An Effective Analysis of Server Log for Website Evaluation

Saritha Vemulapalli[1] and Shashi M.[2]

[1] Department of Computer Science & Engineering,
C M R Institute of Technology, Bangalore, India
saritha_vemulapalli@yahoo.com
[2] Department of CS & SE,
Andhra University College of Engg (A), Vizag, A.P, India
smogalla2000@yahoo.com

Abstract. The Web constitutes huge, distributed and dynamically growing hyper medium, supporting access to data and services. In corporate business due to strong market competition more organizations rely on web to conduct business, website design & management becoming critical issue in web based applications. One of the vital goals of organizations is having attractive & well organized website. Website managers are responsible to take decisions about contents & hyperlink structure in order to capture the attention of visitor's. Visitor's interactions with website are stored in server logs and serves as huge electronic survey of website. In this paper server logs are analyzed using the web log analyzer program to get general statistics about hit's, visitor's, visit's, browsers, operating systems, referring sites, spider URL's, eminent & delicate pages and statistics about error pages, broken links. Obtained results can be useful to website manager to evaluate website, helps in improving the effectiveness of website.

Keywords: Data mining, Web log analysis, Web usage mining, Web usage analysis, preprocessing, Website design & management.

1 Introduction

The Web constitutes a huge, distributed and dynamically growing hyper medium, supporting access to data and services. Effective presence of website is the key to success in global market. One of the vital goals of an enterprises and organizations is having an attractive & well organized website in terms of both content and structure, in case of content based websites such as universities, e-education, e-commerce & newspapers. Usage of an automated tool becomes necessary in order to search, extract, filter, and judge the required information. As a result, in recent time web usage mining has attracted lot of attention [1]. Web based applications generate and collect large volumes of data in their day-to-day activities. Website visitor's actions can be collected in server logs in an unstructured format and later this information can be used for user behaviour analysis. Large quantities of such data are typically

S.C. Satapathy et al. (Eds.): Proc. of Int. Conf. on Front. of Intell. Comput., AISC 199, pp. 133–141.
DOI: 10.1007/978-3-642-35314-7_16

generated by e-commerce web servers. Web mining is the application of data mining which deals with the extraction of interesting knowledge from the web documents and services which are expressed in the form of textual, linkage or usage information [2]. Web mining is divided into web content mining, web structure mining and web usage mining. The process of discovering useful knowledge from the raw information (text, image, audio or video data) available in web pages is web content mining. Analyzing the link between pages of a website using web topology is web structure mining. Cooley et al. [3] introduced the term web usage mining in 1997 and is defined as process of extracting useful information from server logs (i.e. user's history) to improve web services and performance. Source data mainly consist of the (textual) logs stores click stream data, as a result of user's interactions with a website and are represented in standard formats. Obtained user access patterns will be utilized in variety of applications, for example, to keep track of previously accessed pages of a user, to identify the typical behavior of the user [4], making clusters of users with similar access patterns and by adding navigational links [5], reorganization of a website to facilitate clients access to the desired pages more easily and with the minimum delay [6]. In addition to website evaluation, common access behaviours of the users can be used to improve the actual design and for making other modifications to a website [7]. Moreover, usage patterns can be used for business intelligence in order to improve sales and advertisement.

In this paper web log analyzer program is used to analyze the server logs of www.vnrvjiet.ac.in to get general statistics about hit's, visitor's, visit's, browsers, operating systems, referrer sites, spider URL's, eminent & delicate pages and error statistics such as client & server errors, corrupted & broken links, which helps the website manager to improve the effectiveness of the website.

The paper is organized as follows. Section 2 covers overview of web usage mining process and format of web log data. Implementation issues of web log analyzer program are presented in section 3. Section 4 covers experimental results. Conclusion & future enhancements are presented in section 5.

2 Web Usage Mining Process

Web usage mining is the discovery of user access patterns from server logs. The web usage mining process consists of data collection, data preprocessing, pattern discovery & analysis and visualization [8].

Web is interconnection between web documents and these documents are delivered by hypertext transfer protocol. The data collected from server side, client side, proxy servers, topology of website, web page contents, user registration or profile information can be used for mining process. Server logs are the primary source of data for web usage mining that are collected when users access web servers, represented in standard formats (e.g. Common Log Format [9] and Extended Common Log Format [10]). The raw information contained in a web server log file doesn't represent a structured, complete, reliable & consistent data. The quality of data can be improved with preprocessing techniques, such as data cleaning, user

identification, sessionization, session reconstruction and data structurization [11]. Statistical & data mining techniques can be applied to the preprocessed web log data, in order to discover statistics & useful hidden patterns and are represented in visualization techniques such as graphs and reports.

