Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.900 – 907

RESEARCH ARTICLE

A Study of Web Traffic Analysis

Mr. Pratik V. Pande[1], Mr. N.M. Tarbani[2], Mr. Pavan V. Ingalkar[3]

[1]Department of Computer Science & Engineering, Prof. Ram Meghe College of Engineering, Amravati
[2]Department of Computer Science & Engineering, Prof. Ram Meghe College of Engineering, Amravati
[3]Department of Computer Science & Engineering, Prof. Ram Meghe College of Engineering, Amravati
[1] pratik.pande01@gmail.com, [2] niteshtarbani@gmail.com, [3] pvningalkar@gmail.com

Abstract--With the rapid increasing popularity of the WWW, Websites are playing a crucial role to convey knowledge and information to the end users. Discovering hidden and meaningful information about web user's usage patterns is critical to determine effective marketing strategies to optimize the Web server usage for accommodating future growth. Most of the currently available Web server traffic analysis tools explicitly provide statistical information. The web server traffic analysis tools make the 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. The task of web traffic analysis tools becomes more challenging when the web traffic volume is enormous and keeps on growing. In this paper, we propose a various model to discover and analyze useful knowledge from the available Web log data and also provides a comparative study of variety of Log Analyzer tools exist which helps in analyzing the traffic on web server.

Keywords- Web traffic, Logs, Web server log analyzers, Netflow, Hybrid neuro-fuzzy system, LQMs, TDSs

I. INTRODUCTION

The popularity of the World-Wide Web [1], [2] (also called WWW, or the Web) has increased dramatically in the past few years. Data on the web is rapidly increasing day by day. Web is an open medium .Today, WWW traffic is one of the dominating components of Internet traffic. There are many reasons behind this explosive growth in Web traffic. These reasons include: the ease of use of the Web, the availability of graphical user interfaces for navigating the Web, the availability of editors and support tools for creating and "publishing" Web documents, the machine-independent nature of the languages and protocols used for constructing and exchanging Web documents and a continuing exponential increase in the number of Internet hosts and users.

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 process of analyzing server performance begins with the collection of log files spanning some analysis time period. To understand traffic trends there is need to collect logs over a certain period. The phenomenal growth in Web traffic has sparked much research activity on "improving" the World-Wide Web. Much of this recent research activity has been aimed at improving Web performance. The key performance factors to consider are how to reduce the volume of network traffic produced by Web clients and servers, and how to improve the response time for WWW users.

A. *Web Traffic*

The web traffic starts with the high-level activities such as clicking a link and increases with low-level activities such as travelling through network switches and cables. In other words, Web traffic is usually initiated by users via the use of web browsers. It begins with a click to access a URL. Traffic flow starts with a mouse click, which sends browser information to a server that uses predetermined rules and methods to obtain user browser requests. Based on these rules, the server then decides what action is needed. Now a days, the web traffic is increases enormously because there is continuous increase of Internet users each year is motivating online shop, gambling site, and botnet owners to take control of users' moves to point them to their sites. Therefore, there is need of web traffic analysis tools. These tools handle and categorize the traffic and increase the workload handling capacity of the web server.

B. LOGS

Web server logs stores click stream data which can be useful for web traffic analysis purposes [3]. 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 [4]. 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 [5]. If logs are utilized properly, it can be very useful in turning the websites visitors into customers especially in case of an e-commerce website. It guides the analyst in determining the navigational pattern of the user i.e. which pages are frequently visited by the user, the kind of errors that user gets, etc.

A variety of tools are available that take the 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:

1) *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 .For using this tool, installation is not required, only requires a Google account. Email report facility is available in Google analytics.
2) *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.
3) *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. For using this tool, there is no need of creating any account but profile creation is required.
4) *Analog***:** This is 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.

