International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 01, January 2017
www.ijates.com
ijates
ISSN 2348 - 7550

ENHANCING THE ACCESSIBILITY AND USABILITY OF WEBSITE USING WEB LOG ANALYSIS

Satya Prakash Singh[1], Meenu[2]

[1,2]*Dept. of Computer Science & Engineering ,*

Madan Mohan Malaviya University of Technology, Gorakhpur (U.P.) (India)

ABSTRACT

The basic role of web usage mining is to capture, analyze, and the web server logs. Usually it personally discovers the usage behaviour of the Website users. In this paper, we have been implemented a web mining tool to analyze the web server log file of the Website. It evaluates about total hits, page views, visitors, top errors, web browsers used by the website users mostly. The get information shall definitely increase the effectiveness of the website.

Keywords: Web mining, Web Access Logs, Web Usage mining, Access Log Analyzer, World Wide Web, Weblog Expert

I. INTRODUCTION

The internet during the past few years, the World Wide Web (WWW) has become most famous and drastic platform to store, propagate and retrieve information as well as mine useful knowledge. It is a way of communication and information dissemination and it server as a platform for exchanging several types of information. Web usage mining also known as web log mining is the application of data mining techniques of discover interesting usage patterns from web data to understand and better serve the need of web based applications. Usage data hold the identify or filiation of web users along with their browsing conduct at a web site. Web Usage Mining consists of four process, first web server log, second preprocessing, third pattern discovery and fourth pattern analysis. After the consummation of these four process, the user can find the required usage pattern and user this information for the specific need [9] in a variety of way such as improvement of the application, recognizing the visitor's conduct, customer attraction, customer tenure etc. Web usage mining is the third category in web mining. This types of web mining permits for the accumulation of web access information for web pages. Web usage data provides the paths leading to accessed web page. The user logs are collected by the web server. Typical data includes IP address page reference and access time.

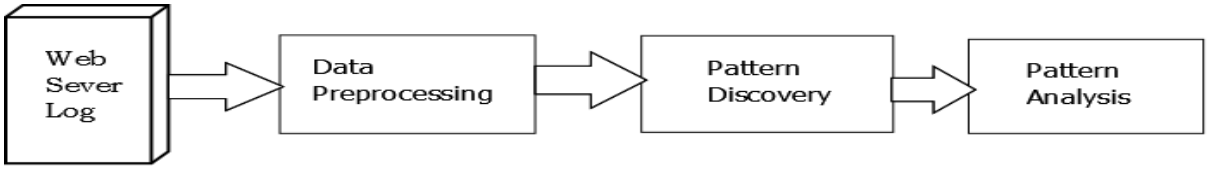


Fig. Phases of web usage mining

Fig.1 Phases of Web Usage Mining

1. Data collection- User log are collected from client and server side servers, proxy server, application etc.
2. Data Preprocessing- Consists phases like data fusion and cleaning, user identification, session identification, path completion.
3. Pattern Discovery- Pattern discovery is the discovering path from preprocessed data using several data mining techniques like statistical analysis, association, clustering and pattern matching etc.
4. Pattern Analysis- Pattern analysis is the once pattern is discovered analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP.

Web usage mining is applying to many actual world problems to discover interesting user navigation patterns form improvement of the website design by complying user or customer conduct from log file [8]. The goal to discover and retrieval useful and interesting pattern from a large dataset. Web usage mining consist of four phases such as Data collection, preprocessing of log data, pattern discovery and pattern analysis shown in fig. 1[7]. In the first phase the data collected from web log files. There are three types of Log data namely web server Logs, web proxy server and client browser. Second phases preprocessing is required to pretermission irrelevant information from original log file and to make the web log file easy for session and user identification process. The main purpose of preprocessing is to improve the quality and accuracy of data. Figure.2 shows the phases data preprocessing in web log mining [14].

