DOI:10.21276/sjeat.2017.2.1.1

Saudi Journal of Engineering and Technology
Scholars Middle East Publishers
Dubai, United Arab Emirates
Website: http://scholarsmepub.com/

ISSN 2415-6272 (Print)
ISSN 2415-6264 (Online)

Research Article

A Brief Investigation on Web Usage Mining Tools (WUM)

Vinod Kumar[1], Ramjeevan Singh Thakur[1]
[1]Dept. Master of Computer Application, MANIT, Bhopal, Madhya Pradesh, India

***Corresponding Author:**
Vinod Kumar
Email: vinodkumarfbkp@gmail.com

Abstract: In the era of World Wide Web, more than one billions of websites are active over the internet. To perform the log analysis on huge number of available websites, although, numerous featured log analysis tools are existing. However, the great difficulty arises in selection of suitable tools. This work provides an investigation of open source and commercial toolsets available for the analysis the study will provide many choices to pick from when deciding a toolset to manage and analyze log data. The paper will help to review the set of tools currently available and positively hook the right tool to get started on analyzing logs in their organization.
Keywords: Web usage mining, Web log analysis,Web log Analyzer, Web Usage Mining Tools.

INTRODUCTION

Web has turned into the atmosphere where folks of all ages, tongues and cultures conduct their daily digital lives. Working or amusing, learning or hang out, home or on the way, discretely or as an assembly, Web users are ubiquitously encircled by an infrastructure of devices, networks and applications. This infrastructure combined with the perpetually growing amount of information supports the user's intellectual or physical activity. Whether searching, using or creating and disseminating the information, users leave behind a great deal of data disclosing their information requirements, attitudes, private and environmental facts. Web designers assemble these artifacts in a variety of Web logs [39, 41] for subsequent analysis. Hence, it is interesting and required to study the user actions and its result on analysis task one of the key tasks. Web log mining is to discover the hidden facts, such as user browsing habit and deliver it to the website managers or service providers for improving content organization. Today, Web log mining [32, 38] is being performed at its peak over World Wide Web. Web mining is categorized into three basic class- web content mining, web structure mining and web usage mining (WUM) [29, 36]. Here, web usage mining focuses on determining the user's behavior [48] from web log data. A variety of tools are available that can perform web usage mining taking web access logs as an input and generate the reports as an output. Thus, a log analysis tool [1] is defined as a piece of software that allows analytical processing of log files according to a user-specified instructions. The paper is organized in four major section where section 1 briefly introduces about the various popular tools for web log analysis, here only salient and steering features of each tools are included to help in selecting right tools. It also includes the quick comparison table of 20 web log analysis tools to assist in fast selection of tools. Section 2 describes the whole procedure to perform web usage mining using the web log analyzer. Section 3 presents a real analysis work over real web log data using the web log expert tool to present, how analysis is performed using the web analyzer. Section 4 concludes the overall work covered in the paper.

Some popular Web Usage Mining Tools (WUM)

Since, Web usage mining [42] has been an inevitable task in the World Wide Web era, whether we talk about individual or organization they have the web site as the prime mode of interface for making interaction (online Transaction, knowledge discovery etc.) from the users or customer. To gather the treasure of knowledge about the user's behavior [30, 31, 48] to make the fruitful decision for maximizing and expanding the organizations horizon in the current competitive world, there is a great need of some automated web log analyzing tool which can easily, effectively and efficiently analyses and demonstrate the user's statistics. This section provides the brief and key information about the some popular and widely used WUM tools [29].

Web log Storming

Web log storming [2] is developed by company Dataland Software as first release in August 2003, and the current proprietary version 3.2 is released in May 2016. It is widely used interactive and typical server log analyzer runs on windows platform. It supports IIS W3C Extended log file Apache Combined log file and Nginx file format. This makes it an ideal solution that gives users an insight about both, marketing and technical aspects of the web statistics. It retains log files compressed to saves disk storage space - it uncomporesses zip, gz and tar file automatically to use. It adopts techniques to cache previously read log files for faster analysis. It can interface to FTP or HTTP server and download updated log files automatically. Reports can be exported to HTML files and also may be e-mailed to destined email-id.

Google analytics

Google Analytics [3] is the most world widely used website statistics service. It is a free utility provided by Google since November, 14, 2005. It facilities the real time analytics. It helps to analyze visitor's traffic and provide a complete report about visitors and their requirements by tracing their path. It supports different file formats with unlimited size. It also supports mobile app analytics to assist the user effectively.

Web Log Expert

WebLog Expert [4, 44, 45] is a quick and prevailing web access log data analyzer runs on windows platform and available in four editions Standard, Professional, and Enterprise and Lite Editions. It gives information about website's visitors: activity statistics, accessed files, paths over the website, information about referring pages, search engines, browsers, operating systems, and more. Multithreaded DNS Lookup is the amazing feature resided in this software which assistances to extract domain name of the source IP addresses found in logs. It has IP addresses to country/state/city mapping built-in database. It is equipped with built in scheduler and command line interface. It can generate web site's statistics at very granular level in HTML, PDF, and CSV in tabular and graphical form.