II. WEB SERVER LOG ANALYZER

Automating the analysis of server logs is essential to allow using the logs as a proactive administrator tool. Log analyzer is 3-tier architecture based software that parses through the log files generated by whichever web server follows the standard web server logging standards. Analyze parsed data and categorize them into meaningful reports that can be read by the user for administration or monitoring purpose. A software application designed to parse a log file, which typically contains raw collected data, and convert it to an easy-to-read and understand form**.**

These log files can be very large and are very detailed on which files were requested from our web server. Every time a page, or image, or movie, or any other kind of file is downloaded off of your web server, the date/time and IP address of the requestor is logged in the web server's log file. Web logs can provide information about which user groups access the website, which resources are viewed the most, and the links users follow to arrive at the site. The web server log analyzer produces reports with summary statistics about website. It uses the information in server's log files and also helps to handle the large amount of traffic on web servers.

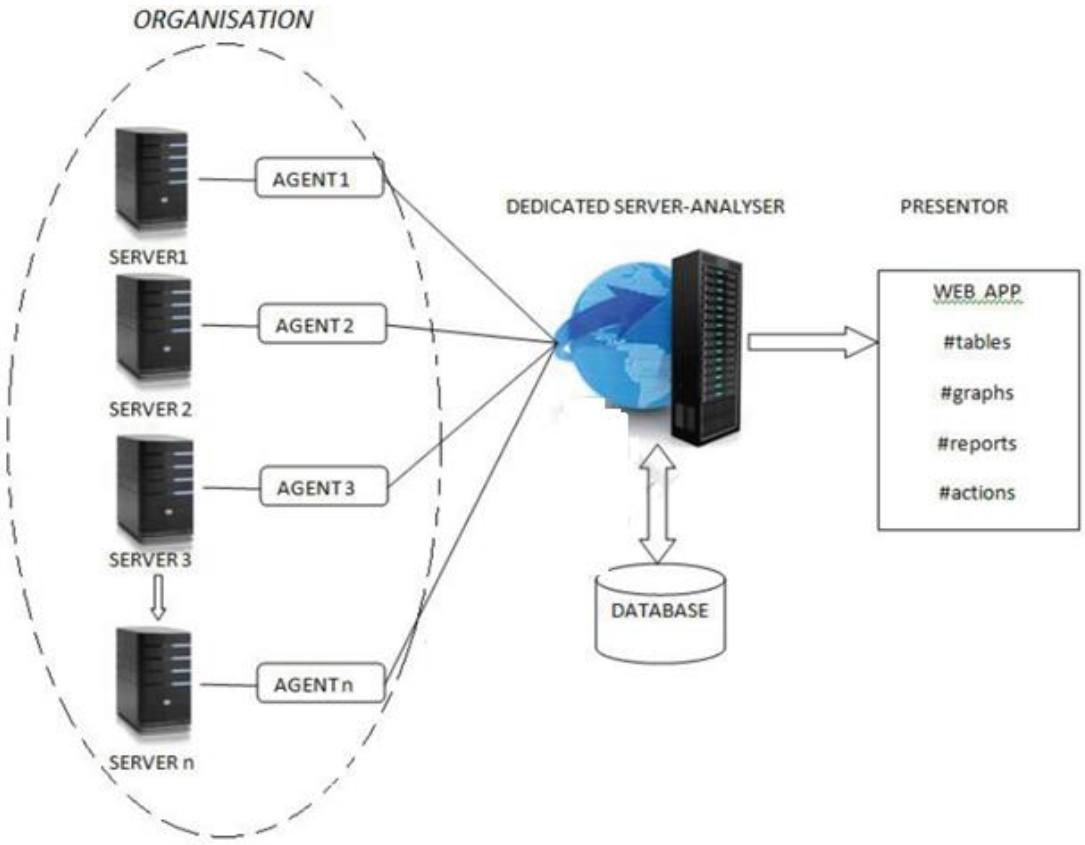

Fig 1: 3-tier system Architecture

In the above architectural diagram fig. 1, system is consist of 3 tier architecture which is mainly design to handle the group of servers of an organization.

