International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 4, April - 2013

Use Of Web Log File For Web Usage Mining

Savita Devidas Patil
Assistant Professor
Department of Computer Engineering
SSVPS's B.S.Deore College of Engineering
Dhule, INDIA

Abstract

Many web page designers may be unaware that web servers record transaction information each time they send a file to a browser. Others may know that a server log exists but they may see it only as a source of general statistical information such as site use distributed over time or counts of the number of times that each page was served. This paper describes how server logs can be used to give designers a much more detailed view of how users are accessing their site. Server logs can be used to monitor use patterns and employ them to improve the design and functionality of the web site using web usage mining. Web log data has been used to analyze and redesign a wide range of web-based material, including: online tutorials, databases, fact sheets, and reference material. Web usage mining an application of data mining can be used to discover user access patterns from weblog data.

1. Introduction

Server log files are records of web server activity. Provides details about file requests to a web server response to those requests, Web logs are maintained by Web servers and contain information about users accessing the site. World Wide Web is the biggest and most widely known information source that is easily accessible and searchable. Computers that deliver the web pages are called as web servers. The Web server stores all of the files necessary to display the Web pages on the user's computer. All the individual web pages combines together to form the completeness of a Web site. When user accesses website, log file are created. Log file recorded information about each user. Tremendous uses of web, web log files are growing at faster rate. Web servers log files that can collect activity information when a user accesses to a Web Server. Log files are files that list the actions that have been occurred. These log files reside in the web server. These web log files give the information about the behavior of user[7]. In business area, extracting information about user's behavior plays an important role. Web log files are reference data on web pages. Many commercial web log analyzer tools are available in the market that analyzes the web server log data to produce different kinds of statistics.

2. Web Log File

Web log files give the information about the behavior of user. A Web logs files contains requests addressed to web servers. These are recorded in a chronological order. Web Server logs are plain text (ASCII) files, that is independent from the server platform.

2.1 Web log file data

The Log files in different web servers maintain different types of information. Web log contains basic information shown in following table1

Table 1: Web log file data

User's IP address	IP address of user
User's authentication name	Username and password if the server requires user authentication
The date-time stamp	The time spent by the user in each web page while surfing through the web site.
The HTTP request	The HTTP status code returned to the client,
The Response status	Response status of requested page
The size of requested recourses	Requested resource size
The reference URL	The resource accessed by the user
User's browser identification	Browser from where the user sends the request to the web server.
Visiting Path	The path taken by the user while visiting the web site
Request type	The method used for information transfer is noted

2.2 Types of web server logs

- **Access log file:** Data of all incoming request and information about client of server. Access log records all requests that are processed by server.
- **Error log file:** list of internal error. Whenever an error is occurred, the page is being requested by client to web server the entry is made in error log. Access and error logs are mostly used, but agent and referrer log may or may not enable at server.
- **Agent log file:** Information about user's browser, browser version.
- **Referrer log file:** This file provides information about link and redirects visitor to site.

2.3 Web log entries

Generally there are two types of log entries such as simple log entry and complex log entry.

- **Simple log Entry**

03/08/12<tab>00:53:39<tab>OK<tab>wwta03.prox.aol.com<tab>/MinuteSeven.html<tab><tab>http://drott.cis.drexel.edu/MinuteSix.html<tab>

The tab character with <tab> In the original log, the character is actually a tab. For other web servers the separating character (delimiter) may be a blank space.

