International Journal of Engineering & Technology, 7 (4.41) (2018) 15-20

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Analysis of Web Server Logs to Understand Internet User Behavior and Develop Digital Marketing Strategies

Saleh Mowla[1]*, Nisha P. Shetty[2]

[1,2]Department of Information & Communication Technology,
Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India.
**Corresponding author E-mail: nisha.pshetty@manipal.edu*

Abstract

With the advancement of technology and widespread availability of information through the internet and the World Wide Web, it has become possible for small-scale industries to convert and expand themselves by increasing awareness and exposure of their businesses. One of the most popular and easiest way to increase a customer base is by creating websites and thus engaging with customers online. However, with the increase in the number of websites and other available source of advertisements, it has become a need to design websites in a way that keeps the customers and viewers engaged and interested so that they return with positive expectations. One way of analyzing website popularity and online customer behavior is by analyzing web server logs with the help of web usage mining techniques to find unknown patterns and generate insights about different aspects of the website. The paper discusses a live scenario where logs of a blogging website have been mined and analyzed to better improve the structure of the site as well as understand the behavior of the viewers who have visited the site. The paper also provides recommendations on improving the structure of the website to adopt effective digital marketing strategies.

Keywords: *digital marketing; pattern discovery; server logs; web analytics; web usage mining*

1. Introduction

The impact of the internet has crept on to every sector of human life be it education, business, shopping and so on. E-commerce has completely encased this user dependency on the web by effectively tracking customer browsing preferences precisely by capturing every mouse click and user input. Pursuing data such as user subscription, websites visited, etc. can be beneficial for companies to improve their sales by incorporating suitable marketing strategies. While Search Engine Optimization (SEO) techniques are beneficial to increase the traffic of websites by augmenting its page rank [1], it has become a need to adopt web usage mining techniques to better understand the analytics of website data and thus gain valuable insights. These insights can prove to be profitable to the owner in the long run. Web usage mining techniques can also help the owner to predict the visitor's behavior and accordingly recommendation systems are built to ease the visitor's navigation and experience [2].

1.1. Web usage mining and its phases

Web Usage Mining is a category of web mining where the user's interaction with the web is analyzed and studied by first retrieving data such as user transactions, site hits, etc. [3] Some of the sources of data required for Web Usage mining include access logs, user ratings and profiles, database transactions on websites as well as website structure and connectivity [4]. The process of web usage mining can be broadly classified into three phases: [5]

1. Pre-Processing Phase

2. Pattern Discovery Phase

3. Pattern Analysis Phase

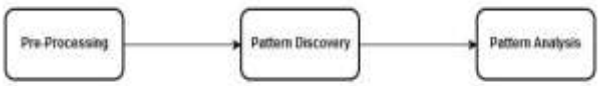

Figure 1: Phases of Web Usage Mining

1.1.1. Pre-processing phase

This phase mainly deals with converting the available raw data into a form which can be easily comprehended by the tools and algorithms which are applied during the pattern discovery phase. Data thus needs to be 'cleaned' and this process involves removal of noisy and irrelevant data [6]. Missing or incorrect values also need to be filled in and modified accordingly.

1.1.2. Pattern discovery

In this phase, we apply different algorithms or use tools to determine patterns which were difficult to predict initially. Depending on the business requirement, we apply different techniques such as clustering, classification, etc. On the basis of efficiency and accuracy, different algorithms are used to get the best possible result. This phase assumes that the result obtained is based on correct and clean data without which we may end up discovering patterns that actually do not exist. Once the patterns have been discovered, one can proceed with analyzing them in the next phase.

1.1.3. Pattern analysis

This phase is particularly important for business intelligence and analytics. The patterns discovered in the previous phase are analyzed to decide which patterns are significant and not irrelevant. Most of the time, the patterns discovered are in a mathematical form which is why interpretation of the patterns is a key aspect to pattern analysis. Based on the analysis, we can determine how to restructure our website to achieve better results in terms of traffic, convenience in navigation and performance optimization.