Webalizer

Webalizer [5] is a fast command line operable open source developed in c language web log analysis tool which runs on Linux environment. It supports Common logfile format, server logs, wu-ftpd/proftpd xferlog (FTP) format logs, Squid proxy server native format, and W3C Extended log formats, several variations of the NCSA Combined logfile format. It can directly execute the compressed gzip (.gz) and bzip2 (.bz2) files without uncompressing. It has in built geo-location services. It can be configured to scheduled analysis and automatic report generation. It generates very in depth, easily configurable usage reports in HTML format can be viewed with any standard web browser in different languages.

PIWIK

PIWIK [6] is the fastest open source log analysis tool released in June, 2007. It is compatible with Windows, MacOS, Linux, Solaris environment. Apart from the web analysis, Piwik has a set of plug-in to enhance the reporting formats. It has own interface using python to get the reports. It is very flashy with an Ajax or Web 2.0 feel to it. One of the finest features is user can develop own widgets to monitor whatever data want to track. It is also available as Cloud-hosted Piwik i.e. Piwik PRO Cloud without own technical setup and analytics will be hosted on reliable and secure servers, while still giving full ownership of data.

Open Web Analytics

Open Web Analytics (OWA) [7] is an open source web analytic software and licensed under GPL. It is built on technology PHP, MySQL, JavaScript. It is proficient of processing really huge logs and can optionally fetch those directly from a database format too. Contrasting many other professional tools, OWA can provide a click-stream report. This aids website code troubleshooter, to recognize exactly what the website visitor performed, and can go repeating those steps to reproduce the problem. It can also create heat map type of report whereby the website statistics is separated into most-hit and least-hit pages. It also facilitates the Content Management System (CMS) integration like Drupal, Mediawiki, and WordPress.

AWStas

AWStats [8] is a free powerful and featureful tool, available absolutely free with sources (GNU General Public License) that generates advanced web, streaming, FTP or mail server statistics, graphically. This log analyzer works both as a browser CGI and from command line interface (CLI) and demonstrations all possible facts that log contains, in few graphical web pages. It can analyze the IIS log files (W3C), Apache NCSA combined log files (XLF/ELF) or common (CLF), WebStar native log files and other web, proxy, WAP or streaming servers log files. It uses a partial information file to be able to process large log files, often and quickly.

W3Perl

W3Perl [9] is a CGI based free open source web analytics software tool, distributed under the GPL and it can be installed on platforms Linux, Mac and Windows. It offers the ability to make use of a page bug to track page data without looking at log files or the ability to

read the log files and report across them. It can parse WWW / FTP / Squid / CUPS / DHCP / SSH and Mail log files. It also helps admin to manage and control remotely. It display statistics form hours to years, hosts/pages to pages/hosts. It is known for producing user's activity at granular level and generates reports in html as well in pdf and also can email to target email address.

Visitors

Visitors [10] is a command line free log analysis tool distributed under the terms of the GPL license. It can run on variety of platforms – Windows, UNIX / LINUX and its various flavors. It can generate both HTML and text reports by simply running the tool over log file. To specify the log format is not required at all. It works out of box with apache and most other web servers with a standard log. One interesting feature is the real time streaming data can set up. It requires no installation, can process up to 150,000 lines of log entries per second in fast computer. It support for real time statistics with the visitors Stream Mode introduced with version 0.3.The current version is 0.7.

RealTracker

RealTracker [11] uses a code that is placed on Web pages to track your pages, analogous to Google Analytics. It provides a bunch of different reports but the real benefit to tool is that it's easy to add to pages and easy to read the results. If more features are needed, one can switch to the professional or enterprise versions of the tool.

Analog

Analog [12, 13] is a very old and widely used free Web log analysis tool. It was originally came in market on June 21, 1995, by Stephen Turner as generic freeware; the license was transformed to the GNU General Public License in November 2004. It works on any Web server and can process various kinds of web log files, and also it is quite easy to install and run on Windows, Mac OS, Linux, and most Unix-like operating systems. It has provision for 35 languages, and offers the ability to perform reverse DNS lookups on log files, to point out where web site hits generates. It can generates various statistics of user's activity at granular level.

Dailystats

Dailystats[14] is freely distributed under the GNU General Public License. Web analysis program that is not aimed to be complete analytics package. Instead, Dailystats wants to give a small sub-set of statistics that are useful for reviewing on a regular basis - such as daily. It provides information on entry pages, page views of each page, and referrer log analysis. It offers to generate monthly/daily reports, traffic breakdown per document, referrer log analysis, gateway analysis, and multiple logs - Even if your site is being delivered by multiple web servers the program can keep track of multiple logs and merge the data appropriately. It is fast and logs at speeds close to 3000 lines per second on a Pentium II-class machine.

Relax

Relax[15, 14] is an open source web analytics tool which runs on GNU Linux with limited features that tells just who is referring people to the website. It looks at search engines and search key words as well as specific referral URLs to give the precise information on who is directing customers towards website. It's not a complete analytics package, but it works healthy for referral information. Produced HTML reports can be set up to include links to other web-based keyword analysis tools, making it at ease to further mend the ranking of pages in search engines.

