International Journal of Science, Environment and Technology, Vol. 2, No 5, 2013, 1008 –1016
ISSN 2278-3687 (O)

ANALYSING SERVER LOG FILE USING WEB LOG EXPERT IN WEB DATA MINING

[1]V. Jayakumar and [2]Dr. K. Alagarsamy
[1]Assistant Professor in Computer Science
Ayya Nadar Janaki Ammal College, Sivakasi
[2]Associate Professor in Computer Science
Madurai Kamarajar University, Madurai
E-mails: [1]kumarsep10th@gmail.com / [2]alagarsamymku@gmail.com

Abstract: Web Log Expert is a fast and powerful access log analyzer. It will give you information about your site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. The program produces easy-to-read reports that include both text information (tables) and charts. View the Web Log Expert sample report to get the general idea of the variety of information about your site's usage it can provide.

1. Introduction

Web log analysis tools help you to see a wide variety of statistical information about visitors and traffic to your website. These tools turn basic server text log files into graphical formats that are easy to understand. Most web log analysis tools tell you how many visitors your website receives in a given period of time, what web browsers are used most often by your visitors and which pages on your website the visitors viewed. These types of statistics are valuable, because they can help you to learn how your website is performing and determine the positive and negative results of recent design changes or marketing efforts.

2. WebLog Expert

WebLog Expert is a purchased application that comes in a "Lite" and "Full" version. WebLog Expert works on both Apache and IIS web servers. The software features HTML reports that include multiple graphical charts to show the number of visitors, how they found your site, and what pages they viewed after arriving. WebLog Expert also includes a click overlay report, which shows you specific links visitors click on when they're viewing a given page on your website.

*Received August 20, 2013 * Published October 2, 2013 * www.ijset.net*

2.1 Log Files

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.

Most hosters provide access to log files via FTP or control panel so you can download them and analyze with WebLog Expert Lite. If you have your own web server you can find and change log location in the server settings. WebLog Expert Lite supports log files of the most popular web servers: **Apache** and **IIS**.

2.2 Apache Log Format

The program supports combined and common log formats of Apache web server. We recommend you to use the Combined log format because the Common log format doesn't contain information about referrers and user agents (OS, browsers, spiders). By default Apache uses the Common log format but the majority of hosting providers set the Combined log format for Apache on their servers. Log format can be configured by editing the "httpd.conf" file in the Apache conf directory (if you have access to this file).

2.3 IIS Log format

WebLog Expert supports the W3C Extended log format that is the default log format of IIS 4/5/6/7/8. By default IIS log files contain only few fields but you can configure IIS to show other fields. Using IIS 4/5/6 First open the Properties" dialog for your web server, then on the "Web Site" property sheet choose "W3C Extended Log File Format" as the active log format, click the "Properties" button to open the "Extended Logging Properties" dialog and use the “Extended Properties” sheet to set the logged fields. By using IIS 7/8, first open your server or site in the IIS manager and Double –click the “Logging” icon. Then choose “W3C” as the active log format, click the “Select Fields” button to open the “W3C Logging Fields” dialog and use it to set the logged fields.

3. Web Log Fields

This weblog Expert software contain various fields Error, Bandwidth, Cache Request, Entry Page, Exit Page, Failed Request, Hit, Host, Incomplete Request, Page View, Spider, Total Unique IPs and Visitors.

The 404 Error will be displayed if a user requested a file, which doesn’t exist in the site. Bandwidth can be calculated by the amount of traffic transmitted from the site. A request that was cached on a client. If a browser has a cached copy of the requested file, it sends

special request to a server so it sends the file only if it hasn't been modified. Otherwise the browser uses the cached copy of the file, and the request is logged on the server as the cached one. Then the entry page and exit page will be counted by the first page visited by a user on the site and the last page visited by a user on the site. If a request which caused an error, it might be considered as Failed request. A request for any file (page, image, etc) is to be determined in the name of the field Hit. A request in response to which the server sent only a part of a file. Many download managers download files using several threads each of which downloads a part of the file, so it is logged as several incomplete requests. Incomplete requests may also occur if the files (pages, images) are too large and/or users have problems with getting them from the site. The Hosts are nothing but a computer connected to Internet. User hosts are shown in the reports as IP addresses or domain names. Spider is a program which automatically gets information from sites. Spiders gather information for search engines, extract emails, check links, etc. A number of different user IP addresses or domain names are called Total Unique IPs. The program determines number of visitors by the IP addresses. If a request from an IP address came after 30 minutes since the last request from this IP, it is considered to belong to a different visitor. Requests from spiders aren't used to determine visitors.

