ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Web Log Files in Web Usage Mining Research – A Review

[1] Dr. S.Vijayarani, [2]E. Suganya, [3]M. Prakathambal
[1]Assistant Professor, [2]Ph.D Research Scholar, [3]P.G. Student
[1] [2] [3]Department of Computer Science, Bharathiar University

***Abstract* - The World Wide Web has a vast amount of information resources and services. Every website is comprised of a number of web pages. Whenever, a user access the websites, the server saved this information in web log files which is a plain text (.txt) file. Web log files contain unnecessary and noisy data. It can be preprocessed using web mining techniques. Data preprocessing is the process of selecting standardized data from the original log files. Data cleaning, user identification, session identification and path completion are different stages of data preprocessing. Log files contain the information about the users like user name, visiting path, the path traversed, time stamp, page last visited, success rate, user agent and URL. The log files are stored in different locations like web server, web proxy server and the client browser. This paper has provided a detailed review of web log files; i.e. concepts of web server data, application server data, application level data, web server logs, log file parameter, types of log file format, various locations of web log files and the different types of web log files. In addition to this, we also surveyed the existing research works and given the information about how web log files are used in web usage mining research.**

Index Terms— Web Usage Mining, Web Server Data, Web Log File, Log File Formats.

I. INTRODUCTION

website is an essential platform for the web users to obtain required information such as education, entertainment, news, health, e-commerce, business, etc. Today the Internet is the most emerging technology in the world [2]. The World Wide Web is also known as 'Information Superhighway'. It is a system of interlinked hypertext documents which is called web documents which is accessed through the Internet. The Internet is a global system of interconnection of computer networks. The WWW is one of the services which run on the Internet [1]. In short, World Wide Web is also known as Web considered as an application 'running' over the Internet. It is a large and dynamic domain of knowledge and discovery. It has become the most popular services among other services that the Internet provides. The numbers of users as well as the number of websites have been increasing dramatically in the recent years. A huge amount of data is constantly being accessed and shared among several types of users, both humans and intelligent machines. Extracting the required information from the web is a difficult task, but it is done by web mining techniques.

Web mining is used to extract meaningful information from web. It can be classified into three kinds: content mining, structure mining and usage mining [8][9]. Web content mining is used to extract the content i.e. text, image, audio, video, etc. and structure mining are used to extract the information based on the link [18]. Web usage mining is used to extract user behavior on the web [4]. It is an emerging research area in web mining concerned about the online user's behavior [3]. Web log files are located in three different locations they are web server logs, web proxy servers and client browsers. It provides more complete and accurate usage of data to the web server, but the log file does not record cache files [3]. Web proxy server takes the HTTP request from the user and sends to the web server which passed the result and finally return to the same server. Web mining is a technique used to analyze the online web contents, navigate between various websites and perform transaction of data across the Web.

A. Web Usage Mining

Web Usage Mining deals with the extraction of useful knowledge from web data which includes the information about the user [5]. Web usage data are stored in three different locations such as web server, proxy server and the client browser. The web usage data include registration data, online user profiles, user sessions/transactions, cookies, queries, bookmark data,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

mouse clicks and scrolls and etc. [4]. The scope of web usage mining is local, which means that the scope of web usage mining spans an individual website [16]. It discovers and predicts the behavior of the user, in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service.

Web usage mining helps to discover interesting usage patterns from web usage data [12]. Web logs are preprocessed using web usage mining techniques such as data cleaning, user identification, session identification and path completion which are represented in Figure I.

