P: ISSN NO.: 2394-0344
E: ISSN NO.: 2455-0817
Remarking An Analisation
Title Code : UPBIL04788
Vol-I * Issue- VI* July - 2016

# Web Log File for Effective Web User Management : A Review

**Gudakesh Sharma**
Research Scholar,
Deptt.of Computer Science and Information Technology,
Sam Higginbottom Institute of Agriculture, Technology and Sciences, Allahabad

**Jeberson**
Head of Department,
Deptt.of Computer Science and Information Technology,
Sam Higginbottom Institute of Agriculture, Technology and Sciences, Allahabad

**Abstract**

Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. By analysing these log files gives a neat idea about the user. This paper gives a detailed discussion about these log files, their formats, their creation, access procedures, their uses.

**Keywords**: Log Files, Web Server

**Introduction**

Web log file is a log file which is automatically created and maintained by a web server for various purpose. Whenever a client or user of the web site "hit" to the Web site, the web log file records all the information related to the client.The format of the raw web log file is only one line of text for each hit to the web site. This web log file contains all the critical information about the visitors of the site, from which server the request came from and exactly the activity they performed are recorded.

**Aim of the Study**

The aim of the study is to analyse the characteristics of web log files for efficient web user management.

**Review of Literature**

Grace et al. (2011) has published a paper which gives a detailed discussion about these log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in the log files which in turn gives way to an effective mining. It also provides the idea of creating an extended log file and learning the user behaviour.

Suneetha et al. (2009) has concerned with the in-depth analysis of Web Log Data of NASA website to find information about a web site, top errors, potential visitors of the site etc. which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web using mining. The obtained results of the study will be used in the further development of the web site in order to increase its effectiveness.

Revathi et al.( 2015) proposed an effective and enhanced data preprocessing methodology which produces an efficient usage patterns and reduces the size of weblog down to 75-80% of its initial size.

Sheshasaayee et al. (2016) has proposed an improved K-means clustering algorithm for identifying internet user behaviour. Web data analysis includes the transformation and interpretation of web log data find out the information, patterns and knowledge discovery. The efficiency of the algorithm is analyzed by considering certain parameters. The parameters are date, time, S_id, CS_method, C_IP, User_agent and time taken. This dataset provides a better analysis of Log data to identify internet user behaviour.

Elhiber et al.(2013) in their paper they discusses the process of Web Usage Mining with few steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has also presented several approaches such as statistical analysis; clustering, association rules and sequential pattern are being used to discover patterns in web usage mining.

**Web Server**

At the most basic level, a web server is simply a computer program that dispenses web pages as they are requested. It is a computer system that processes requests via HTTP, the basic network protocol used to distribute information on the World Wide Web and the server delivers the data back to the browser that had requested the web page. The term web server can refer to the entire system, or specifically to the software that accepts and supervises the HTTP requests[1] Web server stores the files

necessary to display the Web pages on the user or client computer. The web pages may be in the form of ASP, PHP, JSP etc. server side script files. Whenever the client requests the server for some task the web server software executes the clients request and generate the response in HTML format and transmit to the respective client.

**Log File**

A server log is a log file (or several files) automatically created and maintained by a server consisting of a list of activities it performed. A typical example is a web server log which maintains a history of page requests. The W3C maintains a standard format (the Common Log Format) for web server log files, but other proprietary formats exist.The activity entries are updated at the end of the log file. The content of each updated entry consists of the information about the request, including client IP address, request date/time, page requested,HTTP code, bytes served, user agent, and referrer are typically added. All this data may be stored in a single file or in various other separate files such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information.[2]

These log files contains confidential information which cannot be given access to the public or internet user but can only be given access to the web administrator or any other administrator who is granted the privilege to access the server. This server log files are used to analyse the user traffic patterns daily, weekly or even monthly.Using statistical techniques and web mining techniques the web administrator can generate various types of prediction for the future needs. The marketing department manager can predict the customer behaviour of online shopping so that they can tune or optimise the business strategy for better results. Marketing
departments of any organization that owns a website should be trained to understand these powerful tools.[2]

Website statistics are based on server logs. A server log is a simple text file which records activity on the server. There are several types of server log — website owners are especially interested in access logs which record hits and related information. [6] Access logs come in several different formats but they all tend to look something like this:

