Indian Journal of Science and Technology, *Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/105824, December 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Analyzing HTTP Traffic Patterns for Monitoring and Analyzing User Behavior

Shilpa Mahajan* and Shilpa Yadav

Department of Computer Science, The North Cap University, Gurgaon, Haryana, India;
shilpa@ncuindia.edu, yadav.shilpa93@gmail.com

Abstract

Objective: This paper presents a method for analyzing user behavior pattern by evaluating what web users are looking for in websites. **Methods/Statistical Analysis:** There are various approaches available in diverse fields for analyzing human behavior. In today's generation web usage has increased tremendously due to wide variety of information and communication facility. The main source of data in web browsing is the web logs that stores user actions on web pages. The generated logs are analyzed in phases and then classification techniques are applied to predict future behavior of the user. **Findings:** This information can be used by E-commerce companies to know about their customer requirements and can later improve their websites information and structure as per results. Similarly, same analysis technique can be used by organizations to know about employee requirement. **Applications/Improvement:** This analysis method can be used by E-commerce companies and organizations to predict their customers and employee needs and behavior.

Keywords: Behavior, Cluster, Pattern, Traffic, Web Log

1. Introduction

User behavior analysis involves tracking, collecting and analysis of data through monitoring devices. There are many things we get to know by analyzing the behavior of a user. The web data is very vast, formless and messy in nature. In addition to this, due to varying and diverse nature of data, web searching is a tedious task for users. It becomes essential to analyze this disordered data to retrieve useful information from it. This information can be used for further predictions. For example, in an E-commerce business, every company has a website with a list of all products it sells. A customer can search for particular items which a customer wants to purchase. From a database, a company can track each user interest and explore it. This would results in knowing the liking of that particular customer. Later, this information can be used to send further notifications on the kind of products which a customer can buy or interested in. This analysis can be done on the number of hits on the products which a customer made and also on the products added on to its cart.

An analysis on the collected data can be used for varied purposes. It can be used for predicting the future behavior of the user, current market needs, employee satisfaction surveys and others. It can be used to find any deviation from the regular behavior pattern and enables to detect outliers. In an organization, web logs can be used to determine, when user log on to the Internet, the web pages often visited by user, traffic pattern and the contents in which user is interested in viewing like advertisements, job sites etc.

**Author for correspondence*

They proposed a method for analyzing the pattern based on mobile phone sensors using a MAST model[1]. This method explores directions and also identify technologies requirement. Various actions like movement on location change and talk, press on a key are considered. Phone sensors are analyzed using MAST. The analyzed results are in terms of time (in seconds) and location (where event happens).

They described a method to analyze user behavior using click stream data. It describes a method to reach the number of potential customers through online media like face book, twitter, Gmail, advertisements etc. It is a process of collecting and reporting the data about the pages visited by the user and succession of mouse clicks by the user[2]. The click stream analysis is classified in two parts i.e. Traffic analysis and E-commerce analysis. Traffic analysis operates at server level and it tracks the path taken by the user while the user is navigating through the website. It also tracks how long it takes to load a page and how many times user goes back by hitting back arrow on the screen and reload the page. In E-commerce analysis it keeps track of which items are seen by user. The user adds into the shopping cart and removes out of the shopping cart. This way they keep track of what items are liked by the user.

They proposed a method in which users with similar behavior are grouped in clusters. These clusters are formed and categorized on the basis of user clicks[3]. The user clicks are traced. In this paper, click stream data method is used to reach potential customers through various online media outlets and their collected data is then analyzed for future.

In fine and compared various log analyzer tools that can be used for analyzing weblogs. In their work, they have analyzed astrology website logs[4]. It helped them to add various features on to the existing web pages based on user behavior access pattern. This also helped them to know about customer behavior, interest and its requirements.