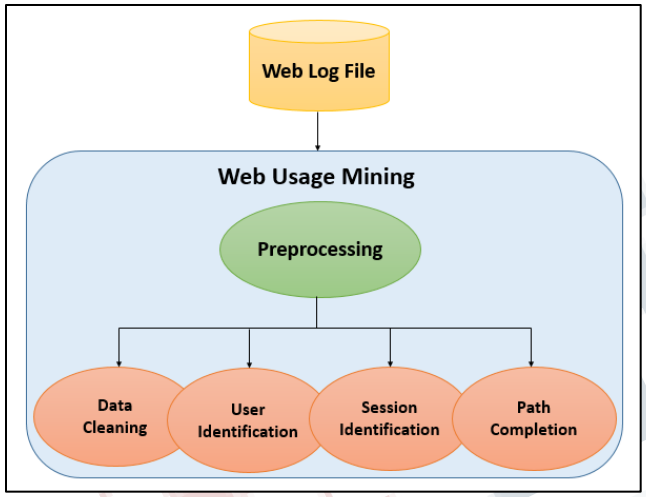


FIGURE I. PREPROCESSING IN WEB LOG FILES

II. WEB LOG FILES

Web log file is a file that is automatically created and maintained by a web server. Each time a visitor request any file such as page, image, video, etc. from that website information on their request is appended to a current log file [6]. It contains the information about the user like time span, URL, IP address, etc.

The host is an IP address of the system. A user Id is the unique name which is used to identify who visit a particular web page [7]. It is displayed when the user would like to make any transactions on the website and URL is a website address. The sample web log files are given in Table I.

TABLE I

SAMPLE WEB LOG FILE

Host	User Id	URL
117.197.6.155	1	/images/pic010.jpg
131.253.41.47	2	images/chemlab_d.jpg
95.108.158.238	3	/images/pic8.jpg
117.201.98.145	4	/images/Result_Scan.jpg

A. Log File Parameters

Log files contain various parameters which are very useful in recognizing user browsing patterns. Table II shows the list of parameters [10].

TABLE II

LIST OF PARAMETERS

S. No	Log File Parameters	Description
1.	User Name	It identifies the user who has visited the website and this identification normally is the IP address
2.	Visiting Path	It refers the visiting time of the user when they visit and which website.
3.	Path Traversed	It includes the information about the user path within the website
4.	Time Stamp	It is also known as session which is the time spent by a user on each page.
5.	Page Last Visited	This parameter has the information about the page last visited by the user while leaving the particular website.
6.	Success Rate	This parameter has measured by user activity like downloads, copying the information from the website
7.	User Agent	It is the browser that the user uses to send the request to the server

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

8.	URL	It is the resource that is accessed by the user and it may be of any format like HTML, CGI etc.
9.	Request Type	It is the method that is used by the user to send the request to the server and it can be either GET or POST method.

B. Types of Log File Format

There are mainly three types of log file formats that are used by a majority of the servers.

- Common Log File Format
- Combined Log Format
- Multiple Access Logs

a. Common Log File Format

It is the standardized text file format that is used by most of the web servers to generate the log files [10]. The configuration of the common log file format is

"%h %l %u %t \"%r\" %>s %b" common CustomLog logs/access_log common

For example:

127.0.0.1 RFC 1413 frank [10/Oct/2000:13:55:36 -0700] "GET/apache_pb.gif HTTP/1.0" 200 2326

b. Combined Log Format

It is same as the common log file format but with three additional fields i.e., referral field, the user_agent field, and the cookie field [14]. The configuration of combined log format is given below in the box.

LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{Useragent}i\"" combined CustomLog log/access_log combined

For example:

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET/apache_pb.gif HTTP/1.0" 200 2326

"http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I;Nav)"

c. Multiple Access Logs

It is the combination of the common log format and combined log file format but in this format multiple directories can be created for access logs [13]. Configuration of multiple access logs is given below in the box.

LogFormat "%h %l %u %t \"%r\" %>s %b" common

CustomLog logs/access_log common

CustomLog logs/referer_log "%{Referer}i -> %U"

CustomLog logs/agent_log "%{User-agent}i"

III. LOCATION OF WEB LOG FILES

If a user visits many times on the website then it creates entry many times on the web server [14]. A log file is located in three different places which is shown in figure II.