1.2. Web server logs

Typically, a log file stores all the activities performed on the server [7]. Depending on the condition, the server maintains these records in separate log files namely access logs, error logs, piped logs, script logs, etc. *Access Logs* store all the requests which have been processed by the server. *Error Logs* record the errors which arise during the processing of requests and can be used later for diagnostic purposes. *Piped Logs* is a mechanism through which the server can write both access and error logs to a process instead of writing it to a file directly. *Script Logs* records all the inputs and outputs to and from CGI scripts and is useful for debugging and testing purposes.

1.2.1. Web access log format

In order to perform web usage mining and analyze the mining patterns of users, we first start with looking into the access logs of the server from where the user requests a service or a web page. The Apache access logs are popularly stored in two different formats:

Common Log Format: This format gives information regarding the client's IP Address, Host Name, Username, Timestamp and zone of the request, Request line (method, URL of resource and protocol), status code sent by the server and the amount of bytes of the object returned in response to the client's request.

Combined Log Format: In addition to the information provided by the Common Log Format, it also provides information about the previously process request (Referred URL of the current request) and the User Agent (the client's browser information).

The access logs are stored in a file where each line represents a request made to the server and provides information related to the identity of the user, the resource or service requested, the time the request was made, etc. [8] Each aspect of an access log is separated by a whitespace as can be seen clearly from a sample access log file in Figure 2. The combined log format is explained in Table 1.

1.3. Related works

With respect to web analytics and digital marketing, many have proposed different models and techniques correlating the two. Chaffey and Patron described techniques that can be used to set up a digital marketing optimization program [9]. Xun analyzed the visit duration of visitors on e-commerce websites and tested a model using observed e-commerce data from multiple stores in the United Kingdom [10]. A method has also been introduced using social media analytics to assess and understand different issues that revolve around a marketing campaign [11]. There has also been a study in [12] that describes a web analytics and visualization tool which collects analyses and visually represents the data collected in an e-learning environment. The goal of the tool was to allow the teacher to better understand his students' behavior in the e-learning environment. Hasan, Morris and Probets used Google Analytics to evaluate the usability of e-commerce websites in their study [13].

In addition to identifying patterns and establishing the behavior of visitors, this paper also suggests how approaches can be changed by using newly discovered information and design websites accordingly.

Table 1: Description of the combined log format

Log Parameter	Description	Example
IP Address	Unique to every client (multiple requests can be made by the same IP address)	117.195.11.154
Host Name	Gives the identity of the user (usually ignored since it is unreliable)	-
User Name	Determined by HTTP authentication. This should be ignored when the status code of the request is 401 as it signifies that the user has not yet been authenticated	-
Timestamp	Determines which date and what time the request was made by the user	02/Jul/2017:15:44:26
Offset	Time zone of the request can be determined	+0000
Method	The method of the request made by the user	GET or POST
Path	The URL of the web page or service currently requested by the user	www.expressionselectic.com/articles
Protocol	Determines how the request was transmitted	HTTP/1.1
Status	This is sent by the server back to the client to determine if the request was successful, redirected, etc.	200
Bytes	Determines how many bytes of data was sent back by the server to the client in response to the request	408
Referral URL	Used to determine from which resource or webpage the client had requested before making the current request (refer to 'Path' above).	www.expressionselectic.com/home
User Agent	Identifies and gives information about the client browser.	"Mozilla/5.0 … Firefox/54.0"

```
117.195.11.154 - - [02/Jul/2017:15:44:26 +0000] "POST /wp-admin/admin-ajax.php HTTP/1.1" 200 408
"http://expressionselectic.com/wp-admin/post.php?post=70&action=edit" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:54.0) Gecko/20100101 Firefox/54.0"
117.195.11.154 - - [02/Jul/2017:15:44:58 +0000] "GET /wp-admin/images/loading.gif HTTP/1.1" 200 2254
"http://expressionselectic.com/wp-admin/load-styles.php?c=1&dir=ltr&load%5B%5D=dashicons,admin-bar,buttons,media-views,common,forms,admin-menu,dashboard,l
ables,edit,revisions,media,themes,about,nav-menu&load%5B%5D=s,widgets,site-icon,l10n,wp-auth-check&ver=4.8" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:54.0)
Gecko/20100101 Firefox/54.0"
117.195.11.154 - - [02/Jul/2017:15:44:59 +0000] "POST /wp-admin/admin-ajax.php HTTP/1.1" 200 6
"http://expressionselectic.com/wp-admin/post.php?post=70&action=edit" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:54.0) Gecko/20100101 Firefox/54.0"
117.195.11.154 - - [02/Jul/2017:15:45:49 +0000] "POST /wp-admin/admin-ajax.php HTTP/1.1" 200 47
"http://expressionselectic.com/wp-admin/post.php?post=70&action=edit" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:54.0) Gecko/20100101 Firefox/54.0"
```

