International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 4, Issue 1, 2013, pp1-8
http://bipublication.com

ANALYSIS OF WEB SERVER LOG FILES TO INCREASE THE EFFECTIVENESS OF THE WEBSITE USING WEB MINING TOOL

Arvind K. Sharma[1] and P.C. Gupta[2]

[1]School of Computer and System Sciences
Jaipur National University, Jaipur, Rajasthan-India
[2]Department of Computer Science
University of Kota, Kota, Rajasthan-India
*Corresponding author: Email: arvindsharma133@gmail.com

[Received-24/10/2012, Accepted-30/11/2012]

ABSTRACT:
The fundamental role of Web usage mining is to capture, analyze, and model the Web server logs. Usually it automatically discovers the usage behaviour of the Website users. In this paper, we have been implemented a Web mining tool to analyze the Web server log files of the Website. It evaluates the important information about visitors, top errors, web browsers and different platforms used by the Website users mostly. The obtained information shall definitely increase the effectiveness of the Website.

Keywords: Web Usage Mining, Web Server Log Files, WebLog Expert

[I] NTRODUCTION

Due to the volatile growth of information available over the Internet during the past few years, the World Wide Web has become most popular and powerful platform to store, propagate and retrieve information as well as mine useful knowledge. It is a way of communication and information dissemination and it serves as a platform for exchanging various kinds of information. The volume of information available on the internet has been increasing rapidly with the explosive growth of the World Wide Web. While users are provided with more information and service options, it has become more difficult for them to find the 'right' or 'interesting' information, the problem commonly known as information overload. The primary goal of web usage mining is to find out the useful information from web data or web log files. To do this task, web usage mining focuses on investigating the potential knowledge from browsing patterns of the users and to find the correlation between the pages on analysis.

The rest of paper is organized as follows: Section-II explains web usage mining. Section-III discusses about web server log files. Section-IV summaries related works. Section-V demonstrates Web Log Expert tool. Section-VI contains experimental results. Conclusion is shown in section-VII while references are mentioned in the last section.

[II] WEB USAGE MINING

Web usage mining is the application of data mining techniques to discover the knowledge

hidden in the web server log file, such as user access patterns from web data and for analyzing users' behaviour patterns. Web mining deals with the data related to the web, which may be the data actually present in web pages or the data concerning the web activities. Web mining refers to overall process of discovering potentially useful and previously unknown information from web documents and services[1]. This is to ensure an improved service of web-based applications. The user access log files present very significant information about a web server. It is applied to fix several world problems by discovering the interesting user navigational patterns. The web information is categorized into two categories: deep web and shallow web. The deep web includes information stored in searchable databases often inaccessible to search engines and it is accessed only by Website's interface. In other hand, the shallow web information can be accessed by search engines without accessing the web databases[2].

[III] WEB SERVER LOG FILES

The server log files are simple text files which records activity of the users on the server. These files reside on the server. If user visits many times on the Website then it creates entry many times on the Server. The main source of raw data is the web access logs which are known as web server log files. The log files can be analyzed over a time period. The time period can be specified on hourly, daily, weekly and monthly basis. The typical web server log files contain such type of information: IP address, request time, method (e.g. GET), URL of the requested files, HTTP version, return codes, the number of bytes transferred, the referrer's URL and user agents.

3.1 Taxonomy of Web Server Logs

Web server logs are plain texts i.e. ASCII files which are independent from the server[11]. There are some distinctions between server softwares but traditionally there are four categories of web server logs, which are shown in fig.1.

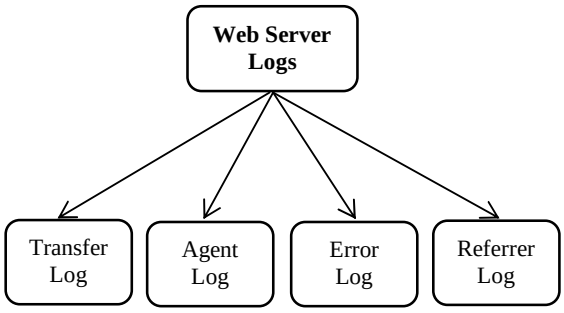

Fig.1: Taxonomy of Web Server Logs

The first two web logs such as Transfer Log and Agent Log are standard. The Referrer and Agent Logs may or may not be Turned On at the Server or may be added to the Transfer Log file to create an Extended log file format.

3.2 Sample of Web Server Log Files

We have been shown a sample of web log data in table-1 below. A user Id is the unique name that uses to identify. User Id is displayed when the user would like to make any transactions on the website or any other means.

