Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Usage Mining Using Web Log Expert Tool

Arti Tyagi
AIM&ACT, Banasthali University
Rajasthan, India

Sunita Choudhary
Computer Science, Banasthali University
Rajasthan, India

Abstract— Web Usage Mining is the part of the web mining, which uses data mining techniques to extract and discover knowledge from the web data (web structure, web content and web usage data).Web Usage Mining uses data mining techniques for the discovery of usage patterns from data extracted from Web Log files. It discovers the usage behaviour of the Website users. In our paper analysis of web server log files is done to increase the effectiveness of the website by using web usage mining tool. In this paper, the complete analysis of web server log files has been done by using WebLog Expert tool.

Keywords— Web Mining, Web Usage Mining, Web Logs, Web Log Analyzer, Web Log Expert

I. WEB USAGE MINING

Web usage mining is the field of web mining, which is also the part of the data mining. Web usage mining is the application of data mining techniques to discover and extract the information

hidden in the web server log file. The extracted information is user access patterns and used for analysing users behaviour patterns. Understanding the frequently access patterns of the users allows the website owners to manage and improve the website accordingly in order to improve web based applications. Analysing the web usage log data web mining systems can discover knowledge about users' interest and systems usage characteristics.

When users uses a website, their browsing behavior is stored in web server logs which are stored in the server. These log files contain important information regarding users experience in the site. This information is crucial for companies and their internet/intranet based applications. They use the analyzed reports of those patterns for different purposes. The applications generated from this analysis can be classified as personalization, system improvement, site modification, business intelligence and usage characterization. The work of Web usage mining is to capture, analyze and model the Web server logs.

Steps followed in web usage mining are

1. Data collection – Collection of different web log files from the server.
2. Data Integration – Integrating different web log files into one file.
3. Data preprocessing – Cleaning and structuring data for the processing.
4. Processing – Interesting patterns are extracted from the web log file with the help of the web log analyzing tool.
5. Pattern analysis and visualization – Analyze the extracted pattern.

II. APPLICATIONS OF WEB USAGE MINING

The results produced by the mining of web logs can used for various purposes .They are listed below:

1. To improve the website design or structure.
2. To satisfy the user requirements.
3. Personalization of Web Content: Web Usage Mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behavior by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users. [1]
4. Prefetching and Caching: The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. Typically, Web Usage Mining can be used to develop proper prefetching and caching strategies so as to reduce the server response time as done in. [1]
5. Support to the Design: Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications. [1]

easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

III. WEB SERVER LOG FILES

The server log files are simple text files which records activity of the users on the server. These files reside on the server. If user visits many times on the Website then it creates entry many times on the Server. The main source of raw data to the web usage mining process is the web access logs which are known as web server log files. The log files can be analyzed over a time period. The time period can be specified on hourly, daily, weekly and monthly basis. The typical web server log files contain such type of information: IP address, request time, method (e.g. GET), URL of the requested files, HTTP version, return codes, the number of bytes transferred, the referrer's URL and user agents.[2]

Contents of a Log File

Web log file reside on the server. If user visits many times on the website then it creates entry many times on the server. The Web log file has been containing the following key fields:

(i) Visiting Path– Paths which follow by the user to visit on the Website.

(ii) User Name– Identify the user through IP addresses which provide by ISP. It is temporary address.

(iii) Success Rate– It is user activity which is done on the Website that is number of downloads and number of copies.

(iv) Path Traversed– The path identifies who is visit on the website through user.

(v) Last visited Page– It stores the last record that is visited by the user.

(vi) URL of the Web page accessed– It may be HTML page and CGI program. This is accessed through the user.

(vii) Request Method (GET or POST): This is a method which is performing on the Website like GET and POST.