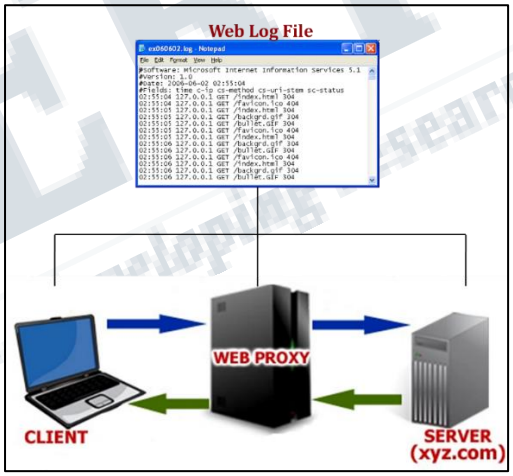


FIGURE II. LOCATION OF WEB LOG FILES

A. Web Servers

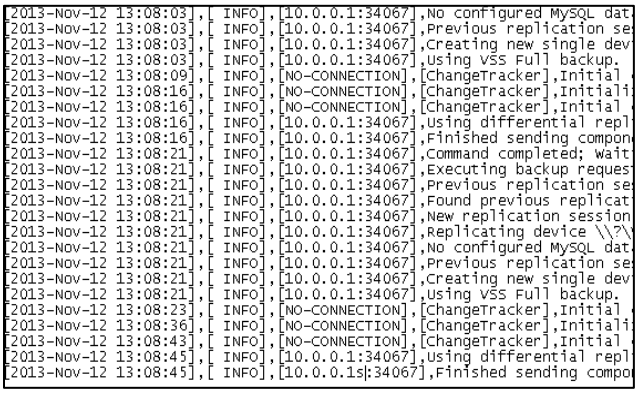
The web logs which usually supply the most complete and accurate usage data. These log files reside in web server and activity of the user browsing website [15]. There are four types of web server logs, they are, access logs, agent logs, error logs and referrer logs.

A. Access Log

It is one of the major web log server, it will record each click, hits and access of the users [17]. For capturing information about the user which have the number of attributes such as Client IP, Client name, Date and Time, Server site name and Server IP.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

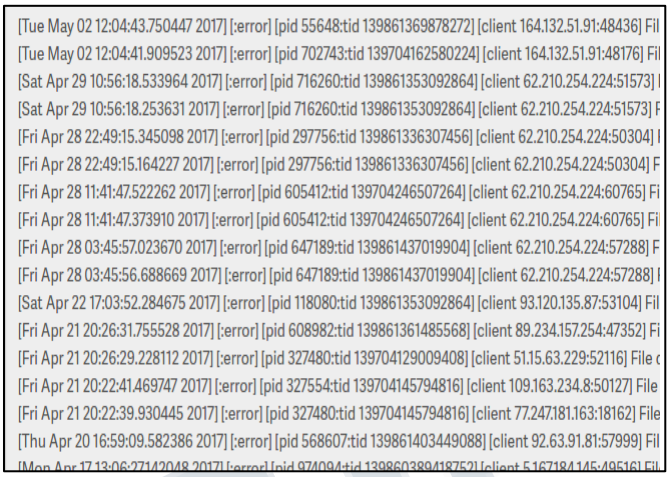


B. Agent Log

The file which is used to record the information about user's browser, browser version and operating system. Different versions of different user's browsing history are very helpful for designer and web site changes are made accordingly [13].

c. Error Log

The error log file is used to record the error found on web sites, especially when the user clicks on a particular link and the browser does not display the particular page or web site and the user receives error 404 not found [17].



d. Referrer Log

It is used to record the referrer log that a user came from the particular website by using the user's page link. Google has implemented the page-rank algorithm for assigning the weight to referrer sites.

B. Web Proxy Servers

A proxy server takes the HTTP requests from users and passes them to a web server then returns to users the results passed to them by the web server [11]. These log files contain information about the proxy server from which user request came to the web server.

C. *Client Browsers*

Participants remotely test a web site by downloading special software that records Web usage or by modifying

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

the source code of an existing browser. HTTP cookies could also be used for this purpose [17]. These log files reside in client's browser.