StatCounter

Stat Counter [15, 16] is a Web analytics tool that uses a small script that you place on each page. It can also go as a counter and show the count right on page. The free version only counts the last 100 visitors, then it resets and starts the count over again. But within that limitation, it provides a lot of features like- Invisible and Configurable Counter Option, Drill Down, Configurable Summary Stats, Magnify User, Recent Keyword Activity, Search Engine Wars, Popular Pages, Entry Pages, Exit Pages, Visitor Paths, Visit Length Returning Visits, Recent Page load Activity, Recent Visitor Activity, ISP Stats, Browser Stats, O.S. Stats, Resolution Stats, Email Reports, User Access Management, Country/State/City Stats Recent Visitor Google Map, Public Stats, Blocking Cookie, HTTPS Tracking, Multiple Site Management. Thus, it is a packed full of advantageous and great tools to aid in making better judgments about the website.

MyBlogLog

MyBlogLog [17] is a tool with many different features. The analytics are not very healthy, but they are not envisioned to be. Actually, the objective of the MyBlogLog analytics is to provide you with evidence about where your visitors are going when they leave your site. This can help you to mend the website so they don't leave as quickly. It is not recommended MyBlogLog as your only analytics tool, but it does a good job on the stats it provides.

Webtrax

Webtrax [18] is a free Web analytics tool that is very customizable, but not as good programmed as it could be. This tool is developed in Perl and is therefore portable to many platforms. It provides all the general statistics of user's activity using the web log file. In brief, it support a number of reports and provides good

information from log files. It works best on logs that include the "referrer".

SiteMeter

The free version of SiteMeter [19] offers a variety of good statistics and reviews in your web site. Web page Meter makes dynamic 3D charts displaying viewers, page views, visit periods, nation maps, and many more! It simplest supplies knowledge on the first a hundred viewers, after which after that it resets and starts over. But if wanted more knowledge than that, it can be upgraded to the paid version of SiteMeter. Like different non-hosted analytics instruments, SiteMeter works by using inserting a script on each page of the website. This gives the actual-time visitors.

Sawmill

It is a powerful web log statistical analysis software package [20] written in C language and developed by Flowerfire Inc. in 1998. It support multi-platform windows/Unix/Linux based operating system. It provides dynamic, customizable user interface with real time, contextual filtering reports. Moreover, it also offers support for almost every server log file formats, with new formats added on request. This also contains a page tagging server and JavaScript page tag for the analysis of client side client's click requests providing a total view of visitor traffic and on-site behavioral activity. It facilitates the user in three modes - as a software package for user deployment, as a turnkey on-premises system appliance, and as a SaaS. It can analyze any device or software package creating a log file and that may be proxy servers, firewalls, mail servers, web servers, syslog servers, databases, networking devices and so forth.

GoAccess

This visual web log analyzer [21] is an open source real-time and interactive viewer written in C language that runs in a terminal in *nix systems or via web browser. It offers fast and valuable HTTP statistics for system administrators that need a visual server report on the run. GoAccess was aimed to be a prompt, terminal-based web log analyzer. Its main idea is to speedily analyze and visualize web server statistics in real time without requiring to use web browser great. While the terminal output is the default output, this analyzer is capable to analyze almost all web log formats (Apache, Amazon S3, Elastic Load Balancing, Nginx, CloudFront, etc.), it has the ability to produce a complete real-time HTML report, as well as a JSON, and CSV report.

Nihuo Web Log Analyzer

Nihuo Web Log Analyzer [22] is a fast and good log analyzer software tool for small and average size websites. It can run on various platforms - Windows, Linux, MacOS and FreeBSD. Nihuo Web Log Analyzer can analyze logs produced by Apache, Ngnix, lighttpd and IIS web servers. It can even read BZIP, ZIP, GZIP, and ZIP64 compressed log files so no need to decompress them. Nihuo Web Log Analyzer is extremely configurable. With Nihuo Web Log Analyzer one can create one's own custom reports or tailor standard reports to meet specific needs.

HTTP-Analyzer

HTTP-Analyzer [23] is a very old log analyzer which may run on different platforms -Windows/UNIX /Linux /MacOS. It has easy navigation throughout the statistics report with an intuitive-to-use interface. It contains the 3Dstats log file analyzer, which creates a 3D view of hits by month and hits by hour/weekday. Computes Hits, File count, cached file count, page views (Textual pages), sessions (unique hosts per 24 hours) and data sent in KB. It has capability to produce summaries up to 3 levels of detail. It also partakes fully customizable layout of the statistics report and able to analyze various log formats like- NCSA Common and common Log Format, W3C Extended Log Format. Due to the massive demand of service providers to be capable to practice the analyzer with altered statistics reports and an individual guide to the statistics report in native language.

Deep Log Analyzer

Deep Log Analyzer [49] is an advanced user friendly and affordable web analytics software solution for small and medium size websites. This software helps to analyze web site users' behavior and provide the complete website usage statistics in many ways and easy step. Contrasting the other tools, it equips with extensible capability to analyze various kinds of logs including FTP logs. It can create a list of keywords and the hits on web pages that holds keywords. It is very valuable for search engine optimization (SEO). With Deep Log analyzer website statistics and web analytics software helps to know exactly where website's visitors come from and how they move through the website i.e. their location and the access path in the website.