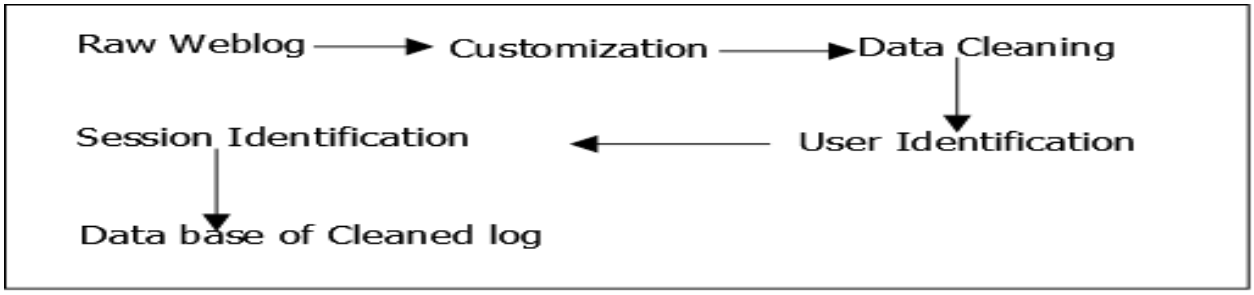


Fig. Phasecof data Preprocessing in web usage minig

Fig2. Phase of Data Preprocessing in Web Usage Mining

II. PROPOSED METHODOLOGY

There is much necessity to study web user conduct to better serve the users and increase the value of institution or enterprises. The web site design is currently based on entire investigation about the interest of web site visitors and investigated supposition about their shrewd conduct. The analysis of web log data allows to identify useful pattern of the browsing conduct of users, which exploited in the process of navigation of conduct. Web log data holds web browsing conduct of user from a web site. Educational institution is good example that develop web site. This paper present visitor pattern analysis performed through educational institution web log data.

Determine the usability of the website, including the-

1. Visitor pattern analysis
2. Page view analysis

3. Time analysis
4. Origin of the web site

In this work, the web log expert reports of the time analysis and page view analysis. The time analysis at the different time of day, day of week, and days of month that the website receive the most visitor.

2.1 Data Collection

In this study, the server access log data has been collected from an educational institution. The web log data contains the information of five days from: 2016-12-31 00:22:33. During this period, 1.80MB data was transferred. There are 7957 entries in log file. In our work, we have a server log file of a college, which is of a common log format. Figure 3 shows the example of log file entries.

```
#Software: Microsoft Internet Information Services 7.0
#Version: 1.0
#Date: 2016-12-31 00:22:33
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(User-Agent) cs(Cookie) cs(Referer) cs-host sc-status sc-substatus sc-win32-status sc-bytes cs-bytes time-taken
2016-12-31 00:22:33 W3SVC772 SVMSVR1 184.172.12.53 GET / - 80 - 157.55.39.115 HTTP/1.1 Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm) - - www.alumni.mmmut.ac.in 200 0 0 26191 277 1877
#Software: Microsoft Internet Information Services 7.0
#Version: 1.0
#Date: 2016-12-31 01:54:01
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(User-Agent) cs(Cookie) cs(Referer) cs-host sc-status sc-substatus sc-win32-status sc-bytes cs-bytes time-taken
2016-12-31 01:54:01 W3SVC772 SVMSVR1 184.172.12.53 GET /Pdf/attendance-09.pdf - 80 - 180.76.15.14 HTTP/1.1 Mozilla/5.0+(compatible; +Baiduspider/2.0;++http://www.baidu.com/search/spider.html) - - www.alumni.mmmut.ac.in 200 0 0 15471 245 594
#Software: Microsoft Internet Information Services 7.0
#Version: 1.0
#Date: 2016-12-31 02:12:57
```

Fig.3. Sample of Log File

2.2 Data Selection

The data is collection from web log file. There are three types of log data first is web server logs, second is web proxy server and third is client browser [6].

2.3 Web Log Data

Web log data is a listing of page reference data sometime it is referred to a clickstream data [7]. The web plays an important role and medium for extracting useful information. The web server log data contains various attributes. These attributes are as follows-

1. **Date-** The date from Greenwich mean time (GMT×100) is recorded for each hit. The date format is (YYYY/MM/DD).
2. **Time-** Time refers of transactions. The time format is the (HH:MM: SS).
3. **Client IP Address-**Client IP Address is the number of computer who access or request the website.
4. **Server IP Address-**Server IP Address is a static IP provided by internet services provides. The IP will be a reference for access the information from the server.
5. **Server Port-** It is a port used for data transaction.
6. **Server Method (HTTP Request)-**The term of request refers to an image pdf, txt, HTML file, movies sound and more.
7. **URL-** URL is a path from the host. It is representing the structure of the web site.
8. **Agent Log-**Agent log is providing data on a user's browser, browser version and operating system.

