

Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool

Neha Goel
Research Scholar
Banasthali Vidyapith, India

C.K. Jha, PhD.
Associate Professor,
Banasthali Vidyapith, India

ABSTRACT

Internet is acting as a major source of data. As the number of web pages continues to grow the web provides the data miners with just the right ingredients for extracting information. In order to cater to this growing need a special term called Web mining was coined. Web mining makes use of data mining techniques and deciphers potentially useful information from web data. Web Usage mining deals with understanding the behavior of users by making use of Web Access Logs that are generated on the server while the user is accessing the website. A Web access log comprises of various entries like the name of the user, his IP address, number of bytes transferred timestamp etc. A variety of Log Analyzer tools exist which help in analyzing various things like users navigational pattern, the part of the website the users are mostly interested in etc. The present paper makes use of such log analyzer tool called Web Log Expert for ascertaining the behavior of users who access an astrology website. It also provides a comparative study between a few log analyzer tools available.

Keywords

Web Access Logs, Web usage Mining, Web server, Log Analyzer.

1. INTRODUCTION

A revolution has been observed in the way people work on the internet. People are making use of this important tool for disseminating their ideas, conducting business and most important entertaining themselves. Data on the web is rapidly increasing day by day [1]. Web is an open medium [2]. And due to this openness it has become real tough for users to plough through this abundant information. In order to provide a solution to this problem the term Web mining was coined. Web mining refers to the use of data mining techniques to automatically retrieve, extract and analyze information for knowledge discovery from Web documents and services [3]. Web mining is categorized into namely: Web Content Mining, Web Structure Mining and Web Usage Mining. Web Usage mining itself can be classified further depending on the kind of usage data considered into: Web Server Data, Application Server Data and Application Level data [4].

The Web Server data is actually the user logs that are generated on the Web Server. These logs enable the analyst to keep a track of and analyze the user's behaviors who visit the website. The present paper begins with an insight of what logs are, their typical structure and why log analysis should be done. It also enlists the various tools available for performing analysis of log data along with a comparative study of some tools. In the end the results and interpretations obtained after the application of log analyzer tool have been provided.

2. LOGS

Web server logs stores click stream data which can be useful for mining purposes [5]. They are plain text (ASCII) files which contain information about User Name, IP Address, Time Stamp, Access Request, URL that Referred, error codes (if any) etc. and generally reside in the web servers. Traditionally there are four types of server logs: Transfer Log, Agent Log, Error Log and Referrer Log [6]. The Transfer and the Agent Log are said to be standard whereas the error and referrer log are considered optional as they may not be turned on. Every log entry records the traversal from one page to another, storing user IP number and all the related information [3].

A sample log taken from an astrology website is given below:

```
2010-12-27 14:36:09 W3SVC1 H-ASTROSCIENCES
69.16.245.169 GET /Index.html - 80 - 69.16.234.108 HTTP/1.0
check_http/1.81+(nagios-plugins+1.4.2) - - 69.16.245.169 200 0
64 0 91 93

2010-12-27 15:02:49 W3SVC1 H-ASTROSCIENCES
69.16.245.169 GET /Index.html - 80 - 69.16.234.108 HTTP/1.0
check_http/1.81+(nagios-plugins+1.4.2) - - 69.16.245.169 200 0
64 0 91 0

2010-12-27 15:07:49 W3SVC1 H-ASTROSCIENCES
69.16.245.169 GET /Index.html - 80 - 69.16.234.108 HTTP/1.0
check_http/1.81+(nagios-plugins+1.4.2) - - 69.16.245.169 200 0
64 0 91 15
```

2.1 Why Log Analysis is required?

The mere complexity of the data volume paves way for the requirement of hybrid intelligent systems for intelligent information analysis and trend prediction [7]. But this complex data needs to be handled very carefully. For analyzing the data of web access logs of a website one needs to preprocess it. Preprocessing involves Cleaning, User/Session Identification and Transaction identification. A variety of data cleaning algorithms have been proposed in this context.