General Features

General features that almost every web log analysis tools includes whether it is proprietary or open source some common feature are mentioned here as the general information that even a simple typical web log analyzer tool will include general Information like - Hits Detail, Visits Detail, Page Views Detail, Referral Detail, Search Engines Detail, Technical Detail, Web Traffic/Load Detail. Figure 1 shows the important common points under which the analysis is done and almost every web log analyzer possesses it. For the sake of convenience to understand, the statistics that can be

obtained from the log file using the log analyzer is categorized as five major points and sub points. These points are concisely described here –

Search Engines

This report shows a list of search engines used by the visitors to find website ranked by the number of referrals (Number of Hits column) from each search engine. It displays the information under the heads- Popular Search Engines, Popular Search Pages, Key Phrases and Keywords, Web Compression Statistics. It Benefits in optimizing the website for search engines.

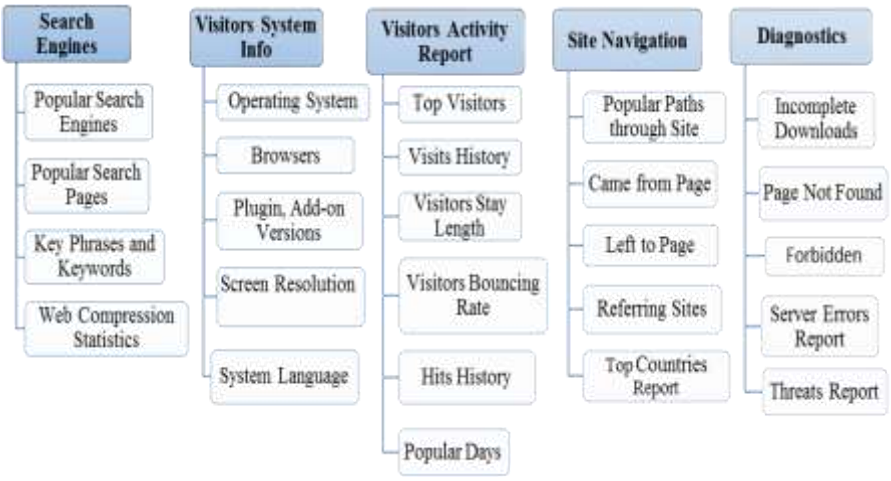

Fig-1: General features of a web log analyzer

Visitors System Info [49]

This portions collects the information related to visitors system such as – operating system, Browser used, presence of plugins and add-on versions, JavaScript Support and System language in use, it also reports screen resolution of computer used by the visitors.

Visitors Activity Report

This is very important and most focused part of any web log analyzer, here, effort is made to record the user's activity and impact of visitor's hits on website and web server at granular level. It record the visitor's information under headings -Top Visitors, Visits History, Visitors Stay Length, Visitors Bouncing Rate, and Hits History, Popular Days etc.

Site Navigation

This part is responsible for reporting the information about the significant locations of the resources within the website. It records the visitor's information under labels - Popular Paths through Site, Came from Page, Left to Page, Referring Sites, Top Countries Report. It benefits in determining the user's behavior on one's website. It helps to effective designing of website web pages.

Diagnostics [49]

This report shows the different errors occurring on your website. It records the visitor's information under the points -Incomplete Downloads, Page Not Found, Forbidden, Server Errors, report Threats Report. A large number of errors can decrease website's usability and reputation. It assists in diagnosing and fixing the errors occurring on the website.

Summarized Comparison of various web log Analyzers

It is quite obvious from the study that large set of web log analyzer software tools are available in the market for performing the analysis from web log to get the fruitful information in simple and effective way. Among the large set of tools available today, it is a tedious and time consuming task to find the suitable tool to perform the analysis economically, efficiently and effectively. The summarized and comparative[27][28] representation of study in Table 1, positively discovers the right tool for the job to help the reader get started on analyzing logs in their organization or in individual. Although, there are a great number of features available in each web analyzer tools, however, some of the significant and steering features are included to perform comparison.

