Exploitation of Server Log Files of User Behavior in Order to Inform Administrator

Hamed Jelodar
Computer Department,
Islamic Azad University,
Science and Research Branch,
Bushehr, Iran

ABSTRACT

All requests that are send to website server are stored in a file called Server File Log, all pages users observe or all user's requests to the Server are stored in that File. In this paper we are going to investigate the user's behavior by web usage mining through Server File Log in order to improve the website better service and at the end by analyzing Download software website we present the result in illustration.

General Terms

Web Mining

Keywords

Web Mining, Web Log File, Web Usage Mining.

1. INTRODUCTION

Most managers of websites that are active in the fields of business is usually looking for ways to improve their website and the quality of the service to have a better productivity .To achieve this goal, we examine the behavior of customers we need to be able to identify them. When a user sends a request to a Web site or browse pages of a website all the events are stored in a file called the web server log file. Then all the requests sent by the users to the website server will be sent all stored In a file, this paper, Web Usage Mining via server log files, we decided to investigate and identify the behavior of users. All this is described in Section second of this paper to review the classification of Web mining, and it is spoken, the third section of the web server log file is described in the fourth and fifth explores web usage mining and the process of that is discussed in Section fifth discuss the data preparation phase is analyzed in the sixth section of a website and its results are shown the seventh part of the material conclusions is said.

2. RELATED WORK

All The most important data sources through which one can study the behavior of Web users can check the log file to the server. In this file all requests that are sent to the server on behalf of the users in this file will be saved and it can be said that all the interactions between the user and the server is recording. So far a lot of research to investigate users' behavior through log files that the server has already is referred some of them. Inbarani and et al , the paper he K-Means algorithm with records in server log files are divided into clusters [1]. K.Etminani and et al This paper presents a new method for mining navigation patterns from web logs. For this purpose, the ant-based clustering is used [2].

D. Singh and et al In the article he Analytics Server log file has been academic institution. The results Analytics can be aware of things such as popularity among visitors to your site, diagnose server error and so on. His research identified the main challenges of web robots enter the human users [3]. Chitraa and et al , In his article on Web Usage Mining preprocessing steps were discussed and the conclusion was that the importance of this step is very important [4]. In this article the technique of Web Usage Mining Corporation to analyze the log file server is a website to download the software will be paid according to results achieved what a positive impact for the administrator in order to better services to your users and enhance the Web site services.

3. WEB MINING

Web data mining is called Web mining, in other words Web mining a series of Data Mining Techniques to extract knowledge And Useful Information from web data [1]. Other helpful resources can be extracted in the types of text, video or multimedia. Web mining can be classified into three types including Web Content Mining, Web Structure Mining and Web Usage Mining.

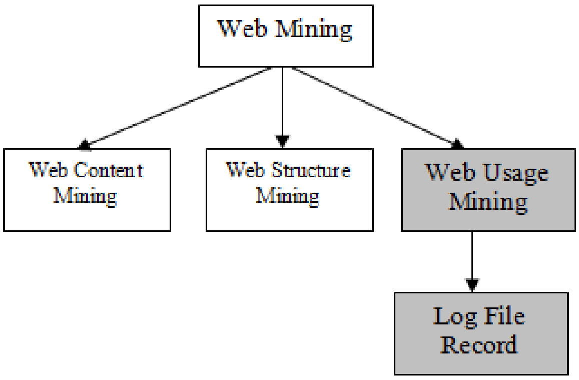


Fig 1: Classification of Web Mining

3.1 Web Content Mining

Web content mining discovers useful information from the website. Namely, that is a kind of web mining that Analyses Web content Such as Text, charts, graphics, etc [2].

3.2 Web Structure Mining

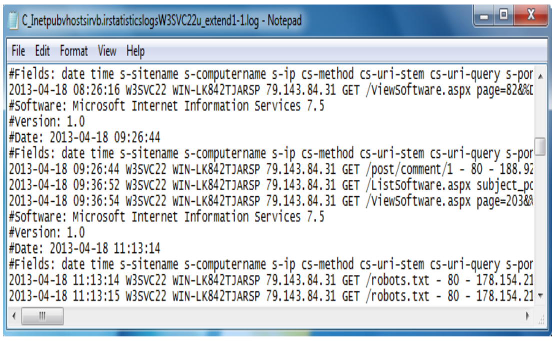
Web structure mining process using graph theory for analysis node and link structure of a website. This type of web mining mostly concentrates on web structure and connection of web pages can be identified by that.

