Identifying System Errors through Web Server Log Files in Web Log Mining

[1]Arjun Ram Meghwal, [2]Dr. Arvind K Sharma
[1]Ph.D CSE Scholar, Career Point University, Kota, Rajasthan, India
[2]Research Supervisor, Career Point University, Kota, Rajasthan, India

Abstract

Web log mining is an application of data mining techniques to discover usage patterns from web data, in order to better serve the needs of web based applications. The user access log files present significant information about a web server. Websites have been playing vital role in providing an information and knowledge to the end users. Web usage patterns are an important aspect to discover hidden and meaningful information. It will be a big challenge in web log mining when the volume of traffic is large and the volume of web data is still in the growing phase. To face the challenges, an approach for web log mining is shown. In this paper we have been identifying system errors through web server log files for web log mining with the help of web log analyzer.

Keywords

Web Log Mining, Web Server Log Files, WebLog Expert Lite

I. Introduction

Website is an essential tool for the web users to obtain vital information such as education, entertainment, health, e-commerce etc. Today the Internet is most emerging technology in the world. The terms Internet and World Wide Web are generally used in everyday in life speech without much distinction. The World Wide Web is also known as 'Information Superhighway'. It is a system of interlinked hypertext documents accessed via Internet. However, the Internet and the World Wide Web are not one and the same. The Internet is a global system of interconnection of computer networks. While World Wide Web is one of the services which run on the Internet [1]. It is a collection of text documents and other resources, linked by hyperlinks and URLs, usually accessed by Web browsers. In short, World Wide Web (is known as Web) considered as an application 'running' over the Internet [2]. It is a large and dynamic domain of knowledge and discovery. It has become the most popular services among other services that the Internet provides. The numbers of users as well as the number of website have been increasing dramatically in the recent years. A huge amount of data is constantly being accessed and shared among several types of users, both humans and intelligent machines.

Paper is organized into different sections: Section-II describes literature survey. Section-III explains Web log mining and its essential phases along with taxonomy. Proposed methodology is mentioned in Section-IV. Section-V provides experimental evaluation. Conclusion is shown in section-VI while references are mentioned in the last section.

II. Literature Survey

In this section we discuss related works in the web mining domain, now days, web log mining is one of the emerging areas where data analysis is most important to track user behaviour in order to better serve users.

In one of the work a novel approach was introduced for classifying user navigation patterns and predicting user's future request [6].

In another work a methodology was proposed and web log data was used to improve marketing activities [7].

Valter Cumbi et al. have done a case study of e-government portal initiative in Mozambique for visitor analysis[8].

A work is done on mining interesting knowledge from web logs which presented in [9].

Ramya et al. have proposed a methodology for discovering patterns in usage mining to improve the quality of data by reducing the quantity of data[10].

Maheswara Rao et al. have introduced a research frame work capable of preprocessing web log data completely and efficiently. This framework helps to mine usage behavior of the users [11].

One of work specifies a recommender system that was able to online personalization for user patterns [12].

III. Web Log Mining

Web Log Mining is a part of Web Mining (also called Web Usage Mining) which, in turn, is a part of data mining. As data mining is used to extract meaningful and valuable information from large volume of data, the web usage mining has been used to mine the usage characteristics of the Website users. Web mining refers to overall process of discovering potentially useful and previously unknown information from the web documents and services [3]. This extracted information may be used in a variety of ways such as improvement of the Web application, identifying the visitor's behaviour, checking of fraudulent elements etc. Web access patterns mined from Web log data have been interesting and useful knowledge in practice. Examples of applications of such knowledge include improving design of the websites, analyzing system performance to understand user's reaction and motivation, build adaptive websites [4]. The ultimate aim of web log mining is to discover and retrieve useful and interesting patterns from huge dataset.

A. Phases of Web Log Mining

Web log mining contains three phases such as Preprocessing of web data, Pattern discovery, and Pattern analysis [5]. Preprocessing is a primary task in web log mining process. The main phases of web log mining are shown in fig. 1.