In put forth web usage mining based on fuzzy clustering. This paper aims to categories users on similar behavior pattern based upon their preferences, searches and requirements[5]. This paper provides an effort to cluster similar Web users, considering two factors page-click number and web browsing time that are stored in Web log and their degree of impact. An effort has been made to organize web pages on browsing history. This information is used further for improvement and addition of information based on customer interest.

In describes process for mining E-learning data and text mining of moodle data. The paper also makes use of free data mining tools and techniques for finding informative pattern from web access data[6]. The paper also describes a new model based on visual clustering for discovering hidden pattern from web access log files.

In discuss about the use of web page accesses from the different server logs to discover the frequent usage by the client and also from the experimental study, finds some interesting patterns through association rule mining algorithm and compares pattern mining algorithm i.e. Apriori and FP growth algorithm[7].

In define pattern extraction algorithms for web usage mining. Author proposed a method to break down a session into various transactions using forward reference approach[8]. This approach refers last visited page before directing to a new page.

This paper aims to collect web log data from clients over a period of time, analyze this data using tools. The behavior analysis of different users on the basis of collected web access log can be made. This information can be used further for future prediction based on access pattern of the users. This data is further grouped and classified based on machine learning algorithms.

2. Proposed Methodology

This paper aims to analyze web access logs for obtaining useful information. For implementation, the process has been divided into various phases as explained in Figure 1. These phases involves creation of web server, collection of web access log files, cleaning of data using data mining tools and then classification and clustering of data using machine learning.

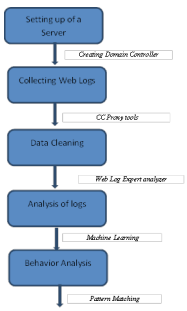

Figure 1. Proposed project model.

2.1 Setting up of Server

Initially, it is required to setup a Domain Controller. It is a server which responds to authentication requests for logging in and getting security permissions within Window domain. A domain is introduced to grant access to the number of resources with the help of username and password. There could be primary domain controller as well as backup domain controller. Primary domain controller is for maintaining user and group information. Backup domain controller keeps account of a user database. Steps for setting up a domain controller are defined below.

- Insert CD of Windows Server 2003 in CD-ROM of computer.
- Click on Start, then Run, and type DC promo.
- Start Active Directory Installation Wizard.
- Click on Domain controller to start a new domain.
- Select Domain in a new forest.
- Specify full DNS name for new domain.
- Provide NetBIOS name.
- Assign location of database and log files.
- Assign location of SYSVOL folder.
- Configure DNS server on the system by clicking on Install.
- Click on Permissions with OS.
- Give password.
- Confirm the selected options.
- Installation proceeds.
- When user is prompted, then restart the system.

After configuring a Domain Controller, clients get connected to the server. Server log is a log file or files created and managed by the server. It lists all tasks which took place on client machines. A classic example is web server log that maintains history of pages which are visited by client. There is a standard format specified for web server log files, but different proprietary formats also exist. The most recent entries are appended at the end of a log file. User specific information is not present in the server logs. The files are available to the administrator only. There are number of tools available to get http access logs like AWStats, Analog, Deep Log Analyzer. In this paper, logs are generated using CCProxytool. The CCProxy tool environment can be seen in Figure 2.

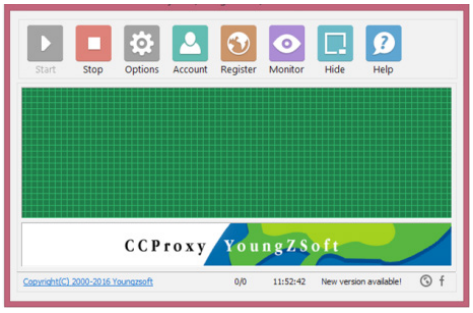

Figure 2. CCProxy environment.

It is required to configure CCProxy tool. For configuration, the user needs to have Network Interface Card (NIC). NIC connects a system with the network. In Figure 3 shows depicts the same.