III.TOOL FOR EXPERIMENT

There is multiplicity of tools available for analyzing a log file and generating the reports. Some are freely available and some are paid. They are of two kinds some of the tools are taking log file as input and other does not import raw logs, they take information of website as input and directly access the visitor's information from website. These tools generate report are provide us with all sorts of information starting from how many hits the site is getting to the number of visitors accessing the site, the IP address, time, zone, URL, OS, browsers of the visitor. Some of the tools are: Google analytics, Stat Counter, Deep Log Analyzer and Web Log Expert [8]. Web log Expert which as freeware. In this paper, we are discussing the result of web log expert lit 9.3 version [11]. Weblog expert is a fast and powerful access log analyzer. The web log expert is installation is quite easy and GUI provided by the tools is highly user friendly. It will be give the information about the site's visitors: activity statistics, accessed file, paths through the site, information about referring pages, search engines, reports that include both text information and charts. Feature of web log expert lite 9.3 are shows below [10].

1. It supports IIS and apache logs.
2. Automatically detects log format.
3. Can read GZ and ZIP compressed logs.
4. Create reports that include text information and charts.
5. It gives the information of General activity, Activity statistics and access activity.
6. Provides a log information a visitor, Browser, errors and referrers.

IV.EXPERIMENTAL RESULTS

In this work, the web log data collected the information of 1 Jan 2017. Data is collected from the web server of the web site of an educational institution. In this paper, we have analyzed the log file by using web log Expert lite 9.3version [11].

4.1 General Activity

The general activity statistics of web site are shown in the table-1 results of general statistics shows that there are 596hits, 521 visitors, 40IPs, 62 page views.

Table-1: General Activity Statistics of the Website Usage

Hits	
Total Hits	596
Visitor Hits	521
Spider Hits	75
Average Hits per Day	298
Average Hits per Visitor	11.84
Cached Requests	27
Failed Requests	59
Page Views	
Total Page Views	62
Average Page Views per Day	31
Average Page Views per Visitor	1.41
Visitors	
Total Visitors	44
Average Visitors per Day	22
Total Unique IPs	40
Bandwidth	
Total Bandwidth	13.14 MB
Visitor Bandwidth	10.63 MB
Spider Bandwidth	2.51 MB
Average Bandwidth per Day	6.57 MB
Average Bandwidth per Hit	22.58 KB
Average Bandwidth per Visitor	247.45 KB

These tools give valuable information in this table which is easy to read and understand. Table shows the information in four types: first Hits, second page view, third visitors and fourth bandwidth. The first of the number of namely Hits: the number of visitor Hits, spider Hits, Average Hits per Day, Average Hits per visitor,

Cached Requests, Failed Requests. Second is general statistics the shows table the number of Average page view per day and per visitor. Third is visitor and which also give the information of Average visitor per Day. Fourth is the Bandwidth visitor, Spider Bandwidth, Average Bandwidth per Day. This state render the information of overall usage accessibility of website.

4.2 Activity Statistics

Activity statistics is shows the daily and hourly activity of the log file

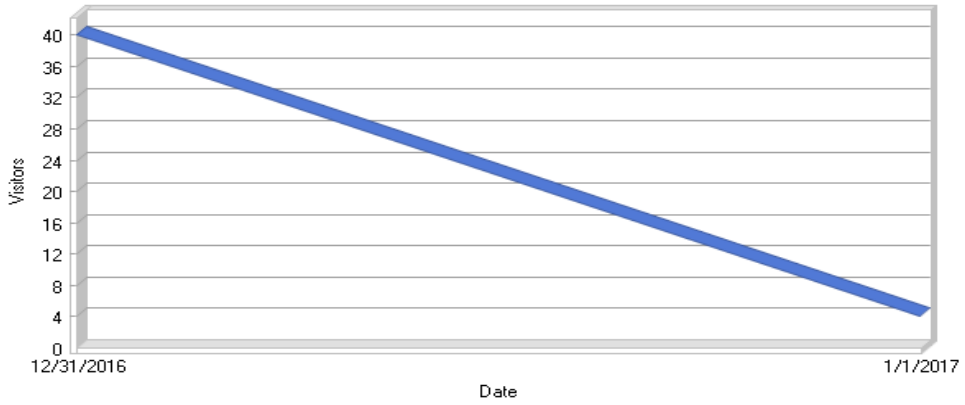


Fig.4. Daily Website Visitors Report

The daily activity of website shown in table 2. The number of hits **596,** page view 62, visitor44, Average visit Length 01:08 and bandwidth, 13456.

