Identifying the Number of Visitors to improve Website Usability from Educational Institution Web Log Data

Arvind K. Sharma
Dept. of CSE
Jaipur National University, Jaipur,
Rajasthan,India

P.C. Gupta
Dept. of CSI
University of Kota
Kota, Rajasthan-India

Abstract: Web usage mining deals with understanding the Visitor's behaviour with a Website. It helps in understanding the concerns such as present and future probability of every website user, relationship between behaviour and website usability. It has different branches such as web content mining, web structure and web usage mining. The focus of this paper is on web mining usage patterns of an educational institution web log data. There are three types of web related log data namely web access log, error log and proxy log data. In this paper web access log data has been used as dataset because the web access log data is the typical source of navigational behaviour of the website visitor. The study of web server log analysis is helpful in applying the web mining techniques.

Keywords: Web Usage Mining, Web Log Data, WebLog Expert Lite7.8

1. INTRODUCTION

Website is an important tool for web users to obtain information such as education, entertainment, health, e-commerce, etc. Today, the Internet is most emerging technology in the world. The terms Internet and World Wide Web are often used in everyday speech without much distinction. The World Wide Web is also known as 'Information Superhighway'. It is a system of interlinked hypertext documents accessed via Internet. However, the Internet and the World Wide Web are not one and the same. The Internet is a global system of interconnected computer networks. In other hand, the World Wide Web is one of the services that run on the Internet[1]. It is a collection of text documents and other resources, linked by hyperlinks and URLs, usually accessed by Web browsers from web servers. In short, the World Wide Web is also considered as an application 'running' on the Internet[2]. It is a large and dynamic domain of knowledge and discovery. It has become the most popular services among other services that the Internet provides. The number of users as well as the number of website has been increasing dramatically in the recent years. A huge amount of data is constantly being accessed and shared among several types of users, both humans and intelligent machines.

Paper is organized in different sections: Section-II explains Web usage mining. Proposed methodology is shown in Section-III. Section-IV contains Experimental results. Conclusion is shown in section-V while references are mentioned in the last section.

2. WEB USAGE MINING

Web Usage Mining is a part of Web Mining, which, in turn, is a part of data mining. As data mining has been used to extract meaningful and valuable information from large volume of data, the web usage mining has been used to mine the usage characteristics of the Website users. Web mining refers to overall process of discovering potentially useful and previously unknown information from the web document and services[3]. This extracted information can be used in a variety of ways such as improvement of the Web application, identifying the visitor's behaviour, checking of fraudulent elements etc. Web access patterns mined from Web log data have been interesting and useful knowledge in practice. Examples of applications of such knowledge include improving design of the websites, analyzing system performance to understand user's reaction and motivation, build adaptive websites[4]. The aim in web usage mining is to discover and retrieve useful and interesting patterns from a large dataset.

2.1 Phases of Web Usage Mining Process

Web usage mining process consists of three phases such as Preprocessing of web data, Pattern discovery, and Pattern analysis [5]. Preprocessing is a primary work in web mining process. The main phases in Web usage mining process are shown in fig.1 below.

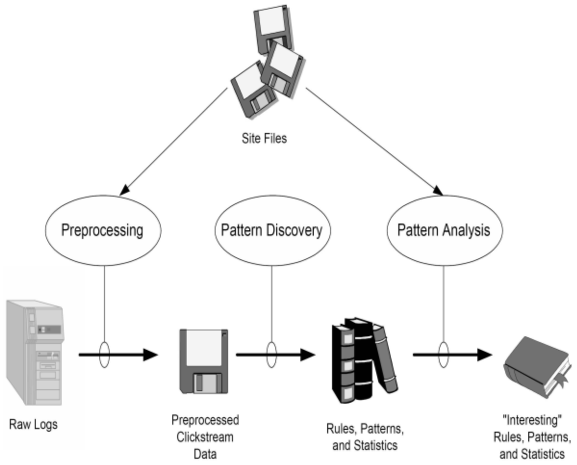
Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Fig.1: Phases of Web Usage Mining Process

2.1.1 Preprocessing: Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

2.1.2 Pattern Discovery: Web usage mining can be used to uncover patterns in web server log data but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data.

2.1.3 Pattern Analysis: This is the final step in the Web Usage Mining process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information.