Table-1: Summarized and comparative demonstration of various WUM Tools

Serial No.	Name of Web Analysis Tool	Summarized Features															
		Vendor	Current version	Log file Formats	Website linkage	License	Language	Database	User interface	Platform	Report format	Availability of Dynamic Reports	E-Mail facility	Report Scheduler	Ability to handle compressed files	Real Time Analysis	Mobile tracking	Source URL
1	Google Analytics	Google Inc.	Single	CLF,XLF, ELF	Yes	Free	JavaScript	In-Built	GUI	Windows,Mac, Linux, Solaris	HTML,PDF,CSV	Yes	Yes	Exterior	No	Yes	Yes	http://www.google.com/analytics
2	Deep Log analyzer	Deep Software Inc.	7.0	Apache, IIS, CLF,XLF,ELF	Imported	Proprietary	In-Built	MS Access	GUI	Windows	HTML /Ms-Excel	Yes	No	In-Built	YES	Yes	No	http://www.deep-oftware.com/
3	Web Log Expert	Alentum Software	9.3	Apache, IIS	Imported	Proprietary	In-Built	In-Built	GUI	Windows	HTML,PDF,CSV	Yes	Yes	In-Built	YES	No	No	http://www.weblogexpert.com/
4	Visitors	Salvatore Sanfilippo	0.7	CLF,Apache,IIS,W3C	Imported	GNU GPL	C	In-Built	CLI	Unix/Linux/Mac	HTML	No	No	Exterior	No	Yes	No	http://www.hping.org/visitors/
5	Webalizer	Webalizer	2.23 - 08	CLF,XLF,ELF, FTP	Imported	GNU GPL	C	GeoDB	CLI	Unix/Linux Mac/	HTML	No	No	Exterior	Yes	No	No	http://www.webalizer.org/
6	Analog	Community Development	6.0.	CLF,IIS W3C	Imported	GNU-GPL	C	Logfile-based	CLI	Windows/Unix/ Linux/Mac	HTML	No	No	Exterior	Yes	No	No	http://analog.gsp.com/
	Piwik	Piwik Inc.	2.16.5	Apache, IIS, Ngnix	Imported	GNU-GPL	Python/ Ruby	MySQL,WAMP	GUI	Windows/Mac/ Linux/Solaris	HTML/PDF	Yes	Yes	Exterior	No	Yes	Yes	http://www.piwik.org/
8	Open Web	Open web	1.5.7	CLF,XLF,ELF	Imported	GNU-GPL	PHP	MYSQL	GUI	Windows	HTML	No	No	Exterior	No	No	No	http://www.openwebanalytics.com

	Analytics	analytics																
9	AWStats	AW Stats Inc.	7.5	CLF,XLF,ELF, W3C,etc.	Imported	GNU GPL	Perl	Logfile-based	GUI	Windows	HTML/PDF	Yes	Yes	Exterior	No	Yes	No	http://www.awstats.org/
10	Weblog Storming	Dataland Software	3.2	IIS W3C Extended log file format Apache	Yes	Proprietary	C	In-Built	GUI	Windows	HTML/PDF	Yes	Yes	Exterior	Yes	Yes	No	http://www.weblogstorming.com
11	Webtrax	John Callender	23	NCSA Combined Format	Imported	GNU-GPL	Perl	In-Built	CLI	Windows/Unix/ Linux/Mac	HTML	No	No	Exterior	No	No	No	http://multicians.org/thvv/webtrax-help.html
12	Dailystats	Perlfect Solutions	3.0	CLF,XLF,ELF, W3C,etc.	Imported	GNU - GPL	Perl	In-Built	CLI	Windows/Unix /Linux/Mac/	HTML	No	No	Exterior	No	No	No	http://www.perlfect.com/freescripts/dailystats/
13	Relax	Free Software Foundation,Inc.	2.80	Apache combined, NCSA extended/CTL, WebSTAR	Imported	GNU - GPL	Perl	In-Built	CLI	Linux	HTML/CSV/Text	No	No	Exterior	No	No	No	http://ktmatu.com/software/relax/
14	StatCounter	StatCounter	3	Embedded in web page	In-Built	Proprietary	-	In-Built	GUI	Windows	CSV	YES	No	In-Built	No	Yes	Yes	https://statcounter.com/
15	SAWMILL	Flowerfire Inc.	8.7.8	IISW3C,ELF format, Apache	In-Built	Proprietary	C	MS-SQL	GUI	Windows/Unix/ Linux/Mac	HTML	YES	Yes	In-Built	No	Yes	Yes	http://sawmill.net/
16	GoAccess	MIT Licensed	1.0	Apache, ginx, CLF CloudFront, ELF	In-Built	Open Source	C	In-Built	GUI-CLI	Unix/Linux/Mac	HTML, CSV,JSON	YES	No	Exterior	No	Yes	No	https://goaccess.io/
17	Nihuo Log Analyzer	Nihuo Software Inc	4.19	Apache, Zeus, Lighttpd/ NCSA, IIS 4/5/6/7 logs	Impported	Proprietary	JavaScript, HTML	In-Built	GUI - CLI	Windows /Linux /Mac OS/ FreeBSD	HTML	YES	No	In-Built	Yes	Yes	No	http://www.loganalyzer.net
18	HTTP-ANALYZE	RENT-A-GURU, Inc	2.4	NCSA CLF, W3C ELF	In-built	Proprietary	C	In-Built	GUI - CLI	Windows/Linux/ Unix Mac OS/	HTML	YES	No	Exterior	No	Yes	No	http://http-analyze.org/index.php
19	Log Analytics Sense	Statspire Software	2.3	More than 40 formats	imported	Proprietary	-	In-Built	GUI - CLI	Windows	HTML	YES	No	Exterior	Yes	No	No	http://www.statspire.com/
20	AlterWind Log Analyzer	Alterwind Software	4.0	ApacheCommon, Combined, and IIS log file formats	Imported	Proprietary	-	SQL	GUI - CLI	Windows	HTML	YES		No	Yes	No	No	http://www.alterwind.com/

Abbreviations that are used – GUI-Graphical user interface, CLI-Command Line Interface, CLF-Common log format, ELF-Extended log File.

