Web Usage Mining: Users Behavior in Web Page Based on Web Log Data

S.Kamalakkannan
Research Scholar, Department of BCA and IT, Vels University, Chennai, India.

Dr.S.Prasanna
Associate Professor, Department of MCA, Vels University, Chennai, India.

Abstract – The main purpose of this paper is to Study the User Behavior using the Web Log. This paper presents the theory and knowledge related to Web Logs and then presents a Web Log mining process, discusses about user browsing behavior and interest, and Web Mining Technology. The paper also discuss about how to estimate of user behavior is carried out based on the study of web logs. The work can provide a theoretical basis for the management and optimization of the site for site managers. Experiments have showed that the work is effective and efficient.

Index Terms – Web Log, Web Usage Mining, user behavior analysis.

1. INTRODUCTION

Internet is acting as a major source of data. As the number of web pages continues to grow the web and provides the data miners with just the right ingredients for extracting information. In order to provide this growing need a special term called Web mining was coined. Web mining is the Data Mining technique that automatically discovers or extracts the information from the web documents .The process of Web mining is divided into four stages: source data collection, data preprocessing, and pattern discovery and pattern analysis.

Source Data Collection:

In mining of Web data, Web log files on the Web server are the main source of data. Web log files contain the history of the visitor's browsing behavior. Web log files include the server log, agent log and client log.

2. WEB MINING

Web mining can be classified into web content mining, web structure mining and web usage mining.

2.1. WEB CONTENT MINING

Web Mining is basically extracts the information on the web. Which process and access the information on the web. It is web content mining. Many pages are open to access the information on the web. These pages are content of web. Searching the information and open search pages is also content of web. Last accurate result is defined the result pages content mining.

The various contents of Web Content Mining are

- Web page
- Search page
- Result page

2.2. WEB STRUCTURE MINING

We can define web structure mining in terms of graph. The web pages are representing as Nodes and Hyperlinks represent as edges. Basically it’s shown the relationship between user and web. The motive of web structure mining is generating structured summaries about information on web pages. It is shown the link on one web page to another web page. The various contents of Web structure mining are

- Links Structure Mining
- Internal Structure Mining
- URL Mining

2.3. WEB USAGE MINING

It is the discovery of meaningful pattern from data generated by client server transaction on one or more web localities. A web is a collection of inter related files on one or more web servers. It is automatically generated the data stored in server access logs, refers logs, agent logs, client sides cookies, user profile, meta data, page attribute, page content and site structure. Web mining usage aims at utilize data mining techniques to discover the usage patterns from web based application. It is technique to predict user behavior when it is interacting with the web. Web usage mining deals with understanding the behavior of users by making use of Web Logs that are generated on the server while the user is accessing the website. Web Usage Mining itself can be classified further depending on the usage data considered. It can be classified into Web Server Data, Application Server Data and Application Level Data.

3. OVERVIEW OF WEBLOG

A Web Log file records activity information when a web user submits a request to a Web Server. The main source of raw data is the web access log which we shall refer to as a log file. Log files are originally meant for debugging purposes.

3.1. LOCATION OF A LOG FILE

A Web Log is a file to which the Web server writes information each time a user requests a website from that particular server. A log file can be located in three different places:

- Web Servers
- Web proxy Servers
- Client browsers

3.2. Web Server Log files

The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser. In the server which collects the personal information of the user must have a secured transfer.

3.3. Web Proxy Server Log files

A Proxy server is said to be an intermediate server that exist between the client and the web server. Therefore web server gets a request of the client via the proxy server then the entries made to the log file, the information of the proxy server is not of the original user. These web proxy servers maintain a separate log file for gathering the information of the user.

3.4. Client Browsers Log files

These kinds of log files can be made to reside in the client's browser window itself. Special types of software exist which can be downloaded by the user to their browser window. Even though the log file is present in the client's browser window the entries to the log file is done only by the Web server.

3.5. Web Log Structure

Web Server Logs are plain text files, that is independent from the server platform. There are four types of server logs: Transfer Log, Agent Log, Error Log and Referrer Log. The first two types of log files are standard. The Referrer Log and Agent Log may or may not be "turned on" at the server or may be added to the Transfer log file to create an "Extended" Log File format. 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. Web log file is server recording information of user's requests for resources to a specific site each time. Most logs use common log format. The following is a log fragment taken from a web server:

```
68.249.65.107 - - [12/Oct/2015:04:54:20 -0400] "GET /support.html HTTP/1.1" 200 11179 "-" "Mozilla/5.0(compatible;Googlebot/2.1;+http://www.google.com/bot.html)".
```

This reflects the information as follows:

Remote IP address or domain name: An IP address is a 32-bit host address defined by the Internet Protocol; a domain name is used to determine a unique Internet address for any host on the internet. One IP address is usually defined for one domain name.

Authorized User: Username and password if the server requires user authentication

Entering and exiting date and time.

Modes of request: GET, POST or HEAD method of CGI (Common Gateway Interface).

Status: The HTTP status code returned to the client, e.g., 200 is "ok" and 404 are "not found".

Bytes: The content-length of the document transferred.

Remote log and agent log.

Remote URL.

"request:" The request line exactly as it came from the client.

Requested URL

4. STATUS CODES OF HYPER TEXT TRANSFER PROTOCOL

The Hypertext Transfer Protocol (HTTP) is an application-level protocol has been in use by the World-Wide Web since 1990. The first version of HTTP, referred to as HTTP/0.9, was a simple protocol for raw data transfer across the Internet. HTTP/1.0, as defined by RFC 1945. Status codes of Hypertext Transfer Protocol are shown in Table 1 to indicate error conditions as well as successful transmission of data.