3. PROPOSED METHODOLOGY

As those trends become stronger and stronger, there is much need to study web user behaviour to better serve the users and increase the value of institutions or enterprises. Website design is currently based on thorough investigations about the interests of website visitors and investigated assumptions about their exact behaviour. Today, understanding the interests of users is becoming a fundamental need for Websites owners in order to better serve their users by making adaptive the content and usage, structure of the website to their preferences. The analysis of web log data permits to identify useful patterns of the browsing behavior of users, which exploited in the process of navigational behavior. Web log data captures web-browsing behaviour of users from a Website. Academic institutions are good examples that develop website. One such institution of the education sector has been considered in our work. This paper presents visitor pattern analysis performed through educational institution web log data. We have been performed different analysis on a sample of Web log data to–

➢ Determine the usability of the Website, including the-

- Visitor Pattern Analysis
- Page View Analysis
- Time Analysis
- Origin of the Website Visitors
- Portions of the Website that are accessed
- Number of document downloads(both hits & accesses)

In this work, WebLog Expert reports undergo a time analysis and page view analysis. The time analysis looks at the different times of day, days of week, and days of month that the Website receives the most visitors. The page view analysis provides which website pages are most viewed by the visitors. The combination of these statistics will help us to predict the attributes of the Website user and the Website usability.

3.1 Data Collection

In this study, the user access web log data has been collected from the Educational Institution Website's 0server www.davkota.org which stores normally secondary data source in view of the fact that web log keeps every activity of the user regarding to visit of the Website. The web log data contains the information from 31 October 2012 to 30 November 2012 of one month period. During this period, 1.01 GB data had been transferred for the complete work.

3.2 Data Selection

At present, towards Web Usage Mining technique, the main data origin has three kinds: Server-side data, Client-side data, and Proxy-side data (middle data). In this work, we use the case of the Web server.

3.2.1 Web Log Data

A Web log data is a listing of page reference data sometimes it is referred to as click stream data[6]. The web plays an important role and medium for extracting useful information. There is a need for data log to track any transaction of the communications. This data can offer valuable information insight into website usage. It characterizes the activity of many users over a potentially long period of time. The web server log data contains several attributes. These attributes are as follows:

Date-The date from Greenwich Mean Time(GMTx100) is recorded for each hit. The date format is YYYY/MM/DD. The example above shows that the transaction was recorded at 2012/11/01.

Time-It refers Time of transactions. The time format is HH:MM:SS.

Client IP Address-It is the number of computer who access or request the website.

User Authentication-Some websites are set up with a security feature that requires a user to enter username and password. Once a user logs on to a website, that user's 'username' is logged in the log file.

Server IP Address-It is a static IP provided by Internet Service Provider. This IP will be a reference for access the information from the server.

Server Port-It is a port used for data transmission. Usually, the port used is port 80.

Server Method(HTTP Request)-The term request refers to an image, pdf, .txt, HTML file, movie, sound, and more.

URL-It is a path from the host. It represents the structure of the websites.

Agent Log-It provides data on a user's browser, browser version, and operating system. This is the significant information, as the type of browser and operating system determines what a user is able to access on a website.

3.3 Tool for Experiment

There are various commercial and freely available tools exists for web mining purposes. WebLog Expert Lite7.8 is one of the fast and powerful Web log analyzer tool[7]. This tool helps to reveal important statistics regarding a web site's usage such as activity of visitors, access statistics, paths through the website, visitors' browsers, etc. It supports W3C extended log format that is the default log format of Microsoft IIS 4.0/.05/6.0/7.0 and also the combined and common log formats of Apache web server. It reads compressed log files (.gz, .bz2 and .zip) and can automatically detect the log file format. If necessary, log files can also be downloaded via FTP or HTTP. We have been used a web log analyzer WebLog Expert Lite7.8 web mining tool. It is one such program and used to produce highly detailed, easily configurable usage reports in Hypertext Markup Language (HTML) format, for viewing with a standard web browser[7]. Using this web mining tool we have been identified Hits statistics like Total Hits, Visitors Hits, Average Hits per Day, Average Hits per Visitor, etc., Page View Analysis like Total Page views, Average Page Views per Day, Average Page Views per Visitor, total Visitors, Total Visitors, Average Visitors per Day, Total Unique IPs, Bandwidth, Total Bandwidth, Visitor Bandwidth, Average Bandwidth per Day, Average Bandwidth per Hit, and Average Bandwidth per Visitor of the Website on monthly and day of the week basis.

