ISSN 2250-3153

Extracting Knowledge from User Access Logs

Aditi Shrivastava, Nitin Shukla

Shri Ram Institute of Technology, Jabalpur, India

Abstract- As the size of web increases along with number of users, it is very much essential for the website owners to better understand their customers so that they can provide better service, and also enhance the quality of the website. To achieve this they depend on the web access log files. . It is is a file to which the Web server writes information each time a user requests a resource from that particular site. In this paper we will study web access log files and the information we can mine from logs which is useful in understanding user behavior. This information is used in restructuring and redesigning of the website.

Index Terms- web mining, usage mining, server logs, pattern mining.

I. INTRODUCTION

In this world of Information Technology, accessing information is the most frequent task. Every day we have to go through several kind of information that we need and what we do? Just browse the web and the desired information is with us on a single click. Today, internet is playing such a vital role in our everyday life that it is very difficult to survive without it. The World Wide Web (WWW) has influenced a lot to both users (visitors) as well as the web site owners. The web site owners are able to reach to all the targeted audience nationally and internationally. They are open to their customer 24X7. On the other side visitors are also availing those facilities. the growth in number of web sites and visitors to those web sites has increased exponentially Due to this growth a huge quantity of web data has been generated. To mine the interesting data from this huge pool, data mining techniques can be applied. But the web data is unstructured or semi structured. So we can not apply the data mining techniques directly. Rather another discipline is evolved called web mining which can be applied to web data. Web mining is the use of data mining techniques to automatically discover and extract information from

Web mining is categorized into 3 types.

1. Content Mining (Examines the content of web pages as well as results of web Searching)
2. Structure Mining (Exploiting Hyperlink Structure)
3. Usage Mining (analyzing user web navigation)

A) Web usage mining

Web usage mining is a research field that focuses on the development of techniques and tools to study users web navigation behavior. Understanding the visitors navigation preferences is an essential step in the study of the quality of an electronic commerce site. In fact, understanding the most likely access patterns of the users allows the service provider to customize and adapt the site's interface for the individual user, and to improve the site's static structure within the underlying hypertext system.

When web users interact with a site, data recording their behavior is stored in web server logs. These log files may contain invaluable information characterizing the users experience in the site. In addition, since in a medium size site log files amount to several megabytes a day, there is a necessity of techniques and tools to help take advantage of their content.

Five major steps followed in web usage mining are

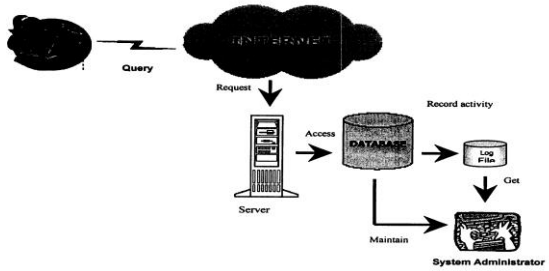
1. Data collection – Web log files, which keeps track of visits of all the visitors
2. Data Integration – Integrate multiple log files into a single file
3. Data preprocessing – Cleaning and structuring data to prepare for pattern extraction
4. Pattern extraction – Extracting interesting patterns
5. Pattern analysis and visualization – Analyze the extracted pattern
6. Pattern applications – Apply the pattern in real world problems

II. SERVER LOGS

A **server log** is a log file (or several files) automatically created and maintained by a server of activity performed by it. It is a file to which the Web server writes information each time a user requests a resource from that particular site. Log file data can offer valuable information insight into web site usage. It represents the activity of many users over a potentially long period of time .A Web server when properly configured, can record every click that users make on a Website. For each click in the visit path, the server adds to the log file information about user request.

The logs collect data on the server in the files of specific format. Measures hold information about web site usage by recording how users visit the web site and how active they are. Depending on the log format structure, different data is stored. Usually logs contain data such as: client's IP address, URL of the page requested, time when the request was send to the server etc. This data is used later as the basis of usage behavior discovery. Depending on server settings log files format can differ. The form of web logs files standard changes over years as there was more requirements for the web log processing.

A. How Log Files are Created



When a user sends queries to the server, requested databases will be retrieved. At the same time ,the user session including the URL, Client's IP address, accessing date and time , query stem will be recorded in the server logs. These server logs can be preprocessed and mined in order to get some insight into the usage of a server site as well as the user's behavior.