The above mentioned are the key fields present in the log file. This log file details are used in case of Web usage mining process. According to Web usage mining, it mines the highly utilized Website. The utilization may be the frequently visited Website or the Website being utilized for longer time duration. Therefore the quantitative usage of the website can be analysed, if the log file is analysed. [3]

IV. WEB USAGE MINING WITH WEB LOG ANALYZER TOOL

Web log analysis tools help you to see a wide variety of statistical information about visitors and traffic to your website. These tools turn basic server text log files into graphical formats that are easy to understand. Most web log analysis tools tell you how many visitors your website receives in a given period of time, what web browsers are used most often by your visitors and which pages on your website the visitors viewed. These types of statistics are valuable, because they can help you to learn how your website is performing and determine the positive and negative results of recent design changes or marketing efforts. [4]

Information Obtained By Analysis Tools

i) Number of Hits: This number usually signifies the number of times any resource is accessed in a Website. A hit is a request to a web server for a file (web page, image, JavaScript, etc.). When a web page is uploaded from a server the number of "hits" or "page hits" is equal to the number of files requested. Therefore, one page load does not always equal one hit because often pages are made up of other images and other files which stack up the number of hits counted.

ii) Number of Visitors: A "visitor" is exactly what it sounds like. It's a human who navigates to your website and browses one or more pages on your site.

iii) Visitor Referring Website: The referring website gives the information or URL of the website which referred the particular website in consideration.

iv) Visitor Referral Website: The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

v) Time and Duration: This information in the server logs give the time and duration for how long the Website was accessed by a particular user.

vi) Path Analysis: Path analysis gives the analysis of the path a particular user has followed in accessing contents of a Website.

vii)Visitor IP address: This information gives the Internet Protocol (I.P.) address of the visitors who visited the Website in consideration.

viii) Browser Type: This information gives the information of the type of browser that was used for accessing the Website.

ix) Cookies: A message given to a Web browser by a Web server. The browser stores the message in a text file called cookie. The message is then sent back to the server each time the browser requests a page from the server. The main purpose of cookies is to identify users and possibly prepare customized Web pages for them. When you enter a Web site using cookies, you may be asked to fill out a form providing such information as your name and interests. This information is packaged into a cookie and sent to your Web browser which stores it for later use. The next time you go to the same Web site, your browser will send the cookie to the Web server. The server can use this information to present you with custom Web pages. So, for example, instead of seeing just a generic welcome page you might see a welcome page with your name on it.

x) Platform: This information gives the type of Operating System etc. that was used to access the Website.

V. WEB USAGE MINING WITH WEB LOG EXPERT TOOL

A. Web Log Expert Tool

WebLog Expert is a software application whose sole purpose is to help individuals analyze log files of web servers (such as Apache and IIS) and generate reports regarding specified web pages. **WebLog Expert Lite** is software for Windows-based computers that provides specific, precise information about the site's visitors. Its purpose is to reveal statistics about your website activity.

WebLog Expert is a fast, powerful and feature rich web server log analyzer that will provide you with detailed information about your site visitors. The information include general statistics, activity statistics, accessed files ,statistics about paths taken through the site, information about referring pages, search engines, browsers, operating systems, errors, and much more. The log analyzer features intuitive interface. Built-in wizards will help you quickly and easily create a profile for your site and analyze it. The user interface is straightforward and offers quick access to the main functions of the program. In order to make WebLog Expert analyze the log files, you are required to create a profile. It is possible to add multiple profiles, and to edit or delete them. It is also possible to edit copy and delete profiles and generate a report by clicking the "Analyze" button. After completing this process, a HTML file is going to be opened in a new tab in your default web browser. It will display information such as total hits, average page views per day, graphs displaying daily visitors, top hosts and daily referring sites.

Requirements

- It works under Windows 95/98/Me/NT/2000/XP/2003/Vista/2007/8 and server 2012 versions of operating system.
- It supports Apache and IIS 4/5/6/7 log files.

Features

- It support for custom reports.
- It supports page title retrieval.
- It gives general statistics from web log analysis.
- It can read and analyzes GZ and ZIP compressed logs.
- It generates information about visitors: hosts, top-level domains.
- It generates report about referring sites, URLs and search engines.
- It gives information about browsers, operating systems used by the visitors.
- It also gives information about errors, error types and detailed error information.
- It gives details about activity statistics of the website daily, by hours of the day, by days of the week, by weeks and by months.
- It generates reports about access statistics like statistics for pages, files, images, directories, view time, entry pages, exit pages, bounces, paths through the site, file types.