3.3 Web Usage Mining

It is the process which gives the information on how to use the web. In another word it extract the behavior of users.in this paper, we discuss this type of web mining.

4. LOG FILE SERVER

When a request sent by a client or client's request to the server, it is stored in a file called Log File Server. Requests can be on the behalf of human or the spider engines. If a user requests the visit of a page, it would error if the server did not contain that, which these events are stored in Log File Server. Thus, all the activities that users do in the website or wholly all the requests given to the server is stored in this file.


C_Inetpubvhostsirvb.irstatisticslogsW3SVC22u_extend1-1.log - Notepad
File Edit Format View Help

```
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-por
2013-04-18 08:26:16 W3SVC22 WIN-LK842TJARSP 79.143.84.31 GET /ViewSoftware.aspx page=82&%D
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2013-04-18 09:26:44
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-por
2013-04-18 09:26:44 W3SVC22 WIN-LK842TJARSP 79.143.84.31 GET /post/comment/1 - 80 - 188.92
2013-04-18 09:36:52 W3SVC22 WIN-LK842TJARSP 79.143.84.31 GET /ListSoftware.aspx subject_pc
2013-04-18 09:36:54 W3SVC22 WIN-LK842TJARSP 79.143.84.31 GET /ViewSoftware.aspx page=203&%
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2013-04-18 11:13:14
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-por
2013-04-18 11:13:14 W3SVC22 WIN-LK842TJARSP 79.143.84.31 GET /robots.txt - 80 - 178.154.21
2013-04-18 11:13:15 W3SVC22 WIN-LK842TJARSP 79.143.84.31 GET /robots.txt - 80 - 178.154.21
```

Fig 2: A view of a web server log file

Figure 2 is view of a W3C log file format that is downloaded from the Web site server. The requests and events occurred for the server has been recorded as illustrated .When a user sends a request for visiting a page, it's like that several requests have been sent to the server because by loading a page containing pictures a series of request must be sent to the server with the aim of depicting pictures. Below is the example of a request of all the requests sent by the user which has been recorded in File Server Log, the concepts of recorded field can be understood by referring to table of 1 and 2.

Example Log File Record:
Software: Microsoft Internet Information Services 7.5
# Version: 1.0
# Date: 2013-05-04 03:43:27
# 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
2013-05-04 03:43:27 W3SVC22 WIN-LK842TJARSP 79.143.84.31 GET / ViewSoftware.aspx page = 82 80 - 5.114.1.34 HTTP/1.1 Mozilla/5.0 + (Windows; + U; + Windows + NT +6.1;

+ En-US; + rv: 1.9) + Gecko/2008052906 + Firefox/3.0 ASP.NET_SessionId = dco13zr4xxcs3cutsxkzfhci - irvb.ir 200 0 0 35289 445 62

Table 1. Meanings of the fields

Meanings	Prefix fields
Server	S-
Client	C-
Server to Client	Sc-
Client to Server	Cs-

Table 2. List of Field Server log file format W3C

Description	Example	Fields
Historical activity has occurred.	2013-05-04	date
The time which the activity occurred.	03:43:27	time
The Internet service name and instance number that was running on the client.	W3SVC22	s-sitename
Name of the server where the file log Created.	WIN-LK842TJARSP	s-computername
Address IP Server where the log file is created.	79.143.84.31	s-ip
Client-server type of application method.	GET	cs-method
The page requested by the user.	/ViewSoftware.aspx	cs-uri-stem
Parameter single, convenient in demand.	page = 82	cs-uri-query
Server port number	To 80	s-port
That is a valid user name to access the server.Anonymous users is shown with a dash.	-	cs-username
Address IP Client	5.114.1.34	c-ip
The protocol version the client uses.	HTTP/1.1	cs-version
Type of client browser.	Mozilla/5.0 + ... + Firefox/3.0	cs (User-Agent)
The content of the cookie sent.	dco13zr4xxcs3cutsxkzfhci	cs (Cookie)
Last page the user is directed to a page through it now.	http://www.google.com/search	cs (Referrer)
If there is a header name host.	irvb.ir	cs-host
Status Code HTTP.	200	sc-status
Error status code HTTP.	0	sc-substatus
The Windows status code.	0	sc-win32-status
The number of bytes sent by the server.	35289	sc-bytes
The number of bytes received from the server.	445	cs-bytes
Field shows the length of time that it takes for a request to be processed and its response to sent.	62	time-taken