METHODOLOGY FOR WEB LOG ANALYSIS

This section focuses and describes entire Methodology for web log analysis [43] how the web log analysis function is performed. The web log analysis in the lack of analyzer is quite complex and time consuming and difficult task. Now, the analysis work is simplified by using the automated software tools shows in table 1. Fig -2 presents the simplified basic stages which is adopted in web log analysis process. The stages are detailed below-

Log Data Collection

First step in the analysis process is the collection of the log data [40] from the sources, the source may be your computer system, proxy server, web server etc. The log data exists in the various sources as Web server log, application software log, System log. Further, the prepressing of the log file is done. Because, in the original database file extracted, not all the information are valid for web usage [47] mining, we only need entries that contain relevant information. The original file is generally prepared of text files that has large size of information concerning queries fired to the web server in which in most cases contains irrelevant, extraneous, partial and misleading information for web mining purpose.

Preprocessing phase

This is the preliminary task in the web log analysis task. The preprocessing of log encompasses the following subtasks –

Data cleaning- During data cleaning, Irrelevant and extraneous data are eliminated. Since, the major aim of web usage mining is to find the navigational pattern, the given record of following type must be removed.
The records of graphics, video and format information,The records with failed HTTP status code.

User Identification

The task user identification [35] is find the different user who has visited the web site.

- The different IP addresses distinguish different users.
- If the IP addresses are same, the different browsers and operating systems indicate different users which can be obtained by client IP address and user agent.
- If all three things IP address, Browser and Operating Systems are same, the referrer information should be considered for user identification.

Session Identification

According to [25] [26] a session can be described as a sequence of activities carried out by a user between the entry and exit from the website. Moreover, one session can be made up of two clicks, if the time interval between them is less than a specific period.

Path Completion

The Local caching and proxy servers poses the problems for path completion [34] [35] because users may reach the pages in the local caching or the proxy servers caching devoid of leaving any trace in server's access log. As a result, the user access paths are incompletely stored in the web access log. The purpose of the path completion is to accomplish task by appending the discovered user's travel pattern, the missing pages in the user access path.

Pattern Discovery

Pattern discovery [35] is the crucial course of action in web usage mining which comprises grouping of users centered on similarities in their profile and search behavior. There are diverse web usage data mining methods and algorithms that can be embraced for pattern discovery, which includes, path analysis, clustering, and associate rule.

Pattern Analysis

Pattern analysis [37] is the final phase in web usage mining which is aimed at mining interesting rules, pattern or statistics from the result of pattern discovery step, by discarding unrelated rules or statistics. The pattern analysis phase offers the tool for the transformation of information into knowledge.

Finally, the result obtained from pattern analysis is the main crux of the web log analysis task. Thus report generated from the web log analyzer may be saved as the HTML, CSV, and PDF etc. Some of the tool provide the facility to configure the target email address in the tool to automatically email the analysis report at particular scheduled time. One can see the tools which provide the automatic email facility from Table 1.