Benefits

- This software tool supports logs in formats such as LOG, ZIP, GZ and BZ2.
- Web Log Expert has a simple and easy to use and user-friendly interface with wizards.
- WebLog Expert Lite creates an easy-to-read HTML files reports which include text and charts.
- CPU and memory usage is minimal at all times and thus, the system's performance is not going to be hampered in any way.

B. Information Collected by Web Log Expert[2]

- *Number of Hits*– This number usually signifies the number of times any resource is accessed in a Website. A hit is a request to a web server for a file i.e. web page, image, JavaScript, Cascading Style Sheet, etc.
- *Number of Visitors*– A visitor is exactly what it sounds like. It is a human who navigates to the website and browses one or more pages on the website.
- *Visitor Referring Website*– The referring website gives the information or URL of the website which referred the particular website in consideration.
- *Visitor Referral Website*– The referral website gives the information or URL of the website which is being referred to the particular website in consideration.
- *Time and Duration*– This information in the web server logs give the time and duration for how long the website was accessed by the particular user.
- *Path Analysis*– Path analysis gives the analysis of the path to a particular user has followed in accessing contents of a website.
- *Visitor IP Address*– This information gives the IP address of the visitors who visited the website.
- *Browser Type*– This information gives the information of the type of web browser that was used for accessing the website.
- *Platform*– This information provides the type of operating systems or platforms etc. which has been used to access the website.

• *Cookies*– **A** message given to a web browser by a web server. The browser stores the message in a text file called cookie. The message is then sent back to the server each time the browser requests a page from the server. The main purpose of cookies is to identify users and possibly prepare customized web pages for them.

Hit– Each file sent to a web browser by a server is known as an individual hit.

Visit– A visit happens when someone visits the website. It contains one or more page views/hits. One visitor can have many visits to the website. A unique visitor is determined by the IP address or cookie. By default, a visit session is terminated when a user falls on inactive state for more than 30 minutes. If the visitor left the website and came back 30 minutes later, then WebLog Expert will report 2 visits. If the visitor came back within 30 minutes, then WebLog Expert will still report 1 visit.

Page View– A page view is each time a visitor views a web page on the website, irrespective

Of how many hits are generated. Pages are comprised of files. Every image in a page is a separate file. When a visitor looks at a page i.e. a page view, they may see numerous images, graphics, pictures etc. and generate multiple hits. For example– if a web page contains 5 images, a 'hit' on that page will generate 6 'hits' on the web server, one page view for the web page, 5 hits for the images.

C. Analysing Web Server Log with Web Log Expert Lite

Now a day's different web server log analyzers are available but we are using Web Expert Lite 8.5 in our work to analyze sample web server log obtained. The information obtained by analyzing web logs with the help of web log expert lite was as follows:

Total Hits, Visitor Hits, Average Hits per Day, Average Hits per Visitor, Failed Requests, Page Views Total Page Views, Average Page Views per Day , Average Page Views per Visitor, Visitors Total Visitors Average Visitors per Day, Total Unique IPs , Bandwidth, Total Bandwidth , Visitor Bandwidth , Average Bandwidth per Day, Average Bandwidth per Hit, Average Bandwidth per Visitor. Access Data like files, images etc., Referrers, User Agents etc.

IV. EXPERIMENT

In this work, we have been used web log data from December 8, 2007 to December 15, 2007 collected from the web server of the website *www.smsync.com* have been analyzed by using WebLog Expert Lite 8.5 web mining too. The complete experiment has been done on the basis of web log data. The general activity statistics of the website usage is shown in Figure-1.