Table -2: Daily Activity Statistics of the Website Usage

Date	Hits	Page Views	Visitors	Average Visit Length	Bandwidth (KB)
Sat 12/31/2016	526	30	40	01:14	12,187
Sun 1/1/2017	70	32	4	00:15	1,268
Total	**596**	**62**	**44**	**01:08**	**13,456**

The number of hits, number of page viewers and in rate of data transferred the best day out of this day of log data in sat 12/31/2016 and sun 1/1/2017.

Activity by Hour of Day

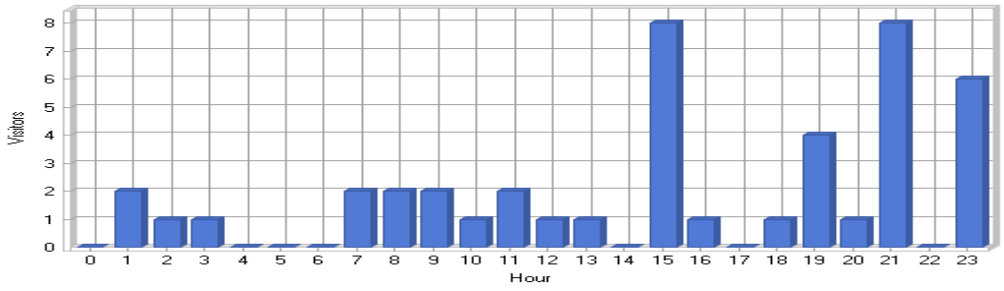


Fig.5. Hourly Website Visitor Report

The hourly basis of report of website visitor in the shown figure.5. It shows number of hits per hour, page views per hour, visitor per hour and bandwidth in KB per hour.

Table-3: Hourly Activity Statistics of the Website Usage

Hour	Hits	Page Views	Visitors	Bandwidth (KB)
00:00 - 00:59	3	0	0	39
01:00 - 01:59	33	1	2	74
02:00 - 02:59	2	1	1	44
03:00 - 03:59	30	30	1	1,083
04:00 - 04:59	2	0	0	26
05:00 - 05:59	1	0	0	25
06:00 - 06:59	0	0	0	0
07:00 - 07:59	35	1	2	761
08:00 - 08:59	19	2	2	607
09:00 - 09:59	15	6	2	101
10:00 - 10:59	35	2	1	750
11:00 - 11:59	35	2	2	748
12:00 - 12:59	33	1	1	726
13:00 - 13:59	52	1	1	1,008
14:00 - 14:59	1	0	0	29
15:00 - 15:59	73	3	8	2,610
16:00 - 16:59	32	1	1	746
17:00 - 17:59	1	0	0	25
18:00 - 18:59	34	1	1	748
19:00 - 19:59	33	1	4	753
20:00 - 20:59	1	1	1	0
21:00 - 21:59	37	6	8	659
22:00 - 22:59	20	0	0	365
23:00 - 23:59	69	2	6	1,518
Total	**596**	**62**	**44**	**13,456**

Show tables the accurate information of hourly activity of web site.

4.3 Access Activity

It is provider of information of the most popular page, most downloaded files and most requested image Figure6. Is the daily page access and Figure 7 shows the result of most popular page of web site after analyzing log file.

