A Survey on Web Log Analysis Tools

Khaing Phyo Wai, Yi Yi Aung
University of Computer Studies, Mandalay
khaingphyowai.mm@gmail.com, yiyiaung123@gmail.com

Abstract

Web log data or click stream data has a lot of valuable information about a web site or web site visitors' online behavior. Web usage analysis or web log mining is the process of capturing, analyzing and modeling the Web server logs. Analyzing of web usage mining has been made with the help of the web log analyzer tool to find out the abstract information from particular server and also tried to find out the user behavior. This paper survey the most famous log analyzer tools based on their features and difference between them. The aim of this survey is to help the web analysts to determine the best tool for what the need.

Keywords*: Web Log Data, Web Log Mining, Web Usage Analysis, Web log Analyzer Tool*

1. Introduction

Advancement in technology and growing use of the internet has opened up different study areas for statisticians. Every time users visit website, clicks are saved that can be used for extracting useful patterns. But due to unstructured and semi-structured data in webpage, it has become a challenging task to extract relevant information. Its main reason is that traditional knowledge based technique are not correct to efficiently utilization the knowledge, because it consist of many discovered pattern, contains a lots of noise and uncertainty. It is a complex task of searching and retrieving data or information from log data. It is also difficult to analyze since it is available as unstructured data and many different formats depending on the web server.

Web mining is the appliance of data mining functionality which is used to mine relevant information from web log data [6], [7]. Whatever interesting data has to retrieve from Web, it is possible through web mining [8]. Today huge amount of data is available on the web to extract data from the vast collection is a complex task. By applying some data mining method, we can find out useful pattern using web mining.

Web analytical tools can also be used to analyze the unstructured and semi-structured data in web log data [5]. These tools range from simple reporting applications to much advanced analytical software applications. There are many popular sophisticated tool which help analytics in analyzing and visualizing web log data. This paper surveys the powerful tools to analyze the raw web log data from the various webs.

2. Literature Review

Due to huge volume of web data, web data is not available in proper structure, due to which the searching result is irrelevant and also consist noise. The user access log files present very significant information about a web server [10]. This is especially true in regard to user

identification. IP addresses, which are all that really identifies the web user, cannot be traced back to an individual as the use of proxy servers. The primary goal of web mining is to find out the useful information from web data or web log files [11]. 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 [12]. Web mining refers to overall process of discovering potentially useful and previously unknown information from web documents and services [14].

3. Web Analytic Data

A web log is the recording of the parts of the screen a computer user clicks on while web browsing or using another software application [9]. As the user clicks anywhere in the webpage or application, the action is logged on a client or inside the web server, as well as possibly the web browser, router, proxy server or ad server. There are four main ways that this sort of data is captured, web logs, web beacons, JavaScript tags, and packet sniffing.

3.1. Web Log File

A web log file is a file produced by a Web server to record activities on the Web server. The log file is text file. Its records are identical in format. Each record in the log file represents a single HTTP request. A log file record contains important information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information.

```
02:49:12 127.0.0.1 GET / 200
02:49:35 127.0.0.1 GET /index.html 200
03:01:06 127.0.0.1 GET /images/sponsered.gif 304
03:52:36 127.0.0.1 GET /search.php 200
04:17:03 127.0.0.1 GET /admin/style.css 200
05:04:54 127.0.0.1 GET /favicon.ico 404
05:38:07 127.0.0.1 GET /js/ads.js 200
```

Figure 1. Some Sample Records from an IIS Server Log File

```
192.168.198.92 - - [22/Dec/2002:23:08:37 -0400] "GET
   / HTTP/1.1" 200 6394 www.yahoo.com
   "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1...)" "-"
192.168.198.92 - - [22/Dec/2002:23:08:38 -0400] "GET
   /images/logo.gif HTTP/1.1" 200 807 www.yahoo.com
   "http://www.some.com/" "Mozilla/4.0 (compatible; MSIE 6...)" "-"
192.168.72.177 - - [22/Dec/2002:23:32:14 -0400] "GET
   /news/sports.html HTTP/1.1" 200 3500 www.yahoo.com
   "http://www.some.com/" "Mozilla/4.0 (compatible; MSIE ...)" "-"
192.168.72.177 - - [22/Dec/2002:23:32:14 -0400] "GET
   /favicon.ico HTTP/1.1" 404 1997 www.yahoo.com
   "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; rv:1.7.3)..." "-"
192.168.72.177 - - [22/Dec/2002:23:32:15 -0400] "GET
   /style.css HTTP/1.1" 200 4138 www.yahoo.com
   "http://www.yahoo.com/index.html" "Mozilla/5.0 (Windows..." "-"
192.168.72.177 - - [22/Dec/2002:23:32:16 -0400] "GET
   /js/ads.js HTTP/1.1" 200 10229 www.yahoo.com
   "http://www.search.com/index.html" "Mozilla/5.0 (Windows..." "-"
192.168.72.177 - - [22/Dec/2002:23:32:19 -0400] "GET
   /search.php HTTP/1.1" 400 1997 www.yahoo.com
   "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; ...)" "-"
```

Figure 2. Some Sample Records from an Apache Server Log File

A browser may fire multiple HTTP requests to Web server to display a single Web page. This is because a Web page not only needs the main HTML document; it may also need additional files, like images and JavaScript files. The main HTML document and additional files all require HTTP requests.