Apart from making use of cleaning algorithms and then applying various mining algorithms to the cleansed data, yet another way by which we can extract useful information from this log data is by making use of automated log analysis tools. Web access patterns mined from Web logs are interesting and useful knowledge in practice [8]. This extracted data can be highly useful if dealt with carefully. It guides the analyst in determining the navigational pattern of the user i.e. which pages are frequently visited by the user, after which page is he losing his interest in the website, which browser is being used by the people who access the website, the kind of errors that they get etc.

Logs if utilized properly can prove to be very useful in turning the websites visitors into customers especially in case of an e-commerce website. The personalization that is offered on various websites these days often makes use of this aspect of web usage mining.

3. TOOLS AVAILABLE

A variety of tools are available that take these web access logs as an input and generate the reports as an output. These tools provide us with all sorts of information starting from how many hits the site getting to the number of visitors accessing the site, the browsers that they use the length of their stay, and much more. Some of the tools that are available are:

- **Google Analytics** – It is a free utility provided by Google which helps in keeping a track of unique visitors. It also helps in determining which marketing packages are offering the best.
- **AWStats**- It is available free of cost. This tool works as a CGI Script or from command line. It displays all sorts of information that the log contains.
- **WebLog Expert**- Yet another log analyzer tool that provides thorough analysis of the web access logs. It provides the users with specific and precise information about user’s statistics. It supports log files extracted from Apache and IIS.
- **Analog** – An easy to use and install freely available log analysis tool. It is extremely fast, highly scalable, works on any operating system and an easy to install tool.

The installation of Awstats and Analog log analyzer tools is bit typical. Awstats requires PERL and Analog works on Command line interface which makes it difficult to understand and interpret. Hence Web Log Expert was taken as a tool to analyze the logs of the website.

Web log Expert is a freeware and comes in four editions: Enterprise, Professional, Standard and Lite. It is a fast and powerful access log analyzer. The installation is quite easy and GUI provided by the tool is highly user friendly. It gives information about our site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. It produces easy-to-read reports in the form of charts and tables.

Google Analytics is yet another log analyzer tool which provides standard and custom reporting services for people accessing our website. It helps a lot in measuring the sales and conversions of a website.

A comparative study between Google Analytics and Web Log Expert Lite was done and the following features were noticed.

Table 1: Comparison between Google Analytics and Web Log Expert Lite

Feature	Google Analytics	Web Log Expert
Vendor	Google	Alentum Software Ltd.
Edition	Single	Enterprise, Professional, Standard, Lite
Installation	No installation required, only requires a Google account	Easy to install
Import Raw Logs	No	Yes
User Interface	Easy	Easy
Need for creating an account	Yes	No, But profile creation is required
Can be linked to website	Directly	Logs can be imported and then used
Email report facility	Available	Not available in Lite edition
Export report facility	Available, Report can be exported in a variety of formats e.g. CSV, TSV, PDF	Lite Version does not support this utility. It cannot generate reports in PDF or CSV. Supports HTML reports only.
Freeware	Yes	Yes, only Lite edition. Trial version of 30 days available for the rest.
Read Compressed files	No need as directly accesses website statistics.	Yes
Advanced Reporting Options	Available	Not supported by every version
Website	http://www.google.com/analytics/	http://www.weblogexpert.com/

As discussed in the table above Google Analytics is a powerful log analyzer tool which is freely available and is highly user friendly. The major difference that lies between Google Analytics and Web Log Expert is that importing of logs utility is missing in case of Google Analytics. One needs to own a website and moreover needs to sign up in Google Analytics for tracking the user’s behavior. It does not work on raw logs. But the features provided by Google Analytics far exceed than those provided by Web Log Expert (Lite) despite both being freely available.

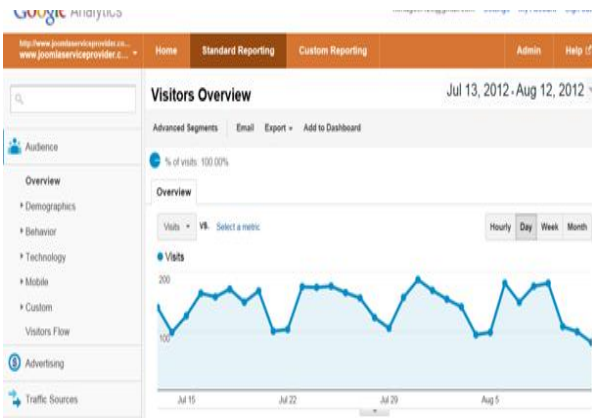


Figure 2: Snapshot of Google Analytics

4. RESULTS AND INTERPRETATIONS

Log analyzer tools are required as they help extensively in analyzing the information about visitors, top errors which can be utilized by system administrator and web designer to increase the effectiveness of the web site [5].