Figure 2: Sample Access Log File

2. Methodology

This section describes the manner in which data has been gathered from a blogging website and was processed thereafter to discover patterns and generate insights about the website and behavior of the visitors.

2.1. Data Gathering

In order to perform web usage mining, a new website was created (www.expressionseclectic.com) to analyze live web logs. The website is a blogging website filled with articles, poems, videos, etc. on different themes. Using different SEO tools and techniques, the site attracted visitors and logs were collected for a period of two weeks before they could be used for analysis.

2.2. Data cleaning

The requests in the access log files were stored in the combined log format [14]. Requests with status code '200' and method 'GET' were considered for pattern discovery and analysis. In addition, all requests made by clients involving images i.e. requests with extensions .gif, .jpg, .ico, .png, etc. have also been discarded. Requests made by robots as well as admin have also not been considered since the objective is to analyze patterns of the visitors. Users are identified according to the IP addresses mentioned in the access logs [15]. The data presented in Figure 3 shows a subset of clean and organized data obtained from web server logs in Figure 2 which includes noisy data mentioned above.

2.3. Pattern discovery and analysis

After cleaning the data, it becomes easier to analyze the access patterns of the user. For the purpose of detecting patterns, trends and popularity of the pages, visualization tools are helpful for any non-technical user [16]. Various graphs have been created using R viz. a data analysis and visualization tool, in order to understand and detect patterns which could not be understood by simply viewing the access log files of the server. Pattern discovery is facilitated with the help of access logs, referral logs and answers questions like "how many people have accessed this page this week?" or "how many people from a particular country are viewing and accessing the services from my website?" Pattern analysis requires that the structure of the website (pages as nodes and hyperlinks as edges) as well as its contents are analyzed first. Figure 4 shows the structure of www.expressionsecelctic.com as a network of webpages connected by inbound and outbound links. Inbound links are the hyperlinks which connect to a web page whereas outbound links are those which leave a webpage to connect to another. It then helps the business owner to gain insights about his website's strengths and weaknesses based on the popularity of the web pages after tracking the access pattern and behavior of the user.