Each Web server has its own log file format. If your Web site is hosted by an ISP (Internet Service Provider), they may not keep the log files for you, because log files can be very huge if the site is very busy. Instead, they only give you statistics reports generated from the logs files.

4. Web Log Analysis Tools

To deal with the growth of log data, log analysis tools have been built over the last few years to help developers and web analysts [13]. In order to get the most out of the log files, there are some steps you have to do. First, you need to collect log files. Then use a web log analysis tool to analyze the log file. Web log analysis tools create a set of reports that synthesize the information contained in your log files. Typically, web log analysis tools generate statistics about numbers of hits and visitors for the whole site and individual pages, what browser was used to view the site, what

operating system the user had when they visited the site, and where the user came to the site from. And then, analyze the information in the reports from your web server log analysis tool. Finally, make change to your site based on the information garnered from your web server log files.

4.1. AWStats

AWStats is a free powerful tool that generates advanced web, streaming, ftp or mail server statistics, graphically [1]. This log analyzer works as a CGI or from command line and shows all possible information the log contains, in few graphical web pages. It uses a partial information file to be able to process large log files, often and quickly. It can analyze log files from all major server tools like Apache log files (NCSA combined/XLF/ELF log format or common/CLF log format), WebStar, IIS (W3C log format) and a lot of other web, proxy, wap, streaming servers, mail servers and some ftp servers.

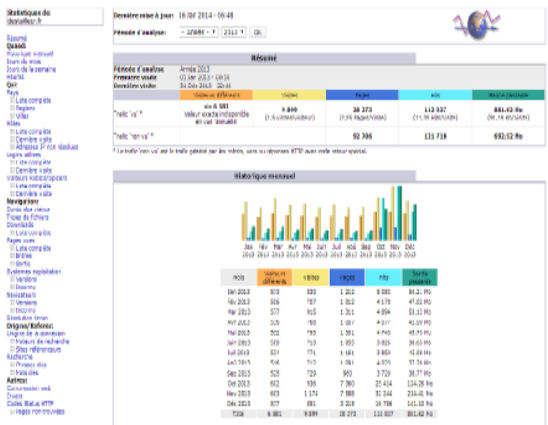

Figure 3. AWStats Log Analyzer Tool

4.2. Deep Log Analyzer

Deep log analyzer can analyze web site visitors' behavior and get complete Web site usage statistics [2]. It features hierarchical reports presentation with interactive navigation, convenient charts, and open database format. It can also view how statistics change over time, compare reports for different time intervals, dig deeper into the site data with the unique hierarchical reports and reveal hidden patterns in the site stats. It can give reports about accessed site resources, visitors' activity, referral sites, search engines and keywords, browsers and OS, search spiders, server errors. Moreover, it can create your own custom reports, access database in other applications. It can analyze log files from all popular Web servers including IIS and Apache and it can download log files via FTP. Beside Deep Log Analyzer can process logs archived in GZ or ZIP on the fly without extracting them.

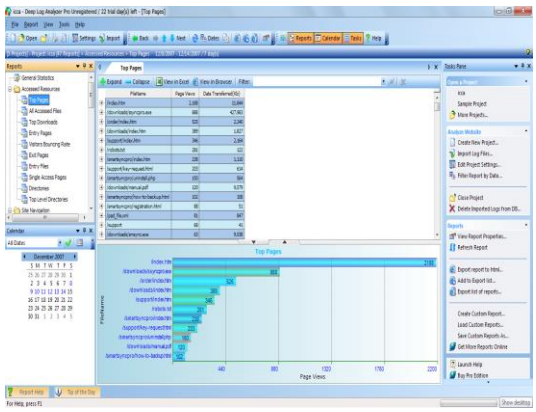

Figure 4. Deep Log Analyzer Tool

4.3. WebLog Expert

WebLog Expert is a fast and powerful access log analyzer [3]. It can give the information about web 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. The log analyzer can create reports in HTML, PDF and CSV formats. It also includes a web

server that supports dynamic HTML reports. WebLog Expert can analyze logs of Apache and IIS web servers. It can even read GZ and ZIP compressed log files so you won't need to unpack them manually. Built-in wizards can help you quickly and easily create a profile for your site and analyze it.

