Web Log Analysis for Identifying the Number of Visitors and their Behavior to Enhance the Accessibility and Usability of Website

Navjot Kaur
Assistant Professor
Department of CSE
Punjabi University Patiala

Himanshu Aggarwal, Ph.D.
Professor
Department of CSE
Punjabi University Patiala

ABSTRACT
Web usage mining is crucial for the CustomerRelationship Management (CRM) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned. Web usage mining is also helpful for identifying or improving the visitors of a particular Website by accessing the log file of that site. In this paper the focus is on Web usage mining of Log data of an educational institution . Web logExpert Lite 8.6 tool has been used for analysis results are shown.

General Terms
Web Log Analyzer, Web Log,Web Usage Mining.

Keywords
Web Mining, Web Usage Mining,WWW, WebLog Expert Lite8.6 , Log File.

1. INTRODUCTION
Web usage mining(WUM) also known as Web Log Mining is the application of Data Mining.WUM techniques are applied on large volume of data to extract useful and interesting patterns from Web data, specifically from web logs, in order to improve web based applications. Web usage mining consists of four phases, data source, pre-processing, pattern discovery, and pattern analysis. After the completion of these four phases the user can find the required usage patterns and use this information for the specific needs[3] in a variety of ways such as improvement of the Web application, identifying the visitor's behaviour, customer attraction, Customer retention etc.

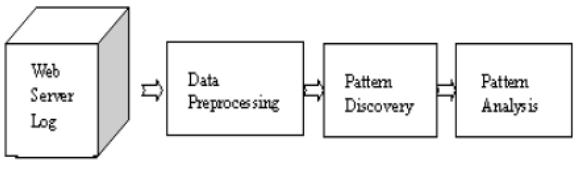

Fig 1: Phases of Web Usage Mining[1]

Web Usage mining is applied to many real world problems to discover interesting user navigation patterns for improvement of website design by observing user or customer behaviour from Log File[5]. The aim is to discover and retrieve useful and interesting patterns from a large dataset. Web Usage mining consist of four phases such as Data collection, Pre-processing of log data, Pattern Discovery and Pattern analysis shown in figure 1[1]. In first Phase the data is collected form Web Log files. There are three types of Log data namely Webserver Logs, Web Proxy Server and Client Browser . In Second Phase Pre-processing is required to eliminate irrelevant information form original log file and to make the web log file easy for Session and user identification process. The main purpose of pre-processing is to improve the quality and accuracy of data.Figure 2 shows the phases of Data Pre-processing in web log mining[9].

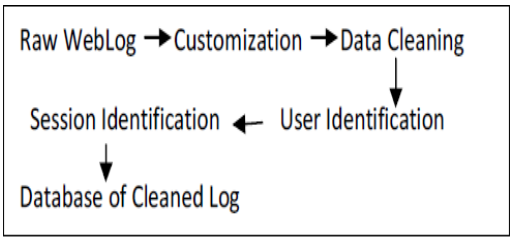

Fig 2:Phases of Data Pre-processing in Web Usage Mining

Next and Third Phase of Web usage mining is Pattern Discovery which means discovering patterns from preprocessed data using various data mining techniques like association, clustering , Statistical analysis and so on[10]. The In the last phase of WUM , Pattern analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations[6].

2. DATA SOURCE
As already mentioned the data is collected form Web Log files . There are three types of Log data namely Webserver Logs, Web Proxy Server and Client Browser[2] . 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 26 Aug 2013 to 30 Aug 2013 . During this period, 1.80 MB 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 our log file entries.

```
117.96.61.194 - - [26/Aug/2014:06:03:30 +
0530] "GET /misc/drupal.css HTTP/1.1" 200
9315 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/537.1 (KHTML, like Gecko)
Chrome/21.0.1180.83 Safari/537.1"
```

Fig3. Sample Of Our Log File

Figure 3. reflects the information of first entry as follows:

- **117.96.61.194** : It is the Remote IP address or domain name which is 32bit host address defined by the internet Protocol.
- **- :** This is remote user. Usually the name of remote user is omitted and replaced by hyphen("-").

