Comprehensive Analysis of Web Log Files for Mining

Vikas Verma [1], A. K. Verma [2], S. S. Bhatia [3]

[1] M. M. Institute of Computer Tech. & Business Management , M. M. University, Mullana, Ambala, Haryana-133203, India.

[2] Dept. of Computer Sc. and Engg., TIET, Thapar University, Patiala, Punjab-147004, India.

[3] School of Mathematics and Computer Applications, Thapar University, Patiala, Punjab-147004, India.

Abstract

World Wide Web is a global village and rich source of information. Day by day number of web sites and its users are increasing rapidly. Information extracted from WWW may sometimes do not turn up to desired expectations of the user. A refined approach, referred as Web Mining, which is an area of Data Mining dealing with the extraction of interesting knowledge from the World Wide Web, can provide better result. While surfing the web sites, users' interactions with web sites are recorded in web log file. These Web Logs are abundant source of information. Such logs when mined properly can provide useful information for decision making. Mining of these Web Logs is referred to as Web Log Mining. This paper analyses web log data of NASA of the month of August 1995 of 15.8MB and depicts certain behavioral aspects of users using web log mining.

Keywords: *Web Mining, Data Mining, Web Log Mining.*

1. Introduction

The expansion of the World Wide Web has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized such that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. In accordance with Kosala, Blockeel and Neven [1], the term 'Web Mining' is defined as the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services. Web mining research, is an integrate research from several research communities such as: Database (DB), Information retrieval (IR), The sub-areas of machine learning (ML) and Natural language processing (NLP).

An important constituent category of Web Mining is Web Log mining also known as Web Usage mining, is the process of extracting interesting patterns from web access logs [6]. The different techniques are represented through Figure 1. However, not much concentration is done on techniques, since the focus of this paper is exclusively on web logs. Web usage data can include a variety of data from different sources. These sources can include web server access logs, proxy server logs, browser logs or any other data that is generated by users interacting with a website. The issues are outlined by Linoff and Berry [3].

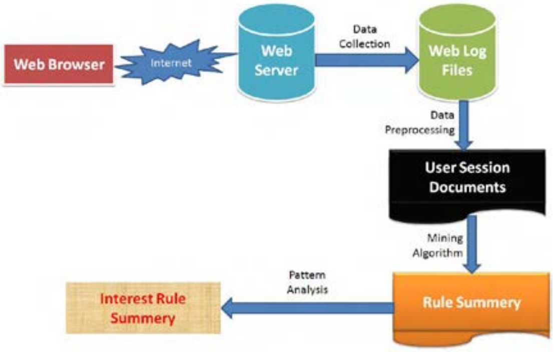

Fig. 1: Web Mining Techniques

There are several general challenges associated with obtaining due results from the data. Firstly, extraneous information is mixed with useful one. Secondly, multiple server requests may be generated by a single user action. Thirdly, multiple user actions may generate the same server request. Fourthly, local activities (for example browser navigation using 'back', and 'forward' buttons) are not recorded.

The paper is organized as follows: In section 2, we emphasize on web logs; in section 3, latest developments in the field web usage mining are presented; in section 4, visualization is depicted through a tool; section 5, focus on the future prospects and conclusion.

II. Motivation

Web Log mining is the process of identifying browsing patterns by analyzing the user's navigational behavior. A Web log file [4] 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. Web server logs are plain text (ASCII) files, independent of server platform. There are some differences between server software, but traditionally there are four types of server logs [5, 14]:

1. Transfer (access) log
2. Error log
3. Referrer log
4. Agent log

The first two types of log files are “standard / common”. 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 transaction, whether completed or not, is recorded in the logs, and some transactions are recorded in more that one log. For example, most (but not all) HTTP errors are recorded in the transfer log and the error log.

A transfer access log typically is a long line of ASCII text, separated by tabs and spaces. A sample log is considered below.