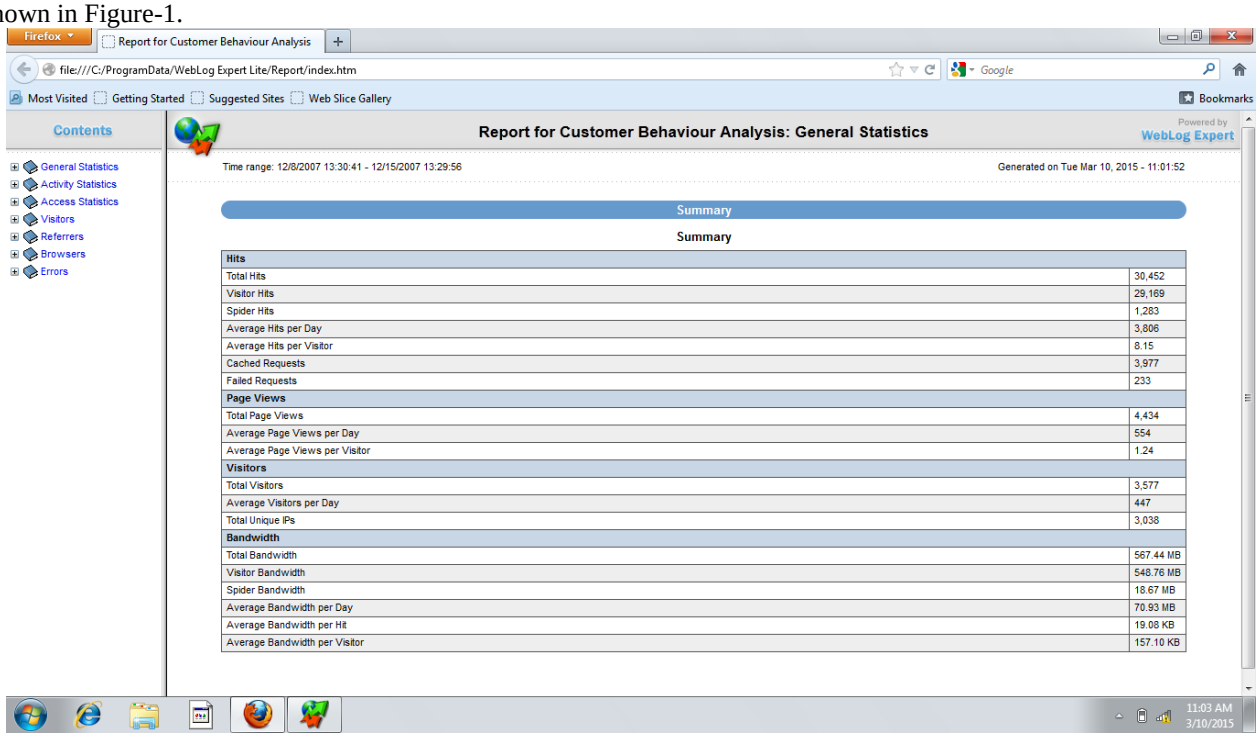
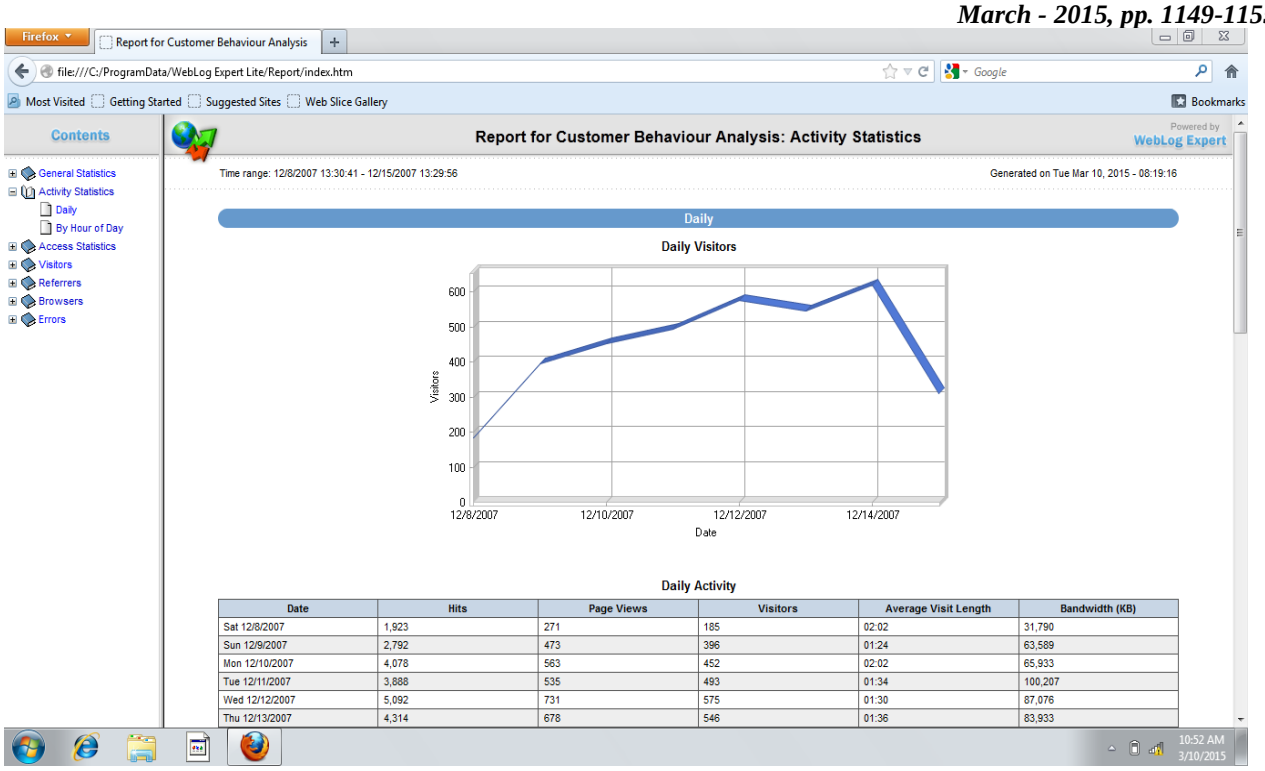


