ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(42), DOI: 10.17485/ijst/2016/v9i42/99300, November 2016* ISSN (Online) : 0974-5645

Discovering User Pattern Analysis from Web Log Data using Weblog Expert

K . Dharmarajan[1*] and M. A. Dorairangaswamy[2]

[1]Department of Information Technology, Vels University, Chennai - 600117, Tamil Nadu, India; dharmak07@gmail.com
[2]St. Peter's University, Chennai - 600054, Tamil Nadu, India; drdorairs@yahoo.co.in

Abstract

Objective: This article tries to discover the hidden knowledge and identifying user behavior on the web by using the web data sources. With the help of this knowledge, the overall performance of future accesses, the typical browsing behavior of a user and subsequently to predict desired pages, a user wants to access in future. **Methods/Statistical Analysis:** The user pattern is analyzed by using the modified Web Log Expert tool from the web access log file collected from the organization. This modified tool tries to conduct a web mining in a domain independent manner. This algorithm consists of three parts: 1. Given an input entity, extracting a set of IP addresses and visor lists and then ranking them according to comparability, 2. Extracting the domains in which the given entity takes part and 3. Identifying and summarizing the competitive evidence that details the organization's strength. **Findings:** The main aim of the research work is extracting of user frequent access page using web log data, which is based on user session time, IP addresses, browser details, operating system, top user. This complete analysis work has been implemented in the Web Log Expert tool. The experimental results provide an easier way to navigate the website and improve the website design architecture. This work deliberates the detailed results of a website in a specific education domain application. We investigate the statistics of hourly based, daily based, week and monthly based report of the web usage patterns. The goal is to capture, model and analyze the behavioral patterns and profiles of users interacting with a website. The knowledge about users and their behavior on the web helps the organization benefits and leads directly to profit increase.

Keywords: Frequent Patten, Session Identification, User Behavior, Web Usage Mining

1. Introduction

The Internet is one of the greatest inventions and has become a most popular source, disperse and improve information as well as extract useful information[1]. Web data mining is a powerful technique in data mining and an emerging research area in which different techniques have been used to solve the various issues related to analyzing the pattern of the web usage of user available in the web server. This is used to identify user behavior, assess the usefulness of a particular website and help compute the achievement of visitor. Web log mining implemented to identify the real world problems like frequent accesses, most visited pages, interesting user and user behavior emerging patterns so it will help for developer to develop more themes[1]. Web server log file automatically created at web server and collected from www.velsuniv.ac.in. The main web data mining analysis of navigations acts as a key factor in educational domain which provides the user behavior of original implementation towards real-time data with different level of implications[2]. Our Web Log Expert experiment results initially stresses with retrieval web pages as nodes and hyperlinks as edges in order to classify the webpage as a frequent access webpage or similar webpage. It manipulates the favorite access information about user, top view, error page, browser and various platforms used by the website users' frequently[2]. This work focuses on web log mining and in extracting emphases on analyzes the web user structural patterns of websites from the server log data.

** Author for correspondence*

2. Web Usage Mining Phases

It is part of data mining technique on huge web log files sources to discover automatically and analyses wealth of useful emerging and user's behavioral patterns[3]. To identify the task of website users is a tremendous task. The stages in web usage mining listed in Figure 1.

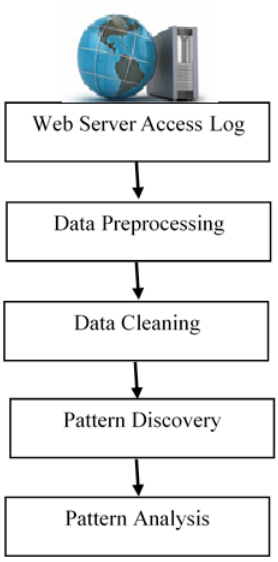

Figure 1. Web usage mining phases.

2.1 Data Collection

User web log dataset is gathering from many sources like server side, client side, proxy servers and so on. Web log server is definitely the easy and the most common source of data[6]. When a user request a resource form a website, the information is stored in web log data which is a normal plain text file format[7].

2.2 Preprocessing

This technique involves removing of the click stream data and splitting of data into a set of user transactions with their respective visits to the website[3]. All server log files do not have the right format, so there is a necessary for preprocessing technique.

2.3.Data Cleaning

Data cleaning is one of the main complex phases of the web usage mining process. So it is important to reduce noise and unrelated data. The process of minimizing data size for the data cleaning phase is to remove the jpeg, word, gif, sound, animation files. User identification is the phase of identifying users by using IP address and user agent fields of log entries. A user session is measured by the user visit and duration of access of a website.