IV. RESEARCH ASPECT

Web log files plays an important role in web usage mining research because it contains the user information on the web. Web usage mining is used to extract the user behavior on the web. Web log files should be preprocessed using data preprocessing techniques because it contains relevant and also irrelevant information like error messages. Research aspect of web log files in web usage mining is to find who all are visits a particular website/web page. For example, in amazon website (https://www.amazon.in/) has more number of web pages which includes product details and how many users visits the particular product. In table 2 shows the related works.

TABLE III

RELATED WORKS

Author	Title	Algorithms/ Techniques used	Inference
Arjun Ram Meghwal and Dr. Arvind K Sharma (2016) [21]	Identifying System Errors through Web Server Log Files in Web Log Mining	WebLog Expert Lite tool	Authors proposed a methodology to identify the system errors by using web server log files has been investigated. WebLog Expert tool is used in the complete log mining process. The findings of this work would be helpful and useful for the System Administrators, Web Masters, Web Analysts, Website Maintainers, Website Designers and Web Developers to manage their systems by identifying occurred errors, corrupted and broken links. This work will also improve the loyalty and reliability of the web sites.
Ripal Patel, Mr. Krunal Panchal, and Mr. Dushyantsinh Rathod (2015) [22]	A Survey on Web Mining From Web Server Log	Web mining algorithm, E-web log miner	The author overviewed various approaches of Web Mining techniques. They conclude that by using Pre-processing we can process the unstructured data. We can use different log files and combine them in one file and then they use the mining on the integrated file. It will decrease the time and increase the efficiency.
R. Lokeshkumar, R. Sindhuja and Dr. P. Sengottuvelan (2014) [20]	A Survey on Preprocessing of Web Log File in Web Usage Mining to Improve the Quality of Data	Preprocessing techniques	The authors surveyed different preprocessing techniques to identify the issues in web log file and to improve the accuracy.
Sana Siddiqui and Imran Qadri (2014) [19]	Mining Web Log Files for Web Analytics and Usage Patterns to Improve Web Organization	Log File Integration	The authors reviewed the discovering method which helpful for patterns from the online server log file of an educational institute. The results are employed in totally different applications like net traffic analysis, economical web site administration, website modifications, system improvement and personalization and business intelligence etc.
Nanhay Singh, Achin Jain and Ram Shringar	Comparison Analysis of Web Usage Mining Using Pattern	Web Log Explorer (WLE) tool, filtering	The authors studied about the web usage mining. Web log data collected from NASA web server to find out

Raw (2013) [23]	Recognition Techniques	technique	useful browsing pattern. NASA website visitors is image file with extension ".gif" and on Thursday at hour 12. From the comparison between JPG and GIF image files it was clear that if the web administrator uses GIF files for the image media than bandwidth of the server will be saved. It is useful for web administrator in order to improve web site performance through the improvement contents, structure, presentation and delivery.

V. CONCLUSION

Web is an interface which is used to access remote data, commercial and non-commercial services. Web log file is a file that is automatically created and maintained by a web server. Log files contain the information about the users like user name, visiting path, the path traversed, time stamp, page last visited, success rate, user agent and URL. The log files are stored in different locations, different types of log format and several log files. This paper discussed a detailed review of web log files like web server data, application server data, application level data, web server logs, log file parameter types of log file format, various locations of web log files and types of web log files.

REFERENCES

[1] Roop Ranjan, Sameena Naaz and Neeraj Kaushik, "Web Miner: A Tool for Discovery of Usage Patterns From Web Data", International Journal on Computer Science and Engineering (IJCSE), Vol. 5 No. 05 May 2013, pp. 286-293, ISSN: 0975-3397.

[2] The W3C Technology Stack; "World Wide Web Consortium", Retrieved April 21, 2012.