Figure-1

Based on the analyzer report, we have been found several necessary records like total hits, total cached hits, average hits per day, average hits per hour, average hits per visitor, average data transfer per hits, total visitors, average visitors per day, average time spent, average page views per visitors, average downloads per visitors, average data transfer per visitor, visitors who visit once, visitors who visit more than once, average page views per day, total files downloads, average files downloads per day, total data transferred and average data transfer rates have been found.

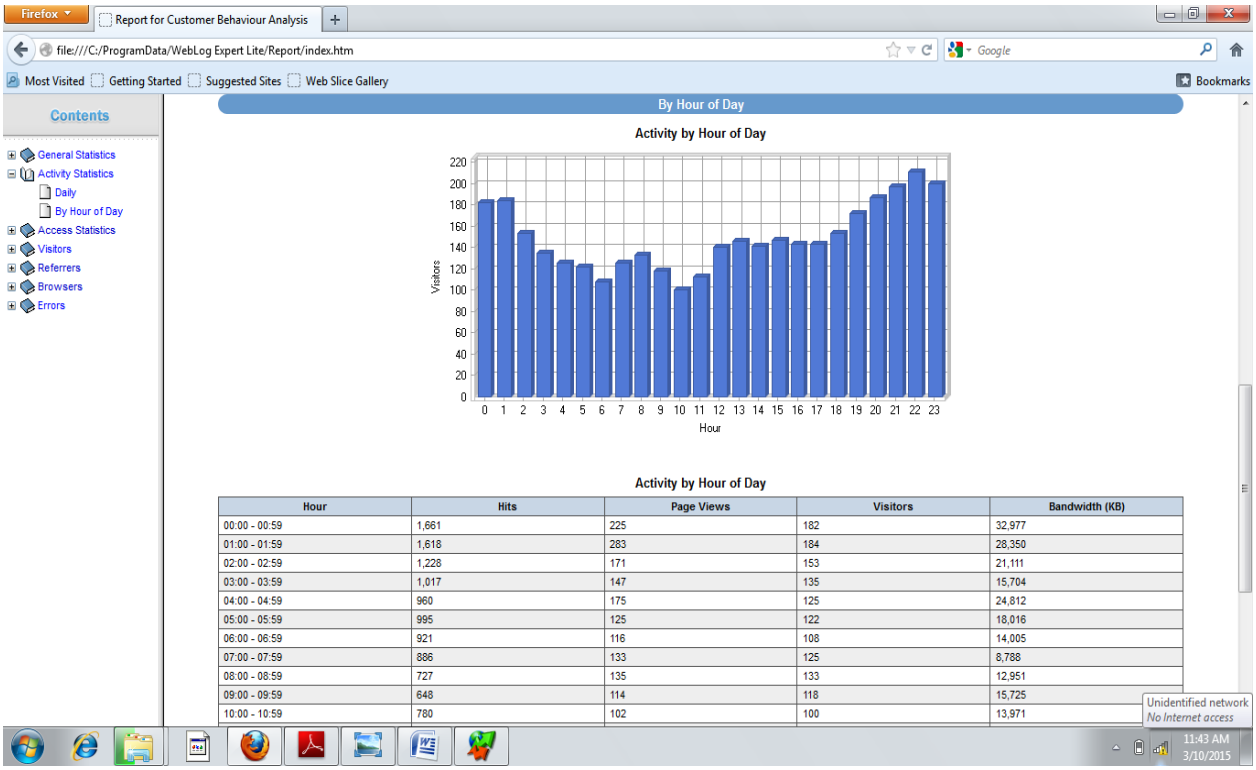
Figure 2 shows the daily visit report of the website visitors. This summary report produced daily usage activity such as total hits of website visitors, hits per day, total page views, page views per day, total visitors, visitors per day, total time spent, data transfer per day and total data transfer on the Website. This report shows day wise total number of visitors or users who are visited the Website. From this statistics, it will be helpful to identify the number of visitors of the Website.