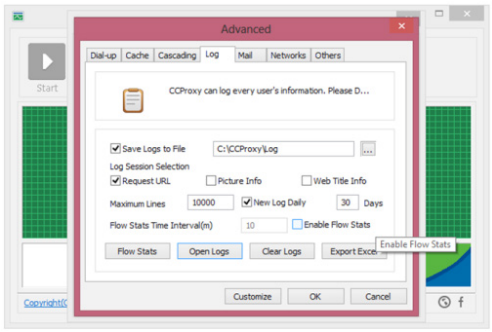

Figure 3. Configuring CCProxy.

2.2 Generation of Web Log Files

Sample Server http log file generated can be seen in Figure 4. This is a text (ASCII) file that contains information about User Name, IP Address, Time Stamp, Access Request, URL that Referred, error codes etc. There are four types of server log like Transfer Log, Agent Log, Error Log and Referrer Log.

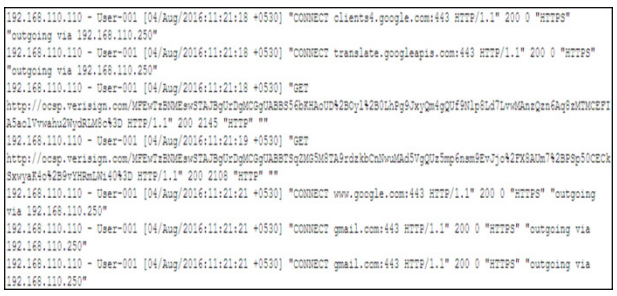

Figure 4. Sample log file.

2.2.1 Error Logs

It is the most crucial log file. The name and location of this file is set by Error Log directive. Here Apache HTTP server will send information and record errors which are encountered while processing requests. Whenever a user encounters an error then Error Log is the first place where problem could be tracked since it contains the details of error and fixing the error.

Format for error log is given below:

[Thu Oct 28 16:31:42 2000] [error] [client 127.0.0.10] client denied by server configuration: /export/home/live/ap/htdocs/test

The first entry contains date and time. The second entry tells about the type of error and then come error log directives.

2.2.2 Access Logs

Access log[9] records all the requests which are processed by server. Custom Log directive take care of contents and location of a log file. Log Format directive could be used for simplifying selection of contents of logs. Initialization of Log management is achieved by storing information in access log. After initialization, the information is analyzed to produce statistics useful for a user. The format for access logs is configurable. The format is listed using format string which looks similar to printf() format string used in C language.

2.2.3 Common Log Format

Configuration for access log is given below:-

LogFormat"%h %l %u %t \"%r\" %>s %b"commonCustom Log logs/access_log common

2.3 Data Cleaning

Log data collected from CCProxy cannot be used as it is. This data needs to be cleaned in order to remove unwanted details and also for preprocessing to obtain useful information from it. Numbers of tools are available that takes log files as input and produces desirable output file[3].

There are different log analysis tools present are[10]

- **AWStats** – It is an analysis open source, platform independent tool that can analysis web server log files comprises of thousands of users. This tool generates structured reports and can also analyze FTP files. The reports can be generated in various formats as required by the user.
- **Analog** - A powerful user friendly and open source tool, which is machine independent. This tool can generate reports in almost 24 languages. Features of this tool can be enhanced further for detailed analysis and output report can be generated graphically as well as statistically way.
- **Deep Log Analyzer –** A user friendly customizable interface that can analyze even Apache and IIS web servers logs. The resultant reports can generate information in Excel and HTML format. It supports execution of script for automatic generation of reports. Webmasters could also generate report which focuses on optimization of search engine.
- **WebLog Expert** – An easy to install and a user friendly tool. This tool generates detailed reports and can analyze results on number of different attributes. The information regarding visitors, Accessed files, paths, referred pages, browser, search engines and operating system. It generates easy to read report which includes both text and charts. The reports can be generated in CSV, PDF and HTML formats. This tool can even support dynamic HTML reports. It can also read ZIP and GZ compressed log files. These attributes can be used for performing clustering. Using these attributes clustering is performed. In this work, WebLog Expert tool is used for data cleaning.