Daily Page Access

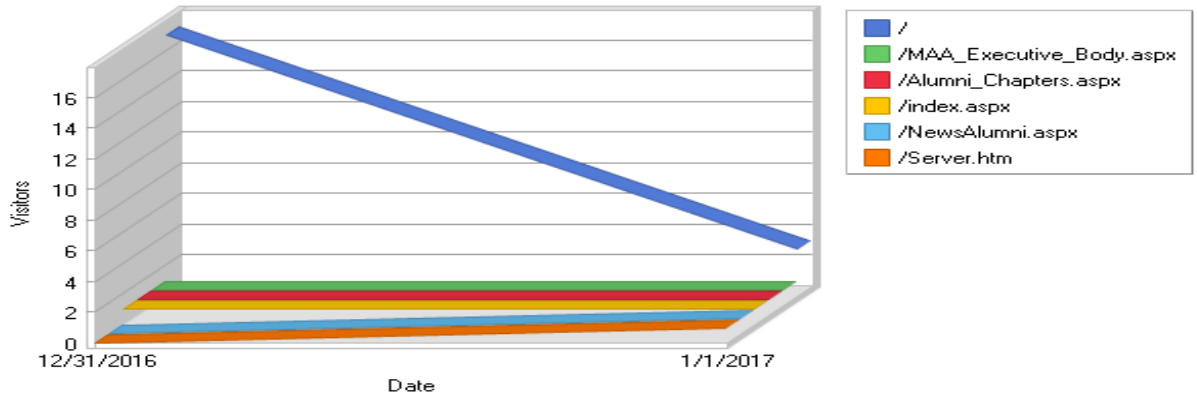


Fig6. Daily Page Access

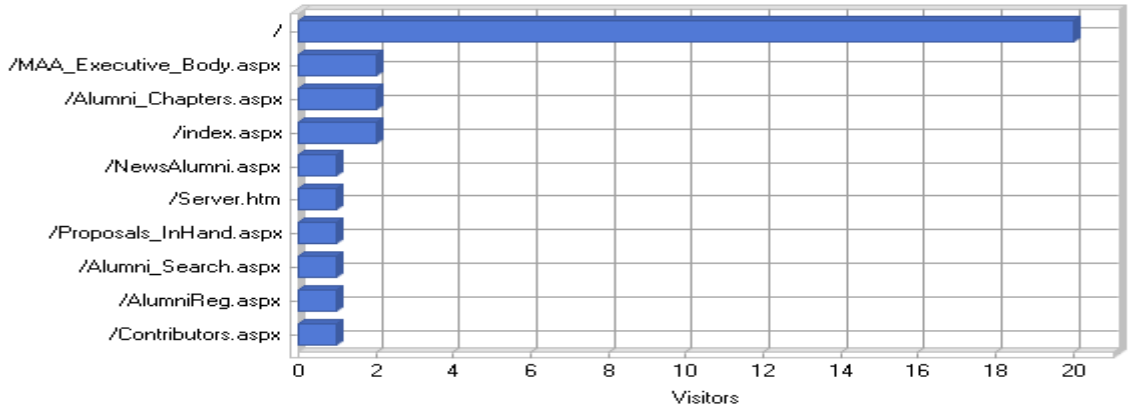


Fig7. Most Popular Page

The topmost files that are downloaded the most number of time from the file page of the web site shown in figure 8.

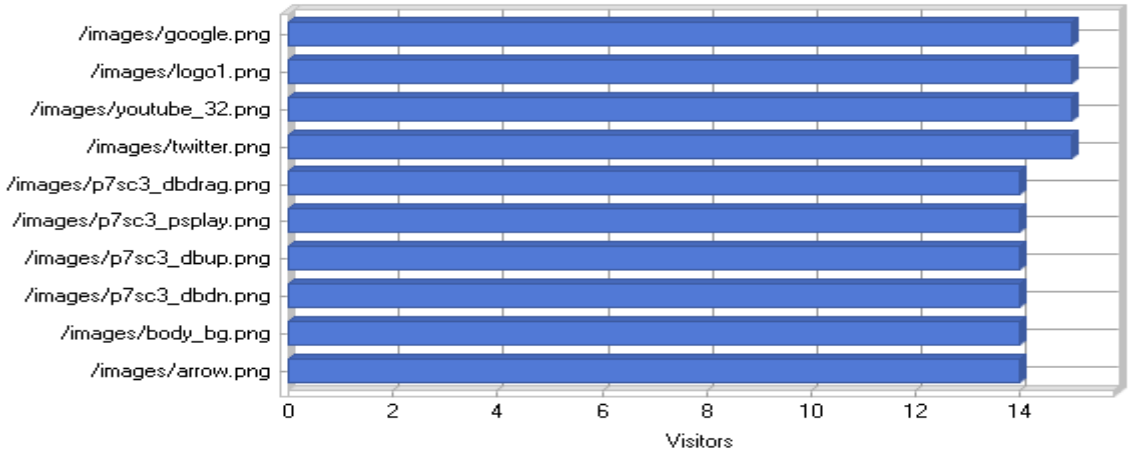


Fig8. Most Downloaded Files

The figure 9 shown the image those have been downloaded most during this month. The most. Next image logo.png.