B. Information in Server Logs

Server logs can be used to glean a certain amount of quantitative usage information. Compiled and interpreted properly, log information provides a baseline of statistics that indicate usage levels and support and growth comparisons among parts of a site or overtime. Such analysis also provides some technical information regarding server load, unusual activity, or unsuccessful requests, as well as assisting in marketing and site development and management activities.

Web server logs are plain text(ASCII) files ,that is independent from server platform. There are some differences between server software, but traditionally there are four types of server logs:

1. Common log or access log
2. Agent Log
3. Error Log
4. Referer Log

The first two types of log files are standard. The referrer and agent logs 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. Each HTTP protocol transacxtion ,weather completed or not, is recorded in the logs and some transactions are recorded in more than one log.

1. Access Logs

The first of the three logs is Common log, sometimes referred to as the access log, which is identical in format and syntax to the NCSA Common log format. There are 19 attributes in this type of log file.

a) Date – the date from GMT are recorded for each hit.
b) Time – the time of transaction.
c) Client IP Address – Client IP is the number of computer who access or request the site.
d) User Authentication – some web sites are set up with a security feature that requires a user to enter username and password. Once a user logs on to a website, that user's "username" is logged in the fourth field of the log file.
e) Server name – Name of the server.
f) Server IP address – server IP address is a static provided by Internet Service Provider. This IP will be a reference for access the information from the server.
g) Server Port – This is used for data transmission, usually port 80.
h)Server Method (HTTP Request) – the word request refers to an image, movie sound, pdf, txt, HTML file and more. Currently, there are three formats that Web servers send information in GET, POST, and Head.
i) URI Stem – URI Stem is path from the host. It represents the structure of the websites.
j) Server URI Query – URI-Query usually appears after sign "?". This represents the type of user request and the value usually appears in the Address Bar.
k) Status – This is the status code returned by the server. There are four classes of codes:
Success(200 series) Redirect(300 series) Failure(400 series) Server Error(500 series
l) Bytes Sent – amount of data returned by the server, not counting the header line.
m) Bytes Received – amount of data sent by the client to the server.
n) Time stamp – is used to determine how long a visitor spent on a given page.
o) Protocol version – HHTP protocol being used.
p) Host – is either the IP address or the corresponding host name of the remote user requesting the page.
q) User Agent – is reported by the remote user's browser. Typically is the string describing the type and version of browser software being used.
r) Cookies – can be used to track individual users thus make the sessionizer task easier.
s) Referer – the referring page, if any, as reported by the remote user's browser. It is possible to analyze the following variables in the access log:

The data from Access Logs provides a broad view of a Web server's and users. Such analysis enables server administrators and decision makers to characterize their server's audience and usage patterns.

2. Agent Log

The agent log provides data on a user's browser, browser version, and operating system. This is the significant information, as the type of browser and operating system determines what a user is able to access on a site. Sample is presented below:

```
Mozilla/3.0 (Win 95; 1)
```

Browser: The type of browser used to access a website. There are several different Web browsers on the market today, each of which has different viewing capabilities.
Browser version: Each browser has its own capabilities.

Operating System: The type of computer and operating system used to determine the Graphical User Interface (GUI) of a website depending on the computer platform.

The Agent log information is essential for the design and development of Websites. Without such information, server administrator could design sites that require viewing capabilities that vast majority of the site's users do not posses. This could lead to wasted effort by the server administrators. Worst still, this can lead to improperly displayed web content, thus effectively rendering the site useless to the user.

3. Error Log

The average Web user will receive an "Error 404 File Not Found" message several times a day. When a user encounters this message, an entry is made in the Error Log. The analysis of Error log data can provide important server information such as missing files, erroneous links, and aborted downloads. This information can enable server administrators to modify and correct server content, thus decreasing the number of errors users encounter while navigating a site.

4. Referer Log

The referrer log indicates what other sites on the Web link to a particular server. Each link made to a site generates a Referer Log entry, The Referer Log entry provides the following data:

Example:-

[10/Oct/1999:21:15:05 +0500]
"http://www.ibm.com/index.html"

a) date:time timezone ([10/Oct/1999:21:15:05 +0500] in the example)

The date and time stamp of HTTP request

b) Referral: **referrer** ("http://www.ibm.com/index.html" in the example)

The referrer is the URL of the HTTP resource that referred the user to the resource request If a user is on a site, and clicks on a link to another site, then another entry will receive an entry in their Referer Log. The log will show that the user came to the other site via first link.