2.1 Common Log Format (CLF)

Each entry of log file represented in the common log format has the following syntax. [Host/IP Rfcname Userid [DD/MMM/YYYY: HH:MM:SS -0000] "Method /Path HTTP/version" Code Bytes]

The "-" shown in a field indicates missing data.

- Host/IP is the IP address of the client (remote host), which made the request to the web server.
- Rfcname returns user's authentication. It operates by verifying specific TCP/IP connections and returns the user identifier of the process who owns the connection.
- Userid is the user id of the person requesting for the document.
- [DD/MMM/YYYY: HH:MM:SS -0000] is the date, time, and time zone when the server completed processing of the request.
- "Method /Path HTTP/version" is the request line from the client. Method is the request method, /path is the requested resource, and HTTP/version is the HTTP protocol.
- Code is the HTTP status code returned to the client.
- Bytes are the size of the object returned to the client, measured in bytes.

2.2 Extended Common Log Format (ECLF)

It's an extension to CLF, having some additional information about user_agent, cookie and referrer. User_agent is the visitor's browser & O.S version. Cookie is a persistent token, which defines the cookie sent to a visitor. Referrer defines the URL from where the visitor came from. Each entry of log file, represented in the ECLF has the following syntax. Fig. 1 shows log file represented in ECLF.

[s-computername s-ip s-port c-ip rfcname cs-userid date time cs-method cs-uri-stem cs-uri-query cs-version sc-status time-taken sc-bytes cs(user-agent) cs(cookie) cs(referrer)]

```
74.110.62.155 – user1 [10/Mar/2011:13:55:34 -0700] "GET /www.vnrvjiet.ac.in/home.html
HTTP/1.0" 200 2326 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+98) – –
74.110.62.155 – user1 [10/Mar/2011:13:55:36-0700] "GET /VNRInfrastructure/index3.html
HTTP/1.0" 200 2326 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+98)
ASPSESSIONIDASRDQQCA=NDBJHEFADELJLEAGJOPIEFBP
www.vnrvjiet.ac.in/home.html
```

Fig. 1. Example of Typical Extended Common Log Format Web Server Log

2.3 Http Protocol Status Codes

Http status codes returned by the server are classified into five classes [12]. i) Continue (100 series) ii) Success (200 Series) iii) Redirect (300 Series) iv) Failure (400 Series) v) Server Error (500 Series)

A status code of 100 series means that server has received the request, continuing process. A status code of 200 series means that the transaction was successful. A status code of 300 series means that the transaction was redirected. A status code of 400 series means that the transaction was failed due to error at client side. The most common failure codes are 401 (failed authentication), 403 (Forbidden request), and 404 (file not found). A status code of 500 series means that the transaction was failed due to server side error. The most common failure codes are 503 (Out of Resource).

3 Implementation of Web Log Analyzer Program

The Web Log Analyzer Program consists of components such as data collection, pre-processing, pattern discovery & analysis, and visualization is shown in Fig. 2. The implementation details of various components are explained below.

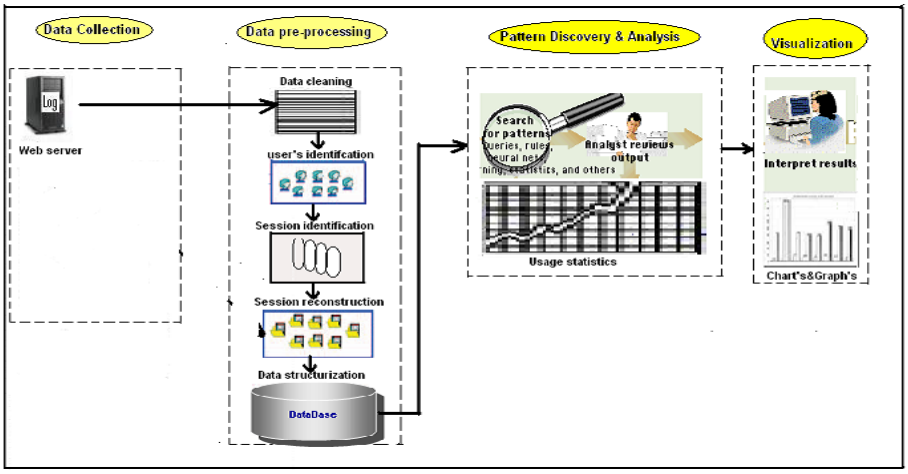

Fig. 2. Components of Web Log Analyzer Program

3.1 Data Collection

The web log analyzer program uses server logs of www.vnrvjiet.ac.in, represented in extended common log format (ECLF).

3.2 Data Preprocessing

The raw data contained in a server log file doesn't represent structured, complete, reliable & consistent information. As the web server logs aren't designed for data mining, preprocessing should be carried out in order to get reliable and accurate information. Low-quality of the data will produce low-quality mining results. The quality of the data can be improved with data preprocessing techniques, thereby helping to enhance the accuracy and efficiency of the subsequent mining process. Nearly 80% of mining efforts are required to improve the quality of data [13].