Summary

Hits	
Total Hits	30,474
Visitor Hits	29,191
Spider Hits	1,283
Average Hits per Day	4,353
Average Hits per Visitor	8.18
Cached Requests	3,979
Failed Requests	233
Page Views	
Total Page Views	4,435
Average Page Views per Day	633
Average Page Views per Visitor	1.24

Visitors	
Total Visitors	3,570
Average Visitors per Day	510
Total Unique IPs	2,970
Bandwidth	
Total Bandwidth	567.48 MB
Visitor Bandwidth	548.81 MB

4. Activities

The web log experts is used to perform various activities like activity statistics, access statistics, information about visitors, referrers, browsers and information about errors. The activity statistics is use to give the details about the daily user access by hour, day, week and month. Whether the particular user has accessed the data to calculate the statistics in order to find the value of pages, files, images, directories, queries, and view time. Then calculate the entry and exit pages, paths through the site, file types, virtual domains and load balanced servers details. There are two types of filters available in this web log experts.

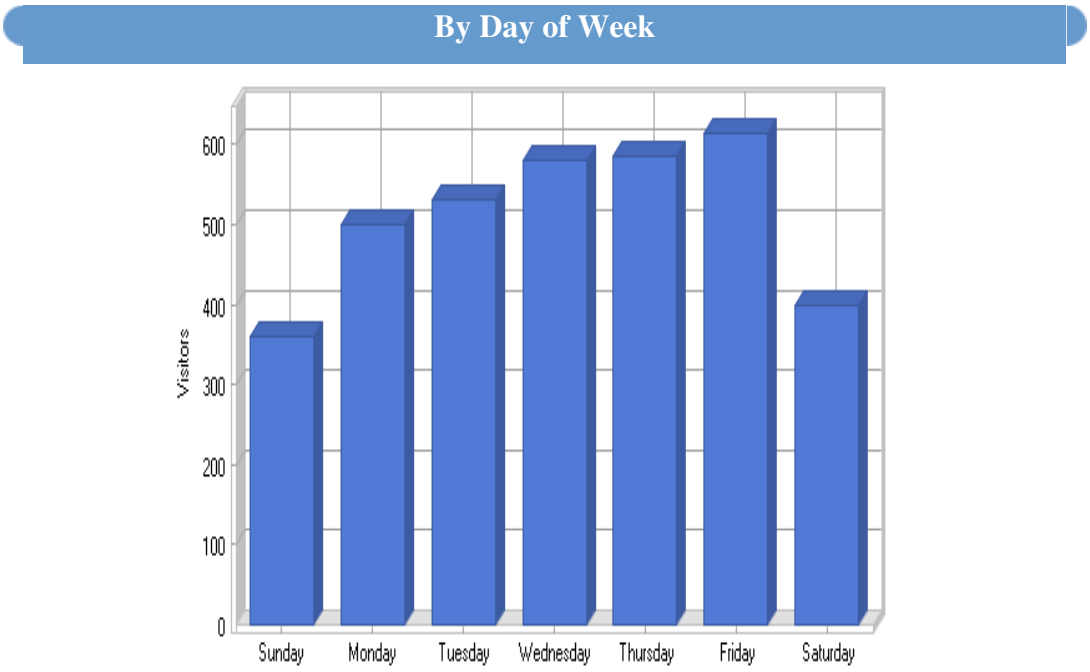
Hit (log entry) filters: host, requested file, query, referrer, status code, method, OS, browser, spider, user agent, day of the week, hour of the day, country, state, city, organization, authenticated user , virtual domain.

Visitor filters: visitors who accessed a specific file, visitors with a specified entry page, visitors with a specified exit page, visitors who came from a specific referring URL,

Date	Hits	Page Views	Visitors	Bandwidth (KB)
Sat 2/09/2013	3,306	485	399	64,767
Sun 2/10/2013	3,065	490	361	51,544
Mon 2/11/2013	4,301	583	500	84,284
Tue 2/12/2013	4,243	599	532	100,983
Wed 2/13/2013	4,897	759	579	86,360
Thu 2/14/2013	4,893	714	585	105,799
Fri 2/15/2013	5,769	805	614	87,362
Total	**30,474**	**4,435**	**3,570**	**581,103**

visitors who came from a specific search engine/phrase

4.1 Activity by Day of the Week



It is permitted to retrieve the details about the visitors' information like hosts, top-level domains, countries, states, cities, organizations, authenticated users, screen resolutions, color depths and languages. The users' referring sites, URLs and search engines (including information about search phrases and keywords) will be collected from the web log files. More over what types of browsers and operating systems they have used to access a web page. Some times user may get the error message when they access the web site, web log is used to determine the types of error and return the detailed error information.

5. Reports

The Web Log Experts used to generate the report for the Accessed pages, Downloaded files, Requested images, Requested Directories, subdirectories are counted separately from parent directories so request for files in /dir/subdir/ are counted for this directory, not for /dir/. It also contains the reports for list of referring sites (domain), list of referring URLs, list of search engines visitors of your site came from and list of search phrases. A search engine/phrase/keyword is counted and shown in reports if a visitor searched in a search engine and clicked a found link to your site. It is not counted if a link to your site was shown in a search engine results but a visitor haven't followed it.

