Discovering User Behaviour Patterns from Web Log Analysis to Enhance the Accessibility and Usability of Website

Harpreet Singh

Abstract**—**Finding relevant information on the World Wide Web is becoming highly challenging day by day. Web usage mining is used for the extraction of relevant and useful knowledge, such as user behaviour patterns, from web access log records. Web access log records all the requests for individual files that the users have requested from the website. Web usage mining is important for Customer Relationship 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 helpful in improving website structure or design as per the user's requirement by analyzing the access log file of a website through a log analyzer tool. The focus of this paper is to enhance the accessibility and usability of a guitar selling web site by analyzing their access log through Deep Log Analyzer tool. The results show that the maximum number of users is from the United States and that they use Opera 9.8 web browser and the Windows XP operating system.

Keywords**—**Web usage mining, log file, web mining, data mining, deep log analyser.

I. Introduction

THE World Wide Web today has become the most extensively used, widely connected knowledge sharing and communication platform. Web mining is the application of data mining techniques which deals with the extraction of intresting and useful information from web [1]. Web mining can be divided into three different types [2], which are Web content mining, Web structure mining and Web usage mining. The process of retrieving relevant information from the contents of web documents is called web content mining [3]. Web structure mining is the structure (graph) of webpages. Web Usage Mining is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect [10]. It is used for deciding business strategies through the efficient use of Web Applications [9].

The major drawback of Web Usage Mining is the nature of the data they deal with. The data is huge and unmanagable due the trillions of users using/updating the data. Also, most of the data available is unrorganised, haphazard, and hence, needs a lot of time and effort to put it in a readily usable form by humans and applications/machines [9].

The aim in web usage mining is to discover and retrieve useful and interesting patterns from a large dataset. Fig. 1 shows that web usage mining consists of four phases such as Data collection, Pre-processing of log data, Pattern discovery and Pattern analysis [4], [6], [7]. Common web usage mining algorithms are association rule generation, sequential pattern generation, and clustering.

II. Data Source and Format

Data is collected form servers in the form of log files, which are automatically created and maintained by the servers. Log files consist of list of activities performed by visitors on web pages. There are three types of servers which act as the sources of log files [5] – Web Server Logs, Web Proxy Server and Client side logs. Web server Log files are the most accurate but these files contain personal information and do not record the visited cached pages. Further, Web server log files are of four types namely Access, Error, Agent and Referrer log files. Proxy servers take an HTTP request from user, gives them to the web server, then the result passed by web server is returned to the user [3]. Web proxy Client side log files can reside in the client's browser window itself. For this, special software is downloaded by the users to their browser window. As mentioned earlier, the log file consists of activities performed by visitors in terms of entries in a log file. Entries of users or visitors are in terms of plain text in some format. Common log formats are [5], [8]:

- Combined Log Format
- Common Log Format
- Multiple Access Logs
- Conditional Logs

In this paper, the data source is a web server access log file of a guitar selling website and in combined Log format.

III. Experiments and Results

The log file needs to be processed to get interesting and useful data from it. There are various tools [3] available for analysis of log files such as Google analytics, Stat Counter [11], Deep Log Analyzer [12], Awstates, Web Log Expert [9], [13] etc. Some of these tools are freely available and some are paid. The Deep Log Analyzer tool is easy to use, web analytics software for small and medium size websites. It analyzes website visitor's behaviour and view statistics in several easy steps. In this, you will get the detailed information about your websites accessed resources, site navigators, visitor's activity, referrals and search Engine and Visitors system information. The system will generate easy to read reports in image and table format.

Harpreet Singh is with the Punjabi University, Department. of Computer Engineering, Patiala, India (e-mail: harpreet_boparai@hotmail.com).

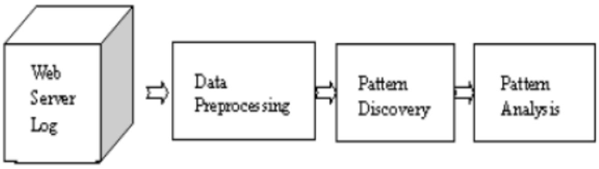

Fig. 1 Phases of Web Usage Mining [9]

Features of Deep Log analyzer are shown below [10].

- Reports can be exported to HTML or Microsoft Excel.
- Provide a lot of information on visitor’s activity, Server errors, Referral Sites, Browsers and OS, Search spiders and Search Engines.
- It supports IIS and apache logs.
- Automatically detects log format.
- Creates custom analytics reports using SQL or Microsoft visual query designer.
- Can read GZ and ZIP compressed logs.
- Reports can be filtered by date.
- Reports can be viewed in hierarchical representation with interactive navigation, convenient charts and open database formats.