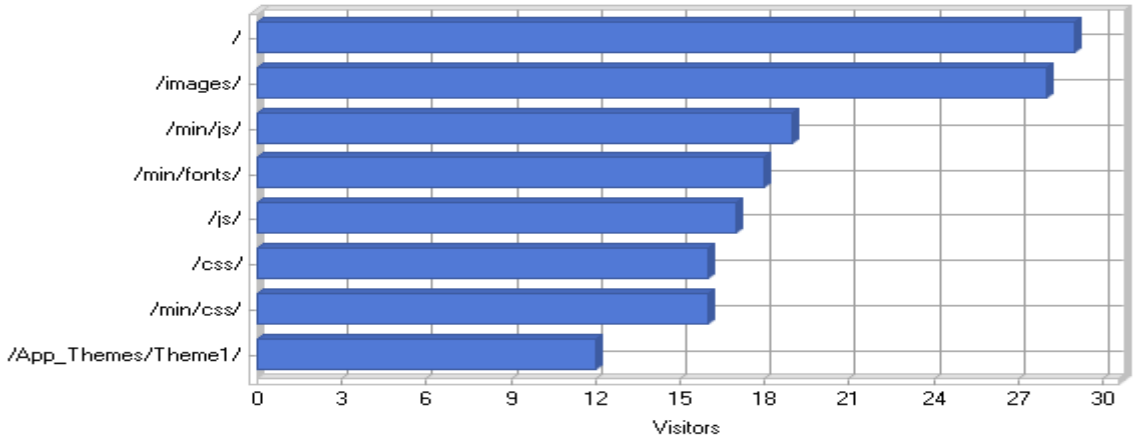


Fig9. Most Requested Image

4.4 Referrer

The search engines used to find the web sites by visitor shown in the figure 10. It shown google is used most out of other search engines.

Fig10. Google Search Engines

Table-4: Search Engines Used to Search the Web Site

	Search Engine	Visitors
1	Google	2
	Total	**2**

4.5 Browsers

This state of analyzer process the information of browser and operating system used by the visitor of the web site.

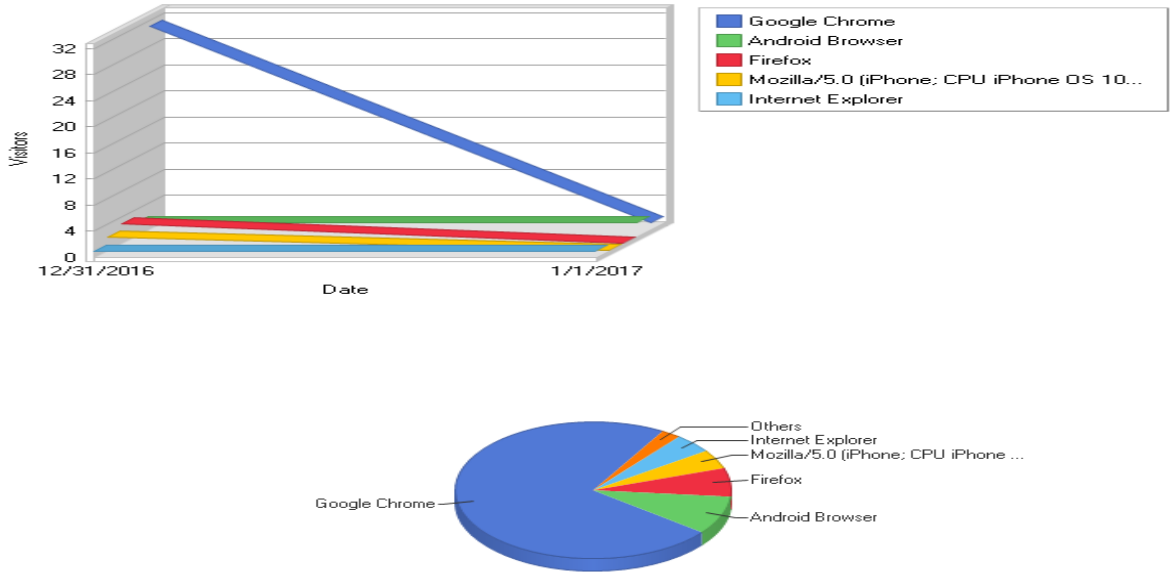


Fig11. Most Used Browsers

The chart of operating system daily used by the visitor. Hits for window7 are more than any other operating system.

