Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 05, May-2018

Usage Analysis of Web Access Behavior

Ankita Gaur
Bhabha Institute of Technology
Kanpur, India

***Abstract*—To identify the user behavior by analyzing access into the web server. In this paper, the various concepts of data mining are used and also some traditional data mining concepts for exploring the web usage. Data mining techniques to discover usage patterns of any user to accessing web server that will be good to understand the web based application. The web log file contains the very significant information about the web server. There are various log file between client and server which hold necessary information for discovering pattern. This will help the system administrator & web designer to improve their effectiveness and performance in order to give the better business outcome. The increase in the popularity of a web usage mining, where data analysis is most important part to analyze the user behavior in order to better serve user. And the web designer must be able to improve their system by determining the failure, system errors and broken links. Our main task is to eliminate unnecessary or irrelevant data from the server logs likes images to make faster and efficiently work of each request and making the website effective. For this purpose, data preprocessing is the method to eliminate irrelevant fields from accessing a log file. There are several methods to identify the statistical information about the web user and how the web server access efficiently through pattern discovery.**

Keywords— Web-Usage Mining, Webserver, Web-Server Behavior, Data Preprocessing.

I. INTRODUCTION

Web mining is the application of data mining technique to discover pattern from the web. It is used to understand the behavior of customer by which evaluating the effectiveness of web sites. And this sometimes also enhances the marketing objective or campaign. The discovery of patterns in data through different strategies which are as follows:

a) Content mining: -Web content mining is the mining, extraction and integration of useful data, information and knowledge from web page contents.

b) Structure mining: - Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

c) Usage mining: - Web usage mining is the process of extracting useful information from server logs i.e. user history. Web usage mining is the process of finding out what users are looking for on different web sites. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

This paper focuses on the data preprocessing for eliminating irrelevant data from the sever log by which we efficiently extract the meaningful pattern from the usage history of the user. This is eliminating the irrelevant data from accessing the log-files. The outcome of the preprocessed data will analyze and this will enhance the performance improvement of the web sites design. There are previously various techniques to analyze access behavior of web usage analysis, which provide the statistical information about the access behavior of the web server. Miri @d server: - This is a computer based system, which produces descriptive statistical information about the Web user's searching behavior [3]. Cobweb:-It is an unsupervised machine learning algorithm. It is belongs to the conceptual clustering family, which is particularly suitable to symbolic training data, as it is the case here, where the training examples are the access sessions. It is used to organize the user of the site who follows similar paths into a small set of communities. [2]

II. SERVER ORGANIZATION

We provide in this section a detailed description of the server structure.

2.1 The Model: - The model implies three families of data which it can exploit on the one hand log-files data and on the other hand bibliographic data and commercial data. Log-data

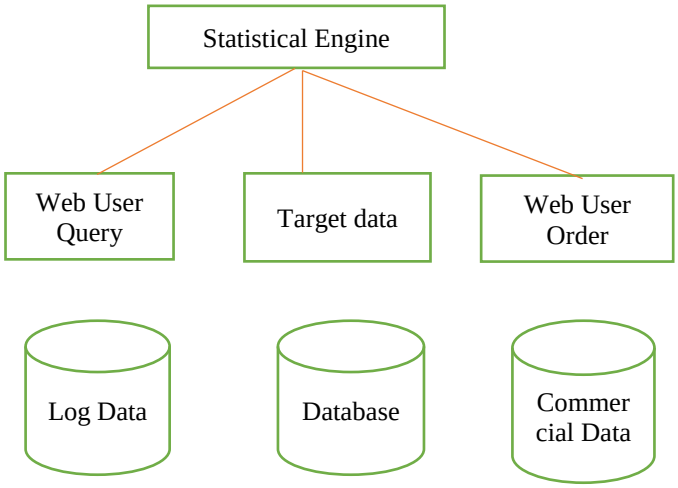

Figure 1. The model [3].

2.2 Web Logs: - When a user or a web browser send a request to the web server activity information stored in the directory called web logs. The log file is generated through the web access and it is the primary source of the raw data. This log file is used for debugging purpose. A log file is simply not a cookie and hence it is located in three different places which are as follows:

a) Server side logs: - These logs generally contain the most complete and accurate usage data and these logs contain the sensitive personal information and which causes the server owner usually keeps them closed.
b) Proxy server side logs: - The proxy server takes the request from the web browser and passes to the web server and after that return the result back to the user by the web server. A

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 05, May-2018

web logging system performance declines if it is employed because each user request processed by the proxy simulator.

c) Client side logs: - The http cookie will also be used for this purpose. And participants remotely test a web-site by downloading special software that records the web usage and modifying the source code of an existing browser. This technique makes hard to achieve compatibility with a different platforms or web browsers.