In this paper, we have analysed the log data of a guitar website from August 31 to Sep 10, 2015. The log file has been analyzed using the Deep Log Analyzer web mining tool. The general activity statistics of the website usage is shown in Table I.

The general activity statistics of the website are shown in Table I. Results of the general statistics show that there are 717 visits and 443 unique visitors. There are 48 average numbers of visits per day. The table also shows that most of the visitors are from the United States.

TABLE I
GENERAL ACTIVITY STATISTICS OF THE WEBSITE USAGE

Hits Summary	Total	Per Day	Visits Summary	Total
Number of Hits:	717	65	Number of Visits:	525
Number of Successful Hits:	619(86%)	56	Average Number of Visits per Day:	48
Outgoing Traffic:	8.55Mb	794 Kb	Average Visit Duration:	0:00 Min
Incoming Traffic:	0 Kb	0 Kb		
Visitors Summary	**Total**	**Page Views Summary**		**Hits**
Number of Unique visitors:	443	Total Page Views:		715
Visitors who visited once:	389 (88%)	Most popular Page:	http://www.guitar-.../index.htm	271
Repeat visitors:	54(12%)	Most popular Download:	http://www..../guitar1-demo.exe	5
Average Visits per visitor:	1.19	Most popular Entry Page:	http://www.guitar-.../index.htm	152
Most visitors from this Country	United States (18% visitors)	Most popular Exit Page:	http://www.guitar-.../index.htm	183
Referral Summary		**Hits**	**Search Engines Summary**	**Hits**
Top Referring Website:	http://statcounter.com	60	Top Search Engine: Bing	6
Technical Summary			Top Key Phrase: guitar online	2
Most Popular Browser:	Opera 9.80		Spiders Requests:	51
Most Popular Operating System:	Windows XP			
Error Hits:	98 (14%)			

TABLE II
TOP TEN ENTRY PAGES

Entry Page	Number of Visits
http://www.guitar-online.com/en/index.htm	152
http://www.guitar-online.com/index.htm	110
http://www.guitar-online.com/en/other-music-software/transcribe-software-to-help-transcribe-recorded-music/index.htm	22
http://www.guitar-online.com/en/other-music-software/metronome-for-simple-compound-and-odd-time-meters/index.htm	19
http://www.guitar-online.com/en/guitar-tutorial-software/index.htm	18
http://www.guitar-online.com/en/information/index.htm	10
http://www.guitar-online.com/en/musical-instruments-online-store/samedaymusic-more-than-240000-items-in-stock/index.htm	9
http://www.guitar-online.com/en/online-shop/index.htm	8
http://www.guitar-online.com/en/other-music-software/index.htm	8
http://www.guitar-online.com/en/tag/note-detection/index.htm	8

Tables II and III show the top 10 entry and exit pages and their number of visits. In 152 visits, the entry page was “guitar-online.com/en/index.htm” and 110 visits to “guitar-online.com”. Similarly, on 183 visits, the exit page is “guitar-online.com/en/index” and on 123 visits, the exit page is “guitar-online.com/index.htm”. From this we can conclude that the most of visitors come from the home page and even exit at the home page.

Table IV shows the report of top 10 visitors. The maximum visitors (10) are from the “A” IP address, while seven visitors are from the “B” IP address. From this result we can find the area of interested visitors and serve them better. Visitor history is shown in Table V, where 79 and the maximum visits are on September 1, 2015. From this table we can find out the day or date on which the frequency of visitors increases or decreases. This data are small; however, if we have a large dataset, we can find the best day (days) of week and month(s) of year.

TABLE III
TOP TEN EXIT PAGES

Exit Page	Number of Visits
http://www.guitar-online.com/en/index.htm	183
http://www.guitar-online.com/index.htm	123
http://www.guitar-online.com/en/other-music-software/metronome-for-simple-compound-and-odd-time-meters/index.htm	24
http://www.guitar-online.com/en/other-music-software/transcribe-software-to-help-transcribe-recorded-music/index.htm	19
http://www.guitar-online.com/en/guitar-tutorial-software/index.htm	16
http://www.guitar-online.com/en/online-shop/index.htm	9
http://www.guitar-online.com/en/tag/note-detection/index.htm	8
http://www.guitar-online.com/en/downloads/index.htm	5
http://www.guitar-online.com/en/guitar-tutorial-software/classical-pieces-for-guitar-volume-1-for-advanced-level/index.htm	5
http://www.guitar-online.com/en/books-and-audio-cd-online-store/yoke-wong-award-winning-piano-improvisation-course/index.htm	5