III. EXAMPLE OF WEB SERVER TRAFFIC ANALYSIS

We have taken an example of the Web user access and server usage patterns of Monash University's main Web server located at http://www.monash.edu.au.This made use of the statistical/text log file data provided by Web log analyzer 'Analog' (Analog, 2002) which is a popular Web server analysis tool. It can generate numerical and text information based on original server log files covering different aspects of the users access log records. The weekly based reports include traffic volume, types of files accessed, domain summary, operating system used, navigation and soon. The typical Web traffic patterns of Monash University in Fig.2 showing the daily Web traffic (request volume and page volume) on the main server site for the week starting from 14-Jul-2002, 00:13 A.M to 20-Jul-2002, 12:22 A.M.

Generally, in a week, Monash University's main Web server (Server Usage Statistics) receives over 7 million hits. Since the data is not only large but also cover different aspects (domains, files accessed, daily and hourly access volume, page requests, etc.), it becomes a real challenge to discover hidden information or to extract usage patterns from such data sets. It becomes more difficult when the traffic volume keeps on growing due to the growth of the organization itself. It is easy to conclude that rather than generating basic statistical data we cannot perform web server traffic analysis.

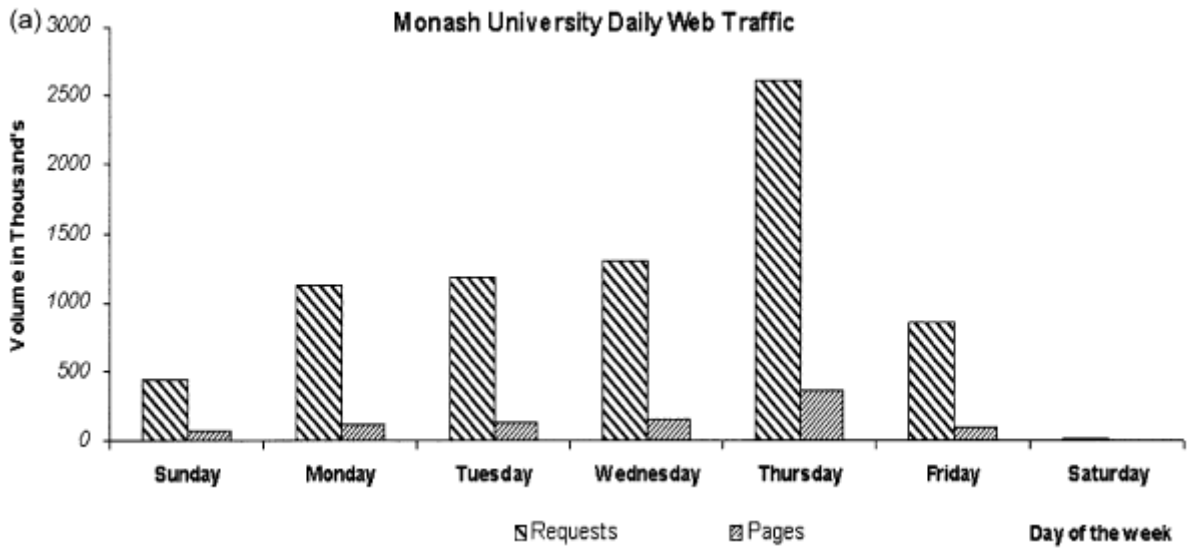

Fig.2: Daily website traffic (request volume and page volume) in a week

IV. METHODS OF WEB SERVER TRAFFIC ANALYSIS

There are different tools, methods and protocol which are available to analyze and handle the traffic on web. Out of these, we have discussed few.

A. Cisco Netflow Configuration for measuring Network traffic rate

The traffic monitoring tools are also useful for security purpose. Netflow is a traffic monitoring protocol developed by Cisco in order to capture and analyze the traffic passing through Cisco devices. This protocol works not just with Cisco devices but also with other brands of network devices. Netflow is one of the essential components adapted by every company to monitor their network traffic.