2.4 Analysis of Logs

The collected log can be used for analyzing the web log. This information can be used for finding number of users visiting particular page, time of visit of any particular web page and bandwidths etc. In this paper, web log files are taken from the server collected over a period of time and these attributes are obtained using WebLog Expert Tool. The results collected are:

2.4.1 General Statistics

General stats give general information like total hits, visitor hits, spider hits, page views and bandwidth. The summary of weblog can be seen in tabular form as shown in Figure 5.

Hits	
Total Hits	39,474
Visitor Hits	29,191
Spider Hits	1,283
Average Hits per Day	3,809
Average Hits per Visitor	8.16
Cached Requests	3,979
Failed Requests	233
Page Views	
Total Page Views	4,435
Average Page Views per Day	554
Average Page Views per Visitor	1.24
Visitors	
Total Visitors	3,577
Average Visitors per Day	447
Total Unique IPs	3,038
Bandwidth	
Total Bandwidth	567.48 MB
Visitor Bandwidth	548.81 MB
Spider Bandwidth	18.67 MB
Average Bandwidth per Day	70.94 MB
Average Bandwidth per Hit	19.07 KB
Average Bandwidth per Visitor	157.11 KB

Figure 5. General statistics showing summarized information of Web Logs.

2.4.2 Daily Visitors

This section defines number of visitors visiting particular website on a particular day. The graph in Figure 6 clearly shows number of visitors and the day of visit. The information can also be used to view particular visitor visiting particular websites and number of times that web page is visited popular pages based on number of hits by the visitor can also be seen as shown in Figure 7.

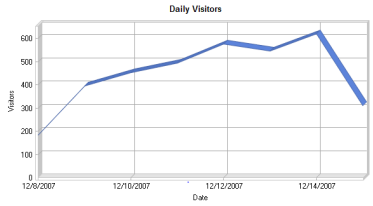

Figure 6. Daily visitor visiting particular webpage.

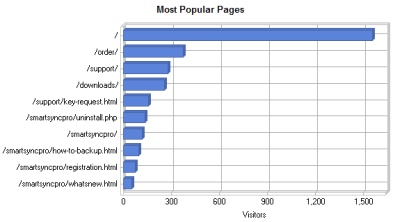

Figure 7. Most popular pages visited by visitors.

2.4.3 Search Engines

Weblog Expert helps to find popular search engines that are used by the users as shown in Figure 8. The results are based on number of hits. Through this tool, the useful information can also be used to find particular websites visited by the user. The results can be further used by E-commerce websites to analyze user behavior and the web pages frequently visited by the customers and the type products looked by the customers.

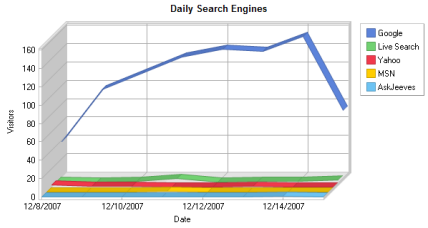

Figure 8. Daily search engines statistics.

2.4.4 Browsers

This tool also enables E-commerce owners to find customers browsing particular websites. Through this, a popular browser used by maximum number of users can also be depicted as shown in Figure 9.

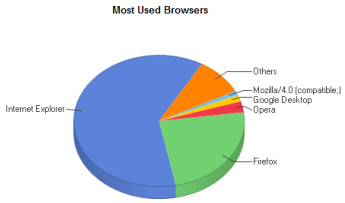

Figure 9. Popular browsers used by the users.

2.4.5 Error Types-

Error types defines the errors which are occurred during browsing of websites. As explained that error log keep accounts of tall the errors encountered by the users. The results are shown as defined in Figure 10.