2.4 Pattern Discovery

The different proposed approaches for pattern discovery are key-matching techniques, statistics, artificial neural network, data mining and Clustering algorithms. This proposed process identifies user's navigational sequential patterns. The emerging pattern discovery includes data mining methods like path identification, Frequent Pattern based Association rule, clustering and classification on preprocessed log data[7]. Clustering approach is used for similarity pattern discovery.

2.5 Pattern Analysis

It is the final stage of web usage mining, which includes the conversion and understanding of the web log data mined pattern. Knowledge discovery technique is used to identify the hidden in sequence or predictive emerging pattern from web log files.

3. Available Tools

Number of experiments is available to analyze these web server access logs as an input. It generates the reports for frequent access and identifies the user access patterns. Table 1 gives the Analysis Report from Web Log Expert. It gives us all sorts of information starting from how many hits, the browser used, length of their stay and much more[4].

Table 1. Analysis report from Web Log Expert

Hits	
Total no. of hits	3,371,252
Total no. of visitor	1,722,744
Web spider hits	1,648,508
Average visitor hits per day	105,351
Average hits per visitor	17.30
Frequent cached requested	89,195
Failed no. of requested	65,732
Web site views	
Sum of page view	267,896
Regular page visited per day	8,371
Average page views per visitor	2.69
Visitors	
Sum of visitors	99,555
Regular visitors per day	3,111
Total unique IP Address	55,656

3.1 Web Log Expert

This web analytics software tool can read log files to analyze website visitor's behavior and get complete website usage statistics, most popular pages. Many Web servers software available such as: Apache, IIS, Glassfish. This tool read ZIP file, GZ and BZ2 constrict log files. So won't need for delimiter the logs manually before analyzing[5]. The User friendly Graphical User Interface of Web Log Expert tool having menu, toolbar and the list of profiles which is shown in Figure 2.

The tools provide easy to understand reports that include picture and table information, charts representation, web pages and pdf formats[4].

4. Tools Analysis Report

This work used the data collected from the web server, www.velsuniv.ac.in website from 01 July 2015 to 31 July 2015. This collected data is analyzed by using Web Log Expert Lite 9.2 web mining tool. The complete experimental analysis was done on the basis of web log data of an educational institutions website. The design and execution of such work is restricted and time consuming.

4.1 Analysis on Apache Log File

The following results were obtained to identify the user behavior and Emerging Pattern of the website users on Apache Web server log data.