- **-** : Login of remote user. Like the name of remote user,Login of remote user is also usually omitted and replaced with hypen("-").
- **[26/Aug/2014:06:03:30 +0530]:** It contain date,time and Zone . First is Date in DD/MM/YYYY] format,Second time which is in HH: MM:SS format and last is zone.
- **"GET/misc/drupal.css HTTP/1.1" :** It contains the Method ,URL relative to domain and Protocol**.** Method could be any one from the "GET or POST or HEAD". "misc/drupal.css" is the URL and HTTP/1.1 is a protocol with version 1.1.
- **200** : This field is for Satus code and 200 code is for sucess.If status code is less then 200 or greater then 299 it means there is error or failure.
- **9315 :** This field shows the content-length of the document transferred in bytes.
- **- :** It is the field of referrer. If person directly accesses the site then this field contain hypen("-").Otherwise it is the URL of the referrer.
- **"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.83 Safari/537.1"** Almost all browsers start with Mozilla Browser type, Netscape Navigator with version 5, "WindowNT 6.1": is the operating system, "WOW64": means a 32-bit Windows is running on a 64-bit processor, "AppleWebKit/537.1": is an unknown fragment, "KHTML": is a free HTML layout engine developed by the KDE project, "like Gecko": is not a Geckeo browser but behaves like a Gecko Browser. Gecko is the open source browser engine designed to support open Internet standards and is used in several browsers like Firefox, SeaMonkey and other, "Chrome/21.0.1180.83": was a Beta Channel Update for Windows or Stable Channel Update for Windows, "Safari/537.1":unknown fragment

3. ACCESS LOG ANALYZER

There are variety of tools available for analyzing a log file and generating the reports. Some are freely available and some are paid. They are of two types. 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 visitors information from website. These tools generate reports and 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. The reports also shows the length of visitor stay on the page and much more. Some of the tools are : Google analytics,Stat Counter,Deep log analyzer and Web Log Expert[5].

Web log Expert which is freeware .In this paper we are discussing the results of Web Log Expert Lite 8.6 version[7]. It is a fast and powerful access log analyzer. The installation is quite easy and GUI provided by the tool is highly user friendly. It will give you the information about your site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, Errors and more. It generates reports that include both text information (tables) and charts. Features of web log expert lite 8.6 are shown below[4]

- It support IIS and apache logs
- Automatically detects log format
- Can read GZ and ZIP compressed logs
- Create reports that include text info and charts
- It gives the information of General Activity ,Activity Statistics and Access Activity.
- Provide a lot information on visitors ,Browsers ,errors and referrers

4. RESULTS

In this work the web log data contains the information of five day period from 26 Aug 2014 to 30 Aug 2014. Data is collected from web server of the website of an Educational institute. As we discussed earlier there are a variety of web log analyzer tools available, some of which are freeware. In this paper we have analyzed the log file by using WebLog Expert Lite 8.6 version[7].

4.1 General Activity

The general activity statistics of the website are shown in table-1. Results of general statistics shows that there are 7960 hits, 852 visitors, 990 IPs, 1662 page views.

Table-1: General Activity Statistics of the Website Usage

Hits	
Total Hits	7,960
Visitor Hits	6,910
Spider Hits	1,050
Average Hits per Day	1,592
Average Hits per Visitor	8.11
Cached Requests	213
Failed Requests	909
Page Views	
Total Page Views	1,662
Average Page Views per Day	332
Average Page Views per Visitor	1.95
Visitors	
Total Visitors	852
Average Visitors per Day	170
Total Unique IPs	990

Bandwidth	
Total Bandwidth	130.78 MB
Visitor Bandwidth	118.96 MB
Spider Bandwidth	11.82 MB
Average Bandwidth per Day	26.16 MB
Average Bandwidth per Hit	16.82 KB
Average Bandwidth per Visitor	142.97 KB

This tool gives valuable information in this table which is easy to read and understand. Tables shows the information in four parts: Hits, Page views, Visitors and Bandwidth. First part of table shows the number of namely Hits: the number of Visitor Hits, Spider Hits, Average Hits per Day, Average Hits per Visitors, Cached Requests, failed requests. Second part of general statistics table shows the number of average page views per day and per visitor. Third part is about Visitors and which also give the information of average visitors per day. Fourth part is about Bandwidth which also shows the visitor Bandwidth, Spider Bandwidth, Average Bandwidth per day, Average bandwidth per hit and average bandwidth per visitor. This phase gives you the information of overall usage accessibility of website.