Table-1: Sample of Web Server Log Files

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

However, other users could not see the real name and other personal information. Each row of web log file represents the URLs that user visits. Attributes of the web log file include Visit Time, Host, URL and other miscellaneous information about user's activity.

[IV] RELATED WORKS

In recent years, web usage mining is one of the favourite area of many researchers. In one of the work a novel approach was introduced for classifying user navigation patterns and predicting user's future request[3]. In another work a methodology was proposed and web log data was used to improve marketing activities[4]. Valter Cumbi et al. have done a case study of e-

government portal initiative in Mozambique for visitor analysis[5]. A work is done on mining interesting knowledge from web logs which presented in[6]. Ramya et al. have proposed a methodology for discovering patterns in usage mining to improve the quality of data by reducing the quantity of data[7]. Maheswara Rao et al. have introduced a research frame work capable of preprocessing web log data completely and efficiently. This framework helps to mine usage behavior of the users[8]. One work specifies a recommender system that was able to online personalization for user patterns[9]. In one more work, a methodology was proposed for interesting knowledge mining through web access logs[10].

[V] WEB LOG EXPERT TOOL

WebLog Expert is a fast and powerful Apache log file analyzer and IIS log analyzer tool. WebLog Expert is a freeware web mining tool[12]. It helps reveal important statistics about the Website usage like: activity of visitors, access statistics, paths through the site, visitors' browsers, and much more. This software tool can read log files of the most popular Web servers such as: Apache (Combined and common log formats) and IIS 4/5/6/7. It can also read ZIP, GZ and BZ2 compressed log files so won't need to unpack the logs manually before analyzing. The GUI Interface of WebLog Expert tool contains menu, toolbar and the list of profiles which is shown in fig.2 below.

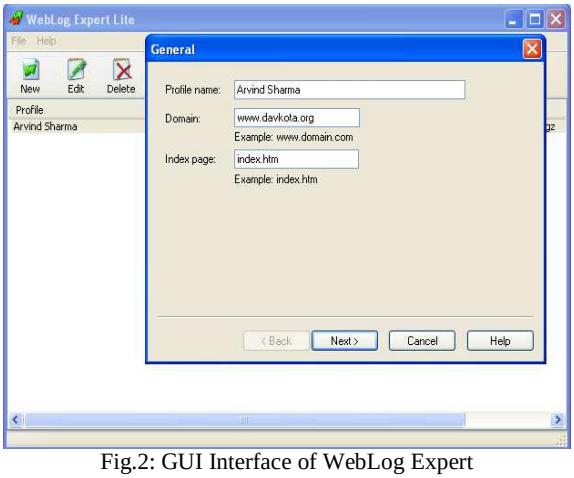

Fig.2: GUI Interface of WebLog Expert

5.1 Information Collected by WebLog Expert

- **Number of Hits–** This number usually signifies the number of times any resource is accessed in a Website. A hit is a request to a web server for a file i.e. web page, image, JavaScript, Cascading Style Sheet, etc.
- **Number of Visitors–** A visitor is exactly what it sounds like. It is a human who navigates to the website and browses one or more pages on the website.
- **Visitor Referring Website–** The referring website gives the information or URL of the website which referred the particular website in consideration.
- **Visitor Referral Website–** The referral website gives the information or URL of the website which is being referred to the particular website in consideration.
- **Time and Duration–** This information in the web server logs give the time and duration for how long the website was accessed by the particular user.
- **Path Analysis–** Path analysis gives the analysis of the path to a particular user has followed in accessing contents of a website.
- **Visitor IP Address–** This information gives the IP address of the visitors who visited the website.
- **Browser Type–** This information gives the information of the type of web browser that was used for accessing the website.
- **Platform–** This information provides the type of operating systems or platforms etc. which has been used to access the website.
- **Cookies–** A message given to a web browser by a web server. The browser stores the message in a text file called cookie. The message is then sent back to the server each time the browser requests a page from the server. The main purpose of cookies is to identify users and possibly prepare customized web pages for them.

5.2 Hits, Visits & Page Views

- **Hit–** Each file sent to a web browser by a server is known as an individual hit.
- **Visit–** A visit happens when someone visits the website. It contains one or more page views/hits. One visitor can have many visits to the website. A unique visitor is determined by the IP address or cookie. By default, a visit session is terminated when a user falls on inactive state for more than 30 minutes. If the visitor left the website and came back 30 minutes later, then WebLog Expert will report 2 visits. If the visitor came back within 30 minutes, then WebLog Expert will still report 1 visit.
- **Page View–** A page view is each time a visitor views a web page on the website, irrespective of how many hits are generated. Pages are comprised of files. Every image in a page is a separate file. When a visitor looks at a page i.e. a page view, they may see numerous images, graphics, pictures etc. and generate multiple hits. For example– if a web page contains 5 images, a 'hit' on that page will generate 6 'hits' on the web server, one page view for the web page, 5 hits for the images.