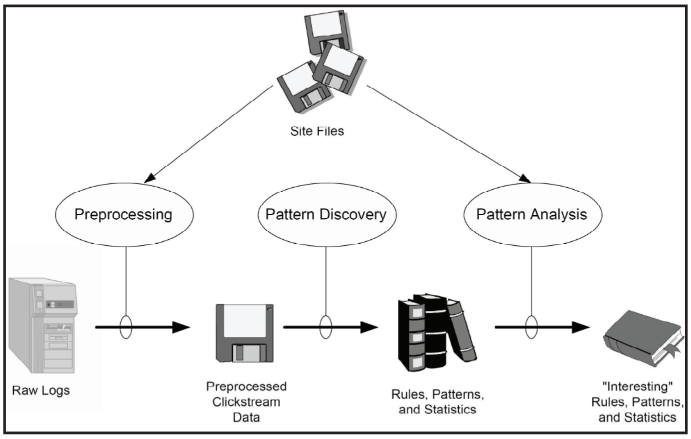

Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Fig. 1: Phases of Web Log Mining

1. Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

2. Pattern Discovery

Web log mining is used to uncover patterns in web server log files but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data.

3. Pattern Analysis

This is the final step in the web log mining process. After the preprocessing and pattern discovery, obtained usage patterns should analyzed to filter uninteresting information and mine the useful information.

B. Nuts and Bolts of Web Logs

The quality of the web patterns discovered in Web usage mining process highly depends on the quality of the data used in the mining processes. Web log files record 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 is known as log file. Log files can be analyzed over a time period. The time period can be specified on hourly, daily, weekly and monthly basis.

C. Taxonomy of Web Logs

Web server logs are plain text (ASCII files) and are independent from the web server. There are some distinctions between server software, but traditionally there are four types of web server logs. The taxonomy of web server logs is shown in fig. 2.

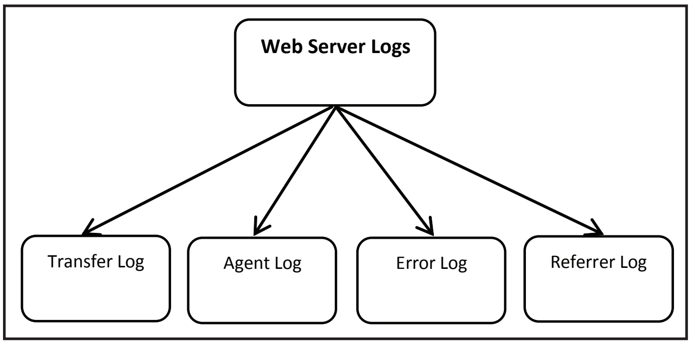


Fig. 2: Taxonomy of Web Server Logs

The first two types of Web logs such as Transfer Log and Agent Log 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 [15].

D. Location of Web Log Files

A web log is a file to which the web server writes information each time a user requests a website from that particular web server. If user visits many times on the website then it creates entry many times on the web server. A log file is located into three different places [16] such as:

1. Web Servers

The web logs which usually supply the most complete and accurate usage data.

2. Web Proxy Servers

A proxy server takes the HTTP requests from users and passes them to a web server then returns to users the results passed to them by the web server.

3. Client Browsers

Participants remotely test a web site by downloading special software that records Web usage or by modifying the source code of an existing browser. HTTP cookies could also be used for this purpose.

E. Contents of Web Log Files

Web log files reside on the server. If user visits many times on the website then it creates entry many times on the server. A web log file contains following key fields:

1. Visiting Path– Paths which follow by the user to visit on the Website.
2. User Name–Identify the user through IP addresses which provide by ISP. It is temporary address.
3. Success Rate– It is user activity which is done on the Website that is number of downloads and number of copies.
4. Path Traversed– The path identifies who visits on the website through user.
5. Last visited Page– It stores the last record that is visited by the user.
6. URL of the Web Page Accessed– It may be HTML page and CGI program. This is accessed through the web user.
7. Request Method (GET or POST): This is a method which is performing on the website like GET and POST.

The above mentioned are the key fields present in the log files. The log file details are used in case of web log mining process.

F. Format of Web Log Files

Raw log files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, text etc.) from the web site information on the request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text. The web logs do not use graphics, such as graphs and charts. Here is a sample of log entry in Apache combined format:

```
213.135.131.79 - [15/Oct/2012:19:21:49-0400] "GET /
faculty.htm HTTP/1.1" 200 9955 "http://www.davkota.
org/download.htm" "Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; Q312461)"
```

G. Sample of Web Log Files

In this section, we are going to present sample of web log file. A user Id is the unique name to use to identify. User Id is displayed when the user would like to make any transactions on the website or any other means. The sample of web log file is shown in Table 1.

Table 1: Sample of Web Log File

Host	User Id	URL
117.197.6.155	1	/images/pic010.jpg
131.253.41.47	2	images/chemlab_d.jpg
95.108.158.238	3	/images/pic8.jpg
117.201.98.145	4	/images/Result_Scan.jpg

However, other users could not see the real name and other personal information. Each row of web log file represents the URLs that user visits. Attributes of the web log file includes Visit Time, Host, URL, and other miscellaneous information about users' actions. Visited URLs of web log file are only records of users' web watching behaviours. In order to get user's interest categories, we should know the categories of web pages that the user visits [17].