1Cust216.tnt1.santa-monica.ca.da.uu.net - - [17/Sept/2011:12:13:03 -0700] GET /gen/meeting/ssi/next/HTTP/1.0 200 9887 http://www.slac.stanford.edu/ Mozilla/3.01-C-MACOS8 (Macintosh; I; PPC) GET /gen/meeting/ssi/next/ - HTTP/1.0

An analysis of each section is done as below

- *1Cust216.tnt1.santamonica.ca.da.uu.net*

This is the address of the computer making the HTTP request. The server records the IP and then, if configured, will lookup the Domain Name Server (DNS).

- *RFC931 (or identification)*

Rarely used, the field was designed to identify the requestor. If this information is not recorded, a hyphen (-) holds the column in the log.

- *Authuser* -

List the authenticated user, if required for access. This authentication is sent via clear text, so it is not really intended for security. This field is usually filled by a hyphen (-).

- *Time Stamp*

[17/Sept/2011:12:13:03 -0700]

The date, time, and offset from Greenwich Mean Time are recorded for each hit. The date and time format is: DD/Mon/YYYY HH:MM:SS. The example shows that the transaction was recorded at 12:13 pm on Sept 17, 2011 at a location 7 hours behind GMT.

- *Request*

GET /gen/meeting/ssi/next/ HTTP/1.0

One of three types of HTTP requests is recorded in the log. GET is the standard request for a document or program. POST tells the server that data is following. HEAD is used by link checking programs, not browsers, and downloads just the information in the HEAD tag information. The specific level of HTTP protocol is also recorded.

- *Status Code 200*

There are four classes of codes

1. Success (200 series)
2. Redirect (300 series)
3. Failure (400 series)
4. Server Error (500 series)

A status code of 200 means the transaction was successful. Common 300-series codes occurs when the server checks if the version of the file or graphic already in cache is still the current version and directs the browser to use the cached version. The most common failure codes are 401 (failed authentication), 403 (forbidden request to a restricted subdirectory), and the dreaded 404 (file not found) messages.

- *Transfer Volume 9887*

For GET HTTP transactions, the last field is the number of bytes transferred. For other commands this field will be a hyphen (-) or a zero (0). The transfer volume statistic marks the end of the common log file. The remaining fields make up the referrer and agent logs, added to the common log format to create the “extended” log file format. An analysis of each section is done as below

- *Referrer URL*

http://www.slac.stanford.edu/

The referrer URL indicates the page where the visitor was located when making the next request. The actual request is shown in the last field of the entry GET /gen/meeting/ssi/next/ - HTTP/1.0 and is duplicated from the HTTP Request.

- *User Agent*

Mozilla/3.01-C-MACOS8 (Macintosh; I; PPC)

The user agent is information about the browser, version, and operating system of the reader.

The description can be generalized through the following table

Table 1: Web log file attributes and their description

Attributes	**Description**
Client IP	Client Machine IP Address
Client Name	Client Name if required by server, otherwise, hyphen
Date	Date when user made access
Time	Time of transaction
Server Site Name	Internet service name as appeared on client machine
Server Computer Name	Server Name
Server IP	Server IP provided by Internet Service Provider
Server Port	Server port configured for data transmission
Client Server Method	Client Method or modes of request can be GET, POST of HEAD
Client Serves URI Stem	Targeted default web page of web site
Client Server URI Query	Client query which starts after “?”
Server Client Status	Status Code returned by the server like 200, 404
Server Client win32Status	Windows status code
Server Client Bytes	Number of bytes sent by server to client
Client Server bytes	Number of bytes received by Client
Time Taken	How much spend by client to perform any action
Client Server Version	Protocol version like HTTP
Client Server Host	Host header name
User Agent	Browser type that client used
Cookies	Contents of cookies
Referrer	Link from where client jump to this site

