
	1.01	192.168.1.145	C	A	IE9;WinXP
User 2	0.18	192.168.1.142	C	-	IE7;Win7
	0.28	192.168.1.142	D	C	IE7;Win7
User 3	1.45	192.168.1.150	F	-	IE6;WinXP;SP1
	1.50	192.168.1.150	G	F	IE6;WinXP;SP1

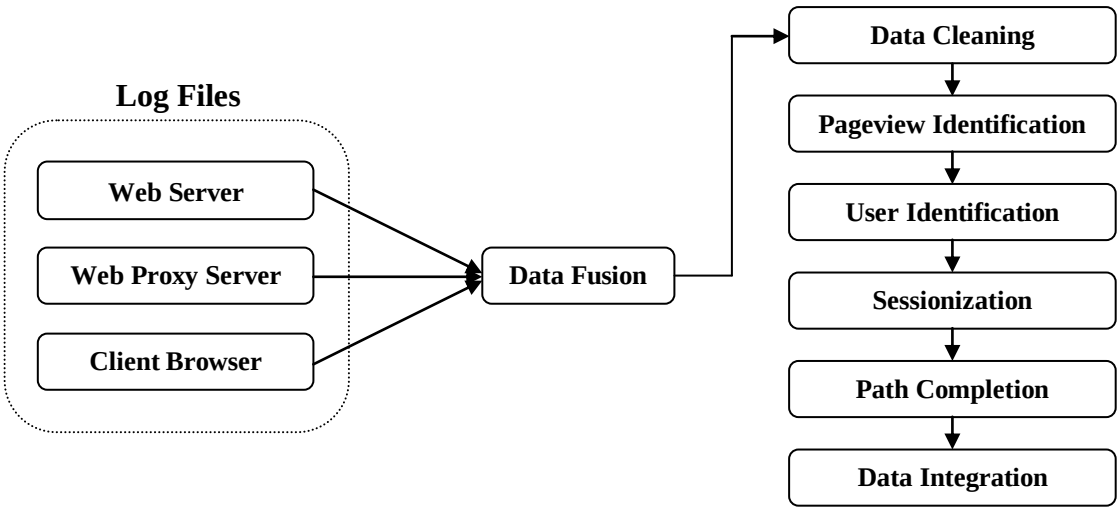

Fig 1: Web Log Data Pre-processing

E. Sessionization

A session is a sequence of page views by a single user during a single visit [5]. Sessionization is the process of segmenting the user activity record of each user into sessions. When a user surfs the internet for a long time, there is a more chance that he might have visited several sites. When the time difference between the visits exceeds certain limit (let's say 30 min of timeout) then we can conclude that the user has started a new session. With the generated user log files, it can be splitted into several user sessions. From table [3] we can split the session of User 1 into two different sessions. Liu Kewen [12] and Nirali [13] has given some notations for identifying user's sessions.

Sessions i={User_ID,Sj,[Urlj1,Urlj2,Urljk])

,O<i<n where n is the total number of sessions. User_ID stands for user's ID that has been identified, Sj stands for one of the user's sessions, Urljk stands for aggregate of Web pages in session.

TABLE III

Sessionization by Time

User 1	**Session 1**	0.01	192.168.1.145	A	-	IE9; WinXP
		0.07	192.168.1.145	B	A	IE9; WinXP
	Session 2	1.01	192.168.1.145	C	A	IE9; Winxp

F. Path Completion

It is an important pre-processing task which is performed after sessonization [3]. When a user browse the websites in this order, let's say

$$\mathbf{A \rightarrow B \rightarrow D \rightarrow E \rightarrow D \rightarrow B \rightarrow C}$$

and when he return back to the website B.html using 'back' button, this particular event will not be stored in the log file since cached version of B.html and D.html is loaded without sending request to the server. This results in the second reference to B.html and D.html not being recorded on the server logs. Missing references due to catching can be inferred through path completion.

The following path and Fig 2 shows the User's actual navigation path

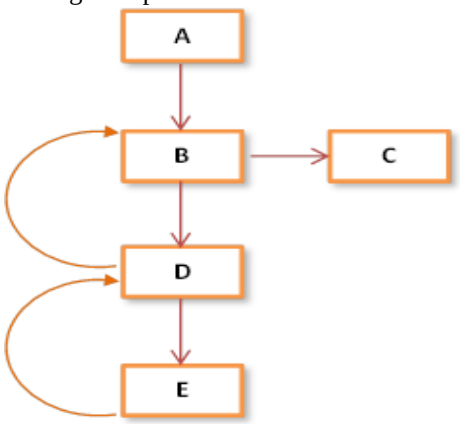

Fig 2: User's Actual Navigation Path

The following Table 4 shows how the User's actual navigation path is stored in Server Log.

TABLE IV

Storation in Server Log

URL	Referrer
A	-
B	A
D	B
E	D
C	B

G. Data Integration

This is the final stage in Data pre-processing. To provide the most effective framework for pattern discovery, data from various sources must be integrated with the preprocessed clickstream data [3]. This mainly helps in e-commerce where the user data and the product attributes are integrated for the discovery of important business intelligence metrics. This widely helps the people in e-marketing to

implement new business strategies for their product sales.

IV. CONCLUSIONS

Data Pre-processing is the major and initial process in Web Usage Mining. Before applying any data mining algorithm for further data analysis and pattern discovery, the raw data has to be converted into a structured data format. This structured data is called processed data. This paper gives an overview of various data pre-processing techniques followed in web usage mining.

REFERENCES

[1]. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, 2000 Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD.

[2]. Priyanka Patil, Ujwala Patil, 2012 preprocessing of web server log file for web mining. WJST, ISSN: 2231-2587

[3]. Web Data Mining – Bing Liu

[4]. Thanakorn Pamutha, Siriporn Chimhlee, Chom Kimpan, and Parinya Sanguansat, 2012 Data Preprocessing on

[5]. Web Server Log Files for Mining Users Access Patterns. IJRRWC Vol.2, No. 2, June 2012, ISSN: 2046-6447

[6]. V. Chitraa, Dr. Antony Selvadoss Davamani, 2010 A Survey on Preprocessing Methods for Web Usage Data. IJCSIS Vol. 7, No. 3,2010

[7]. Marathe Dagadu Mitharam, 2012 Preprocessing in Web Usage Mining. IJSER, Volume 3, Issue 2, February 2012, ISSN 2229-5518

[8]. C.P. Sumathi, R. Padmaja Valli, T. Santhanam, 2011 An Overview of Preprocessing of Web Log Files for Web Usage Mining, JATIT & LSS, Vol. 34 No.2 ISSN: 1992-8645, E-ISSN: 1817-3195

[9]. Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod, 2013 Web Usage Mining: A Review on Process, Methods and Techniques, ICICES, ISBN:978-1-4673-5786-9

[10]. Theint Theint Aye, 2011 Web Log Cleaning for Mining of Web Usage Patterns, ICCRD, ISBN:978-1-61284-839-6

[11]. Sudheer Reddy K, Kantha Reddy M, Sitaramulu V, 2013 An effective data preprocessing method for Web Usage Mining, ICICES,ISBN:978-1-4673-5786-9

[12]. Liu Kewen, 2012 Analysis of preprocessing methods for web usage data , MIC, Volume 1, ISBN: 978-1-4577-1601-0

[13]. Nirali H.Panchal, Ompriya Kale "A Survey on Web Usage Mining". International Journal of Computer Trends and Technology (IJCTT) V17 (4):177-181, Nov 2014. ISSN:2231-2803