IV. Proposed Methodology

The analysis of web log files allows identifying useful patterns of the browsing behavior of users, which exploit in the process of navigational behavior. Web log files capture web-browsing behaviour of the users from a website. Academic institutions are good examples which develop website. One such institution of the education sector is considered in our work.

In this section, we propose a methodology to identify and analyze of system errors by using web server logs. Following steps are included in the methodology:

A. Data Collection

In this step, web access logs have been collected from the web server of an educational institution that normally stores secondary data source in view of the fact. The web log keeps every activity of the user regarding to visit of the Website. The web log data contains the information of one month period for June 2015 during this period, 9.27 GB data is transferred.

B. Data Selection

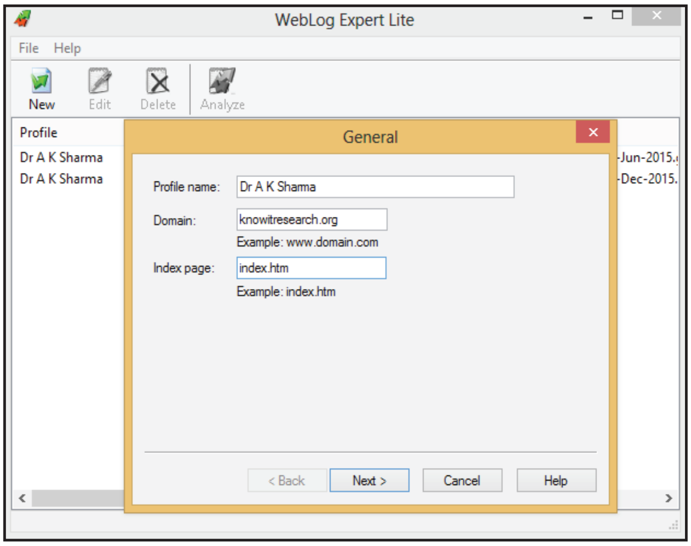
At present, towards web log mining process, the main access data origin has three kinds: Server-side data, Client-side data, and Proxy-side data (middle data). In our work, we use the case of the web server.

C. Web Logs

A web log is a listing of page reference data sometimes it is referred to as click stream data [13]. Web plays an important role for extracting useful information. There is a need for data log to track any transaction of the communications. This data can offer valuable information insight into website usage. It characterizes the activity of many users over a potentially long period of time.

D. Tool Selection-WebLog Expert Lite

There are several commercial and freely available tools exist for web mining purposes. WebLog Expert Lite 7.8 is one of the fast and powerful Web log analyzer tool [14]. It helps to reveal important statistics regarding a web site's usage such as activity of visitors, access statistics, paths through the website, visitors' browsers, etc. It supports W3C extended log format that is the default log format of Microsoft IIS 4.0/.05/6.0/7.0 and also the combined and common log formats of Apache web server. It reads compressed log files (.gz, .bz2 and .zip) and automatically detects the log file format. If necessary, log files can also be downloaded via FTP or HTTP. It is one such web tool used to produce highly detailed, easily configurable usage reports in Hypertext Markup Language (HTML) format, for viewing with a standard web browser [14]. The GUI Interface of WebLog Expert Lite contains menu, toolbars and the list of profiles. The GUI Interface of WebLog Expert is shown in fig. 3.


WebLog Expert Lite

General

Profile name: Dr A K Sharma

Domain: knowitresearch.org

Example: www.domain.com

Index page: index.htm

Example: index.htm

Profile

Dr A K Sharma -Jun-2015.

Dr A K Sharma -Dec-2015.

Fig. 3: GUI Interface of WebLog Expert

In this section, we would be able to identify the Hits statistics like Total Hits, Visitors Hits, Average Hits per Day, Average Hits per Visitor, etc., Page View Analysis like Total Page views, Average Page Views per Day, Average Page Views per Visitor, total Visitors, Total Visitors, Average Visitors per Day, Total Unique IPs, Bandwidth, Total Bandwidth, Visitor Bandwidth, Average Bandwidth per Day, Average Bandwidth per Hit, and Average Bandwidth per Visitor of the Website on monthly.