Table 1. Status Codes of Hypertext Transfer Protocol

Status code	Significance
101	Switching Protocols
200	OK
201	Created
202	Accepted
300	Multiple Choices
304	Not Modified
400	Bad Request
401	Unauthorized
403	Forbidden
404	Not Found

408	Request Time-Out
500	Server Error
502	Bad Gateway
504	Gateway Time-Out
505	HTTP Version not supported.

5. USER BEHAVIOR ANALYSIS BASED ON WEB LOG

The three main steps of Web usage mining are data preprocessing, pattern recognition and pattern analysis. Data preprocessing includes removing unwanted data, in pattern discovery phase, extracting usage pattern from Web data by data mining techniques. Pattern discovery is the key part of Web Mining; it covers a number of research areas, such as data mining, machine learning, statistics and pattern recognition. Statistical analysis, association rules, clustering, classification, sequence pattern and dependence modeling techniques are all used to discover the rules and patterns. The knowledge can be found in the following aspects, performance form of rules, tables, and icons, or performance form of other features, comparisons and predictions, as well as data classified from Web access logs. The purpose of this process is to extract interesting rules or patterns form output of pattern discovery process by eliminating irrelevant rules or patterns.

6. EXPERIMENTAL DATA AND RESULTS ANALYSIS

In the process of this article, we have analyzed 514MB log file of server, do different analyses to identify the user behaviors. By Table 2, we can get the detailed information of user access on day by day, in which the administrator can find the third day should be the most active day for users, at least the user clicks the most on this day, while administrators can analyze the user's basic situation through long-term data analysis, if it needs to shut down the server, they can select the day when less visitors access. Figure 1 shows the overall situation of logged in users, Figure 2 shows the number of all clicks except login failures, and Figure 3 shows the range of independent visitors.

Table 2 information of user record

S.No **(Day)**	**Log record**	**Number of IP address**	**Independent visitor**	**Click rate**	**Number of access failure**
1	62576	7097	8567	27045	2712
2	58198	6140	7778	20373	2045
3	95643	17003	9785	27663	1662
4	67454	15372	8543	21574	1692
5	85255	26143	11143	29074	1369

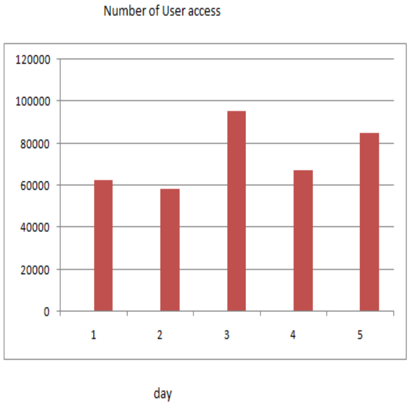

Figure 1

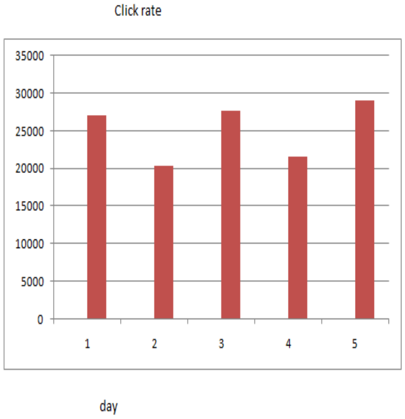

Figure 2

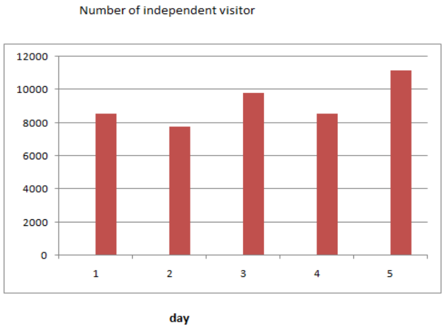

Figure 3

7. CONCLUSION

Based on theory of Web log mining, this paper makes presentation and comparison for sources of a variety of logs, combining with the actual Web server log, analyzes contents of the log record in detail. Finally combining with web server logs, and multifaceted analysis for series of user access situation. The work can give specific guidance for the operators to optimize front page, improve user experience, as well as optimize structure. For example dead chain in the website must be removed, unreasonable page jumps and substandard logic must be fine-tuned in detail, so as to retain the old users and attract new users further. The use of those data is become more and more available and they provide additional information that is very valuable to determine the customer behavior.

REFERENCES

[1] Brijendra Singh, Hemant Kumar Singh: Web Data Mining Research: A Survey, IEEE, 2010.

[2] K.R. Suneetha, R.Krishnamoorthi: Identifying UserBehavior by Analyzing Web Server Access Log File, IJCSNS VOL 9,April2009.

[3] Chen Baoshu, Jing Qimin. Data Preprocessing of Web Data Mining, [J] Computer Engineering, 2002, 28 (7):125-127.

[4] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong- Soo Jeong et al.A Framework for Mining High Utility Web Access Sequences [J]. IETE technical review, 2011, 28 (1) :3-16

[5] Song Jiangchun, Shen Junyi. Efficient and Pluripotent Mining Algorithm of Web Logs [J] Computer Research and Development, 2001, 38 (3):328-333.Commercial College, 2007.

[6] Mahendra Pratap Yadav: Mining the customer behavior using web usage mining in e-commerce IEEE,2012.

[7] Neha Goel, C.K. Jha: Analyzing users Behavior from web access logs using automated log analyzer tool IJCA,2013.

[8] Sanjay Kumar Malik, Nupur Prakash, S.A.M. Rizvi Ontology and Web Usage Mining towards an Intelligent Web focusing web logs 2010 International Conference.

[9] Ashwini ladekarPooja Pawar, Web Log Based Analysis of User's Browsing, International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1680-1684.