[3] Arvind K Sharma, P. C. Gupta,"Enhancing the Performance of the Website through Web Log Analysis and Improvement", International Journal of Computer Science and Technology (IJCST) Vol. 3, Issue 4, Oct-Dec 2012.

[4] Huiping Peng, "Discovery of Interesting Association Rules Based on Web Usage Mining", International Conference 2010.

[5] Cooley, R.,"Web Usage Mining: Discovery and Application of Interesting Patterns from Web data", 2000.

[6] Liu, H., et al., "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting user's future requests", Data and Knowledge Engineering, 2007, Vol. 61, Issue 2, pp. 304-330.

[6] M. Spiliopoulou and L. C. Faulstich. Wum, "A web utilization miner". In Proc. of EDBT WorkshopWebDB98, Valencia, Spain, March 1998.

[7] M. Malarvizhi, S. A. Sahaaya Arul Mary, "Preprocessing of Educational Institution Web Log Data for Finding Frequent Patterns using Weighted Association Rule Mining Technique", European Journal of Scientific Research ISSN 1450-216X Vol.74 No.4 (2012), pp. 617-633.

[8] Sanjay Madria, Sourav s Bhowmick, w. -k ng, e. P. Lim, "Research Issues in Web Data Mining".

[9] A. Jebaraj Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques", Journal Of Theoretical And Applied Information Technology, 2005 - 2010 JATIT

[10] Tsuyoshi, M and Saito, K., "Extracting User's Interest for Web Log Data", IEEE 2006, pp. 343-346, ISBN: 0-7695-2747-7

[11] Cooley, R, Mobasher, B., Srivastava, J., "Web Mining: Information and pattern discovery on the World Wide Web", IEEE 1997, pp. 558-569, ISSN: 1082-3409

[12] Naresh Barsagade, "Web Usage Mining and Pattern Discovery: A Survey", December 8, 2003

[13] Eltahir. M.A. , Dafa-Alla, " mining", IEEE August, 2013, pp.413-417, ISBN: 978-1-4673-6231-3

[14] Yadav, M.P. , Keserwani, P.K, "An efficient web mining algorithm for Web Log analysis: E-Web Miner", IEEE March, 2012, pp. 607-613, ISBN: 978-1-4577-0694-3

[15] Xianjun Ni, "Design and Implementation of WEB Log Mining System", IEEE Jan, 2009, pp. 425-427, ISBN: 978-1-4244-3334-6

[16] Ling Zheng, Hui Gui, "Optimized data preprocessing technology for web log mining", IEEE June, 2010, pp. V1-19 – V1-22, ISBN: 978-1-4244-7164-5

[17] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan, Mohamad Mohsin, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology 48 2008, pp.190-197, DOI: 10.1.1.140.5102

[18] Tamanna Bhatia, "Link Analysis Algorithm for Web Mining", IJCST Vol. 2, Issue 2, June 2011, ISSN: 2319-5940

[19] Sana Siddiqui and Imran Qadri, "Mining Web Log Files for Web Analytics and Usage Patterns to Improve Web Organization", International Journal of Advanced Research in Computer Science and Software Engineering, 4(6), June - 2014, pp. 794-802

[20] R. Lokeshkumar, R. Sindhuja and Dr. P. Sengottuvelan, "A Survey on Preprocessing of Web Log File in Web Usage Mining to Improve the Quality of Data", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 8, August 2014

[21] Arjun Ram Meghwal and Dr. Arvind K Sharma, "Identifying System Errors through Web Server Log Files in Web Log Mining", International Journal of Computer Science And Technology, Vol. 7, Issue 1, Jan - March 2016

[22] Ripal Patel, Mr. Krunal Panchal, and Mr. Dushyantsinh Rathod, "A Survey on Web Mining From Web Server Log", Journal of Emerging Technologies and Innovative Research (JETIR), Volume 2, Issue 10, October 2015.

[23] Nanhay Singh, Achin Jain and Ram Shringar Raw, "Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques", International Journal of Data Mining & Knowledge Management Process, Vol.3, No.4, July 2013