We observed a list of fields recorded in File Log Server Table 1, appropriate patterns resulting users' behavior can be discovered and extracted through this file regarding recorded information. Most of the website that are active in fields of business and publicizing are in search of methods for retaining their clients. One of the ways to achieve this is to analyze the Server log file. By analyzing the server log file information to identify the user's geographical location, days and hours of visited, most visited pages, and more.

5. WEB USAGE MINING

The Web Usage mining can be used to identify and study the behavior of the users. User's behaviors can be utilized for the Improvement of Business Intelligence, sales and publicizing, website designing and more other items [1]. Web Usage Mining, consists of three stages; data preparation, pattern discovery, pattern analysis. In the next section we spell out more about data preparation.

5.1 Data Preparation

Without a doubt the most important step is exploring the use of Web data preparation stage (pre-processing). At this point, we need to purge the gathered data and drag out the needing main data in order to put them in database, so we can carry out the statistical analysis.

5.2 Pattern Discovery

Discover model using statistical methods and data mining techniques (such as Association Rules mining, classification, Clustering) deals with the discovery of patterns [3].

5.3 Pattern Analysis

Pattern analysis is the last stage of Web Usage Mining, the purpose of this process is the identification of accurate required patterns among extracted patterns.

6. DATA PREPARATION

As it is previously mentioned, data preparation stage (pre-processing) is the most significant stage of Web Usage Mining. At this stage collected data need to be purged and drag out the required main data, so we can carry out the statistical analysis. Data preparation incorporates stages such as Data Cleaning, User Identification, Session Identification and Path Completion.

6.1 Data Cleaning

The purpose of purging data is the removal of unnecessary and impertinent items. This technique have a remarkable importance in web analysis. As our intention is to identify user's behavior, it is necessary to filter or remove impertinent items, for instance when spider engines browse a website, their activities are recorded in Server Log File which is of no use, and turn out to be an inhumane behavior, as a result they need to be purged or filtered.

6.2 User Identification

In the previous stage data cleaning function accomplished on the data source (Server File Log). Now we should identify different or specific users, so we are able to extract an accurate statistical result. There are a variety of methods for identifying members; identification of users through IP and Cookies.

6.2.1 User Identification through IP

This kind of Identification is one of the methods by which we can identify the visitors, For example we can be aware of visitor's geographical location by converting the IP and observing the domain suffix (ir,ca,tr,…) But this method has its own drawbacks, a user may enter multiple times, and each entry will also change its IP or users are likely to have come with same IP.

6.2.2 User Identification via cookies

Another way for identifying specific user is use of cookies. In this case, severs creates cookies which are stored on the other side of the client side. Cookies help us identify repetitive user as a specific user. But what makes the problem is the removal of stored cookies due to protection of their privacy [2].

6.3 Session Identification

The collection of pages that a user visits consistently in a particular time are define as user session [4]. The purpose of identifying session is the classification of user's visitors to sessions. Normally, the time which is allocated for this session is 30 minutes. If collections of the user's visits exceeded 30 minutes, it would create a new user session. What makes the problem is that the user tends to spend a long time visiting a website, for instance; 50minutes. Consequently a new user session is created, in another word two user session are recorded for each visitor.

6.4 Path Complete

Another important issue that needs to be resolved is the path completion [5]. Sometimes the user's path are not wholly recorded following the use of cache by browser and are stored in Server File Log incompletely. For instance the user uses back button with aim of visiting previous page and browser utilizes cache for depicting previous page, as a result user's path are not recorded completely [6]. Hence, path completion is utilized for acquiring full path of user's access [7]. There are a variety of methods which are beyond the scopes of this paper.

7. RESULT

With regard to all the topics have been said to want to examine the behavior of the users of a website to download the software. For the analysis of the website of the software "Deep Log Analysis". The data recorded in these files and the date is April 13, 2013 to May 14, 2013 file size is 7 MB. Extracted result of this website analysis is depicted in the form of figure.