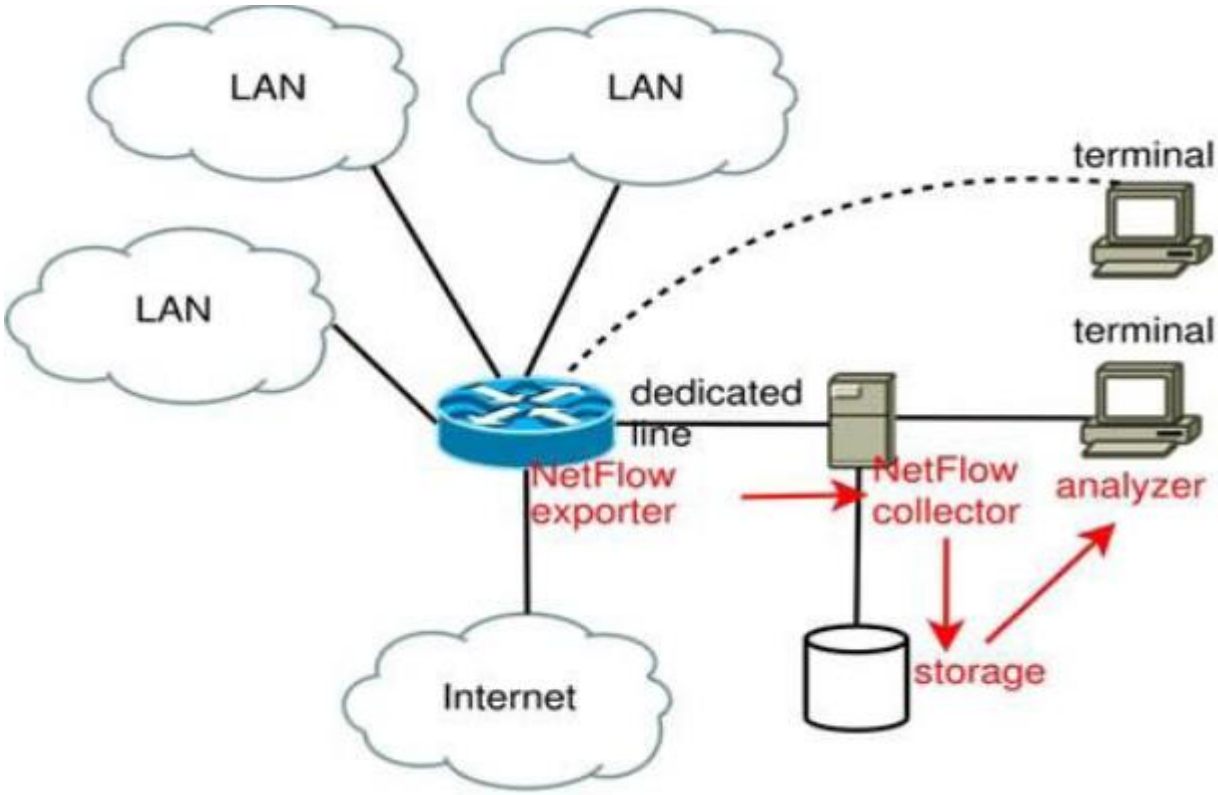

Fig.3: Cisco Netflow architecture to measure network traffic [6]

This software has also done some remarkable work in identifying and mitigating DDoS. Internet Protocol Flow Information Export (IPFIX) is a similar piece of software that was created by the IETF group for the same network monitoring purpose [7]. Netflow registers the details of the packets passing through a router in a log file which can be accessed from a remote system for analysis. It monitors traffic in accordance with various metrics like IP source address, destination address, Type of service, Source port, Destination port etc. This traffic information is gathered and is sent to the traffic collector which then forwards it to the analyzer. In this analysis, Netflow is one of the software that collects traffic information to show the variation between normal and attacked scenario.