V. Experimental Evaluation

In this section, we are going to identify and analyze web server log files collected from the server of a website of an Educational and Research Institution by using WebLog Expert Lite web log mining tool. Statistical or text log files have been used in this experimental work. Many analyses have been carried out to identify the system errors and behaviour of the web users. Web access log contains 25 MB data of one month for June 2015 and we have got 3.5 MB data after applying the preprocessing mechanisms. Web log data received after preprocessing is implemented through WebLog Expert Lite and complete evaluation is presented. The daily numbers of visitors who have accessed the website are identified and shown in fig. 4.

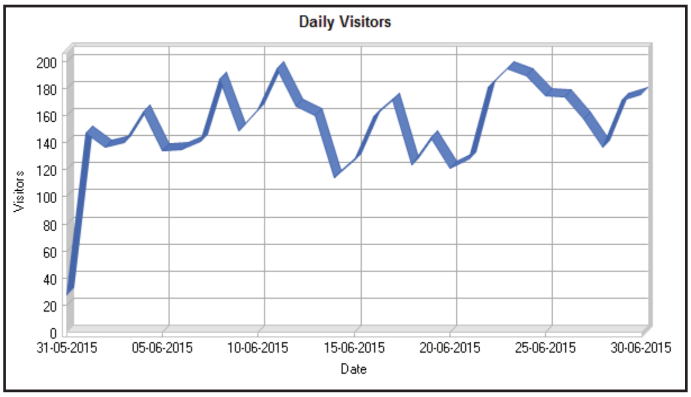


Fig. 4: Graphical Presentation of Visitors

The daily web usage data shows the number of hits occurs, which Files, Pages, Visits, and Kbytes have been visited. In this month, maximum average Hits per Day are 2543, maximum Pages per Day are 194, maximum Visitors per Day are 151 and Average Bandwidth per Day is 306.07 MB which are summarized in Table 2.

Table 2: General Statistics of the Visitors

Hits	
Total Hits	78,861
Visitor Hits	74,681
Average Hits per Day	2,543
Average Hits per Visitor	15.92
Cached Requests	5,140
Failed Requests	3,889
Page Views	
Total Page Views	6,038
Average Page Views per Day	194
Average Page Views per Visitor	1.29
Visitors	
Total Visitors	4,690
Average Visitors per Day	151
Total Unique IPs	4,060
Bandwidth	
Total Bandwidth	9.27 GB
Visitor Bandwidth	8.77 GB
Average Bandwidth per Day	306.07 MB
Average Bandwidth per Hit	123.20 KB
Average Bandwidth per Visitor	1.91 MB

As a result, we found different types of errors occurred during the web surfing by the web users. These errors are shown in Table 3.

Table 3: Error Types (Page Not Found)

S. No	**Error(s)**	**Hits**
1	404 Page Not Found	3,888
2	417 Expectation Failed	1
	Total	**3,889**

It is cleared from the table 3 that 404 is most frequently occurred error. Some other types of errors are also identified. The daily error types are shown in fig. 5.

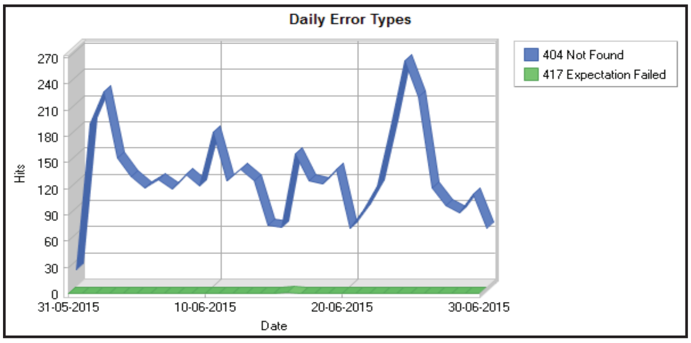


Fig. 5: Identifying System Errors

The graphical presentation of daily system errors is also shown in fig. 6.

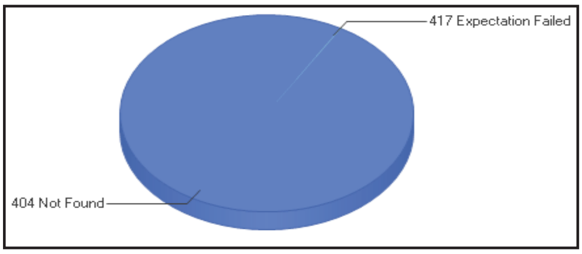


Fig. 6: Graphical Presentation of Errors

VI. Conclusion

Web is one of the most used interfaces to access remote data, commercial and non-commercial services. In this paper, a methodology to identify the system errors by using web server log files has been investigated. WebLog Expert tool is used in the complete log mining process. The findings of this work would be helpful and useful for the System Administrators, Web Masters, Web Analysts, Website Maintainers, Website Designers and Web Developers to manage their systems by identifying occurred errors, corrupted and broken links. This work will also improve the loyalty and reliability of the web sites.