2.2 Web log structure: -

Web Server logs are plaintext (ASCII) files that are independent from the server platform. There are some distinctions between server software, but traditionally there are four types of server logs: Transfer Log, Agent Log, Error Log and Referrer Log. A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. Most logs use the format of the common log format which is given below:

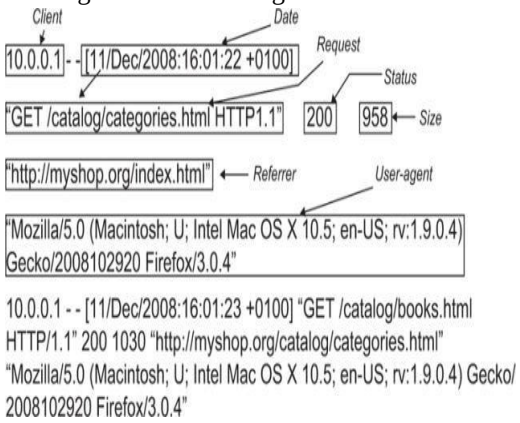

Figure 2.Log File of requested HTTP

This log file contains the following information:

a) Remote IP address and Domain Name both will uniquely define the host. And one IP address assign for only one Domain.
b) Authenticate user: User authentication require through the username and password.
c) Entering and exiting date and time.
d) Mode of request which is either GET, POST, HEAD method.
e) Tell http version.

f) Status code of the http request which return the value e.g. 200 for „ok" and 404 for „not found".
g) Remote URL.
h) Request line exactly as it came from the client.
i) Remote log and agent log.
j) Requested URL.
k) Bytes: the content length of the document transfer.

2.2.1 Status code for HTTP response

The hypertext transfer protocol is an application protocol for distributed collaborative, hypermedia information system. HTTP is the foundation of data communication for the World Wide Web (WWW). HTTP is a stateless protocol because each command it's executed independently. Since 1990, HTTP specifies the common used version HTTP1.1 under RFC2068.

And there are listing some status code for HTTP:

Code	Status	Code	Status
201	CREATED	400	Bad Request
202	ACCEPTED	401	Unauthorized
203	Non-authorative information	402	Payment Required
204	No Content	403	Forbidden
205	Resent Content	404	Not Found
301	Moved Permanently	406	Not Acceptable
302	Found	408	Request timeout
304	Not Modified	100	Switching protocol
305	Use Proxy	200	OK
409	Conflict	502	Bad Gateway

Figure 3.Status code for HTTP request.

III Eliminating the irrelevant details using web usage Mining

There are three stages of web usage mining which are data preprocessing, pattern discovery and pattern analysis. In data preprocessing which removes involves irrelevant details about data. In pattern discovery data mining techniques are used to extract pattern of usage from web data. Pattern discovery is the key process web mining and its covers the different algorithms and techniques from several research areas such as data mining, machine learning, statistics and pattern recognition. Pattern analysis is the rule for extracting pattern from the output of the pattern discovery process by eliminating the irrelevant patterns. It is the final stage of web mining. Our work mainly focuses on the web log file and hence the contents need to be preprocessed. And it's eliminate the irrelevant fields from access log file. Cleaning of web log file through data preprocessed and the following steps used in the removal of irrelevant fields:-

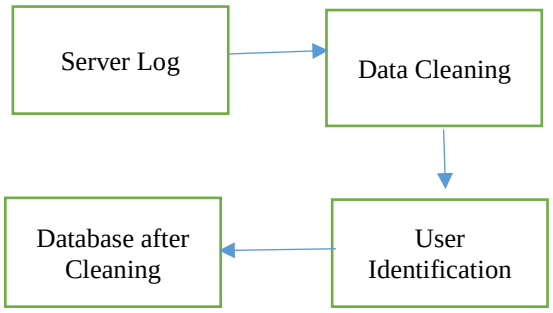

Figure 4. Functioning of data preprocessor.