Date	Hits	Page Views	Visitors	Average Visit Length	Bandwidth (KB)
Sat 12/8/2007	1,923	271	185	02:02	31,790
Sun 12/9/2007	2,792	473	396	01:24	63,589
Mon 12/10/2007	4,078	563	452	02:02	65,933
Tue 12/11/2007	3,888	535	493	01:34	100,207
Wed 12/12/2007	5,092	731	575	01:30	87,076
Thu 12/13/2007	4,314	678	546	01:36	83,933

Figure-2

Figure 3 shows the report of hourly website visitors. The table in the report displays the accurate hourly visitor's activity statistics of the website usage. This summary report produced hourly usage activity such as total hits of website visitors, hits per hour, total page views, page views per hour, total visitors, visitors per hour, total time spent, data transfer per visitor per hour and total data transfer on the Website. From this, it is concluded that the output of this phase plays a major role in predicting the best frequent patterns, which are the foremost information for improving the Website usability and identifying the number of visitors of the Website.



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	8,788
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

Figure-3

Figure-4 and Figure-5 shall definitely help the Website Maintainers, Website Analysts, Website Designers and Developers to manage their System by determining occurred errors, corrupted and broken links. This work will also increase the effectiveness of the Website.

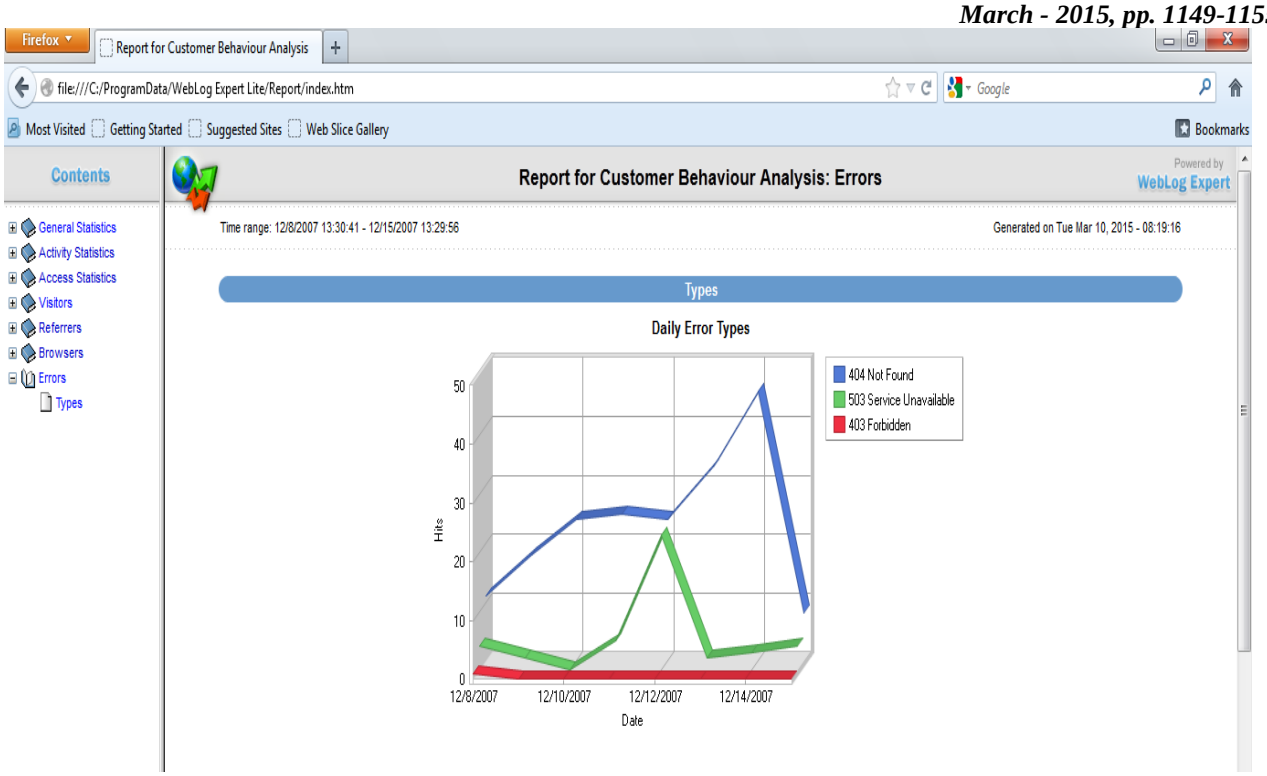


Figure-4

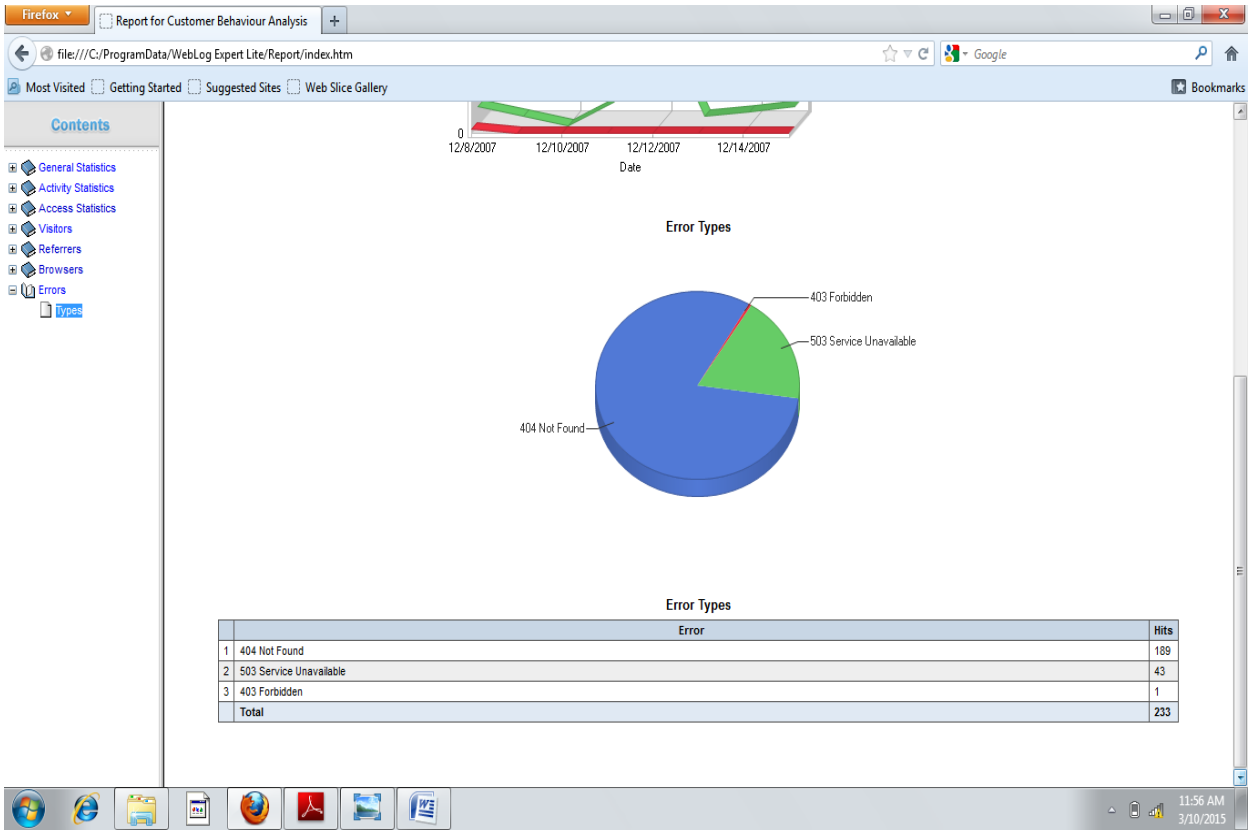


Figure-5

VI. POSSIBLE ACTIONS