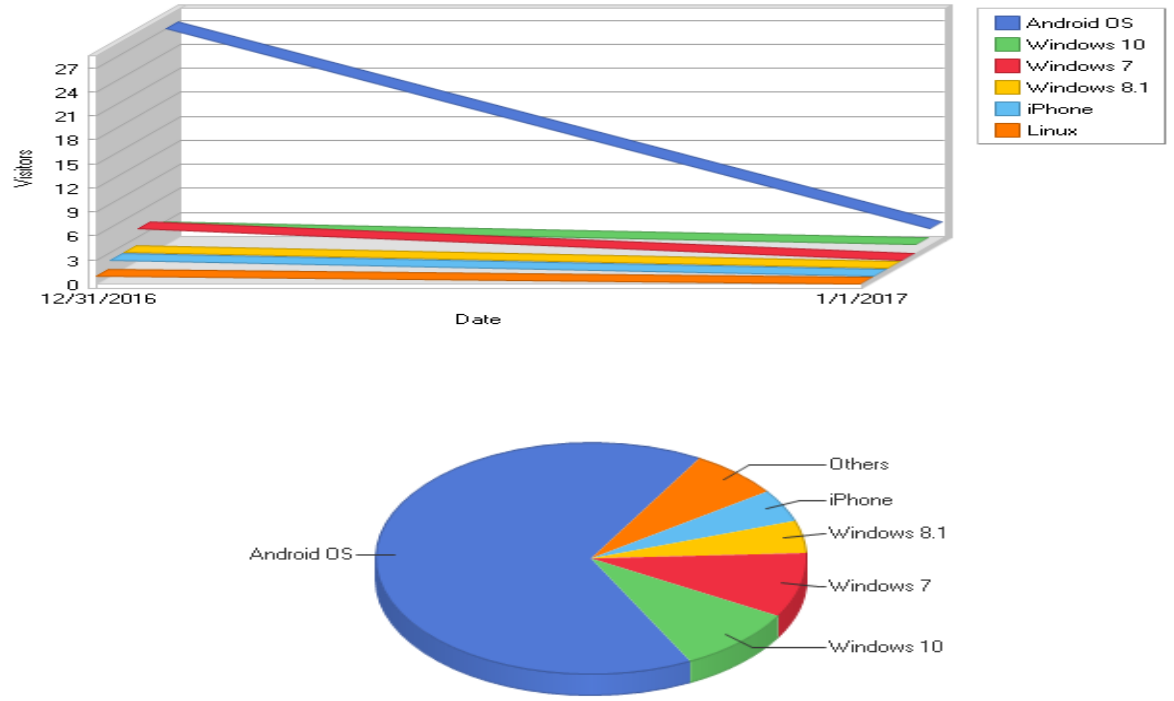


Fig 12 Most Used Operating System

Table 5. Shows the number of visitor who used a particular operating system. The results show that 841 visitors use Android OS,229 use window 7,35 use window xp,78 use window10,124 use other.

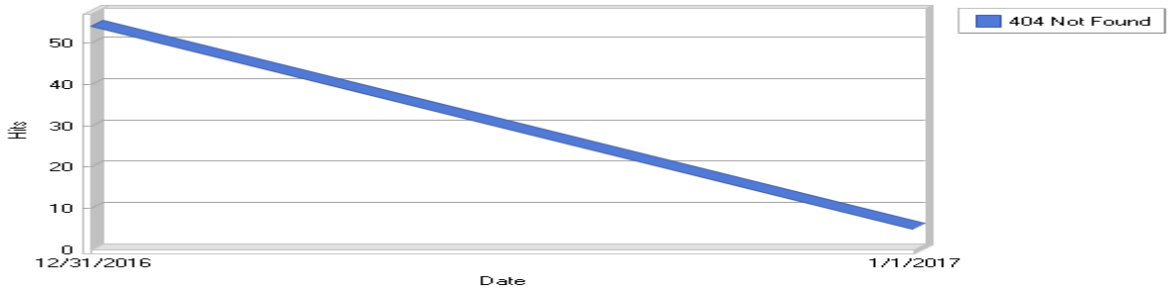
Table5. Most Used Operating System

	Operating System	Hits	Visitors	% of Total Visitors
1	Android OS	269	29	65.91%
2	Windows 10	43	4	9.09%
3	Windows 7	103	4	9.09%
4	Windows 8.1	35	2	4.55%
5	iPhone	34	2	4.55%
6	Linux	1	1	2.27%
7	Windows XP	30	1	2.27%
8	Others	6	1	2.27%
	Total	**521**	**44**	**100.00%**

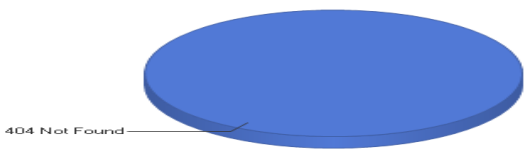
4.6 Errors

When a request is made to server for a page on site, server returns an HTTP statues code in response to the request.