4. EXPERIMENTAL RESULTS

In this work, we have been used web log data from October 31, 2012 to November 30, 2012 collected from the web server of the website *www.davkota.org* have been analyzed by using WebLog Expert Lite7.8 web mining tool[7]. The complete experiment has been done on the basis of web log data of an educational institution’s website. The design and execution of such work is restricted and time consuming. The results had limited in time and space so only a limited period of time is taken to perform the results. The general activity statistics of the website usage is shown in Table-1.

Table-1: General Activity Statistics of the Website Usage

Hits	
Total Hits	23669
Visitor Hits	21744
Spider Hits	1925
Average Hits per Day	763
Average Hits per Visitor	25.46
Cached Requests	2753
Failed Requests	2177
Page Views	
Total Page Views	1517
Average Page Views per Day	48
Average Page Views per Visitor	1.78
Visitors	
Total Visitors	854
Average Visitors per Day	27
Total Unique IPs	935
Bandwidth	
Total Bandwidth	1.01 GB
Visitor Bandwidth	993.44 MB
Spider Bandwidth	42.93 MB
Average Bandwidth per Day	33.43 MB
Average Bandwidth per Hit	44.84 KB
Average Bandwidth per Visitor	1.16 MB

By using the WebLog Expert Lite7.8 web mining tool, we had been found 23669 hits, 854 visitors, 935 IPs, 1517 page views, 1.01 GB data had been transferred and so on. Based on the analyzer report, we have been found several unnecessary records like image files, failed requests and incomplete records and are eliminated and useful information 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. Fig.2 shows the daily visit report of the website visitors.

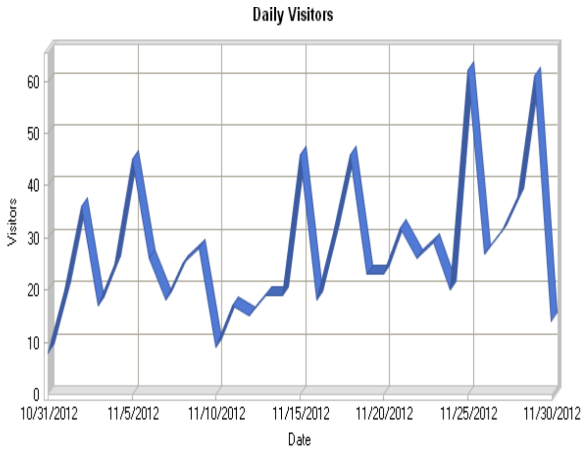

Fig.2: Daily Website Visitors Report

Table-2 shows the accurate daily visitor’s activity statistics of the website usage. 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.

Table-2: Daily Activity Statistics of the Website Usage

Date	Hits	Page Views	Visitors	Band width (in KB)
Wed 31/10/2012	94	6	8	4,691
Thu 1/11/2012	560	28	20	33,309
Fri 2/11/2012	338	44	36	13,894
Sat 3/11/2012	596	25	17	20,454
Sun 4/11/2012	863	105	25	41,395
Mon 5/11/2012	890	71	45	40,398
Tue 6/11/2012	511	27	26	25,445
Wed 7/11/2012	664	29	18	38,742
Thu 8/11/2012	938	44	25	42,123
Fri 9/11/2012	1,044	100	28	60,350
Sat 10/11/2012	611	40	9	28,218
Sun 11/11/2012	860	52	17	35,736
Mon 12/11/2012	935	43	15	17,630
Tue 13/11/2012	1,097	53	19	35,406
Wed 14/11/2012	855	52	19	38,358

Thu 15/11/2012	470	35	46	22,275
Fri 16/11/2012	568	40	18	29,770
Sat 17/11/2012	738	40	31	37,462
Sun 18/11/2012	1,001	76	46	52,991
Mon 19/11/2012	485	26	23	25,025
Tue 20/11/2012	1,261	66	23	24,944
Wed 21/11/2012	822	77	32	38,954
Thu 22/11/2012	1,047	50	26	54,000
Fri 23/11/2012	1,096	61	29	60,547
Sat 24/11/2012	562	42	20	25,553
Sun 25/11/2012	693	41	62	30,171
Mon 26/11/2012	524	35	27	25,890
Tue 27/11/2012	666	44	31	37,516
Wed 28/11/2012	1,417	66	38	57,966
Thu 29/11/2012	1,044	74	61	41,815
Fri 30/11/2012	419	25	14	20,198
Total	**23,669**	**1,517**	**854**	**1,061,241**

The following report produced, total number of hits 23669, total page views 1517, total visitors 854, and total bandwidth 1061241 Kilobytes were found which summarized in table-2. Every day 28 average number of visitors are visited 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 and improve the overall structure of the Website. Fig.3 shows the report of hourly website visitors.