Simple Log entry details

1) **03/08/12 -**The date on which this transaction takes place.
2) **00:53:39 –**The time of the transaction. Both time and date in this particular file are the local time and date at my web server.
3) **OK(result) –** The error status of the transaction. This is not always useful.
4) **Wwta03.proxy.aol.com-** The web address of the requesting browser.
5) **/MinuteSeven.html**- Name of file that was send.
6) **<tab><tab>** - Log file field for which no value reported.
7) **http://drott.cis.drexel.edu/MinuteSix.html<tab>** -The URL of page that provided the link to the file sent

- **Complex Log Entry**

03/08/12<tab>21:09:27<tab>ok<tab>school.eaton.tufts.edu<tab>/nextlessonbutton.Gif<tab><tab>http://drott.cis.drexel.edu/MinuteOne.html<tab>

Complex log entry provides more detail such as in above entry, the files were sent to a computer called "school.eaton.tufts.edu." We can guess that this request originated at Tufts University (tuftsedu.) The specifics of which machine is "school.eaton" is an internal matter at Tufts and may not be available to outsiders.

3. Web Log Analyzer (WLA)

The Web log Analyzer (WLA) performs web log analysis. WLA consist of log purge, session detection term propagation and page merging techniques used for web log analysis shown in following figure1.

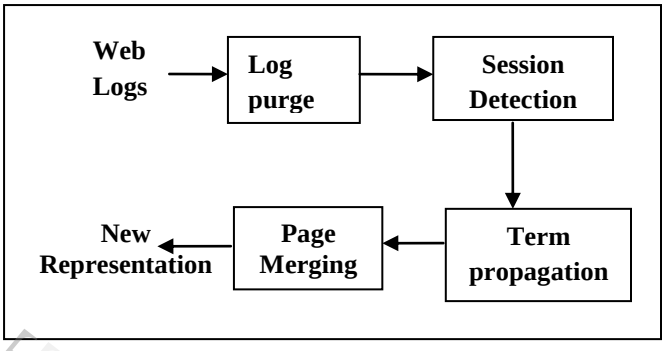

Figure 1: Web log analysis (WLA)

- **Web Log Purge:** Each log entry contains client IP, user name and password ,access time, HTTP request method used, URL, protocol used, status code, number of bytes transmitted, referrer. Since every log entry is not useful, web logs are purged by filtering out all images, video, and audio requests and failed page requests.
- **Session Detection and Representation** A user session is defined as a set of clicks over a limited time by a user. Each user is assumed to be uniquely defined by the IP address recorded in the http request. The whole web server log can be represented as a set of sessions. Entry URL is the first web page visited in a session and entry referrer is the referrer of the entry URL. If the entry referrer is "-", it means that the user typed the URL directly in a browser window. The set of URLs is defined as follows:

$$(url0 , url1 ,----, urlm) \quad (i)$$

And the set of referrers is defined as:

$$(ref0, ref1 ,----, refm) \quad (ii)$$

where $m+1$ is the number of web pages the user visited in a session S_j . Duration is the time staying on each web page. A set of corresponding durations can be represented as:

$$(urld_0, urld_1 ,-----, urld_m) \quad (iii)$$

Duration of each page except the last one can be calculated by:

$$Urld_1 = itim_{e+1} - itime_i \quad \text{(iv)}$$

Where *i* is an integer between 0 to *m*, and $itime_i$ is the time when the user accesses page P_I The duration on a session can be calculated after getting each web page's duration

$$urld_m = \frac{\sum_{i-0}^{m-1} urld_i}{m} \quad \text{(v)}$$

$$SDj = \sum_{i-0}^{m} urld_i \quad \text{(vi)}$$

Where *SDj* is the duration on session *Sj*.

- **Term Propagation.**

To get more terms for each page, propagation is applied within each session. There are two steps in the propagation process: (1) extracting entry terms, and (2) propagating terms along the access path.