TABLE IV
TOP 10 VISITORS

Visitor	Country	Number of Visits
A	United States	10
B	Malaysia	7
C	Sudan	4
D	France	4
E	France	4
F	Germany	3
G	Korea	3
H	Iran	3
I	India	3
G	Pakistan	3

TABLE V
VISITOR HISTORY

Date	Number of Visits
31-08-2015	50
01-09-2015	79
02-09-2015	46
03-09-2015	56
04-09-2015	42
05-09-2015	32
06-09-2015	40
07-09-2015	30
08-09-2015	54
09-09-2015	58
10-09-2015	38

TABLE VI
SEARCH SPIDER VISITS

Spider	Number of Hits	Data Transferred (Kb)
BaiDu	16	46
GoogleBot-Mozilla	14	100
Yandex	8	18
Yahoo	8	203
MSNBot	4	8
GoogleBot	1	97

Spider is a program that automatically fetches web pages. They are used to feed pages to search engines. It is called a spider because it crawls over the web. Table VI shows the search spider visits.

Table VII shows the number of unique visitors from the top 10 countries. The United States has the maximum number of unique visitors (i.e. 78). From this table, we can say that this website is mostly visited by US people. So according to US people's taste, we should make some changes in website to improve the usability and accessibility.

TABLE VII
NUMBER OF UNIQUE VISITORS FROM TOP COUNTRIES

Country	Number of Unique Visitors
United States	78
France	15
United Kingdom (Great Britain)	15
India	15
China	14
Germany	14
Indonesia	11
Spain	11
Brazil	9
Malaysia	9

IV. CONCLUSION

Log data is the exhaustive record of all the 'web activity' of the user as soon as a web user submits a request to a Web Server. This paper analyses the log file of a guitar selling website, using a Deep log Analyzer tool with the aim of increasing the number of visits to the website. In this analysis we find the user interest and behaviour which can make the website more accessible or usable. Experimental results show the general activity statistics of the website. The top 10 entry and exit pages used by visitors during their visit are also shown in Tables II and III, and it was found that maximum clicks and maximum number of unique visitors are from are from IP address "A" and the United States, respectively. All the above results can help us in improving the guitar website in various ways and can help to make the website more accessible and usable. This would help in finding potential customers and may improve the sales and revenues for the seller.

References

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

[2] L. K. Joshila Grace, V. Maheswari, Dhinaharan Nagamalai, "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & its applications, Vol.3, No.1, January 2011.

[3] Navjot Kaur, Himanshu Aggarwal, "A Comparative Study of WUM tools to Analyze User Behaviours Pattern from Web Log Data" International Journal of Advances in Engineering Research, Vol. No. 10, Issue No. VI, December 2015.

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

[5] Navjot Kaur, Himanshu Aggarwal, "A Novel Semantically-Time-Referrer based Approach of Web Usage Mining for Improved Sessionization in Pre-processing of Web Log" , International Journal of Advanced Computer Science and Applications, Vol. 8, No. 1, Feb 2017

[6] G. Castellano, A. M. Fanelli, M. A. Trsello, "Log data preparation for mining Web usage patterns" IADIS International Conference Applied Computing, pp 371-378, 2007.

[7] Doru Tanasa, Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining" Enhancing information, IEEE Intelligent System, pp 59-65,2004.

[8] Aswin G. Raiyani, Sheetal S. Pandya, "Discovering User Identification Mining Technique for Preprocessing Log Data", ISSN: 0975 – 6760, Vol. 2, Issue 2, pp 477-482, Nov 12 to Oct 13.

[9] Navjot Kaur, 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 (0975 – 8887), Vol. 110, No. 4, January 2015.

[10] Yogish H K, G T Raju, Manjunath T N, "The Descriptive Study of Knowledge Discovery from Web Usage Mining", International Journal of Computer Science, Vol. 8, Issue 5, No 1, Sep 2011.

[11] Statcounter, Retrieved data from: https://statcounter.com ON 20 Feb 2017.

[12] Deep Log Analyzer, Retrieved data from: "https://www.deep-software.com/features/" on 20 Feb 2017.

[13] Web Log Expert Lite, Retrieved data from: http://www.weblogexpert.com on 20 Feb 2017.