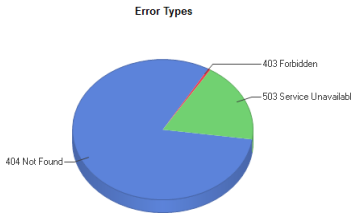

Figure 10. Error types.

2.4.6 Activity by Hour of Day

This is an important statistics that can be used for monitoring the internet usage in particular hours of a day. It also describes in which hour the traffic was peak. Individual users web access history logs in particular

hours can serve as input for further analysis like in an organization, employee state of mind, their behavior analysis can be done through this log. Visitors can be categorized according to hour of the day as shown in Figure 11.

Activity by Hour of Day

Hour	Hits	Page Views	Visitors	Bandwidth (KB)
00:00 - 00:59	1,661	225	182	32,977
01:00 - 01:59	1,618	283	184	28,350
02:00 - 02:59	1,228	171	153	21,111
03:00 - 03:59	1,017	147	135	15,704
04:00 - 04:59	960	175	125	24,812
05:00 - 05:59	995	125	122	18,016
06:00 - 06:59	921	116	108	14,005
07:00 - 07:59	886	133	125	8,768
08:00 - 08:59	727	135	133	12,951
09:00 - 09:59	648	114	118	15,725
10:00 - 10:59	780	102	100	13,971
11:00 - 11:59	918	111	112	27,918
12:00 - 12:59	1,536	176	140	25,377
13:00 - 13:59	1,464	193	146	16,290
14:00 - 14:59	1,581	228	141	26,797
15:00 - 15:59	1,283	174	147	35,298
16:00 - 16:59	1,408	204	143	21,003
17:00 - 17:59	1,408	188	143	33,456
18:00 - 18:59	1,191	181	153	25,608
19:00 - 19:59	1,659	237	172	33,769
20:00 - 20:59	1,621	217	187	36,215
21:00 - 21:59	1,685	267	197	33,832
22:00 - 22:59	1,709	265	211	21,818
23:00 - 23:59	1,630	268	200	37,309
Total	30,474	4,435	3,577	581,103

Figure 11. Table showing activity by hour of day.

2.5 Behavior Analysis

The collected logs are analyzed using data mining tools like Weka. This analysis involves grouping of data into meaning full clusters and classification[11] of data on various attributes. Using machine learning algorithms, outliers can be detected. In Weka, there are built in algorithms like K Mean, Density based clustering, Canopy clustering, Farthest first, Hierarchical cluster and much more for analyzing various attributes.

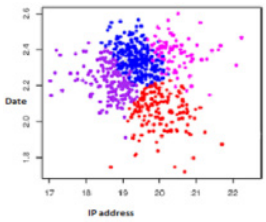

Figure 12. Number of user browsing vs. days.

In this paper, K Mean clustering algorithm[12] is used for creating clusters considering different attributes. The purpose of considering K mean clustering is to perform clustering[13] computationally faster. As K mean produces tighter clusters as compared to any other clustering technique. Figure 12 shows clusters of different colors depict different users accessing web pages in particular days[14,15]. This type of graph can be used to find the type of contents users are interested in browsing.

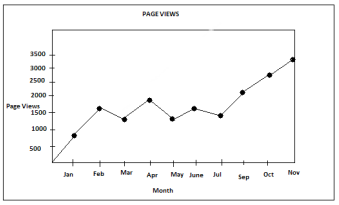

Figure 13. Page view per month.

Figure 13 shows number of web pages accessed in a month. This information can be used to know what type of contents user are interested in. E-commerce companies can use this information to improve organization of their website[16]. This information can be used by organization to know about their employee's behavior also.

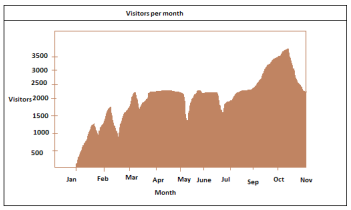

Figure 14. Visitors per month.