- **Page Merging**

In one session, a web page may appear more than once, and in the whole log, a page may appear in different sessions. Because each web page in one session is represented by a term vector and a corresponding weight vector, when merging the pages from different sessions, it is necessary to merge the term vectors and weight vectors into one term vector and weight vector. Therefore, the page is finally represented by one term vector and one corresponding weight vector respectively:

$$Term\_Vector = \{Term_1, Term_2, \ldots, Term_l\} \quad \text{(vii)}$$

$$Weight\_Vector = \{Weight_1, weight_2, \ldots, Weight_l\} \quad \text{(viii)}$$

When we merge the pages from different sessions, the session weight is included in the calculation. The session weight considers two factors: how long the session continues, measured by session duration, and how recently the session happens, and measured by session recency. So the session weight is calculated as

$$SW_j = \log(SD_j \times SR_j \times factor) \quad \text{(ix)}$$

Where SD_j is the session duration, computed (C), *factor* is for scaling purposes (set to 200), and SR_j is the session recency, defined as:

$$SR_j = \frac{1}{\text{Log}(T_{now} - T_{itime\,j})} \quad \text{(x)}$$

Where T_{now} is the current system time, and $T_{itime\,j}$ is the start time of visiting the entry URL.

- **Document Representations**

The goal of the combination approach is to combine document representations into one document representation by merging the terms into one document, and then apply different retrieval models.

4. Web usage mining

Web usage mining is the process of finding out what users are looking for on internet. Some users might be looking at only textual data whereas some other might want to get multimedia data. Web usage mining also helps finding the search pattern for a particular group of people belonging to a particular region[3]. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for the specific needs

4.1 Data Preprocessing

The data present in the log file cannot be used as it is for the mining process. Therefore the contents of the log file should be cleaned in this preprocessing step. The unwanted data are removed and a minimized lo file is obtained[6].

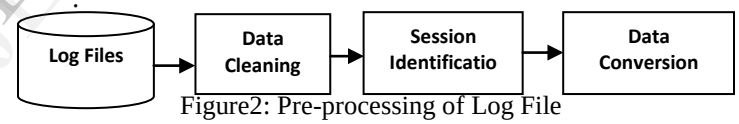

Figure2: Pre-processing of Log File

- **Data cleaning**: In this process the entries made in the log file for the unwanted view of images, graphics, Multi media etc., made by the users are removed. Once these data are removed the size of the file is minimized to a greater extent.
- **Session Identification**: This done by using the time stamp details of the web pages. The total time used by each user of each web page. This can also be done by noting down the user id those who have visited the web page and had traversed through the links of the web page. Session is the time duration spent in the web page.
- **Data conversion**: This is conversion of the log file data into the format needed by the mining algorithms.

4.2 Using log file data in web usage mining

Web usage mining refers to the automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with web resources on one or more Web sites. The data from Web logs, in its raw form, is not suitable for the application of usage mining algorithms.

The contents of the Log files are used in web usage mining. Web usage mining also consists of Navigation Pattern Discovery and evaluating the discovered navigation patterns[1] as shown in figure 3.

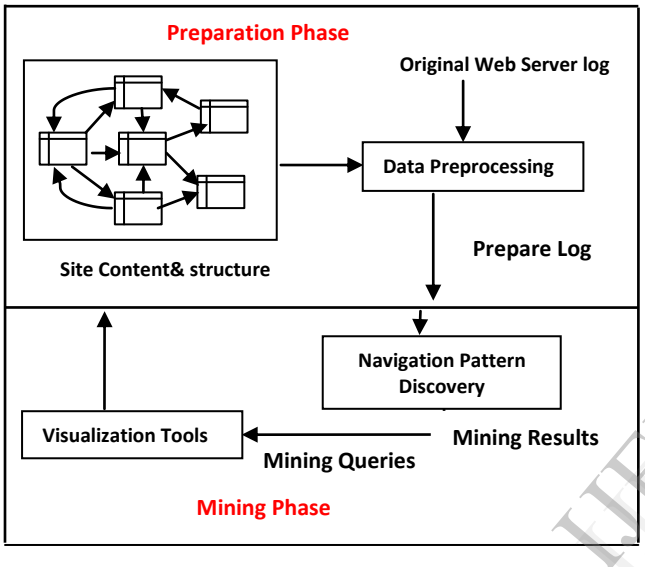

Figure3. Use of Web log data in Web Usage mining

As shown in figure 3, there are two phases, preparation phase and mining phase. In preparation phase using site content and site structure web server log data can be taken which will be preprocess and send to mining phase. In mining phase pattern discovery can be perform as fallows.