Daily Error Types



Error Types



- **Fig.13 Types of Errors**

Table 6. shows in our log data only two types of errors have been deleted one is error no. 404 and another is 400.

Table 6, Types Errors

	Error	Hits
1	404 Not Found	59
	Total	**59**

V. CONCLUSION

Web is one of the extremely used interface to access remote data, commercial and non-commercial services. Web usage mining is a growing area with the growth of the web applications to find the web usage patterns. The web mining usage pattern of an educational institution web data. There are three types of web related data namely web log, access log, error log and proxy log data and collect the data in web server and implemented a web log expert. Our experimental results help to predict and identify the number of visitor for the website and improve the website usability.

REFERENCE

[1] Mahendra Pratap Yadav,Pankaj Kumar Keserwani and Shefalika Ghosh Samaddar (2012),"An EfficientWeb Mining Algorithm for Web Log Analysis: E-Web Miner", IEEE.

[2] R. Cooley, B. Mobasher and J. Srivastava (1997), "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence ICTAI.

[3] Castellano, G., A. M. Fanelli, and M. A. Torsello. "Log data preparation for mining web usage patterns." In *IADIS International Conference Applied Computing*, no. 10000, p. 20000. 2007

[4] Suneetha, K. R., and Raghuraman Krishnamoorthi. "Identifying user behavior by analyzing web server access log file." IJCSNS International Journal of Computer Science and Network *Security* 9, no. 4 (2009): 327-332.

[5] Chitraa, V., Dr Davamani, and Antony Selvdoss. "A survey on preprocessing methods for web usage data." *arXiv preprint arXiv:1004.1257* (2010).

[6] Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." In Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol. 1, pp. 399- 403. IEEE, 2011.

[7] Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." In Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol. 1, pp. 399- 403. IEEE, 2011.

[8] Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." In Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol. 1, pp. 399- 403. IEEE, 2011.

[9] Mele, Ida. "Web usage mining for enhancing search-result delivery and helping users to find interesting web content." In Proceedings of the sixth ACM international conference on Web search *and data mining*, pp. 765-770. ACM, 2013.

[10] S. R. Aghabozorgi and eh Y. Wah, "Using incremental fuzzy clustering to web usage mining," in IEEE International Conference on Soft Computing and Pattern Recognition, 2009, pp. 653–658 .

[11] Kaur, Navjot, and Himanshu Aggarwal. "Web log Analysis for Identifying the number of visitors and their Behavior to Enhance the Accessibility and Usability of Website." *International Journal of Computer Applications* 110, no. 4 (2015).

[12] Marti Punjani, Mr. Vinitkumar Gupta, 2013 "A Survey on data Preprocessing in Web Usage Mining", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 4, PP 76-79.

[13] Arvind K. Sharma. Gupta, 2013 "Identifying the Number of Visitors to improve Website Usability from Educational Institution Web Log Data" International Journal of Computer Applications Technology and Research Volume 2– Issue 1, 22-26.

[14] Goel, Neha, and C. K. Jha. "Analyzing users' behavior from web access logs using automated log analyzer tool." *International Journal of Computer Applications* 62, no. 2 (2013).

[15] Arvind K. Sharma, P.C. Gupta, 2012 "Enhancing the Performance of the Website through Web Log Analysis and Improvement", International Journal of Computer Science and Technology(IJCST) Vol.3, Issue 4.

[16] Grace, L. K., V. Maheswari, and Dhinaharan Nagamalai. "Analysis of web logs and web user in web mining." *arXiv preprint arXiv:1101.5668* (2011).

[17] Cooley, R.2010, "Web Usage Mining: Discovery and Application of InterestingPatternsfromWebdata",http://citeseer.nj.nec.com/426030.html.

[18] Online] http://www.weblogexpert.com [Accessed on 12/11/2014] .