Data Cleaning. The process of removing the irrelevant entry's in pattern discovery. Irrelevant information includes the following [14]:

i) Removing all the attributes with no data at all and are not essential for the analysis.
ii) Removing the log entry's represents image, sound, video, flash animations, frames, pop-up pages, script's and style sheet files.
iii) Removing the access records generated by automatic search engine agents such as crawler, spider, robot, etc.
iv) Removing the access records requested by proxy servers.
v) Removing Log entries that have status of either "error" or "failure".

User Identification. The process of identifying the distinct user's, interacting with a website using the web browser. Users can be identified based on following factors [14]:

i) Different IP address is assumed as new user.
ii) The same IP address, but with different operating system or browser software is assumed as new user.
iii) The same IP address, operating system & browser software, but with different version is assumed as new user.

User's Session Identification. The users will visit the web site more than once. Session identification is the sequence of activities of a single user during a single visit at a defined duration [15]. Since HTTP protocol is stateless and connectionless discovering the user sessions from server log is a complex task.

Session identification can be done with the following rules [14]:

i) When there is a new user, a new session begins.
ii) When the time gap between consecutive requests made by the same user exceeds threshold $\Delta t=10$ minutes and if the referrer is "-", a new session begins.
iii) If the URL in the referrer field has never been accessed before in a current session, a new session begins.

Path Completion. Cache causes some important page requests are not recorded in server log, causing the problem of incomplete path. It is the process of reconstruction of user's navigation path, by appending missed page requests within the identified sessions.

The following rules are used for path completion [14]:

i) If the URL in the referrer field of the page request made is not equivalent to URL of last page user has requested & if the URL in the referrer field is in the user's history of the identified user's session, it is assumed that user uses "back" button. Missed page references that are inferred through this rule are added to the user's session file.

Data Structurization. The web log analyzer program translates the log file in to relational database for input to the pattern discovery & analysis phase. Different tables are designed in the relational database for each object, identified in various stages of preprocessing process.

3.3 Pattern Discovery and Analysis

The process of applying statistical and data mining techniques on the preprocessed web log data, in order to discover useful hidden patterns. This paper concentrates on statistical analysis; web log analyzer program analyzes the server logs of www.vnrvjiet.ac.in to get general statistics about hit's such as total number of hits, successful hit's, spider hits, visitor's such as number of visitor's, visitor's who visited once, repeat visitor's, average visit's per visitor, visit's such as number of session's, average visit duration, browsers, operating systems, top 10 viewed pages, least viewed pages, referrer sites, spider URL's. Error statistics such as server errors, client errors, and page not found errors. Discovered knowledge can be potentially useful in website design & management and providing support for marketing decisions.

3.4 Visualization and Result Presentation

The web log analyzer program generates different types of charts & reports, which represents usage statistics for an easier interpretation of the results.

4 Experimental Results

The web log analyzer program was developed based on IIS web server log represented in ECLF, using java programming language. The server log file of www.vnrvjiet.ac.in of 15th Nov 2010, having 10,375 records is selected for analysis. The results of preprocessing are shown in Table1. After cleaning the No. of records reduces down to 1,220 (12% of original records), 235 unique visitor's & 589 visitor's sessions are identified.

Table 1. The results of data preprocessing

Records in logfile	Records after cleaning	No of unique Visitors	Sessions
10,375	1,220	235	589

General statistics about hit's summary is shown in Table 2. Visitor's & visit's summary are shown in Table 3 & Table 4. 404 error (page not found) URL's and 500 error (internal server error) URL's are shown in Table 5 & Table 6. Eminent pages and delicate pages are shown in Table 7 & Table 8. Referring sites and spiders URL's are shown in Table 9 &Table 10. Browser & Operating system statistics are shown in Fig. 3 & Fig. 4.

Table 2. Hit's Summary

Category	hits
Total No. of Hits	1523
Successful Hits	1460
Spider Hits	293

Table 3. Visitor's Summary

Category	hits
Number of Unique Visitor's	235
Visitor's who visited once	175
Repeat Visitor's	60
Average Visit's per Visitor	2

Table 4. Visit's Summary

Category	hits
No. of Visit's	589
Avg.visit duration	1.6 min

Table 5. 404 Error Statistics

URL
/Annexure%20III.pdf
/adroit/home.html
/alumini_generalinformation.asp
/convergence2k8/imageprocessing.html
/vglug/alternative.html

Table 6. 500 Error Statistics

URL
/btech_mechnicalinfrastructure.asp

Table 7. Eminent Pages

URL
/Index.asp
/contact.asp
/btech_cse.asp
/btech_ece.asp
/placements_selected.asp
/place2009-2010.asp