1) Navigation Pattern Discovery

The data preparation phase restores the users' activities into sequences of page or script accesses. From them, a miner should test whether the site is being used in accordance with the design objectives. Navigation pattern Discovery performs two miner methods as fallows.

- **Sequence miners**

The discovery of typical usage patterns seems to be exactly what sequence miners are built for. They discover events (here: accesses to pages) that occur frequently together in the same order. Hence, a sequence miner can find all sequences of Web pages that have been frequently accessed together. However, most of these sequences would be of a trivial nature.

- **Web log miners**

This system has been designed in accordance with the increased demand for intensive interaction with human users. This interaction is based on a powerful mining language in which expert users can express their background knowledge, guide the miner and gradually refine or refocus the discovery process, according to the mining results obtained after each query. The mining language such as LOGML is used.

2) Evaluating the discovered navigation patterns

Each query to a Web usage miner returns a set of navigation patterns. Then, the analyst faces the nontrivial problem of evaluating these patterns and deriving reliable conclusions from them. Web usage analysis extracts knowledge from a Web server log. Then using visualization tools evaluated patterns are represented in required formats and reports, which generate desired result patterns.

5. Conclusion

Uses of web server logs improve the performance of web site search. Log file data can provide user based measures of web based resources. Web log analysis improves web page content and design which is not an easy task. Web usage mining techniques apply on web log data. By this paper you can know what is in web logs and you can apply only those fields in your web mining techniques. You can also improve your mining results..Using Web server logs we can understand customer behaviour on a web site. The information available about log file analysis is often incomplete or subject to multiple interpretations. Web usage analysis extracts knowledge from a Web server log. Analysis of Web server logs is one of the important challenges to provide Web intelligent service. The extended work is to combine the concept of learning the user's area of interest.

6. References

[1] Mrudang D. Pandya, Prof. Kiran R Amin "Survey on web log data in teams of Web Usage Mining", *International Journal of Engineering Research and Applications (IJERA)*, Vol. 2, Issue 3, May-Jun 2012, pp.1939-1943.

[2] Thanakorn Pamutha, Siriporn Chimphlee, "Data Preprocessing on Web Server Log Files for Mining

Users Access Patterns", *International Journal of Research and Reviews in Wireless Communications (IJRRWC),* Vol. 2, No. 2, June 2012

[3] Jin Zhou, Chen Ding, and Dimitrios Androutsos, "Improving Web Site Search Using Web Server Logs", In Proceedings of the *16th IBM Center for Advanced Studies' Annual International Conference on Computer Science and Software Engineering (CASCON),* 2006.

[4] Miha Grcar, "User profiling: Web usage mining" In SIKDD 2004 at *Multiconference IS 2004*, 12-15 Oct 2004, Ljubljana, Slovenia

[5] Myra Spiliopoulou, "Web usage mining for web site evaluation", *COMMUNICATIONS OF THE ACM*, August 2000/Vol. 43, No. 8,2008

[6] R. Cooley, B. Mobasher, J. Srivastava. Data "Preparation for Mining World Wide Web Browsing Patterns", *Journal of Knowledge and Information Systems*. Vol. 1.No. 1. pp. 5–32. 1999

[7] L.K. Joshila Grace, V.M.a.D.N., "Analysis of Weblogs And Web User in Web User in Web Mining"ANALYSIS OF WEBLOGS AND WEB USER IN WEB MINING," International Journal of Network Security & Its Applications(IJNSA), Vol.3, No.1, 2011.

[8] Natheer, K. and Chan, C.C., "Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining." Proceedings of the *IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*. 2006