4.2 Activity Statistics

Activity statistic shows the daily and hourly activity of the log file. If your log file contains data pertaining to two or more then two months time period then this tool will also shows the weekly and monthly activity of the visitor. Figure 4 shows the daily visit report of the website visitors. The accurate information of daily visitors is shown in table 2.

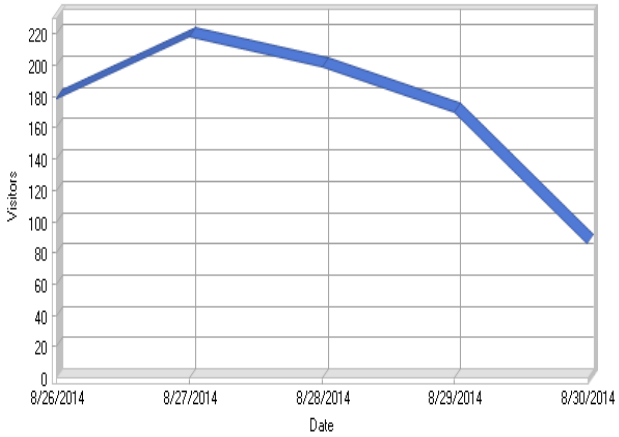

Fig4: Daily Website Visitors Report

In daily activity of website shown in table 2 shows the number of hits per day, page views per day Visitors per day, Average visit length per day and Bandwidth. It also shows the total number of hits 7960, page views 1662, visitors 852 , Average Visit Length 1:20 and bandwidth 133917 kilobytes.

Table -2: Daily Activity Statistics of the Website Usage

Date	Hits	Page Views	Visitors	Average Visit Length	Bandwidth (KB)
Tue 8/26/2014	1,677	259	179	01:28	29,210
Wed 8/27/2014	1,602	311	218	01:17	26,193
Thu 8/28/2014	2,041	443	199	01:12	35,656
Fri 8/29/2014	1,495	327	170	00:44	24,486
Sat 8/30/2014	1,145	322	86	02:41	18,372
Total	**7,960**	**1,662**	**852**	**01:20**	**133,917**

In terms of the number of hits, number of page viewers and in rate of data transferred the best day out of these five days of log data is thursday. Table shows that the number of Visitors are more on tuesday and the average visit length is more on saturday.

Activity by Hour of Day

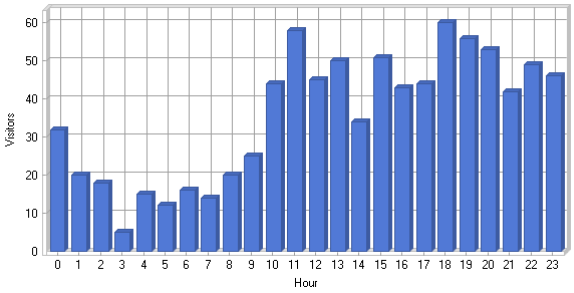

Fig.5: Hourly Website Visitors Report

The report of website visitors on HOURLY BASIS is shown in figure 5. Accurate information of Hourly activity is shown in Table-3 It shows number of hits per hour, page views per hour ,visitors per hour and Bandwidth in KB per hour.It also shows total number of hits 7960,page views 1662,Visitors852 and bandwidth 133917KB.