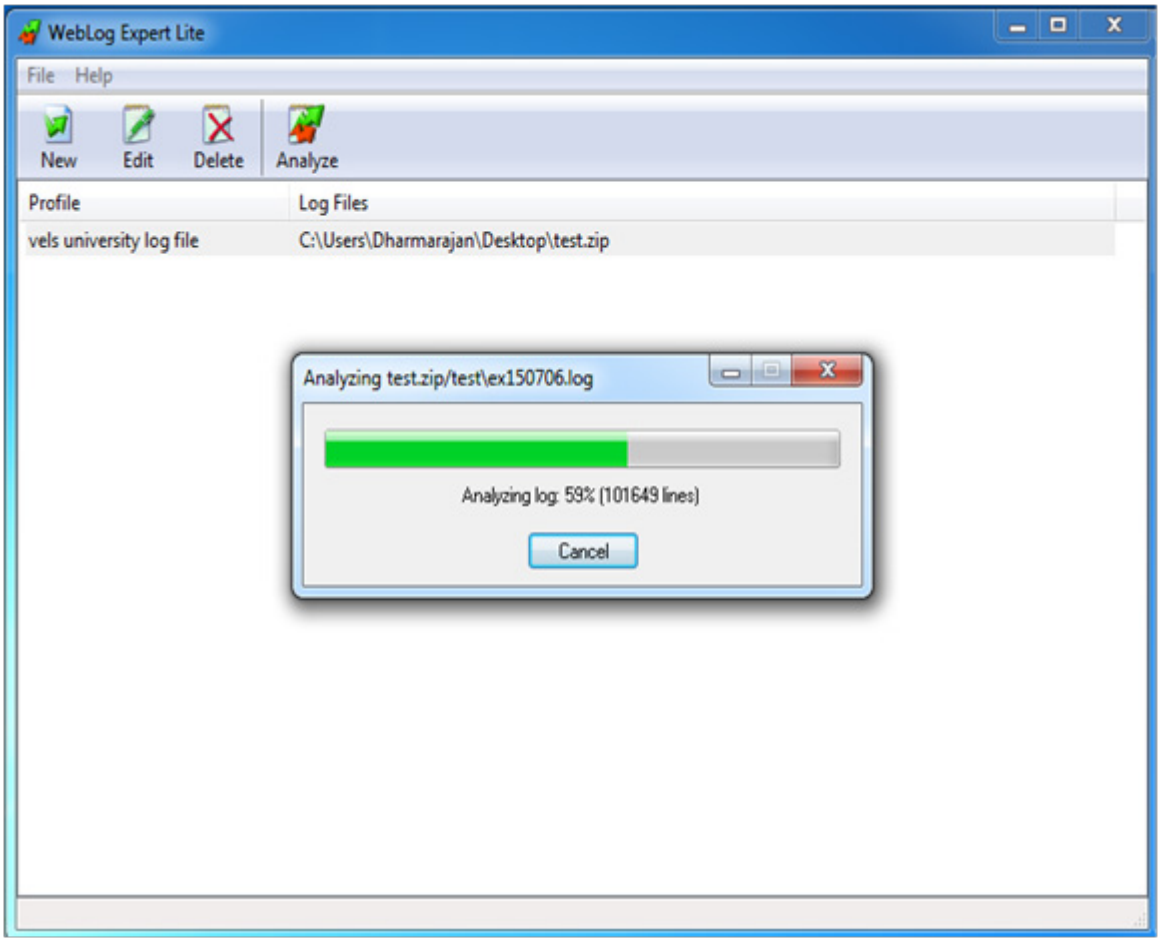

Figure 2. Load web server log file.

- General Statistics: This result shows the Web usage details that also include whole hits, per day average hits, total page views, total bandwidth etc.
- Activity Statistics: This measurement gives the information about the user's activity by date and hour of day. Date and time are essential attributes in web log file because user's activity by date and hour of day provides the details of hits, page views, visitors, bandwidth etc.
- Access Statistics: Here the statistics for most popular website, frequent download files, maximum request images, most request directories, top entry web pages, daily page accessing, image access, directory access, entry pages are shown. Entry page is a first web page visited by visitor on the site. This statistics also provide an idea of the navigational behavior of visitors.
- Referrers: Here the report shows the top indexing sites, top referring links and maximum search engine used.
- Time and Success rate: It is time used by the user in each page while finding complete website. Successor rate calculated by user to find the exact information. This all information stored in web server log file.
- Visitors: It shows the list of IP address/domain names of hosts that visited website along with the number of times hit by a particular user. Visitors section analyses the IP address field of Web log file[8].

The result graph shows the maximum hit page is index.asp and par-time course-offered.asp is least hit page. Table 2 gives the Top Visitor Hosts. This report shows a list of IP addresses, HIT, Frequent visited, particular user hit. Maximum hit. In this work will help identify the user behavior and website maintenance needs a way to track and measure visitor's traffic[8]. This will give us an idea of the navigational behavior of the user and also after visiting which web page, the user loses interest in the website.

Table 2. Top visitor hosts

S. No.	Host	Hits	Visitors
1	64.233.173.243	3,628	327
2	64.233.173.247	3,854	327
3	64.233.173.251	3,694	321
4	50.18.94.121	1,386	182
5	54.151.42.39	1,436	178
6	54.241.198.78	1,548	176
7	37.187.77.43	2,680	159
8	52.6.166.158	395	147
9	91.121.222.130	601	146
10	66.249.82.172	1,061	140

Figure 3 shows the frequent page access graph. This graph shows on daily and hourly basis visitor behavior. This helps the website owner to improve the quality. By utilizing this details particular schemes can be initiated which might aid in improving the user who accesses the web page[6].

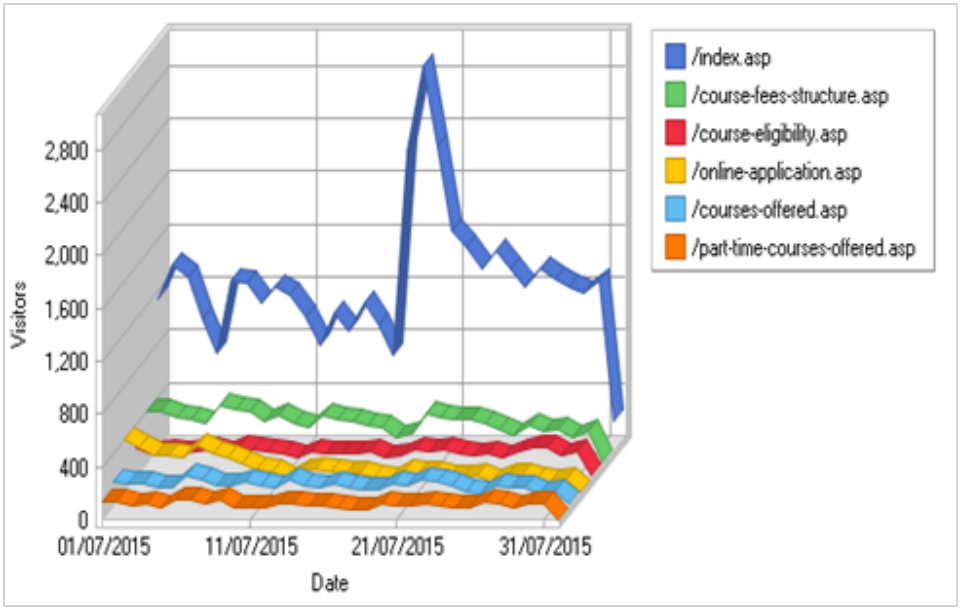

Figure 3. Frequent page access.