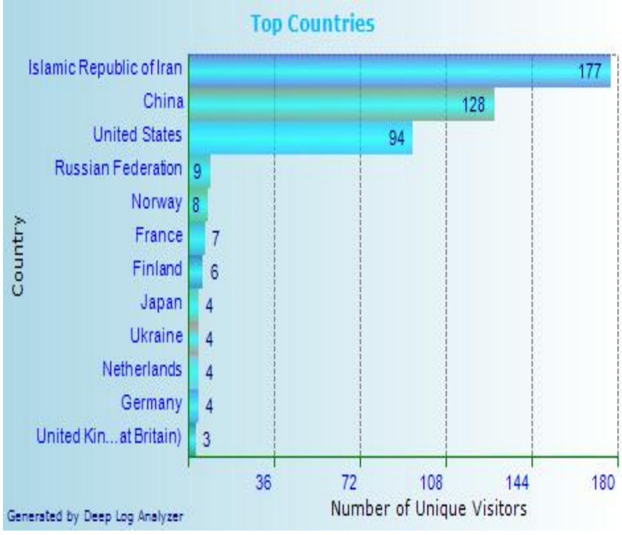


Fig 3: Visitor's geographical location

Figure 3 shows the geographical location. The most visited countries are Iran, China, and America respectively.

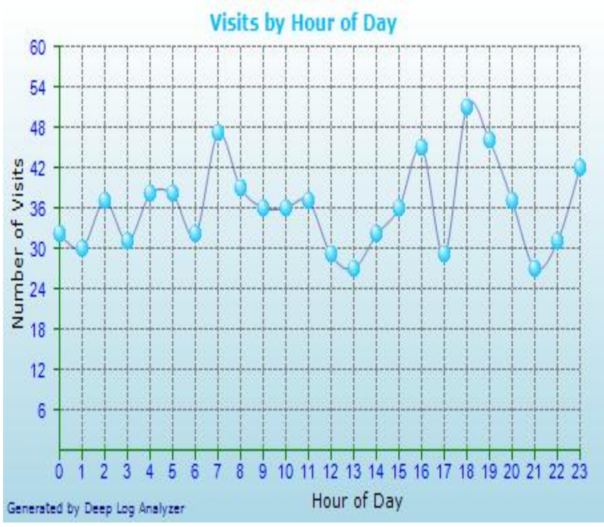


Fig 4: All those visits on weekdays

According to figure 4, the most number of visitors show up at 6 on average and these hours of the day are known as the high traffic time. The website administrator can proceed marketing and publicizing of his products.

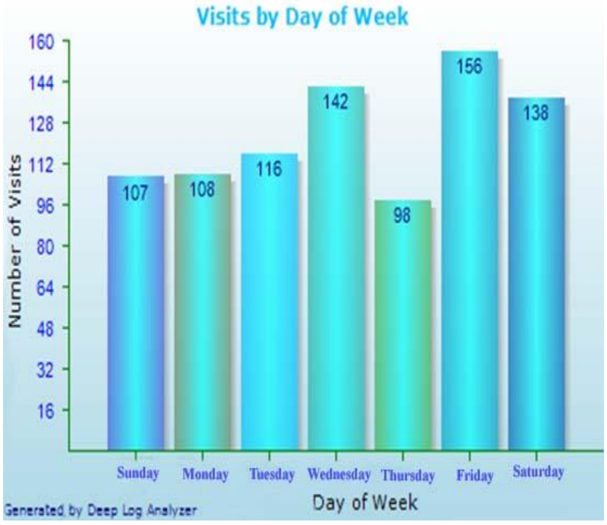


Fig 5: Views on the day of the week

Figure 5 shows user's visit in which Friday is the most visited day of weekdays. The website administrator is able to market his own products by being kept informed.