Table 8. Delicate Pages

URL
/ADROIT/Adroit_mirror/ Index.html
/ADROIT/events.html
/ADROIT/pptresults.html
/ADROIT/shortcuts.html
/Careercounsellingandguidance/ dsc01743.html
/Convergence2k10pics/dsc_ 6968.html

Table 9. Referrer URL'S

REFERRER URL'S
http://adroit2k9.blogspot.com
http://khup.com
http://search.yahoo.com
http://www.facebook.com
http://www.google.co.in
http://www.google.com
http://www.vignanajyothi.com
http://www.way2college.com

Table 10. Spider URL'S

SPIDER URL'S
http://help.soso.com/webspider. htm l
http://search.msn.com/msnbot. html
http://www.bing.com/bingbot. html
http://www.exabot.com/go/robot. html
http://www.google.com/bot.html

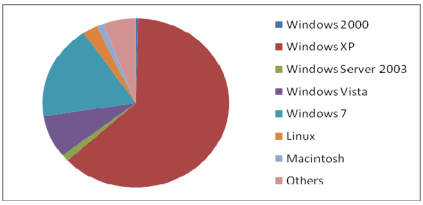

Fig. 3. Browser Statistics

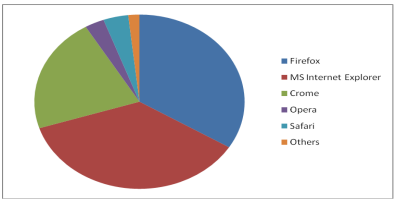

Fig. 4. Operating System Statistics

5 Conclusion and Future Enhancements

The attractiveness of a website in terms of both content & structure is critical for web based applications. Server logs of www.vnrvjiet.ac.in are analyzed using the web log analyzer program to get general statistics about hit's, visitor's, visit's, browsers, O.S, referring sites, spider URL's, eminent & delicate pages and corrupted & broken links. The obtained results can be used by website manager to increase the effectiveness of the website. This can be enhanced in future in order to find association among pages & to relate pages that are most often occur together in a single session by applying association rule generation & clustering algorithms. Such rules can also be helpful to web site managers to restructure the website.

References

1. Cooley, R., Mobasher, B., Srivastava, J.: Web mining: information and pattern discovery on the World Wide Web. In: International Conference on Tools with Artificial Intelligence, pp. 558–567. IEEE, Newport Beach (1997)
2. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Journal of Knowledge & Information System, 1–27 (1999)
3. Cooley, R., Mobasher, B., Srivastava, J.: Grouping Web page references into transactions for mining World Wide Web browsing patterns. In: Knowledge and Data Engineering Workshop, pp. 2–9. IEEE, Newport Beach (1997)
4. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N.: Web usage mining: Discovery and applications of usage patterns from Web data. SIGKDD Explorations 1, 12–23 (2000)
5. Masseglia, F., Poncelet, P., Teisseire, M.: Using data mining techniques on Web access logs to dynamically improve Hypertext structure. ACM SigWeb Letters 8(3), 13–19 (1999)
6. Pirolli, P., Pitkow, J., Rao, R.: Silk from a sow's ear: Extracting usable structure from the web. In: Human Factors in Computing Systems: Common Ground, CHI 1996, Vancouver, Canada, New York (1996)
7. Bosnjak, S., Maric, M., Bosnjak, Z.: The Role of Web Usage Mining in Web Applications. Evaluation Management Information Systems 5(1), 031–036 (2010)
8. Pabarskaite, Z., Raudys, A.: A process of knowledge discovery from web log data: Systematization and critical review. Journal of Intelligent Informatin Systems 28(1), 79–104 (2007)

9. Configuration file of W3C httpd (1995), http://www.w3.org/Daemon/User/Config/
10. W3C Extended Log File Format (1996), http://www.w3.org/TR/WD-logfile.html
11. Cooley, R., Mobasher, B., Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. J. Knowledge and Information Systems 1(1), 5–32 (1999)
12. Hypertext Transfer Protocol Overview (1995), http://www.w3.org/Protocol/rfc2616/rfc216sec1.html
13. Frieder, O., Grossman, D.A.: Information Retrieval: Algorithms and Heuristics, 2nd edn. The Information Retrieval Series (2004)
14. Vemulapalli, S., Shashi, M.: Design and Implementation of an Effective Web Server Log Preprocessing System. In: Satapathy, S.C., Avadhani, P.S., Abraham, A. (eds.) InConINDIA 2012. AISC, vol. 132, pp. 897–905. Springer, Heidelberg (2012)
15. Spiliopoulou, M.: Managing Interesting Rules in Sequence Mining. In: Żytkow, J.M., Rauch, J. (eds.) PKDD 1999. LNCS (LNAI), vol. 1704, pp. 554–560. Springer, Heidelberg (1999)