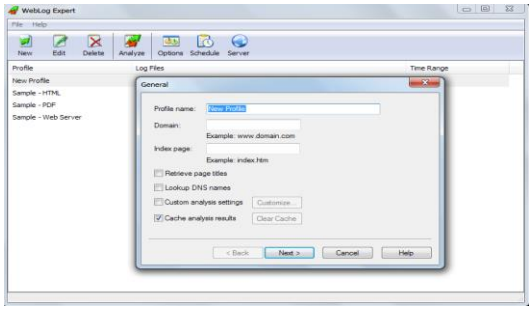

Figure 5. WebLog Expert Tool

4.4. SAWMILL

SAWMILL is a powerful, hierarchical log analysis tool that runs on every major platform [4]. It can handle all major log formats and many minor formats, and you can create your own custom formats. It does not generate static reports and it generates dynamic, interlined reports. It can provide enough flexibility to let you choose the model that works best for you. It can handle a log file with no limits, except those imposed by the limitations of your server.

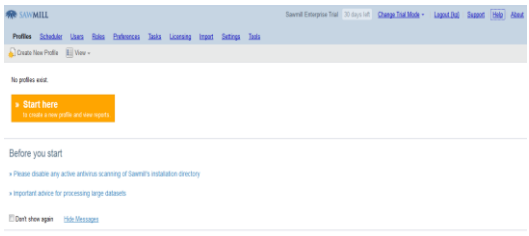

Figure 6. SAWMILL Log Analyzer Tool

5. Comparison of Web Log Analysis Tools

Table 1. Comparison among Tools

Features/ Software	AWSt-ats	Deep Log Analy-zer	WebL-og Expert	SAWML L
Version	7.3	6.0	8.6	8.7
Date	2014	2014	2014	2014
Platform	Perl	Window	Windo w	Window Linix BSD POSIX
Tracking Method	Web log files	Web log files	Web log files	Cookies via JavaScri-pt and Logs
Export statistics to PDF	Yes	No	Yes	Yes
License	Free GPL	Varies per license	Three availa-ble pricing plans	Three availabl-e pricing plans

This survey revealed a few salient points. AWStats can analyze a lot of log formats than the other three tools and it can also work from the command line and all web hosting providers which allow Perl, CGI and log access. Deep Log Analyzer and WebLog Expert can read GZ and ZIP compressed log files without unpacking them manually but Deep Log analyzer cannot export statistics into PDF format. SAWMILL provides extensive log processing and reporting

features to get the best possible insight into network data. Moreover, it can provide real-time reporting and real-time alerting.

6. Conclusion

Analyzing web log data will help to determine the browsing interest of the website users. A various kinds of tools are available which propose huge capabilities in preparing and reporting the results of analysis. These tools take input in terms of web log file, analyze it and generate results. In this paper, a survey was done and four types of web log analysis tools were studied and then the survey results were summarized. This survey could assist in determining the right solution for which web log analysis tool is more useful you.

References

[1] http://www.awstats.org (Last Accessed on December 2014)

[2] http://www.deep-software.com (Last Accessed on December 2014)

[3] http://www.weblogexpert.com (Last Accessed on December 2014)

[4] http://www.sawmill.net (Last Accessed on December 2014)

[5] C. Aggarwal, J.L.Wolf, K-L. Wu, and P.S. Yu, *The Intelligent Recommendation Analyzer*, ICDCS Workshop on Knowledge Discovery and Data Mining, April 2000.

[6] Fang yuan, Li-Juan wang and ge yu, *Study on data preprocessing algorithm in web log mining*, Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.

[7] J. Punin, M. Krishnamoorthy, and M. Zaki, *Web usage mining: Languages and algorithms*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2001.

[8] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, *Web usage mining: Discovery and applications of usage patterns from web data*, SIGKDD Explorations, Vol. 1, No. 2, 2000, pp. 12-23.

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

[10] Monika Yadav, Mr. Pradeep Mittal, *Web Mining: An Introduction*, IJARCSSE, March, 2013.

[11] Nakatani, K. and Chuang, T.-T., *A web analytics tool selection method: an analytical hierarchy process approach*, Internet Research, 21(2), 2011, pp.171–186.

[12] Pani, S.K. et al., *Web Usage Mining: A Survey on Pattern Extraction from Web Logs*, International Journal of Instrumentation, Control & Automation (IJICA), 1(1), 2011, pp.15–23.

[13] Pierrakos, D. et al., *Web Usage Mining as a Tool for Personalization: A Survey User Modeling and User-Adapted Interaction*, 13, 2003, pp.311–372.

[14] S.K. Pani, and et al., *Web Usage Mining: A Survey on Pattern Extraction from Web Logs*, International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.