213.60.233.243 - - [25/May/2004:00:17:09 +1200] "GET /internet/index.html HTTP/1.1" 200 6792 "http://www.mediacollege.com/video/streaming/http.html" "Mozilla/5.0 (X11; U; Linux i686; es-ES; rv:1.6) Gecko/20040413 Debian/1.6-5"151.44.15.252 - - [25/May/2004:00:17:20 +1200] "GET /cgi-bin/forum/commentary.pl/noframes/read/209 HTTP/1.1"2006863"http://search.virgilio.it/search/cgi/search.cgi?qs=download+video+illegal+Berg&lr=&dom=s&offset=0&hits=10&switch=0&f=us" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Hotbar 4.4.7.0)"
151.44.15.252 - -[25/May/2004:00:17:21 +1200] "GET /js/common.jsHTTP/1.1"2002263"http://www.mediacollege.com/cgibin/forum/commentary.pl/noframes/read/209" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Hotbar 4.4.7.0)"151.44.15.252 - - [25/May/2004:00:17:21 +1200] "GET /css/common.cssHTTP/1.1"2006123"http://www.mediacollege.com/cgibin/forum/commentary.pl/noframes/read/209""Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Hotbar 4.4.7.0)" 151.44.15.252 - -[25/May/2004:00:17:21 +1200] "GET /images/navigation/home1.gif HTTP/1.1" 200 2735 "http://www.mediacollege.com/cgi-bin/forum/commentary.pl/noframes/read/209" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Hotbar 4.4.7.0)" 151.44.15.252 - -[25/May/2004:00:17:21 +1200] "GET /data/zookeeper/ico-100.gif HTTP/1.1" 200 196 "http://www.mediacollege.com/cgi-bin/forum/commentary.pl/noframes/read/209" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Hotbar4.4.7.0)"151.44.15.252 - - [25/May/2004:00:17:22 +1200] "GET /adsense-alternate.htmlHTTP/1.1"200887"http://www.mediacollege.com/cgibin/forum/commentary.pl/noframes/read/209" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1;Hotbar4.4.7.0)"151.44.15.25225/May/2004:00:17:39 +1200] "GET /data/zookeeper/status.html HTTP/1.1"2004195"http://www.mediacollege.com/cgibin/forum/commentary.pl/noframes/read/209""Mozilla/4.0(compatible; MSIE 6.0; Windows NT 5.1; Hotbar 4.4.7.0)"

Each line in the log file represents one request (hit). If a visitor requests an HTML page which contains two images, three lines will be added to the log (one for the page and two for the images).

Each line may include some or all of the following information:

1. The IP address of the computer making the request (i.e. the visitor)
2. The identity of the computer making the request
3. The login ID of the visitor
4. The date and time of the hit
5. The request method
6. The location and name of the requested file
7. The HTTP status code (e.g. file sent successfully, file not found, etc)
8. The size of the requested file
9. The web page which referred the hit (e.g. a web page containing a hyperlink which the visitor clicked to get here)

**Contents of a Log File**

The Log files in different web servers maintain different types of information.[3] The basic information present in the log file are

**User Name**

This is used to identify the person who had visited the web site. IP address would be identifier that is assigned by the Internet Service provider (ISP). This may be a temporary address that has been assigned. Therefore here the unique identification of the user is lagging. In some web sites the user identification is made by getting the user profile and allows them to access the web site by using a user name and password. In this kind of access the user is being identified uniquely so that the revisit of the user can also be identified.[4]

**Visiting Path**

This will speak about the user behaviour how the user has navigated the web pages in the web server. This may be by using the URL directly or by clicking on a link or through a search engine.

**Path Traversed**

This identifies the path taken by the user with in the web site using the various links.

**Time Stamp**

The time spent by the user in each web page while surfing through the web site. This is identified as the session.

**Page Last Visited**

The page that was visited by the user before he or she leaves the website.

**Success rate**

The success rate of the web site can be determined by the number of downloads made and the number copying activity under gone by the user. If any purchase of things or software made, this would also add up the success rate.

**User Agent**

This is nothing but the browser from where the user sends the request to the web server. It's just a string describing the type and version of browser software being used.

**URL**

The resource accessed by the user. It may be an HTML page, a CGI program, or a script.

**Request Type**

The method used for information transfer is noted. The methods like GET, POST.

These are the basic contents of in the log file. This log file details are used in case of web usage mining process. According to web usage mining it mines the highly utilized web site. The utilisation would be the frequently visited web site or the web site being utilized for longer time duration. Therefore the quantitative usage of the web site can be analysed if the log file is analysed [4].