i) Most frequently pages accessed by the users follow a particular path. These pages can be put in the website in an easily accessible part which results in the decrease in the navigation path length.

ii) Less frequently accessed pages by users can be either removed or their content can be put in the pages which are highly or moderately accessed by the users..

iii) Redesigning of the Pages to help User Navigation.

iv) Important and business critical advertisements will be put on pages which are frequently accessed by the site's visitors.

v) Removal or updating of broken or corrupted links. This work will also increase the effectiveness of the Website.

VII. CONCLUSIONS

Web is a huge collection of web pages stored in the different web servers. The size of the web increases rapidly as well as the number of its users. So, it's become very important for the website owners to better understand their customers and their needs with the help of this information they are able to give them better service. For this purpose, web logs are proved very useful as they contained the information about the website visitors and their activity stored on the server. Mining the information from the web logs and analysing this information is useful in understanding the user behaviour. Our complete work has accomplished by analysing web log data with the help of Web Log Expert Lite tool. From our work and its results we can say that overall WebLog Expert can be considered as an excellent tool that comes packed with all the necessary features for analysing log files .Our experimental results and their analysis will proved web usage mining as a powerful technique which will help website developers, website designers, website maintainers and website analysts to manage and improve their website by restructuring and redesigning of the website.

REFERENCES

[1] Ankita Kusmakar, Sadhna Mishra ,*Web Usage Mining: A Survey on Pattern Extraction from Web Logs* , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013 ISSN: 2277 128X,Page:-834-838.

[2] Arvind K. Sharma and P.C. Gupta, "*Analysis of Web Server Log Files to Increase the Effectiveness of the Website using Web Mining Tool*, International Journal of Advanced Computer and Mathematical Sciences,ISSN 2230-9624, Vol 4, Issue 1, 2013, pp1-8.

[3] Arvind K. Sharma, *A Comparative Study between Web Mining Tools over some WUM Algorithms to Analyze Web Access Logs*, International Journal of Innovative Technology and Exploring Engineering (IJITEE),ISSN: 2278-3075, Volume-1, Issue-1, June 2012.

[4] V. Jayakumar and Dr. K. Alagarsamy, *Analysing Server Log File Using Web Log Expert in Web Data Mining*, International Journal of Science, Environment and Technology, ISSN 2278-3687 (O),Vol. 2, No 5, 2013, 1008 – 1016.

[5] Yogish H K , Dr. G T Raju, Manjunath T N, *The Descriptive Study of Knowledge Discovery from Web Usage Mining* ,International Journal of Computer Science and Software Engineering, Issues, Vol. 8, Issue 5, No 1, September 2011 ISSN (Online): 1694-0814.