References

[1] Piatetsky Shapiro G. et al.,"Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press, 1996.

[2] The W3C Technology Stack; "World Wide Web Consortium", Retrieved April 21, 2012.

[3] Arvind K Sharma, P. C. Gupta,"Enhancing the Performance of the Website through Web Log Analysis and Improvement", International Journal of Computer Science and Technology (IJCST) Vol. 3, Issue 4, Oct-Dec 2012.

[4] Huiping Peng, "Discovery of Interesting Association Rules Based on Web Usage Mining", International Conference 2010.

[5] Cooley, R.,"Web Usage Mining: Discovery and Application of Interesting Patterns from Web data", 2000.

[6] Liu, H., et al., "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting user's future requests", Data and Knowledge Engineering, 2007, Vol. 61, Issue 2, pp. 304-330.

[7] Arya, S., et al.,"A methodology for web usage mining and its applications to target group identification", Fuzzy sets and systems, 2004, pp. 139-152.

[8] Valter Cumbi et al.,"Mozambican Government Portal Case Study: Visitor Analysis", IST-Africa 2007 Conference Proceedings Paul Cunningham and Miriam, International Information Management Corporation (IIMC), 2007.

[9] F.M. Facca, P.L. Lanzi,"Mining interesting Knowledge from Web Logs: a survey", Elsevier Science, Data and Knowledge Engineering, 2005, 53, pp. 225-241.

[10] G.R.C. et al.,"An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network," Fifth International Conference on Information Processing, 2011; Springer-Verlag.

[11] Maheswara Rao et. al,"An Enhanced Pre-Processing Research Framework for web Log Data Using a Learning Algorithm," Computer Science and Information Technology, DOI, pp. 1-15, 2011.

[12] Mehrdad Jalali et al., "A Recommender System for Online Personalization in the WUM Applications", Proceedings of the World Congress on Engineering and Computer Science (WCECS-2009) Vol. 2, October 20-22, 2009, San Francisco, USA.

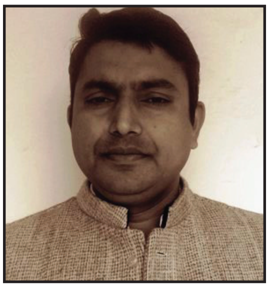
[13] Arvind K Sharma and P.C. Gupta, "Predicting the Behaviour and Interest of the Website Users through Web Log Analysis", International Journal of Computer Applications, Vol. 64, No. 7, February 2013.

[14] [Online] Available: http://www.weblogexpert.com

[15] L.K. Joshila Grace, and et al.; 'Web Log Data Analysis and Mining' in Proc. CCSIT-2011, Springer CCIS, Vol 133, Jan 2011, pp. 459-469

[16] K. R. Suneetha et. al,"Identifying User Behavior by Analyzing Web Server Access Log File", International Journal of Computer Science and Network Security(IJCSNS), Vol. 9, pp. 327-332, 2009

[17] T. Revathi, et. al, "An Enhanced Pre-Processing Research Framework for Web Log Data", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue-3, March 2012.



Arjun Ram Meghwal obtained his B.E. degree in Computer Science and Engineering from M.B.M. Engineering College, Jodhpur in 2005 and M.Tech in Computer Science and Engineering from Jagannath University, Jaipur in 2014. He has 9 years of teaching experience at diploma and degree level in technical education. He is currently pursuing PhD Computer Science and Engineering from Department of Engineering and Technology in Career Point University, Kota. He is also a Lecturer in Computer Science and Engineering at Government Polytechnic College, Tonk, Rajasthan. His research interests include Web Application Security, Web Data Mining.

Dr. Arvind K Sharma holds PhD degree in Computer Science. He has more than 13 years of work experience in academic field. He has published more than 27 Papers in various National, International Journals and Conferences. He has authored and co-authored almost 5 books. He has visited Thailand and Dubai for attending International Conferences. He has participated as Speaker and Keynote Speaker in many National and International Conferences. He is a Member of numerous academic and professional bodies i.e. IEEE, WASET, IEDRC, IAENG Hong Kong, IACSIT Singapore, UACEE UK, ACM, New York. He is a Member of Technical Advisory Committee of many International Conferences in India and abroad. He is also Editorial Board Member and Reviewer of several National and International Journals. His area of interest includes Web Usage Mining, Web Intelligence Applications, Web Data Mining, Big Data Analytics and Machine Learning Tools.