In the current research web access logs were taken from an astrology website for the time period 13/02/2012 to 08/08/2012 and the following results were obtained:

1. **General Statistics:** In this section we get general information pertaining to the website like how many times the website was hit, an average of hits in a day, page views, bandwidth etc. It enlists all the general information which one should know related to a website.

Table 2: General statistics obtained after analyzing web logs

Summary	
Hits	
Total Hits	24,820
Visitor Hits	22,359
Page Views	
Total Page Views	1,515
Average Page Views per Day	8
Average Page Views per Visitor	2.02
Visitors	
Total Visitors	750
Average Visitors per Day	4
Total Unique IPs	1,021
Bandwidth	
Total Bandwidth	305.98 MB
Visitor Bandwidth	270.70 MB

2. **Activity statistics:** The second feature provided by the tool. It provides the statistics on daily and hourly basis. Apart from the tabular report it also gives a graphical chart which helps in ascertaining at which hour of the day along with the days on which the website was hit the maximum. By utilizing this information special schemes can be initiated which might help in increasing the people who access the website.

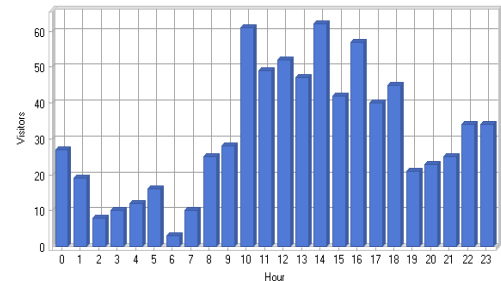


Figure 4 : Activity statistics showing at which hour of the day website is hit maximum.

According to the chart given above the website is hit maximum at 10:00 hrs and 14:00 hrs and is least visited at 2:00 hrs and 6:00 hrs.

3. **Access Statistics:** This part of the analysis can be considered the most important as it not only ascertains which page is hit the maximum number of times rather it also provides an idea of the navigational behavior of users and thus help in restructuring the website to suit the increasing demands of customer. It states that which pages are accessed the maximum amongst all the web pages available in a website.

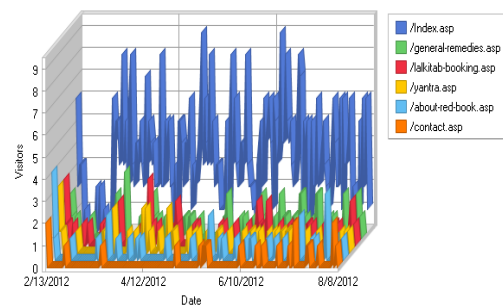


Figure 5: Access statistics of web logs showing which page is hit maximum.

As per the graph given above the page index.asp is hit maximum whereas contact.asp is hit the least. The next table will give us an idea of the navigational behavior of the user or after which web page is the user losing interest in the website.

Table 3: Table showing the number of visitors accessing a web page.

Top Entry Pages	
Page	Visitors
http://www.astrosiences.com/ Index.asp	582
http://www.astrosiences.com/ general-remedies.asp	13
http://www.astrosiences.com/ about-red-book.asp	12
http://www.astrosiences.com/ yantra.asp	9
http://www.astrosiences.com/ lalkitab-booking.asp	6

The table above shows that index.asp receives maximum interest from people. The other pages are hit almost negligibly by users. The page general remedies and about red book are however looked upon by a few users.

- Visitors:** The Visitors section helps in determining that who all accessed the website. The report below shows a list of IP addresses that hit the website along with how many times the website was hit by a particular user.

Table 4: Table showing the details of visitors accessing a web page.