Table-3: Hourly Activity Statistics of the Website Usage

Hour	Hits	Page Views	Visitors	Bandwidth (KB)
00:00 - 00:59	159	25	32	1,646
01:00 - 01:59	195	56	20	3,224
02:00 - 02:59	116	37	18	2,879
03:00 - 03:59	73	6	5	625

04:00 - 04:59	134	23	15	2,542
05:00 - 05:59	59	15	12	1,913
06:00 - 06:59	86	25	16	1,050
07:00 - 07:59	74	10	14	1,909
08:00 - 08:59	232	32	20	2,030
09:00 - 09:59	172	41	25	1,991
10:00 - 10:59	477	102	44	8,485
11:00 - 11:59	878	278	58	18,046
12:00 - 12:59	604	123	45	8,565
13:00 - 13:59	504	99	50	6,087
14:00 - 14:59	418	56	34	6,974
15:00 - 15:59	633	158	51	7,986
16:00 - 16:59	403	107	43	4,998
17:00 - 17:59	344	59	44	6,060
18:00 - 18:59	368	65	60	6,917
19:00 - 19:59	402	49	56	7,824
20:00 - 20:59	558	109	53	11,211
21:00 - 21:59	389	58	42	6,700
22:00 - 22:59	411	63	49	7,204
23:00 - 23:59	271	66	46	7,041
Total	**7,960**	**1,662**	**852**	**133,917**

As table shows the accurate information of Hourly activity of website.It will shows you the best hour in terms of the number of hits,page views, visitors and the rate of data transferred According to the results of activilty statistics we can make some changes in website.

4.3 Access Activity

Access activity provides the information of the most popular pages, most downloaded Files and most requested images. Figure 6 shows the result of most popular pages of website after analyzing our log file. It shows the topmost page is the home page of our educational website and next popular page is the guideline page of the website and so on.

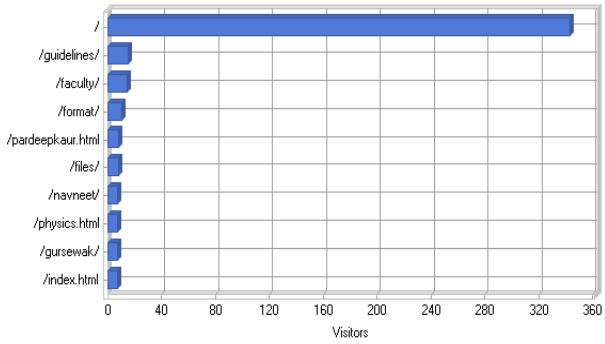

Fig6: Most Popular Pages

Figure 7 shows the topmost files that are downloaded the most number of times in these five days form the files page of the website. It shows that "/files/commm.networks.pdf" files is downloaded almost 32 times and more then any other file. Similarly next file which is in the most downloaded files is "/files/microprocessor qbank.pdf" and so on.

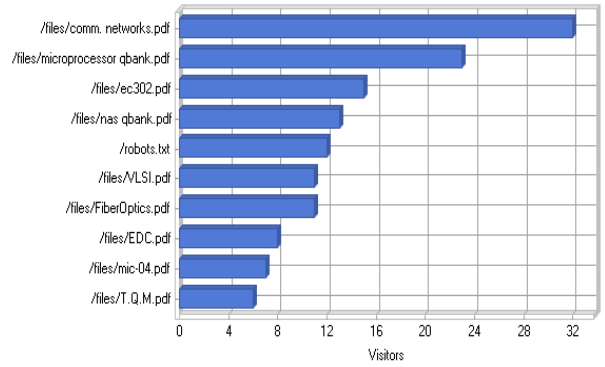

Fig7: Most downloaded Files

Figure 8 shows the images those have been downloaded most during these five days.It shows that from files page the "favicon.gif" image have been downloaded the most.Next image is logo.png" and so on.

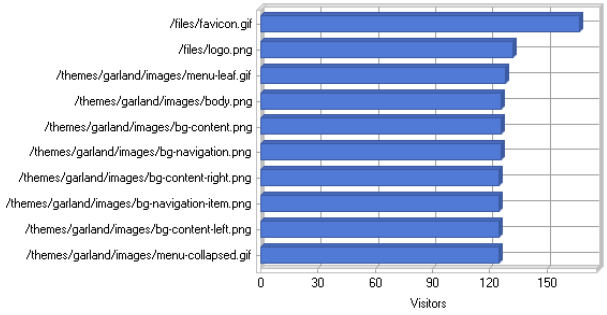

Fig 8: Most requested Images

The information extracted in this phase is very useful or valuable for the improving the usability or the structure of the website

4.4 Referrer

Figure 9 shows the search engines used to find the website by the visitors. It shows google is used most out of other search engines.