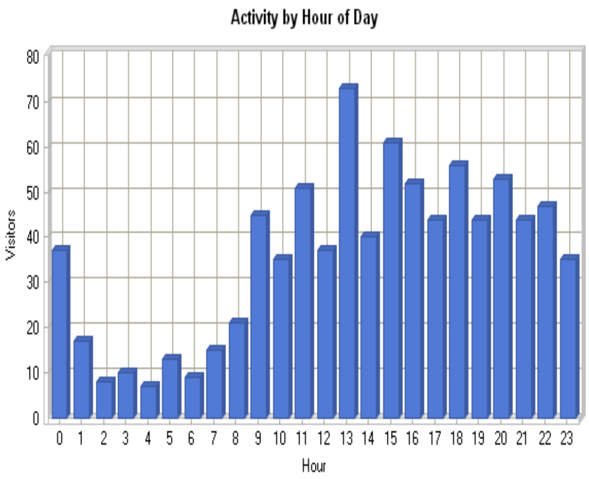

Fig.3: Hourly Website Visitors Report

Table-3 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.

Table-3: Hourly Activity Statistics of the Website Usage

Hour	Hits	Page Views	Visitors	Bandwidth (in KB)
00:00-00:59	997	47	37	28,533
01:00-01:59	221	9	17	11,147
02:00-02:59	225	44	8	9,399
03:00-03:59	129	20	10	6,647
04:00-04:59	74	3	7	1,895
05:00-05:59	110	6	13	3,942
06:00-06:59	169	15	9	3,488
07:00-07:59	287	14	15	11,372
08:00-08:59	803	62	21	35,070
09:00-09:59	1,375	89	45	65,435
10:00-10:59	1,132	55	35	47,489
11:00-11:59	1,322	87	51	72,202
12:00-12:59	1,433	111	37	78,902
13:00-13:59	1,387	140	73	71,503
14:00-14:59	1,201	62	40	59,039
15:00-15:59	1,141	83	61	54,597
16:00-16:59	1,431	100	52	70,286
17:00-17:59	1,670	107	44	52,309
18:00-18:59	1,405	80	56	73,534
19:00-19:59	1,551	86	44	72,728
20:00-20:59	1,848	83	53	84,602
21:00-21:59	1,554	73	44	70,027
22:00-22:59	859	57	47	42,005
23:00-23:59	1,345	84	35	35,078
Total	**23,669**	**1,517**	**854**	**1,061,241**

Total number of visitors found 854 that are shown in table-3. Every day 28 average number of visitors are visited the website. This summary report shows hourly usage of the website and predicts total number of visitors who are accessed 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.

5. CONCLUSION

Web is one of the most used interface to access remote data, commercial and non-commercial services. Web mining is a growing area with the growth of web based applications to find web usage patterns. By using web mining we could found website user's interest and behavior through which we can make our website valuable and easily accessible. The complete work has accomplished by analyzing educational institution web log data for one month period. Our experimental results help to predict and identify the number of visitors for the Website and improve the Website usability.

6. REFERENCES

[1] Piatetsky Shapiro G. et al., “Advances in Knowledge Discovery and Data Mining”, AAAI/MIT Press, 1996.

[2] The W3C Technology Stack; “World Wide Web Consortium”, Retrieved April 21, 2012.

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

[4] Huiping Peng, “Discovery of Interesting Association Rules Based on Web Usage Mining”, International Conference 2010.

[5] Cooley, R., “Web Usage Mining: Discovery and Application of Interesting Patterns from Web data”, 2000, http://citeseer.nj.nec.com/426030.html.

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

[7] [Online] http://www.weblogexpert.com

7. ACKNOWLEDGMENT

Our thanks to DAV Organization, Department of Computer Science and IT, DAV Kota for providing Web server log data.