It is also possible to find how many visitors visited particular website in particular month. This information can be used to determine in which period of session like winter, summer or sale maximum people visited their website. Similarly, organizations can use this information to find out in which type of websites employee visits. This information is based on the number of hits of particular webpages. Figure 14 shows how many visitors visited from January to November.

3. Conclusion

In this work, an effort has been made to analyze web log data to identify similar behavior patterns of a user. For this, the work has been divided into various phases. Starting from collection of data through CCProxy and then mining the data to convert it into useful information. Later machine learning algorithm is applied using data mining tool like Weka to analyze logs. User behavior can be explored and monitored to understand current needs of customers. This information can also be used to find clusters i.e. users showing similar type of data and to identify outliers. In future, organizations

and E-commerce companies can use this information for analyzing user needs and customers' requirements.

4. References

1. Song J, Eugene, Tang Y, Liu L. User behavior pattern analysis and prediction based on mobile phone sensors. NPC'10 Proceedings of the 2010 IFIP International Conference on Network and Parallel Computing; Zhengzhou, China. 2010. p. 179–88.
2. Wang G, Zhang X, Tang S, Zheng H, Zhao BY. Unsupervised click stream clustering for user behavior analysis. SIGCHI Conference on Human Factors in Computing Systems; USA. 2016. p. 1–12.
3. Umamaheswari S, Srivasta SK. Algorithm for tracing visitors' on-line behaviors for effective web usage mining. International Journal of Computer Application. 2014 Feb; 87(3):22–8.
4. Goel N, Jha C. Analyzing users behavior from web access logs using automated log analyzer tool. International Journal of Computer Applications. 2013 Jan; 62(2):29–33.
5. Zhang J, Zhao P, Shang L, Wang L. Web usage mining based on fuzzy clustering in identifying target group. International Colloquium on Computing, Communication, Control and Management. 2009; 4:209–12.
6. Amit V, Nath K. A survey on web log mining pattern discovery. International Journal of Computer Science and Information Technologies. 2014; 5(6):7022–31.
7. Mishra R, Choubey A. Discovery of frequent patterns from web log data by using FP-Growth algorithm for web usage mining. International Journal of Advanced Research in Computer Science. 2012 Sept; 2(9):311–8.
8. Pani S, Panigrahy L, Sankar V, Ratha B, Mandal A, Padhi S. Web usage mining: A survey on pattern extraction from web logs. IJICA. 2011; 1(1):15–23.
9. Joshila G, Maheswari V, Nagamalai D. Web log data analysis and mining. Proceeding of CCSIT-2011Springer CCIS. 2011 Jan; 133:459–69.
10. Kumar P, Iswarya R, Vindhya R. Predictive analysis of user behavior in web browsing and pattern discovery networks. International Journal of Latest Trends in Engineering and Technology. 2014; 4(1):239–45.
11. Aggarwal N, Gaur D. Classification of crime data using rapid miner. International Journal of Applies Engineering Research. 2015; 10(5):27517–21.
12. Mahajan S, Malhotra J, Sharma S. Delay tolerant and energy efficient QoS-based approach for wireless sensor network. International Journal of Systems, Control and Communications. 2014; 6(2):121–35.
13. Mahajan S, Malhotra J, Sharma S. An energy balanced QoS based cluster head selection strategy for Wireless Sensor Network. Egyptian Informatics Journal. 2014; 15(3): 189–99.
14. Livinsa Z, Shri MSJ. Monitoring moving target and energy saving localization algorithm in Wireless Sensor Networks. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–5.
15. Lakshmanan G, Posonia M. A novel analysis on application of neural support on nuclear reactor control process monitoring. Indian Journal of Science and Technology. 2016 Mar; 9(10):1–5.
16. Handa M, Gupta N. A study of the relationship between shopping orientation and online shopping behavior among Indian youth. Journal of Internet Commerce. 2014; 13(1):22–44.