- Bugs: The browser sent requests from a web server, an error may occur. The last feature shows the different kinds of errors occurred while accessing the website for the error feature both a tabular and a graphical form of representation are available. They are listed and explained in reverse order, the six most common HTTP errors.

This error type information given in Figure 4 represents that every web browser shows the words and errors differently. A credential or forbidden error threatening for various web browsers[8].

5. Conclusion

An important research area in web mining is web usage mining which mainly aims on the discovery of patterns in the browsing and navigation data of web users. There are several techniques and tools proposed by different researchers for the web usage mining. In this paper, we discussed about education domain web user pattern discovery and pattern analysis using Web Log Expert. The results obtained should absolutely help the website Analysts, Website Maintainers, Website Designers and Developers. It manages their system by analyzing occurred errors, corrupted and broken links. In our sample dataset we identified the university web portal which is more accentuated on educational links rather than with the individual university links. Since this is a big area and there is lots of work to do be done, this paper could be very useful for as a starting point and in identifying opportunities for the further research.

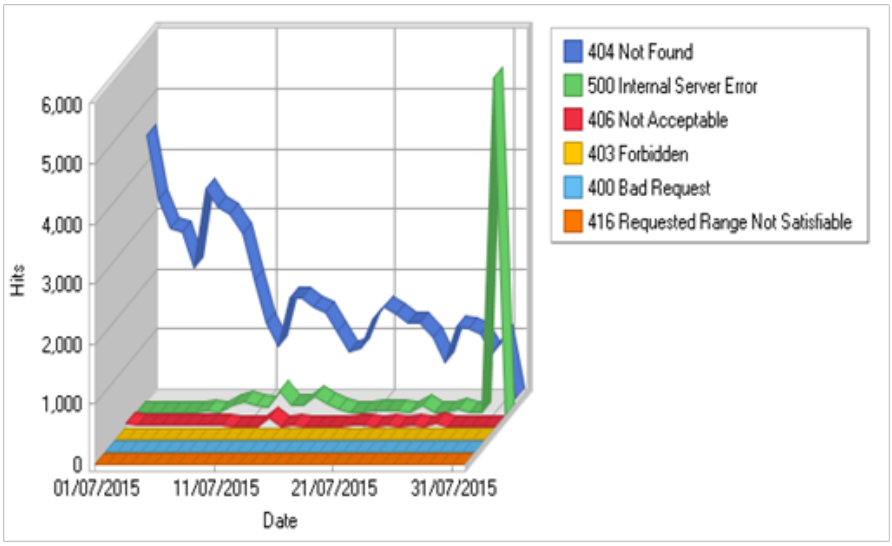

Figure 4. Daily error types.

6. References

1. Victor SP, Rex MMX. Analytical implementation of web structure mining using data analysis in educational domain. International Journal of Applied Engineering Research. 2016; 11(4):2552–6.
2. Asadianfam S, Mohammadi M. Analyzing user behavior from Web Access Logs using automated log analyzer tool. International Journal of Compute Application. 2013; 62(2):1–5.
3. Parthiban P, Selvakumar S. Big data architecture for capturing, storing, analyzing and visualizing of Web Server Logs. Indian Journal of Science and Technology. 2016; 9(4):1–9.
4. Arvind K, Gupta SG. Analysis of web server log files to increase the effectiveness of the website using web mining tools. International Journal of Advanced Computer and Mathematic Sciences. 2013; 4(1):1–8.
5. Russell-Rose T, Clough P. Mining search logs for usage patterns. Text Mining and Visualization: Case Studies using Open-Source Tools. 2016; 40:153.
6. Forsstrom D, Hesser H, Carlbring P. Usage of a responsible gambling tool: A descriptive analysis and latent class analysis of user behavior. Journal of Gambling Studies. 2015; 32(3):889–904.
7. Tandele K, Pansare B. Web usage mining with improved frequent pattern tree algorithms. International Journal of Computer Science and Information Technology Research. 2015 Apr-Jun; 3(2):952–8.
8. Zaiance OR. Web usage mining for a better web-based learning environment. Proceedings of Conference on Advanced Technology for Education; 2001.