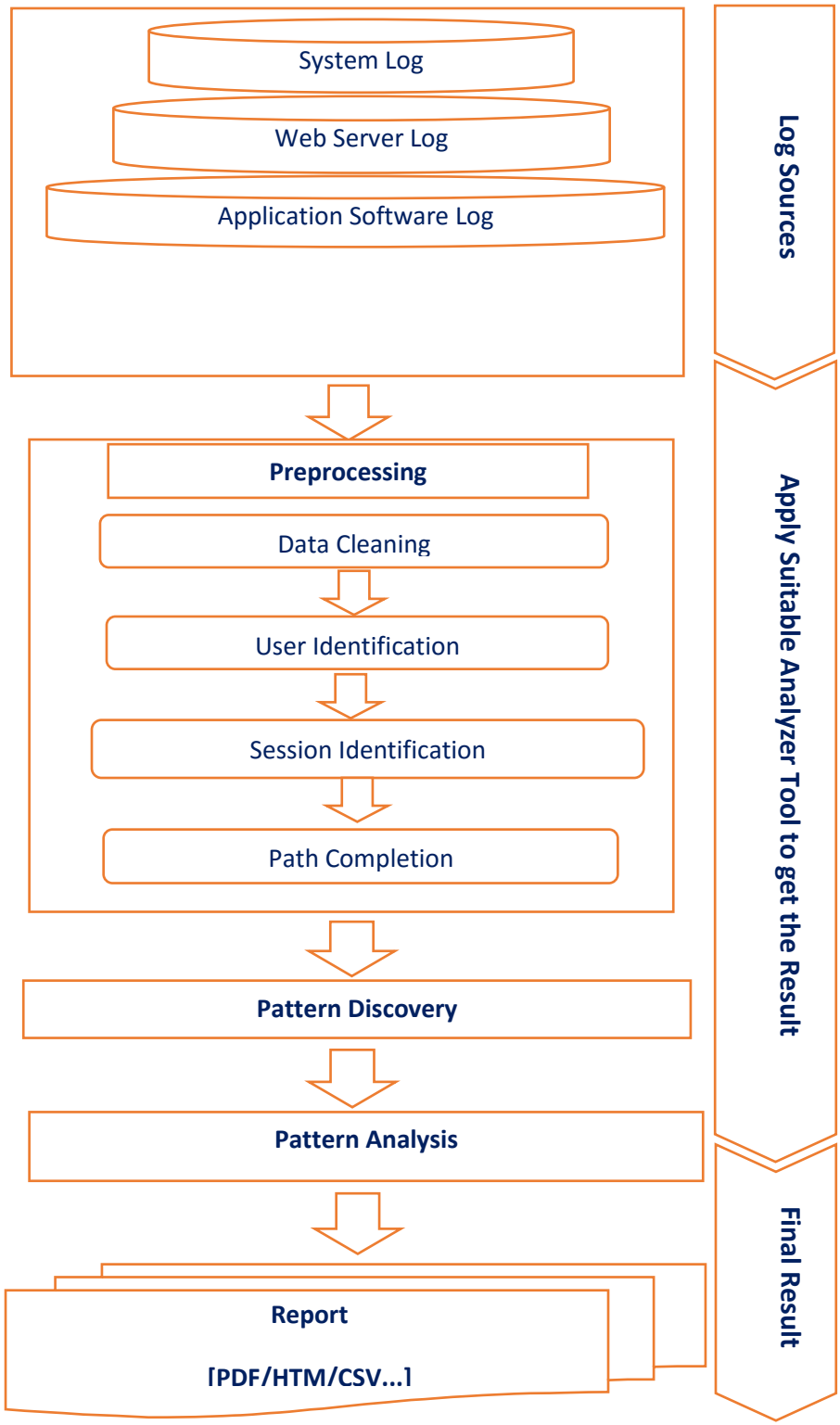

Fig-2: Methodology for web log analysis

CONCLUSION

Analysis of log file supports in determining the navigational pattern of the visitors and his surfing behavior. There is range of tools available online for this type of analysis and producing the reports, some of which are open source and proprietary. A comparative

investigation has be carried out on the widely used analyzing tools. These tools offer a variety of several features which are superior to the other. If someone wish to use proprietary software, the deep web log analyzer, and web log Expert analyzer are very powerful and useful giving the almost similar result. In case of open source, Piwik is very strongly helpful for web log analysis. For online and real time web log analysis, google analytics and statcounter are very useful and powerful web log analyzer. These tools helps in taking informed decisions to incorporate and improve the website as per user requirement. Using the reports and results generated by the tool of the visitors visiting a website one can get a rational and true idea about the behavior of the visitors and their navigational paths and patterns which support one's in determining the influence and popularity of the website.

ACKNOWLEDGEMENT

This work was supported by the Dept. Computer Application, Maulana Azad National Institue of Technology, Bhopal, Madhya Pradesh, India

REFERENCES

1. Jan Valdman, "Log File Analysis", Technical Report No. DCSE/TR-2001-04, University of West Bohemia in Pilsen, July 2001.
2. URL- http://www.datalandsoftware.com/files/WLS_User Guide.pdf (Accessed on 20.8.2016)
3. Google Analytics, URL - http://www.google.com/analytics/ (Accessed on 25.8.2016)
4. Web log expert, URL - https://www.weblogexpert.com/features.htm (Accessed on 25.09.2016)
5. Webilizer, URL- http://www.webalizer.org/webalizer.1.html (Accessed on 28.8.2016)
6. Piwik, URL-http://www.piwik.org/ (Accessed on 05.9.2016)
7. Open Web Analytics, URL- http://www.openwebanalytics.com/ (Accessed on 15.9.2016)
8. Awstats, URL- http://www.awstats.org/ (Accessed on 05.10.2016)
9. W3CPerl, URL-http://www.w3perl.com/(Accessed on 08.10.2016)
10. Visitors web log analyzer, URL- http://www.hping.org/visitors/(Accessed on 15.10.2016)
11. Realtracker Personal, URL- http://www.realtracker.com/website-analytics.asp(Accessed on 15.11.2016)
12. Analog web log analyzer, URL- http://analog.gsp.com/(Accessed on 15.10.2016)
13. Gasson, Gaelyne R. (April 2000). "Web Analysis Using Analog". Linux Journal. 2000 (72es).ISSN 1075-3583.
14. Dailystats Log Analyzer URL- http://www.perlfect.com/freescripts/dailystats/ (Accessed on 16.10.2016)
15. Relax log analyzer, URL- https://directory.fsf.org/wiki/Relax_log_analyzer (Accessed on 18.10.2016)
16. Relax Web Log Analyzer URL- http://ktmatu.com/software/relax/install.html (Accessed on 20.10.2016)
17. MyBlogLog, URL- http://www.webmin.com/ (Accessed on 15.11.2016)
18. Webtrax log Analyzer, URL- http://multicians.org/thvv/webtrax-help.html, WEB TRAXS USER GUIDE version 4.0 (Accessed on 10.11.2016)
19. Sitemeter log analyzer, URL- http://www.sitemeter.com/?a=home(Accessed on 15.11.2016)
20. SAWMILL Web Log Analyzer, URL- http://sawmill.net/index.html (Accessed on 15.11.2016)
21. Goaccess URL-https://goaccess.io/ (Accessed on 15.11.2016)
22. Nihuo Web Log Analyzer, URL- http://www.loganalyzer.net (Accessed on 20.11.2016)
23. HTTP-Analyzer, URL-http://http-analyze.org/index.php (Accessed on 15.11.2016)
24. Web Log Data (1995)," http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html". (Accessed on 20.11.2016)
25. Z. Xuejuu, E. John, H. Jenny, "Personalized online sales using web usage data mining", J. Comput. Ind. Vol.58 pp. 772–782, 2007.
26. L. Habin, K. Vlado, "Combining mining of web server logs and web content for classifying user's navigation pattern and predicting users future request", J. Data Knowledge Eng. 61, pp. 304–330, 2007
27. S. Bhuvaneswari, T. Anand, "A Comparative Study of Different Log Analyzer Tools to Analyze User Behaviors", International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169 Vol. 3(5), pp. 2997-3002, 2015.
28. N. Lakshmi, R. S. Rao, S. S. Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol. 2(4), March 2013.
29. N. Kaur, H. Aggarwal, "A Comparative Study of WUM Tools to analyze User Behaviour Pattern from Web Log Data", International Journal of Advances in Engineering Research (IJAER),