Hosts	Hits	Visitors	Bandwidth
182.71.206.235	1,391	32	15,339
110.234.148.36	337	8	5,542
122.176.196.30	570	6	7,601
94.23.225.68	6	6	165
122.176.202.8	100	4	277
188.138.118.55	3	3	99

- Browsers:** The Web Log Expert Lite tool also helps the website owner in examining which browser is mostly preferred by the users so that the website can be made compatible with that particular web browser. It also gives a tabular description about the total number of hits provided by that particular web browser. Not only browser but it also lists the most preferred Operating System by the users.

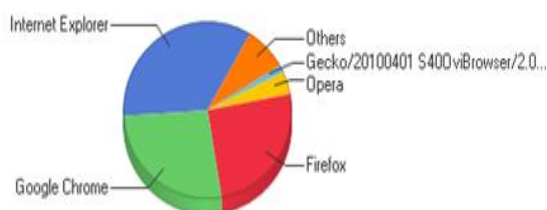


Figure 8: Pie chart showing the most used web browsers by users.

The chart above shows that Internet Explorer and Google chrome are highly preferred browsers.

- Errors:** The last feature provided by this tool is finding out what kind of errors people face when they access the website. For the error feature both a tabular and a graphical form of representation are available.

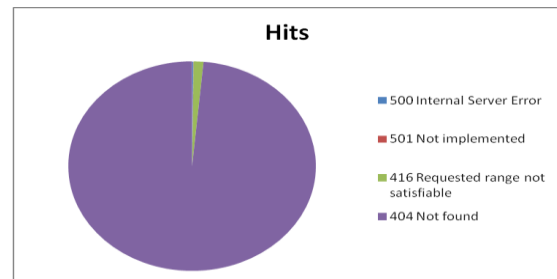


Figure 9: Pie chart showing the errors that frequently occur while accessing the website.

5. CONCLUSIONS

Web Log Analyzer tools are a part of Web Analytics Software. They take a log file as an input, analyze it and generate results. A variety of tools are available which offer great capabilities in reporting the results of analysis. A study was done and some of the tools were studied. Every tool offered some or the other feature which was better than the rest. Finally the tool, Web Log Expert was taken to analyze the web logs of the website as it provided extensive features that too in the free edition. The results were examined and are being tried to incorporate in the website of the user. Such log analyzer tools should be widely used as they help a lot in understanding the customer behavior to the analysts.

6. REFERENCES

- Brijendra Singh, Hemant Kumar Singh: Web Data Mining Research: A Survey, IEEE, 2010.
- G.K. Gupta, Introduction to Data Mining with Case Studies: Web Data Mining, PHI Learning Private Limited, pp. 231-233, 2011.
- Theint Aye: Web Log Cleaning for Mining of Web Usage Patterns, IEEE, 2011.
- Hillool Kargupta, Anupam Joshi, Krishnamoorthy SivaKumar, Yelena Yesha, Data Mining: Next Generation Challenges and Future Directions: Web Mining – Concepts, Applications and Research Directions, PHI Learning Private Limited, pp. 405-409, 2009.
- Navin kr Tyagi ,A.K. Solanki, Manoj Wadhwa : Analysis of Server Log by Web Usage Mining for Website Improvement, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8,pp. 17-21,2010.
- L.K. Joshila Grace, 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.
- Xiaozhe Wang, Ajith Abraham, Kate A. Smith: Intelligent web traffic mining and analysis, ELSEVIER Journal of Network and Computing Applications, 2004.

- [8] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, and Hua Zhu: Mining Access Patterns Efficiently from Web Logs. Application of Hybrid Intelligent Systems, pp. 1034-1043, December 2003.
- [9] K. R. Suneetha, Dr. R. Krishnamoorthi: Identifying User Behavior by Analyzing Web Server Access Log File in IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, pp. 327-332, April 2009.
- [10] Vic Ciesielski , Anand Lalani : Data Mining of Web Access Logs From an Academic Web Site in Proceedings of the Third International Conference on Hybrid Intelligent Systems (HIS'03): Design and
- [11] AWStats log file analyzer 7.1 Documentation , Log AnalyzerComparison:http://awstats.sourceforge.net/docs/awstats_compare.html
- [12] <http://www.google.com/analytics/>