5.3 HTTP Status Codes

The several status codes of hypertext transfer protocol are shown in table-1 below.

Table-2: HTTP Status Codes

Status Code	Description
101	Switching Protocols
200	OK
201	Created
202	Accepted
203	Non-Authoritative Information
204	No Content
205	Reset Content
206	Partial Content
300	Multiple Choices
302	Found
303	See Other
304	Not Modified
305	Use Proxy
306	(Unused)
307	Temporary Redirect
400	Bad Request
401	Unauthorized
402	Payment Required
403	Forbidden
404	Not Found
405	Method Not Allowed
406	Not Acceptable
407	Proxy Authentication Required
408	Request Timeout
409	Conflict
410	Gone
411	Length Required
412	Precondition Failed
413	Request Entity Too Large
414	Request-URI Too Long
415	Unsupported Media Type
416	Requested Range Not Satisfiable
417	Expectation Failed
500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Service Unavailable
504	Gateway Timeout
505	HTTP Version Not Supported
101	Switching Protocols

[VI] EXPERIMENTAL RESULTS

In this work we have been analyzed the web server log files of an Educational Institution's Website i.e. www.davkota.org[13] with the help of WebLog Expert tool. Statistical/text log file data have been used for experimentation provided by WebLog Expert. Various analysis have been carried out to identify the behavior of the Website users.

6.1 Experiment-1

The log files contain the data from October 8, 2012 to October 14, 2012 (Time range: 10/8/2012 18:03:08 - 10/14/2012 17:29:07). In this duration log files have been stored 25 MB data and we have got 3.4 MB data after the preprocessing task for this work. The web log data received after preprocessing has been implemented through WebLog Expert and the complete analysis has been done. Fig.3 shows that Website is accessed every day, it receives far more visits from Monday to Saturday. The increased visits received by the Website on Monday to Saturday, reinforces the

earlier finding that the Website is mainly used by working people, employees or students.

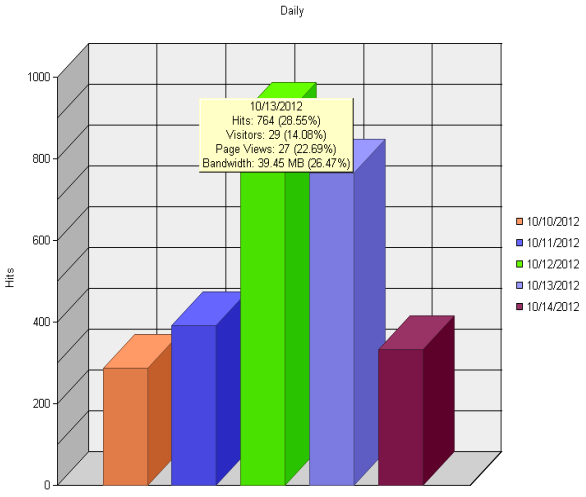

Fig.3: Total Number of Hits Ratio

The daily web data for the entire week from October 8, 2012 to October 14, 2012, tells about the number of Hits occurs, which Files, Pages, Visits, and Kbytes have been visited. For this week, the maximum Hits per Day were 637, the maximum Files per Day were 144, the maximum Pages per Day were 85, the maximum visits per Day were 13 and the maximum Kbytes per Day were 903. (see Appendix-A)

We have found different types of errors occurred during the web surfing. The different types of errors are shown in table-3 below.

Table-3: 404 Errors (Page Not Found)

S. No.	Error	Hits
1.	404 Not Found	368
2.	503 Service Unavailable	31
3.	403 Forbidden	1
	Total	**400**

It is very clear from the table-3 that 404 is most frequently occurred error. Some other types of client and server errors are shown. The graphical presentation of daily error types is shown in fig.4.