B. Hybrid neuro-fuzzy approach for web traffic mining and prediction

The concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data. The hybrid framework combines Self Organising Map (SOM) and Fuzzy Inference System (FIS) operating in a concurrent environment as shown in Fig.4 In this concurrent model, neural network assists the fuzzy system continuously to determine the required parameters especially when certain input variables cannot be measured directly. Such combinations do not optimise the fuzzy system but only aids to improve the performance of the overall system (Abraham, 2001). Learning takes place only in the neural network and the fuzzy system remains unchanged during this phase. The pre-processed data (after cleaning and scaling) is fed to the SOM to identify the data clusters. The clustering phase is based on SOM—an unsupervised learning algorithm (Kohonen, 1990), which can accept input objects described by their features and place them on a two-dimensional (2D) map in such a way that similar objects are placed close together.

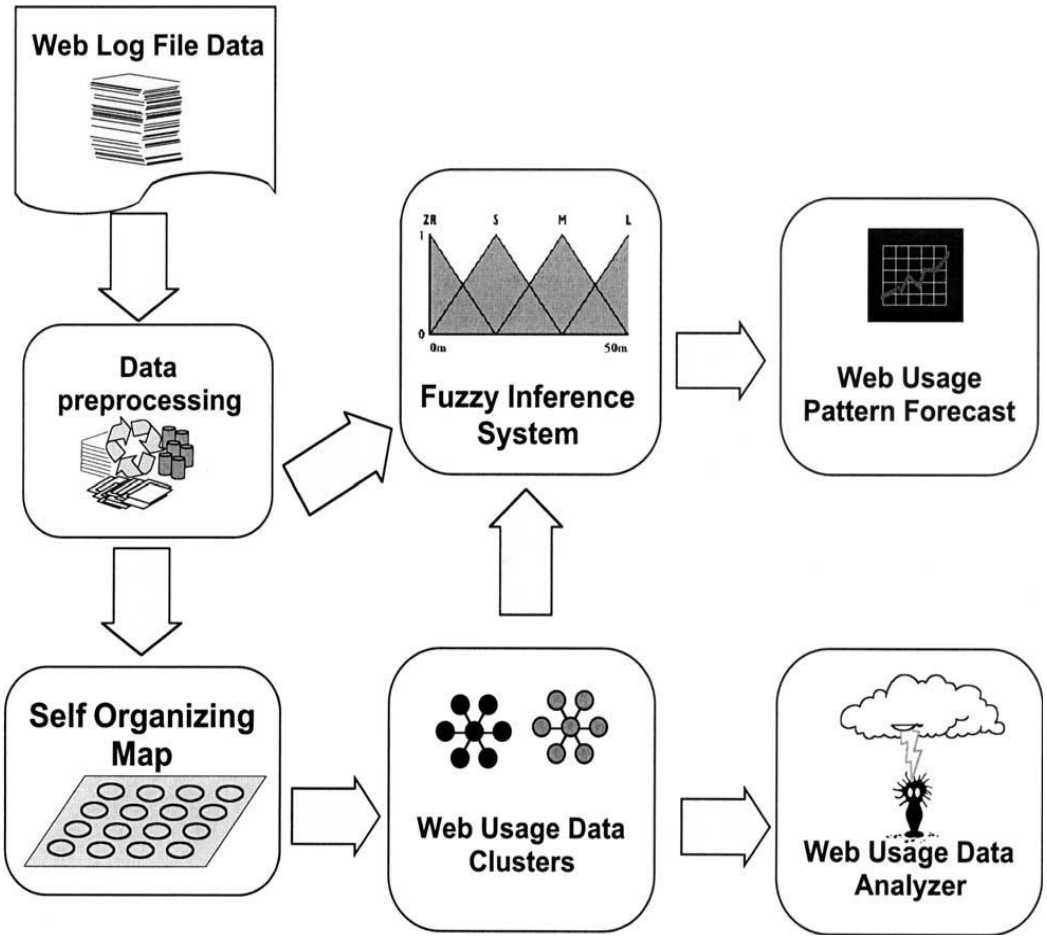

Fig.4: Architecture of the concurrent neuro-fuzzy model for web traffic mining

Fig. 4 Architecture of the Concurrent Neuro-Fuzzy Model for Web Traffic Mining FIS is used to learn the chaotic (Coenen et al., 2000) short-term and long-term Web traffic patterns . FIS is a popular computing framework based on the concepts of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning. The basic structure of the FIS consists of three conceptual components: (i) a rule base, which contains a selection of fuzzy rules; (ii) a database, which defines the membership functions used in the fuzzy

rule and (iii) a reasoning mechanism, which performs the inference procedure upon the rules and given facts to derive a reasonable output or conclusion. The proposed hybrid approach is efficient for mining and predicting Web server traffic.