5.1 Most Requested File Types

S.No	File Type	Hits	Incomplete Requests	Bandwidth (KB)
1	gif	18,403	5	27,744
2	html	4,148	0	19,282
3	jpg	2,186	2	15,781
4	ico	1,746	0	7,170
5	css	1,032	4	6,236
6	exe	890	539	424,439
7	php	224	0	723
8	xml	152	0	1,179
9	pdf	124	103	9,026
10	swf	94	0	47,991
11	txt	32	0	13
12	zip	12	0	2,303
13	asp	7	0	4
14	rtf	1	0	0
	Total	**29,051**	**653**	**561,896**

5.2 Most Used Browsers

S.No	Browser	Hits	Visitors	% of Total Visitors
1	Internet Explorer	21,302	2,260	61.43%
2	Firefox	6,461	883	24.00%
3	Others	275	153	4.16%
4	Opera	400	100	2.72%
5	Google Desktop	60	59	1.60%
6	Mozilla/4.0 (compatible;)	148	38	1.03%
7	ActiveRefresh	19	19	0.52%

8	Gecko/20070308 Minefield/3.0a1	96	10	0.27%
	Total	**29,191**	**3,679**	**100.00%**

5.3 Internet Explorer Versions

S.No	Browser	Hits	Visitors	% of Total Visitors
1	Internet Explorer 7.x	13,045	1,202	53.19%
2	Internet Explorer 6.x	8,012	950	42.04%
3	Internet Explorer 5.x	240	107	4.73%
4	Internet Explorer 2.x	5	1	0.04%
	Total	**21,302**	**2,260**	**100.00%**

5.4 Firefox Versions

S.No	Browser	Hits	Visitors	% of Total Visitors
1	Firefox 2.0.x	6,052	777	88.00%
2	Firefox 1.5.x	275	73	8.27%
3	Firefox 1.0.x	79	23	2.60%
4	Firefox 3.0.x	48	3	0.34%
5	Firefox 0.9.x	3	3	0.34%
6	Firefox 0.10.x	1	1	0.11%
7	Firefox 1.6.x	1	1	0.11%
8	Firefox 0.8.x	1	1	0.11%
9	Firefox 0.5.x	1	1	0.11%
	Total	**6,461**	**883**	**100.00%**

5.5 Most Used Operating Systems

S.No	Operating System	Hits	Visitors	% of Total Visitors
1	Windows XP	21,449	2,383	66.75%
2	Windows Vista	3,921	440	12.32%

3	Others	716	301	8.43%
4	Windows 2000	1,386	181	5.07%
5	Windows Server 2003	890	94	2.63%
6	Mac OS	107	54	1.51%
7	Windows 98	482	50	1.40%
8	Windows NT	55	28	0.78%
9	Linux	174	28	0.78%
10	Windows 95	5	5	0.14%
11	Windows ME	4	4	0.11%
12	FreeBSD	1	1	0.03%
13	Windows 3.x	1	1	0.03%
	Total	**29,191**	**3,570**	**100.00%**

6. Important Features of WebLog

The WebLog Experts can be executed any WINDOWS operation system from Windows 2003 itself. It can support the Apache and IIS server logs. It automatically detects the log format and read the GZ and ZIP compressed logs. We can analyze logs from load balanced servers and download logs via FTP and HTTP. By using this software, it is possible to create a report in HTML, PDF and CSG format perhaps can upload reports via FTP and send via e-mail (SMPT or MAPI). The main role of the IP in this software is to country mapping database with additional city, state and organization database. Finally it supports, date macros, multithread DNS lookup and command line mode.

7. Conclusion

The analyzed log files don't contain information about bandwidth. We should configure our web server so it produces log files that contain such info to generate the personalized data.

References

[1] J. Punin and M. Krishnamoorthy. Log Markup Language (LOGML) Specification. http://www.cs.rpi.edu/_puninj/LOGML/draft-logml.html, 2000.

[2] B. Lavoie and H. F. Nielsen. Web characterization Terminology & Definitions Sheet. http://www.w3.org/1999/05/WCA-terms/, 1999.

[3] B. Masand and M. Spiliopoulou, editors. *Advances in Web Usage Mining and User Profiling:Proceedings of the WEBKDD'99 Workshop*. Number 1836 in LNAI. Springer Verlag, July 2000.

[4] M. Spiliopoulou and L.C. Faulstich. WUM: A Tool for Web Utilization Analysis. In *EDBT Workshop WebDB'98, LNCS 1590*. Springer Verlag, March 1998.

[5] http://cryptome.org/usage-logs.htm

[6] http://www.weblogexpert.com/