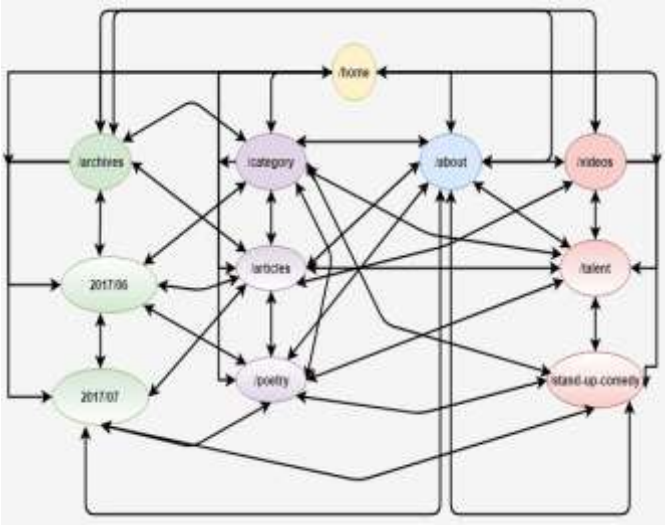

Figure 4: Website structure as a network of web pages and hyperlinks

ID	IP_ADDRESS	HOST_NAME	USER_NAME	DATE	OFFSET	METHOD	PATH	PROTOCOL	STATUS	BYTES	REFERRAL_URL	USER_AGENT	TIME
5	66.249.79.77	-	-	01/07/2017	+0000	GET	/home/	HTTP/1.1	200	54596	-	Mozilla/5.0	10:42:36
12	117.214.115.112	-	-	02/07/2017	+0000	GET	/home/	HTTP/1.1	200	54596	-	Mozilla/5.0	05:30:35
190	117.214.115.112	-	-	02/07/2017	+0000	GET	/home/	HTTP/1.1	200	62859	-	Mozilla/5.0	06:52:00
200	117.214.115.112	-	-	02/07/2017	+0000	GET	/home/videos /stand-up-comedy/	HTTP/1.1	200	63498	http://expressionseclectic.com /home/	Mozilla/5.0	06:52:14
258	117.195.11.154	-	-	02/07/2017	+0000	GET	/home/	HTTP/1.1	200	62859	http://expressionseclectic.com /home/videos/stand-up-comedy/	Mozilla/5.0	08:40:10
259	117.195.11.154	-	-	02/07/2017	+0000	GET	/home/category /travel/	HTTP/1.1	200	69617	http://expressionseclectic.com /home/about/	Mozilla/5.0	08:41:39

Figure 3: Snapshot of dataset after cleaning

3. Experimental Results

This section will summarize the results obtained after following the methodology explained in the previous section.

3.1 Pre-Processing Phase

In a period of two weeks, the total number of requests processed by the server was 5427 (1.36 MB). After cleaning and removing the irrelevant data, the number of requests available for analysis was 427 (71 KB). For the purpose of this experiment, only 7.83% of the contents of the log file were considered fit for pattern discovery analysis.

3.2. Pattern discovery

After the data cleaning process, it is much easier to discover patterns by correlating different variables and aspects of the data to get a new perspective and thus, a new insight. Based on the perspective, the patterns can be analyzed to find out useful information like which page grabs the interest of the user or which page ensures that the visitor continues to browse the site. Finding out which websites on the World Wide Web is directing traffic is invaluable with respect to digital and search engine marketing (SEM). Table 2 shows some of the general statistics gained after processing the data.

Using visualization tools such as R makes it easier for non-technical users to understand the statistics of their website and how well their website is performing. This helps them make decisions as to when to take action or change something to get a better result. Figure 5.1 and 5.2 give a representation of which pages were accessed and visited on which day as well as the number of hits for a particular page.

Figure 6 shows the hourly traffic analysis over the period of two weeks that the site was active. It becomes easier to understand the browsing patterns of the user. For a literature-centric blog which was analysed, it can be seen that a proportion of the visitors prefer reading between 6:00 and 9:00 in the morning whereas a higher proportion of the users prefer reading in the afternoon. The traffic density reported to be the least during the later hours of the evening.

Table 2: General Statistics of Website

S.N.	Website Aspect	Gained Information
1.	Number of Unique Users	52
2.	Total number of Sessions	147
3.	Average Number of pages accessed per session	3 (approximately)
4.	Average time spent per session	2 minutes, 32 seconds

Figure 7 shows which pages were accessed the most by the visitors. Analysis of such patterns can help the website owner to decide which pages pique the interests of the visitors and helps identify their preferences. Similarly, it is possible to find out from which page the visitor has come to the current page as shown in Figure 8. The page named with '-' in Figure 8 means that users visited the site by typing in the URL of the website directly in their web browser. Analysis of such patterns helps in identifying which pages keeps the visitor engaged so that they continue browsing the website. For a more comprehensible analogy, consider online shopping sites. Based on the logs of various customers, the site suggests recommendations to the viewers. For instance, if customer A considered buying mobile phones the site might recommend the same customer to buy additional accessories such as covers, screen-guards, etc. since previous customers have done the same [17].