**Location of a Log File**

A Web log is a file to which the Web server writes information each time a user requests a web site from that particular server. [4-5] A log file can be located in three different places:

1. Web Servers
2. Web proxy Servers
3. Client browsers

**Web Server Log Files**

The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser. The contents of the file will be the same as it is discussed in the previous topic. In the server which collects the personal information of the user must have a secured transfer.[4]

Raw log files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site information on his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text .[8]

Here is a sample of log entry in Apache Combined format:

**213.135.131.79 - - [15/May/2002:19:21:49 -0400] "GET /features.htm HTTP/1.1" 200 9955 "http://www.weblogexpert.com/download.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Q312461)"**

**Web Proxy Server Log files**

A Proxy server is said to be an intermediate server that exist between the client and the Web server. Therefore if the Web server gets a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. These web proxy servers maintain a separate log file for gathering the information of the user .[4-7]

**Client Browsers Log files**

This kind of log files can be made to reside in the client's browser window itself. Special types of software exist which can be downloaded by the user to their browser window. Even though the log file is present in the client's browser window the entries to the log file is done only by the Web server.

This kind of log files can be made to reside in the client's browser window itself. Special types of software exist which can be downloaded by the user to their browser window. Even though the log file is present in the client's browser window the entries to the log file is done only by the Web server.

**Weblog File Formats**

WebLog Expert supports IIS (W3C Extended) and Apache log formats. It also supports log files of Amazon S3 and Amazon ELB, as well as log files of servers that use formats similar to Apache (e.g. Nginx) and IIS (e.g. Amazon CloudFront, Microsoft ISA/TMG, and Windows Azure). The program can also read GZ and ZIP compressed logs[8]. WebLog Expert supports log files of the most popular web servers:

**Apache and IIS**.

It also supports log files of Amazon S3 and Amazon ELB, as well as log files of servers that use formats similar to Apache (e.g. Nginx) and IIS (e.g. Amazon CloudFront, Microsoft ISA/TMG, and Windows Azure).

**Apache Log Format**

The program automatically recognizes Combined and Common log formats of the Apache web server. By default Apache uses the Common log format but the majority of hosting providers set the Combined log format for Apache on their servers [8].

Here is a sample of log entry in Combined format:

**213.135.131.79 - - [15/May/2002:19:21:49 -0400] "GET /features.htm HTTP/1.1" 200 9955 "https://www.weblogexpert.com/download.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Q312461)"**

Log format can be configured by editing the "httpd.conf" file in the Apache conf directory (if you have access to this file). The configuration settings for the log file should look like the following:

# The following directive defines the "combined" nickname

**Log Format "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\"" combined**

# The location and format of the access log file CustomLog logs/access.log combined

If your Apache log files contain information about multiple virtual domains (sites) hosted on the

same server, you can also use the following configuration settings:

# The following directive defines the "combinedvhost" nickname Log Format **"%h %l %u %t \"%r\" %>s %b \"%{Referer}i\"\"%{User-Agent}i\" %v" combinedvhost**

# The location and format of the access log file CustomLog logs/access.log combinedvhost

In this case log entries will have Combined format with one extra field (virtual domain). WebLog Expert needs this information if you use the "virtual domain" filter or need to create the "Virtual Domains" report.

**Web Log Analysis Software**

Web Log analysis software is also called a web log analyzer is a kind of web analytics software that parses a server log file from a web server, and based on the values contained in the log file, derives indicators about when, how, and by whom a web server is visited. Reports are usually generated immediately, but data extracted from the log files can alternatively be stored in a database, allowing various reports to be generated on demand .[7]

**Aim of the Study**

The aim of the study is to analyse the characteristics of web log files for efficient web user management.

**Conclusion**

Web server Log files are very important of web server. This file stores various information related to the web server and it's uses by the clients. This review paper gives a brief idea bout these log files, their formats, their creation and log analysis software.

**References**

1. https://en.wikipedia.org/wiki/Web_server#cite_ref-new_1-0
2. https://en.wikipedia.org/wiki/Server_log
3. Ratnesh Kumar Jain, Dr. R. S. Kasana1, Dr. Suresh Jain, (July 2009 )"Efficient Web Log Mining using Doubly Linked Tree," International Journal of Computer Science and Information Security, IJCSIS, vol. 3.
4. L.K. Joshila Grace1, V.Maheswari, Dhinaharan Nagamalai, Analysis of Web Logs And Web User In Web Mining, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.
5. K. R. Suneetha, and R. Krishnamoorthi,(April 2009 ) "Identifying User Behavior by Analyzing Web Server Access Log File," IJCSNS International Journal of Computer Science and Network Security, vol. 9, pp. 327-332.
6. http://www.mediacollege.com/internet/statistics/logs/
7. https://en.wikipedia.org/wiki/Web_log_analysis_software
8. https://www.weblogexpert.com/faq.htm#q2_1