Data cleaning is a procedure to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. And before all this, data integration task may also perform such as combining data from multiple sources into a coherent store and multiple log file into one. The non-analyzed resources such as images, multimedia files should also be removing. The entries which shoe the status of error and failure may also be removed. By filtering out the request for the particular

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 05, May-2018

web-sites our response is faster as compared to if you include the irrelevant fields such as images or status report. And size is also reduced by their original size after the data cleaning. And log entries must be partitioned into multiple clusters using series of transaction to identify modules. User identification: Identifying the user by their IP address because IP address is unique for the different user. If different user uses the same IP address then the then make a reasonable assumption to identify the different user with in the network. There is IRCTC server log file of seven days which gives the complete idea of the reduction in the percentage compare to original.

Server Log File	IRCTC
Duration	2 Weeks
Original Size	150 MB
Reduced Size after Preprocessing	39 MB
Percentage in reduction	74
No of unique users	185

Figure 5. Preprocessing Data of IRCTC.

IV WEB-USAGE FACTOR

There are two indicators dealing with the usability of the web user based on information retrieval and web customer's orders. The first on is the Web customer order factor and second one is the Web usability factor. These two factors can be used for evaluating online information sources like server log, proxy and web browser log of the user access behavior. Many times an information source is used or displayed by online users and it is well known situation in information retrieval. Now, we are describing the two factors in details for analyzing the usage behavior of the user.
E.g. data log of scientific domain related to each journal by user.

4.1 Customer Order Factor

This is the proportion of articles of a journal ordered by Web customers in a period of time from t_0 to t_1 by the total number of articles published in this journal and stored until t_1.

$$COF_m = \frac{\sum_{i=t_1,t_2} \sum_{m=\pi} ord_i(m,n)}{JT_{t_1}}$$

Where: ordered documents, JT = journal title, N= publication year, JT_{t1} = journal title articles in all publication year which stored until t_1. We can also introduce the notion of information obsolescence in the calculation of COF to observe the Customer Order Factor evolution by publication year.

$$COF(PY)_m = \frac{\sum_{i=t_1,t_2} \sum_{m=\pi} ord_i(m,n)}{JT_{t_1}(PY)}$$

Where, PY publication year, $[t_0,t_1]$= period of time,
JT_{t1} = journal title articles in all publication year which stored until t_1.

4.2 Web Usability Factor

This is the proportion of articles of a journal displayed by Web users in a period of time from t0 to t1 by the total of articles published in this journal and stored until t1.

$$WUF_m = \frac{\sum_{i=t_1,t_2} \sum_{m=\pi} dr_i(m,n)}{JT_{t_1}}$$

Where, JT= journal title, dr= display records, n= publication years, JTt1 = journal title articles in all publication year which stored until t1. The notion of information obsolescence can be also introduced in the calculation of WUF to observe the Web Usability Factor evolution by publication year.

$$WUF(PY)_m \frac{\sum_{i=t_1,t_2} \sum_{m=\pi} dr_i(m,n)}{JT_{t_1}(PY)}$$

Where, PY=publication year,
dr= display record,
[t0,t1]= period of time,
JTt1 = journal title articles in all publication year which stored until t1. [2] These two factors also responsible for the access behavior of the user in any web site like in the website containing the journal related to scientific field during the specified time period.

V SOME RESULTS

We are going to analyze any server log file like IRCTC of size 150 and also various analyses has been carried out to identify the user. And errors which are occurred in accessing web are also examined. By analyzing the server log file, we are able to determine the most active and least active day and most hits on the particular day. And we are going to determine the status code of commonly occurrence of errors, failures, while transmission of information. And also extract information by eliminating the failure. The details of the IRCTC website by the access behavior of user,

Day	Entries	Unique Users	No. of Hits	Failure
1	65026	756	5026	2931
2	62502	625	2502	2645
3	25640	203	5500	2456
4	56452	548	6321	2163
5	59133	608	2536	4561

Figure 5 Log and corresponding days no. of hits

This data tells about the number of hits of any user on the particular day. The web usage access of any user from the server log file we can extract the hit ratio of any web site on a specified day or a time period.

VI CONCLUSIONS

This paper tells about the web usage data that goes beyond widely used web usage statistics. The work presented here belongs in the research area of data mining as applied to data on the web. Through the analysis of the web usage which determine the access behavior of the user to the web server.