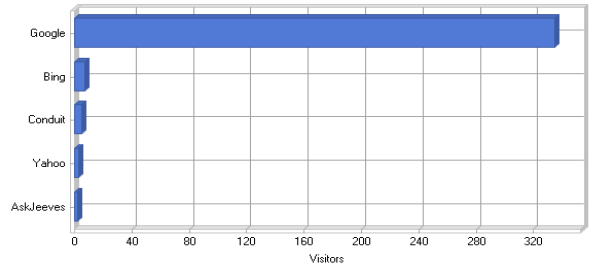

Fig 9: Search Engines Used

As we know the search engines are the best way to get information regarding anything. If user does not know the name of website they usually search for it on a search engine. From our log analysis we find that most people use google for search for this purpose. Table 4 shows the accurate information of top search engine used to search about the website.

Table 4. Search Engines used to search the website

	Search Engine	Visitors
1	Google	335
2	Bing	7
3	Conduit	5
4	Yahoo	3
5	Ask Jeeves	2
	Total	**352**

It shows 352 visitors accessed the website through a search engine and not directly. Out of this 335 used google and very few used other search engine. As there are total of 852 visitors and out of these 352 visitors have google as their referrer. Search engine plays important role in the online reputation of website.

4.5 Browsers

This phase of analyzer provides the information of Browser and Operating System used by the visitor of the website. Figure 10 shows the browsers used to access this website. Experimental results shows that 283 visitors used firefox,257 used google chrome,121 used internet Explorer,121 used opera, 18 used Geck/20100401 S401 S40OviBrowser/2.0.2.62.10 and rest used some other browsers.

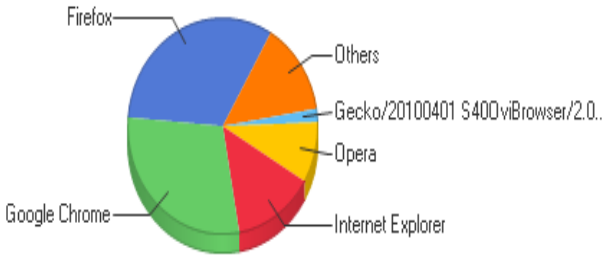

Fig 10: Most used Browser

Fig 11 shows the chart of operating system daily used by the visitors.Hits for window 7 are more then any other operating system.

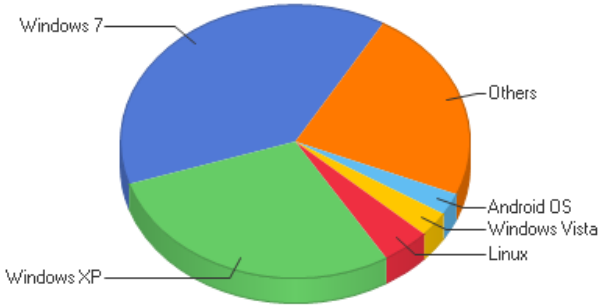

Fig 11: Most Operating System used

Table 5 shows the number of visitors who used a particular operating system. Results shows that 327 visitors use window 7 ,244 use Window XP. Using these results we can find which operating system is mostly used by visitors.