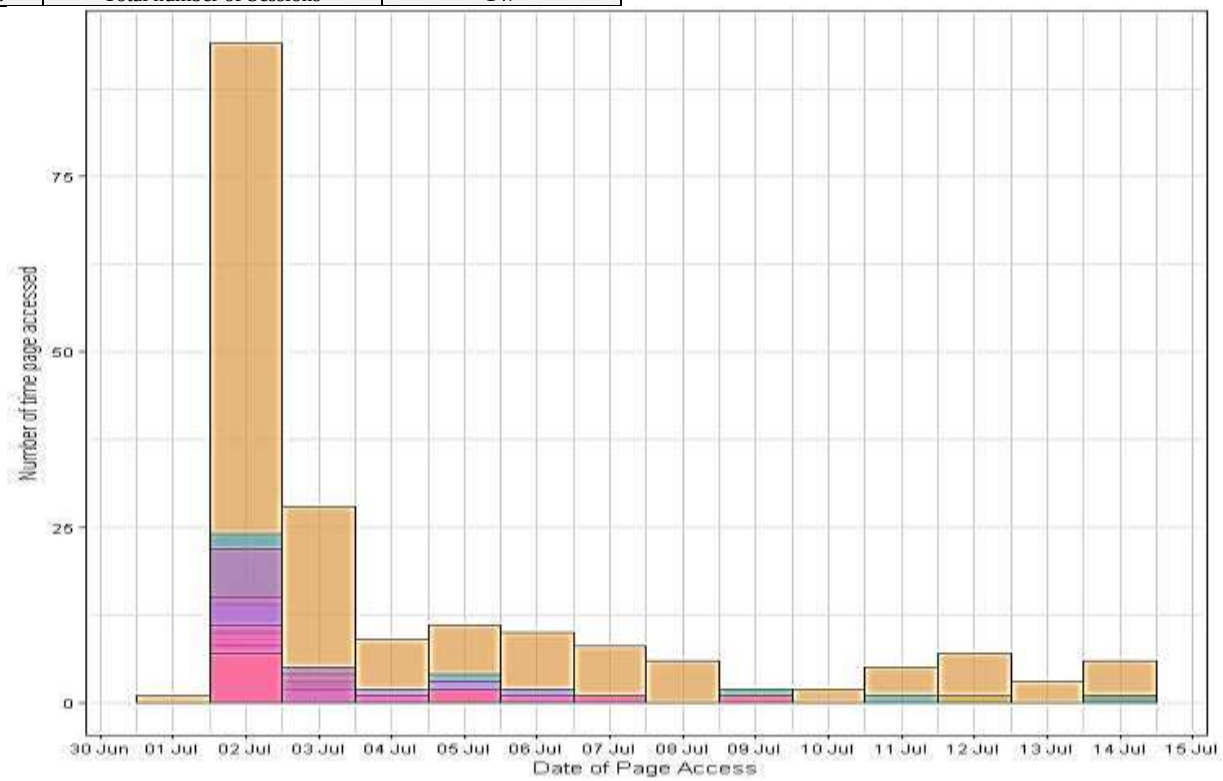

Figure 5.1: Frequency of pages accessed on a daily basis

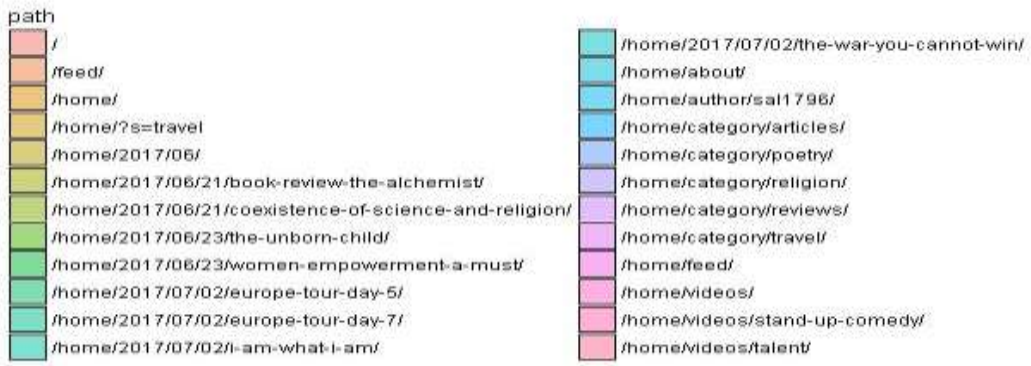

Figure 5.2: Legend for Daily Traffic Analysis

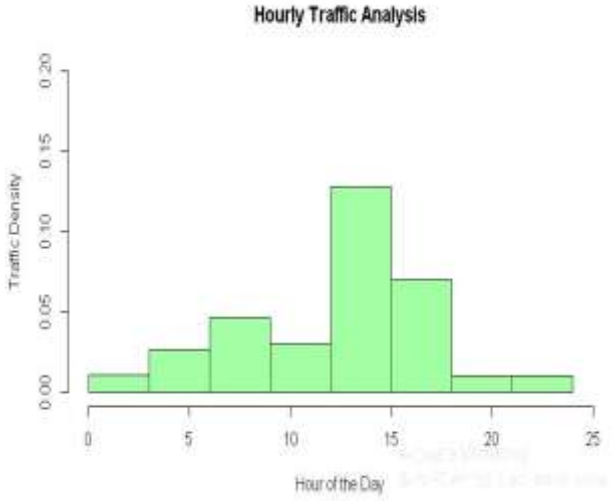

Figure 6: Visitor Traffic recorded on an hourly basis

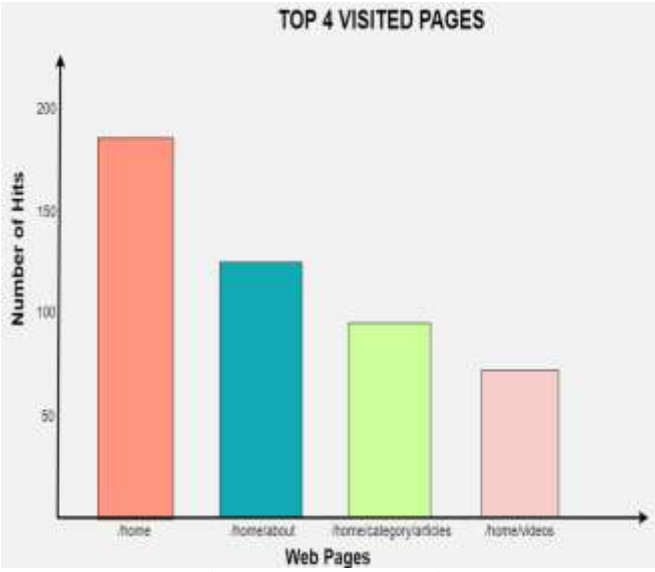

Figure 7: Most frequently accessed pages from the website

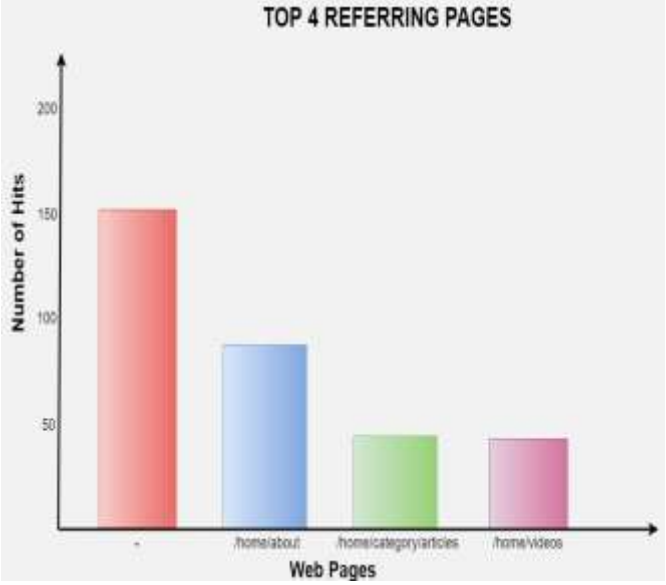

Figure 8: Top 4 Referring Pages from which user has browsed to another page

3.3. Pattern analysis

By analyzing the information obtained in section 3.2, it is now possible to change the approach and adopt different strategies in order to effectively increase the exposure of the website. The information shows that aside from the 'Home' and the 'About' pages, the 'Articles' and 'Videos' pages on the website proved to be more popular than the others. This suggests that the website should focus more on these aspects as they engaged the audience more than the other web pages. Assuming the website was related to e-commerce, it would be beneficial if ad campaigns were based on the popular pages so that the website would attract more potential customers. By analyzing user behaviour on websites, it becomes clear how to adapt and change marketing strategies so as to increase website visibility and quickly achieve the purpose and goals of the website.