C. An LQM for a Web server

The layered queuing models (LQMs) and demonstrated their superiority to traditional queuing network models since they incorporate layered resource demands. The LQM estimate client response time at a Web server. This model predicts the impact on server and client response times as a function of network topology and Web server pool size.

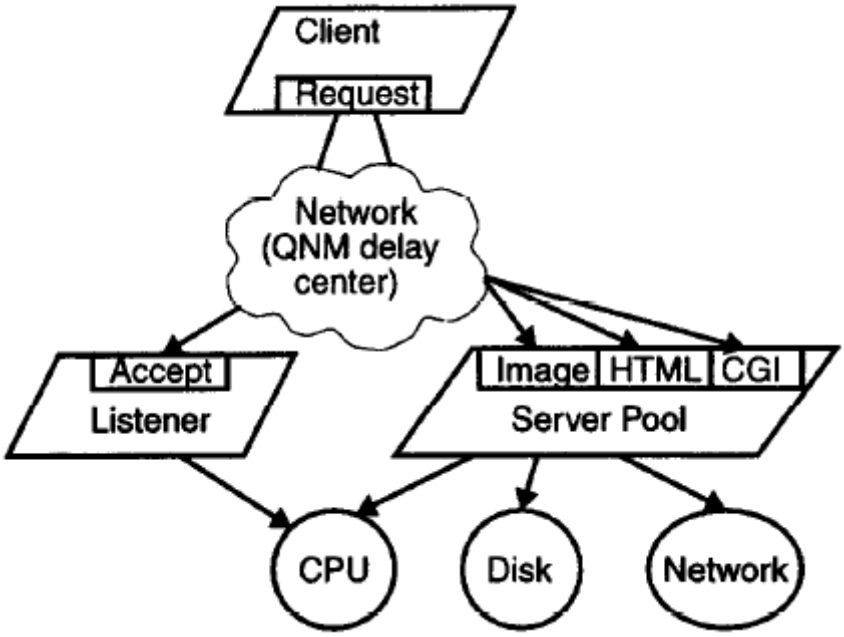

Fig.5: Layered queuing model for a Web server

An LQM for the Web server is shown in Fig. 5. The client class generates the workload for the server. We give it a single service with the name *Request*. It is used to generate the visits to all of the Web server's services. The Listener process has a single service named *Accept* that accepts client requests and forwards them to the server pool. This is reflected in the LQM as a visit by the client class to the Listener process and then a visit to the server pool. With this approach we maintain the visit ratio and blocking relationships
present in the real system.

The server pool offers many services to clients. However, our workload analysis indicated that three services constitute the majority of the resource demand and that each requires significantly different resource demands. These three services supply Image, *HTML*, and CGI objects to the client. The *Image* and *HTML* requests use processor and disk resources of the server process. The CGI service spawns another process to execute a corresponding CGI program. The server pool process waits for the CGI program to complete so it can return results to the client. In general, these CGI programs could exploit middleware platforms such as COBRA, DCOM, RMI or DCE to interact with other layers of servers.