Table 5. Most Used Operating Systems

	Operating System	Hits	Visitors	% of Total Visi tors
1	Windows 7	3,270	327	38.38%
2	Windows XP	1,879	244	28.64%
3	Others	823	154	18.08%
4	Linux	124	38	4.46%
5	Windows Vista	195	25	2.93%
6	Android OS	366	22	2.58%
7	Windows Server 2003	42	13	1.53%
8	Mac –OS	35	5	0.59%
9	Windows 2000	11	5	0.59%
10	IPhone	26	4	0.47%
11	Windows NT	6	3	0.35%
12	Windows 8	15	3	0.35%
13	BlackBerry	20	2	0.23%
14	Windows ME	4	2	0.23%

15	IPad	2	2	0.23%
16	Windows 95	4	1	0.12%
17	Windows 98	3	1	0.12%
18	Windows Phone	22	1	0.12%
19	Chrome OS	63	0	0.00%
	Total	**6,910**	**852**	**100.00%**

Sometimes websites don't look like you expect them to– images might not appear, menus might be out of place, and text could be jumbled together. This might be caused by a compatibility problem between browser and the website. Information provided by analyzer tool will show us the browsers and operating systens which visitors are using to access the website. So the compatibility issues can be resolved and website can be updated.

4.6 Errors

When a request is made to server for a page on your site (for instance, when a user accesses your page in a browser or when Googlebot crawls the page), your server returns an HTTP status code in response to the request. If status code is less then 200 or greater then 299 it means there is error or failure. Error phase shows the type of errors which cropped up during the access of the website. Figure 12 shows the types of errors occurred during access of website. It shows two types of error ,First error "404" which occur when link is not found and second error "403" which is forbidden.This figure shows the error report of three days.

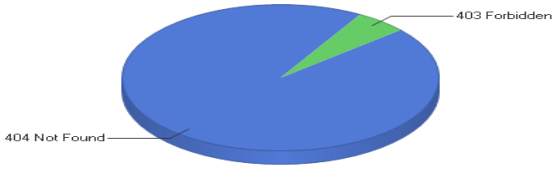

Fig 12: Types of Error Occurred

Table 6 shows in our log data only two types of errors have been detected one is error no 404 and another is error 403. As shown in table total failed requests are 909 and out of these 806 failed request have error number 404 and 48 failed requests have error number 403.From this information the required changes will be made to the code of site, so that in future error rate decreases.

Table 6. Error Types

	Error	Hits
1	404 Not Found	861
2	403 Forbidden	48
	Total	**909**

5. CONCLUSION AND FUTURE SCOPE

Web Log file records activity information when a web user submits a request to a Web Server and we have collected the log file of educational institute for the time period of five days.This log file has been analyzed using the Weblog Expert lite 8.6 analyzer tool ,the results of which have helped to identify the : i) Usage pattern in terms of time(day and hour) during which the number of hits and user activity is maximum. ii) User preferences, like most visited pages, most downloaded files and images , most preferred browsers and operating systems. Thus this valuable information helps in predicting the user behavior and helps in customizing the website accordingly to make it more accessible and user friendly the website. These results can also be used to pin point the area of the website and best time for posting advertisements and information to be conveyed to the visitors so that they can be easily "visible" and hence can be used for effective utilization of the website space and generate revenue. For future work ,if the log file will pertain to a longer time period is used,the more accurate results regarding the user preferences and behavior can be obtained .

6. ACKNOWLEDGMENT

Our thanks to the Educational Institute for providing Web Server Log data.

7. REFERENCES

[1] Murti 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

[2] L.K. Joshila grace, V.Maheswari, Dhinaharan Nagamalai 2011, "Analysis of web logs and web user in web mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1.

[3] 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

[4] Arvind K. Sharma ,P.C. 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

[5] Neha Goel, C.K. Jha, 2013 "Analyzing User Behavior from Web Access Logs using Automated Log Analyzer Tool"International Journal of Computer Applications(0975-8887),Volume 62-No.2.

[6] Cooley, R.2010, "Web Usage Mining: Discovery and Application of Interesting Patterns from Web data", http://citeseer.nj.nec.com/426030.html.

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

[8] V.Chitraa, Dr. Antony Selvdoss Davamani,2010 "A Survey on Preprocessing Methods for Web Usage Data" (IJCSIS) International Journal of Computer Science and Information Security,Vol. 7, No. 3.

[9] Castellano.G et al., 2007 "Log Data Preparation for Mining Web Usage Patterns", International Conference Applied Computing, pp.371-378.

[10] K.R.Suneetha, Dr,R.Krishnamoorthi, 2009 "Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Sceurity,VOL. 9 No.4.