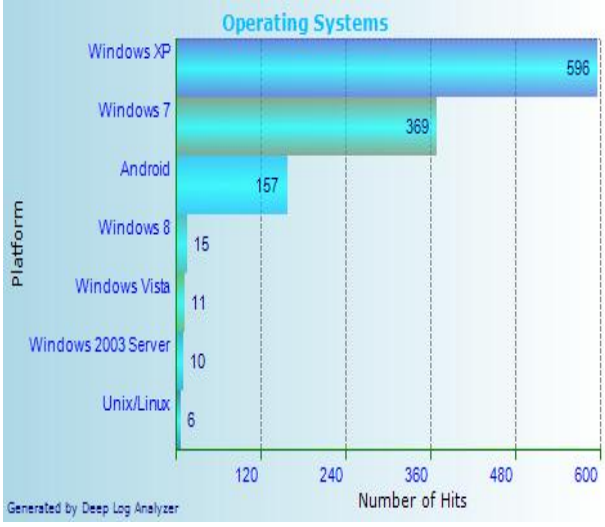


Fig 6: most use Operating Systems

Figure 6 reports show a report of visitor's operating systems. Most website visitors have Windows XP, Windows 7, android. Website administrator is able to provide service according to user's operating system by the below report.

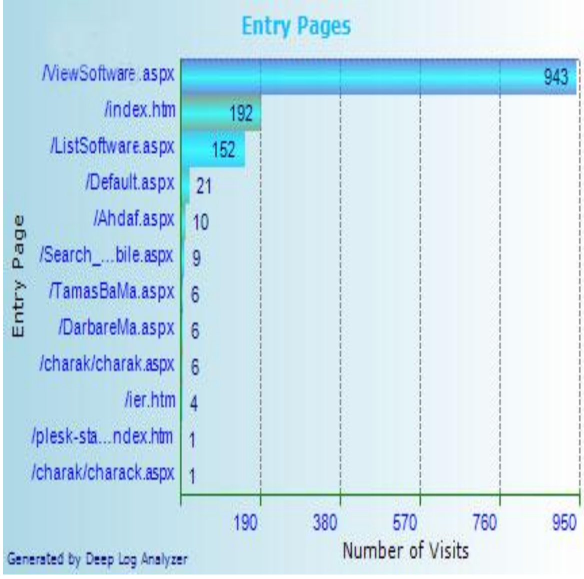


Fig 7: most page views

Figure 7 represents a report of the number of requested pages by visitors. As it's shown, 'ViewSoftware.aspx' page comprises of most visits.

8. CONCLUSIONS

The present study is the use of Web Usage mining through Server File Log with the aim of better service. In the second section of this paper we investigated web mining and its sorts.in the same way, we discussed website Server File Log in the third section and it is stated that all of the given requests to the server is stored in this part. Similarly, It was attempted to explain stages of web usage mining pincipally.in the fifth section, first stage of web usage mining (data preparation) was introduced .then, in the sixth section, to prove all the mentioned subjects, we modeled a download website, and set server log file to the software, so we succeeded to extract the patterns of user's behavior and show it in the form of picture. A website administrator who is active in field of marketing can identify the behavior of users by utilizing this method, so the website administrator is able to maintain clients and proceed better marketing and publicizing.

9. REFERENCES

[1] H.Hannah , K.Thangavel , A. Pethalakshmi, "Rough set based Feature Selection for Web Usage Mining," Conference on Computational Intelligence and Multimedia Applications, 2007, Vol. 1, pp. 33-38.

[2] K.Etminani, M. Akbarzadeh-T and N. Raeeji Yanehsari, "Web Usage Mining: users' navigational patterns extraction from web logs," 2009, IFSA-EUSFLAT, pp.396-401.

[3] D. Singh S, S.Verma, "Web Usage Pattern Analysis Through Web Logs: A Review," Computer Science and Software Engineering, India, 2012. pp.49 -53,

[4] V.Chitraa , A.Selvadoss Thanamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing," International Journal of Computer Applications. India, 2011, Vol 35, No 9, pp. 23-27.

[5] Madhak, Nirali N, Kodinariya, Trupti M., Rathodand Jayesh N, "Web Usage Mining: A Review on Process, Methods and Techniques," International Journal of Computer Applications.India, 2013,

[6] Y Yan LI, Boqin FENG, Qinjiao MAO, "Research on Path Completion Technique In Web Usage Mining", IEEE International Symposium On Computer Science and Computational Technology, 2008,

[7] N.Lakshmi, R. Sekhara Rao, S. Satyanarayana Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis," International Journal of Computer Applications. India, 2013, Vol 2, No 4, pp. 274-279.