4. Conclusion

The World Wide Web has surpassed print and television to become the most common and popular form of media. For online marketing proper analysis of a user's browsing behaviour is invaluable. In this work, hits on www.expressionseclectic.com is scrutinized to closely study the website owner's altercations with the web like time of the day when hits are more, sites frequented more by users and so on. Using data visualization tools, it becomes easier to analyze raw data from web server logs and generate insights that can help achieve the intended results quickly. This work can be extended to improve personalization so as to predict the web page or website the user is likely to visit based on past preferences, semantic analysis and so on.

References

[1] A. Kakkar, R. Majumdar and A. Kumar. Search engine optimization: A game of page ranking. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015 (pp. 206-210).

[2] D. Dixit and J. Gadge. Automatic Recommendation for Online Users Using Web Usage Mining. In International Journal of Managing Information Technology (IJMIT), Vol. 2, No. 3, 2010.

[3] D. Tanasa and B. Trousse. Advanced data preprocessing for intersites Web usage mining. In IEEE Intelligent Systems (vol. 19, no. 2, pp. 59-65), Mar-Apr 2004.

[4] M. Aldekhail. Application and Significance of Web Usage Mining in the 21st Century: A Literature Review. In International Journal of Computer Theory and Engineering, Vol. 8, No. 1, 2016

[5] P. Verma and N. Kesswani. Comparative analysis of algorithms for identification of session on the basis of threshold value. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIA Com), New Delhi, 2016 (pp. 3724-3730).

[6] S. K. Dwivedi and B. Rawat. A review paper on data preprocessing: A critical phase in web usage mining process. 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, 2015 (pp. 506-510).

[7] P. Sharma, S. Yadav and B. Bohra. A review study of server log formats for efficient web mining. 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, 2015 (pp. 1373-1377).

[8] D. S. Sisodia and S. Verma. Web usage pattern analysis through web logs: A review. 2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE), Bangkok, 2012 (pp. 49-53).

[9] Chaffey, D. & Patron, M. "From web analytics to digital marketing optimization: Increasing the commercial value of digital analytics". Journal of Direct, Data Digital Marketing Practice, (2012) 14: 30. https://doi.org/10.1057/dddmp.2012.20

[10] Xun, J. Return on web site visit duration: Applying web analytics data. Journal of Direct, Data Digital Marketing Practice, (2015) 17: 54. https://doi.org/10.1057/dddmp.2015.33

[11] Vorvoreanu, M., Boisvenue, G., Wojtalewicz, C. et al. "Social media marketing analytics: A case study of the public's perception of Indianapolis as Super Bowl XLVI host city" Journal of Direct, Da-

ta Digital Marketing Practice (2013) 14: 321. https://doi.org/10.1057/dddmp.2013.18

[12] Moissa B., de Carvalho L.S., Gasparini I. (2014) A Web Analytics and Visualization Tool to Understand Students' Behavior in an Adaptive E-Learning System. In: Zaphiris P., Ioannou A. (eds) Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences. LCT 2014. Lecture Notes in Computer Science, vol 8523. Springer, Cham. https://doi.org/10.1007/978-3-319-07482-5_30

[13] HASAN, L., MORRIS, A. and PROBETS, S., 2009. Using Google Analytics to evaluate the usability of e-commerce sites. IN: Kurosu, M. (ed.). Human Centered Design, HCII 2009, Lecture Notes in Computer Science 5619, pp. 697-706.

[14] Mo Wang and Juanle Wang. A data preprocessing framework of geoscience data sharing portal for user behavior mining. 2015 23rd International Conference on Geoinformatics, Wuhan, 2015 (pp. 1-5).

[15] G. Neelima and S. Rodda. Predicting user behavior through sessions using the web log mining. 2016 International Conference on Advances in Human Machine Interaction (HMI), Doddaballapur, 2016 (pp. 1-5).

[16] N. Neelima and Syeda Farha Shazmeen. Visual data mining to discover knowledge patterns from Web navigational trends. 2011 International Conference on Recent Trends in Information Systems, Kolkata, 2011 (pp. 117-120).

[17] T. Badriyah, E. T. Wijayanto, I. Syarif and P. Kristalina. A hybrid recommendation system for E-commerce based on product description and user profile. 2017 Seventh International Conference on Innovative Computing Technology (INTECH), Luton, 2017 (pp. 95-100).