D. Traffic direction systems (TDSs)

Sometimes the malwares are distributed by taking the advantage of web traffic. Some sites use more sophisticated methods in order to attract clicks and to direct users to the proper locations via databases and additional interfaces that help control traffic called TDSs. Cybercriminals strongly utilize TDSs to determine traffic type, which will aid them in directing users to certain malicious sites and in determining what malicious payloads to execute on particular systems. Some malware may also be the end result of a particular TDS's series of redirections, making it a malware infection vector.

TDSs present several challenges with regard to malware sample sourcing and malicious URL detection, as these are capable of detecting the use of security tools and often initiate avoidance tactics. A naive approach to looking at TDSs may, therefore, result in bogus findings and possible damage to the systems of innocent users. TDSs serve a wide range of functions such as redirecting unsuspecting users to adult sites, instigating blackhat search engine optimization (SEO) attacks, and others.

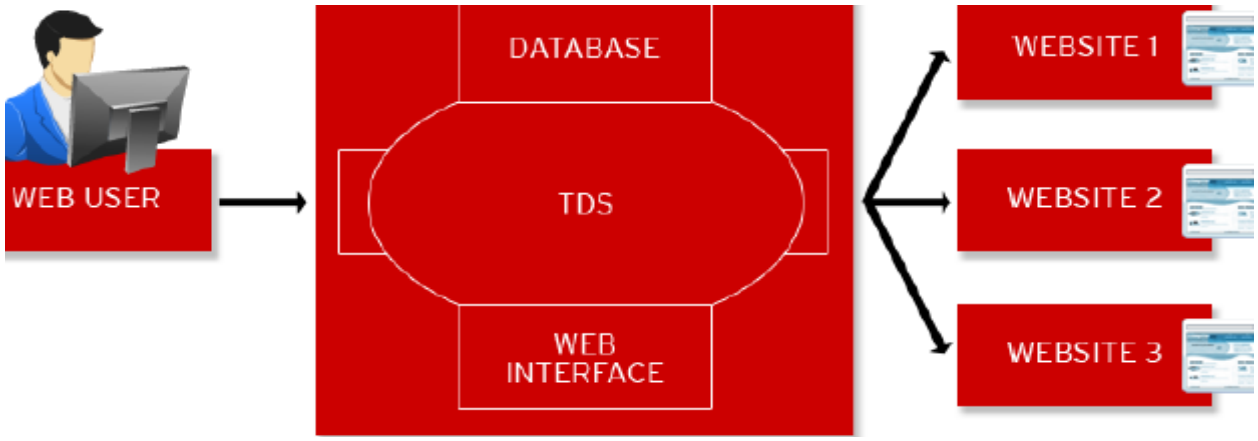

Fig.6**:** Structure of sites that use TDSs

Filtering unwanted traffic is important functionality that TDS owners use to filter unwanted or dangerous web traffic from their sites or to keep their systems hidden from security companies' search engine crawlers. Traffic filtering is implemented based on the same information used for directing traffic and works in almost the same way. Unlike traffic direction, however, traffic filtering directs most of the unwanted traffic to the most popular sites.