III. LOG ANALYSIS

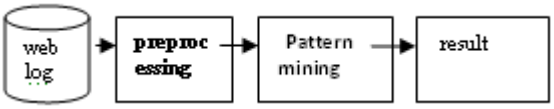


Fig. 3: Steps for realization of logs

Logs are processed to extract statistical information Web logs are link of past activities left behind by the previous users of a website. These historical logs are embedded with significant information about the users and how a website is being used on a day-to-day basis. Such information are invaluable in today's world of customer-oriented businesses, especially for companies that depend on the web to advertise their services and to web designers who wish to maintain a constant flow of visitors to their website. Various steps are involved in identifying the information extracted from the logs are shown below: first data is collected which is known as web logs then preprocessing techniques are applied to that web log so that we can get relevant information.

A. Findings

After pattern mining we get several findings:-

1. General Statistics- This provide the summary of whole log file. Usually it gives provides the total hits,page views and total visitors

2. Access statistics- this provide information such as most popular access page and most downloadable files

3.visitors statistics – provide the information such as most active country which accesses the website

4. Referrer- This will provide information such as the most used search engines and phrases and keyword used

5. Error- Error is important for the system administrator's website in order to improve the site as well as to reduce the error.

IV. CONCLUSION

Web usage mining is the process of applying statistical and data mining methods to Web log data in order to extract useful patterns concerning the users' navigational behavior, user and page clusters, as well as possible correlations between Web pages and user groups. The web is a most important medium to conduct business and commerce. Therefore the design of web pages is very important for the system administrator and web designers. These features have great impact on the number of visitors. So the web analyzer has to analyze with the data of server log file for detecting pattern. In this paper we tried to give a clear understanding of the web server logs,their types and interesting patterns extracted from the web logs. discovering such information that can be used to improve a business's performance or increase the effectiveness of a particular website. By this we can shorten the pages which are not in the user pattern, also we can record the information of the user, This will also help in evaluating address campaigns, restructuring and redesigning of website Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research

REFERENCES

[1] WebTrend, I. (2008). Webtrends visitor intelligence, Online. accessed on 04/2008. *http://www.webtrends.com/Products/WebTrendsVisitorIntelligence.aspx

[2] FlorentMasseglia, Pascal Poncelet, Rosine Cicchetti, "An efficient algorithm for Web usage mining", Networking and Information Systems Journal. Volume X, 2000

[3] R. Pamnani, P. Chawan "Web Usage Mining: A Research Area in Web Mining"

[4] Qiankun Zhao, Sourav S. Bhowmick, "Sequential Pattern Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003118 , 2003.

[5] S. Rawat, L. Rajamani, "Discovering Potential User Browsing Behaviors Using Custom-Built APRIORI Algorithm", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010

[6] Ming-Syan Chen, Jong Soo Park, Philip S. Yu, "Efficient Data Mining for Path Traversal Patterns", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 10, NO. 2, MARCH/APRIL 1998.

[7] Jianhan Zhu, Jun Hong, John G. Hughes, "Using Markov Chains for Link Prediction in Adaptive Web Sites", Soft-Ware 2002, LNCS 2311, pp. 60–73, 2002

[8] WANG Tong, HE Pi-lian, "Web Log Mining by an Improved AprioriAll Algorithm", World Academy of Science, Engineering and Technology 4 2005

[9] Hengshan Wang, Cheng Yang, Hua Zeng, " Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan", Communications of the IIMA 2006 Volume 6 Issue 2

[10] Paola Britos, Damián Martinelli, Hernán Merlino, Ramón García- Martínez, "Web Usage Mining Using Self Organized Maps", International Journal of Computer Science and Network Security, VOL.7 No.6, June 2007

[11] Masseglia, F., Poncelet, P. and Teisseire, M. (2003). Using data mining techniques on Web access logs to dynamically improve hypertext structure, SIGWEB Newsletter 8(3): 13–19

[12] Facca, F. and Lanzi, P. (2005). Mining interesting knowledge from weblogs: A survey, Data Mining and Knowledge Discovery 53(5): 225–241.

[13] www.google.com

[14] www.webmining.com

[15] www.wikipedia.com

[16] www.techfact.com

AUTHORS

First Author – Aditi Shrivastava, ME(C.S.E) IV SEM, Shriram Institute of Engineering and Technology, Jabalpur (M.P), India
Email id – aditi.aditi25@gmail.com

Second Author – Mr Nitin Shukla, Lecturer , M CA Dept, Shriram Instt of Engg and Technology, Jabalpur (M.P) ,India