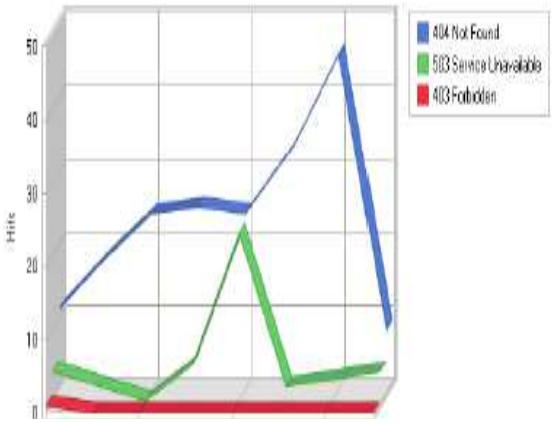

Fig.4: Daily Errors Types

6.2 Experiment-2

In this work, the collected web server log files from October 8, 2012 to October 14, 2012 are experimented through WebLog Expert tool. It has been found that most of the Web browsers are used by the most of the users to visit the Website. Google Chrome is one of mostly used web browser. Which is shown in fig.5.(see Appendix-B)

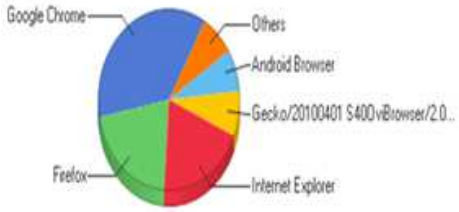

Fig.5: Mostly Used Web Browser

Different versions of the Web browser Internet Explorer are shown in table-4. Most of the users accessed the website through the Internet Explorer version 7.x.

Table-4: Different Versions of Internet Explorer

S. No.	Web Browser	Hits	Visi-tors	% of Total Visitors
1.	Internet Explorer 7.x	119	11	35.48%
2.	Internet Explorer 8.x	140	9	29.03%
3.	Internet Explorer 6.x	30	7	22.58%
4.	Internet Explorer 9.x	54	4	12.90%
	Total	**343**	**31**	**100.00%**

The graphical presentation of this information is shown in fig.6.

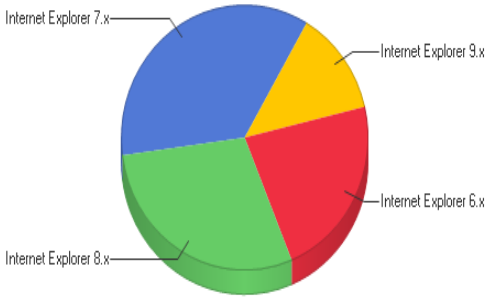

Fig.6: Different versions of Internet Explorer

Several Operating Systems have been used by the website users to access the Website. Windows Operating Systems have been frequently used by the website users. Windows XP operating system has been used by the most of users to access the Website. The mostly used Operating System is shown in fig.7 below. (see Appendix-C)

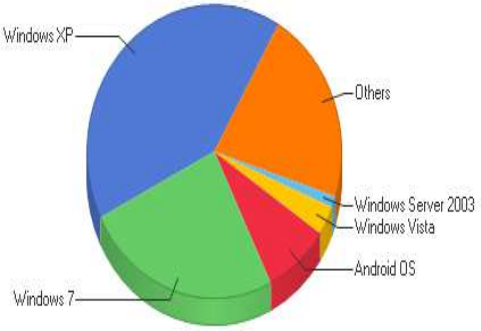

Fig.7: Mostly Used Operating System

[VII] CONCLUSION

Web is a huge storehouse of web pages and links. It offers large quantity of data for Internet users. When users visit the Web they leave copious information in terms of web access logs which is heterogeneous, complex, high dimensional and incremental in nature. Analyzing such type of data will help to determine the browsing interest of the website users. In this paper, the complete analysis of web server log files has been done by using WebLog Expert tool. The obtained results shall definitely help to the Website Maintainers, Website Analysts, Website Designers and Developers to manage their System by determining occurred errors, corrupted and broken links. This work will also increase the effectiveness of the Website.

REFERENCES

[1] Margaret H. Dunham, S. Sridhar, “Data Mining: Introductory and Advanced Topics”, Pearson Education.

[2] Arvind Kumar Sharma, P.C. Gupta, “Exploration of efficient methodologies for the Improvement in web mining techniques: A survey”, International Journal of Research in IT & Management(IJRIM) Vol.1, Issue-3, July 2011.

[3] 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.

[4] Arya, S., et al., “A methodology for web usage mining and its applications to target group identification”, Fuzzy sets and systems, 2004, pp.139-152.

[5] Valter Cumbi et al. “Mozambican Government Portal Case Study: Visitor Analysis”, IST-Africa 2007 Conference Proceedings Paul Cunningham and Miriam Cunningham (Eds) IIMC International Information Management Corporation, 2007.

[6] F.M. Facca, and P.L. Lanzi, “Mining interesting Knowledge from Web logs: a survey”, Elsevier Science, Data and Knowledge Engineering, 2005, 53, pp.225-241.