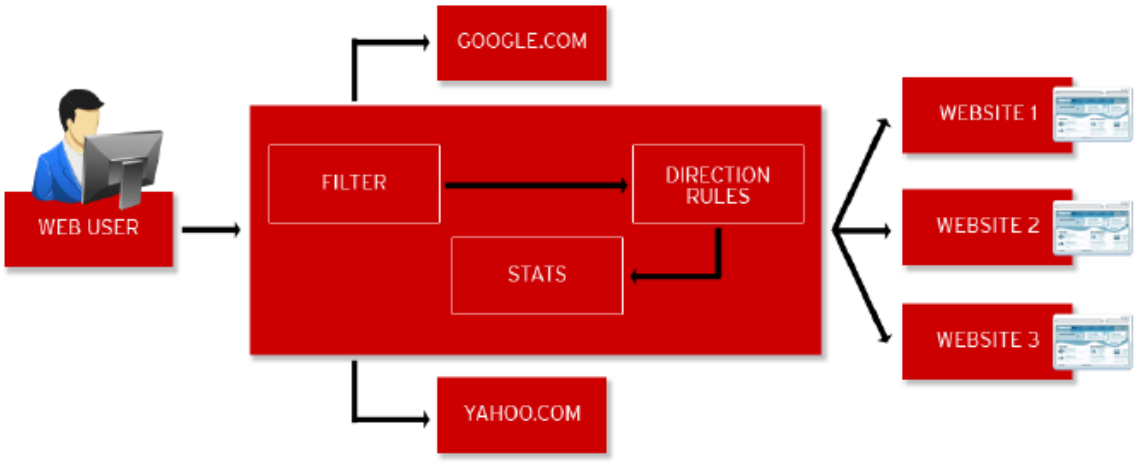

Fig.7**:** Traffic filtering functionality's structure**.**

Fig.7 shows that web traffic is usually filtered then directed from the very beginning while writing information to a database or a text file at the same time for future statistical means. TDS owners that use the traffic filtering functionality usually redirect unwanted requests to the most popular legitimate sites in order to hide their systems' main function. For example, security companies that scan lists of possibly malicious URLs and IP addresses are filtered by malicious TDS owners to block the HTTP requests of antivirus crawlers. This instead directs security experts to non-malicious legitimate sites such as *Yahoo!* and *Google.*

V. CONCLUSION

This paper described various measurement tools and modeling techniques for evaluating Web server performance. 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. Such log analyzer tools should be

widely used as they help a lot in understanding the customer behavior to the analysts. From these logs, we have been able to identify common Web server workloads.

This paper considered Web server traffic data of Monash University's main Web server for a week as an example to illustrate this concept. This suggests that an automated process is necessary for collecting, analyzing and modeling the data required for effective Web server performance management. This paper discussed the different methods and tools of analyzing the traffic on web server and it also suggest how web security is improved by traffic analysis.

REFERENCES

[1] T. Bemers-Lee, R. Cailliau, A. Luotonen, H. Nielsen, and A. Secret, "Tbe World-Wide Web," Colnm!rt. CM vol. 37, no, 8. pp, 76-82,
Aug. 1993.
[2] V. Paxson, "Growth trends in wide area TCP conneclions," IEEE : **Network, vol. S.** pp, 8-17, July/Aug. 1994.
[3] 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.
[4] 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.
[5] Theint Aye: Web Log Cleaning for Mining of Web Usage Patterns, IEEE, 2011.
[6] Brad Reese. (2007 November 11) Cisco invention NetFlow appears missing in action as Cisco invests into the network behavior analysis business [Online]. Available: http://www.networkworld.com/community/node/22284.
[7] Cisco. (2007, October) Cisco IOS Netflow, Introduction to Cisco IOS Netflow – A Technical Overview [Online]. Available: http://www.cisco.com/en/US/prod/collateral/iosswrel/ps6537/ps6555/ps6601/prod_white_paper0900aecd80406232.html
[8] Neha Goel,"Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool," International Journal of Computer Applications (0975 – 8887) Volume 62– No.2, January 2013.
[9] Nisar Nadaf and Pramod Pathak," Web Server Log Analyzer", International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 2, February – 2014.
[10] Xiaozhe Wang, Ajith Abraham, Kate A. Smith," Intelligent web traffic mining and analysis", Journal of Network and Computer Applications.
[11] JohnD illey , Rich Friechich , Tai Jin, Jerome Rolia," Web server performance measurement and modeling techniques", Performance Evaluation 33 (1998) 5-26.