Vol.10 (4), ISSN: 2231-5152/ 2454-1796, December, 2015
30. Neha Goel and C.K. Jha (2013) "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool", International Journal of Computer Applications , Volume 62– No.2, January.
31. B. Bakariya, G. Thakur, "User Behavior Analysis from Web Log using Log Analyzer Tool", IJCSNS, November 2013-Vol. 1.No3 ISSN 2345-3397.
32. Etzioni, O., "The World Wide Web: Quagmine or gold mine. Communications of the ACM, vol. 39, pp. 65–68, 1996.
33. L. K. Joshila Grace, V. Maheswari, Dhinaharan Na gamlai, "Web Log Data Analysis and Mining", Advanced Computing Communications in Computer and Information Science, Vol. 133 pp. 459-469.
34. Y. LI, B. FENG, Q. MAO, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, 2008.
35. [D.A. Adeniyi, Z. Wei, Y. Yongquan ,"Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics, vol. 12, pp. 90–108, 2016]
36. Mobasher, B. "Web usage mining and personalization". CRC Press, LLC, 2005.
37. Liu, B. "Web data mining: Exploring hyperlinks, contents, and usage data", datacentric systems and applications, Springer-Verlag, 2006.
38. Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web." In Tools with Artificial Intelligence, Proceedings. Ninth IEEE International Conference, pp. 558-567. IEEE, 1997.
39. Facca, Federico Michele, and Pier Luca Lanzi. "Mining interesting knowledge from weblogs: a survey.", Data & Knowledge Engineering 53, no. 3 (2005): 225-241.
40. F. Bounch, F. Giannotti, C. Gozzi, G. Manco, M. Nanni, D. Pedreschi, C. Renso, S. Ruggier, "Web log data warehourseing and mining for intelligent web caching", J. Data Knowledge Eng. 36 165–189, 2001.PH:S0169-023x(01)00038-6.
41. M.F. Federico, L.L. Pier, "Mining interesting knowledge from weblog: a survey", J. Data Knowledge Eng. 53 pp.225–241, 2005.
42. C.J. Carmona, et al. Web usage mining to improve the design of an e-commerce website: OrOliveSur.com, Expert Systems with Applications, vol. 39 pp. 11243–11249, 2012.
43. R. Iváncsy et al., "Frequent Pattern Mining in Web Log Data", Acta Polytechnica Hungarica, vol. 3, No. 1, 2006.
44. V. Jaya kumar, Dr. K. Alagarsamy, "Analyzing server log file using web log expert in web, data mining" ,International Journal of Science, Environment and Technology, ISSN 2278-3687(O)Vol.2(5), pp.1008–1016,2013.
45. Arti Tyagi, Sunita Choudhary, "Web Usage Mining Using Web Log Expert Tool", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Vol. 5(3), March 2015.
46. T. Wongsirichot ,S. Sukpisit, W. Hanghu,"A Preliminary Analysis of Web Usage Behaviors from Web Access Log Files", Proceedings of International Conference on Soft Computing Techniques and Engineering Application, vol. 250, pp. 325-332, December, 2013.
47. M. Dhandi and R. K. Chakrawarti, "A comprehensive study of web usage mining," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp. 1-5. doi: 10.1109/CDAN.2016.7570889.
48. V. Anitha and P. Isakki, "A survey on predicting user behavior based on web server log files in a web usage mining," 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, India, 2016, pp. 1-4.
49. Vinod Kumar R S Thakur, "Exploring Behavior of Visitors Activity at Granular Level from Web Log Data using Deep Log Analyzer", International Journal of System and Software Engineering Vol. 4(1), pp. 16-26, Oct, 2016.