[7] G. R.C. et al., "An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network,"Fifth International Conference on Information Processing, 2011. Springer-Verlag.

[8] Maheswara Rao.V.V.R and Valli Kumari.V, "An Enhanced Pre-Processing Research Framework for web Log Data Using a Learning Algorithm," Computer Science and Information Technology, DOI, pp. 1-15, 2011. 10.5121/csit.2011.1101.

[9] Mehrdad Jalali et al., “A Recommender System for Online Personalization in the WUM Applications”, Proceedings of the World Congress on Engineering and Computer Science 2009 Vol. II, WCECS 2009, October 20-22, 2009, San Francisco, USA.

[10] K Sudheer Reddy et al, “An Effective Methodology for Pattern Discovery in Web Usage Mining”, International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012, 3664-3667.

[11] L.K. Joshila Grace et al., “Web Log Data Analysis and Mining” in Proc CCSIT-2011, Springer CCIS, Vol. 133, Jan 2011.

[12] [Online] http://www.weblogexpert.com

[13] DAV Kota website’s server is available at: [Online] http://www.davkota.org

AUTHOR'S PROFILE

Arvind K. Sharma had received his Master's Degree in Computer Science from Maharshi Dayanand University, Rohtak and M.Phil Degree in Computer Science from Alagappa University, Karaikudi. He is pursuing Ph.D in Computer Science from School of Computer and Systems Sciences, Jaipur National University, Jaipur, Rajasthan, India. His areas of interest include Data Mining, Web Usage Mining and Web Applications.

P. C. Gupta had received his Ph.D degree in Computer Science from Bundelkhand University, Jhansi. Presently he is working as Associate Professor in the Department of Computer Science & Informatics, University of Kota, Rajasthan, India. He has published various technical and research papers in National and International Conferences and Journals. His research interest lies in Artificial Intelligence and Neural Networks.

Appendix–A
General Statistics

Time range: 10/8/2012 18:03:08 - 10/14/2012 17:29:07 Generated on Sun Oct 14, 2012 - 23:31:49

Summary

Summary

Hits	
Total Hits	4,114
Visitor Hits	3,814
Spider Hits	300
Average Hits per Day	587
Average Hits per Visitor	24.29
Cached Requests	160
Failed Requests	368
Page Views	
Total Page Views	273
Average Page Views per Day	39
Average Page Views per Visitor	1.74
Visitors	
Total Visitors	157
Average Visitors per Day	22
Total Unique IPs	231
Bandwidth	
Total Bandwidth	187.55 MB
Visitor Bandwidth	180.70 MB
Spider Bandwidth	6.85 MB
Average Bandwidth per Day	26.79 MB
Average Bandwidth per Hit	46.68 KB
Average Bandwidth per Visitor	1.15 MB

Appendix –B

Most Used Browsers

	Browser	Hits	Visitors	% of Total Visitors
1	Google Chrome	2,284	58	36.48%
2	Firefox	514	33	20.75%
3	Internet Explorer	343	31	19.50%
4	Gecko/20100401 S40OviBrowser/2.0.2.68.14	15	13	8.18%
5	Android Browser	453	12	7.55%
6	Opera	187	5	3.14%
7	RockMeltEmbedService	1	1	0.63%
8	SAMSUNG-GT-C6712/C6712DDKK3 SHP/VPP/R5 Dolfin/2.0 NexPlayer/3.0 SMM-MMS/1.2.0 profile/MIDP-2.1 configuration/CLDC-1.1 OPN-B	1	1	0.63%
9	SAMSUNG-GT-C3312/1.0 NetFront/4.2 Profile/MIDP-2.0 Configuration/CLDC-1.1	1	1	0.63%
10	MAUI WAP Browser	1	1	0.63%
11	Google Desktop	1	1	0.63%
12	Mozilla/4.0 (compatible; http://search.thunderstone.com/texis/websearch/about.html)	12	1	0.63%
13	Mobile Safari	1	1	0.63%
	Total	**3,814**	**159**	**100.00%**

Appendix –C

Most Used Operating Systems

	Operating System	Hits	Visitors	% of Total Visitors
1	Windows XP	1,575	64	40.76%
2	Windows 7	1,602	39	24.84%
3	Others	104	31	19.75%
4	Android OS	485	13	8.28%
5	Windows Vista	33	5	3.18%
6	Windows Server 2003	12	2	1.27%
7	iPad	1	1	0.64%
8	Linux	1	1	0.64%
9	Mac OS	1	1	0.64%
	Total	**3,814**	**157**	**100.00%**