III. Work Done

A lot of research projects deal with the Web Log mining. Data Preprocessing, Pattern discovery, and pattern analysis are considered as important phases of Web Log mining process. Most of the efforts focus on extracting useful patterns and rules using data mining techniques in order to understand the users' navigational behavior. Much of the work in this field focuses on user identification, session identification etc. Specifically for web log files [6, 7] has explored certain issues regarding web server log files. Since in Web Log mining several techniques can be used [13], one such technique is Association mining using Web Logs [8, 9 10]. Sequence mining, which is another technique, can be used for discover the web pages which are accessed immediately after another. It is used in [11], using a tree for storing patterns efficiently. [13] Discussed the structure of web log file in detail and performed two preprocessing techniques data cleaning and user identification. [15] Derived the user profiles from the analysis of web log file and Meta data of page contents. [16] Identified that web usage profiles play an important role in web personalization. Profiles were extracted from clusters and clusters were extracted from web usage data after preprocessing the web log file. The navigation pattern can be examined with the data of the server log file by the web analyzer [17].

IV. Visualization

Using weblogexpert [12], software for Web Log mining, NASA server log file is analyzed. Typical behavior based on statistical analysis of the log file is observed and thereafter result is visualized as shown in Figure 2, 3 and 4 depicting Daily Visitors, Daily Error types and Activity-day wise for the month of August 1995. It should be observed that these results can be utilized further for certain web specific applications also.

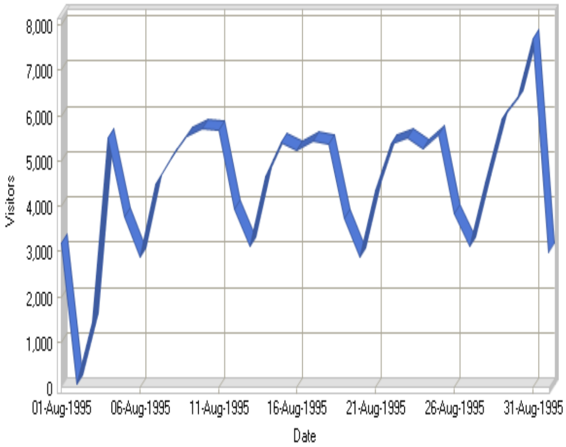

Fig. 2: Daily Visitors schedule

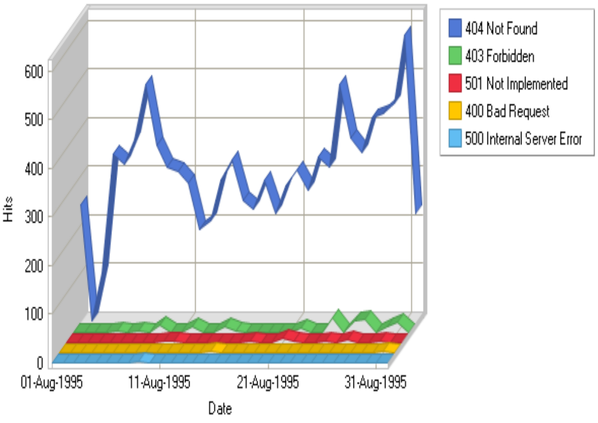

Fig. 3: Daily Error Types

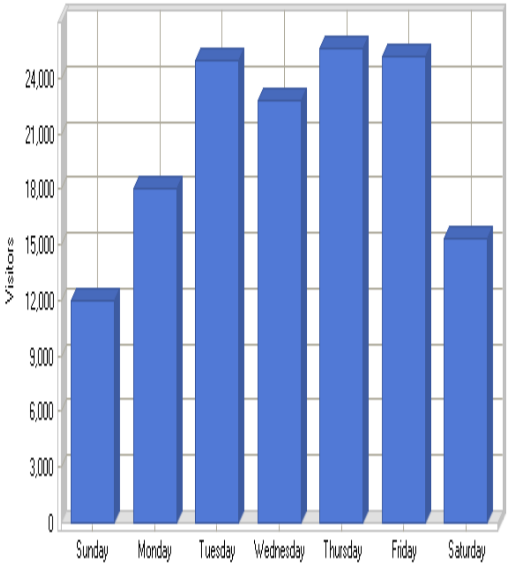

Fig. 4: Activity by Day of Week

V. Conclusions and Scope

The requirement for predicting user needs in order to improve the usability and user retention of a Web site can be addressed by Processing Web Log file efficiently. Future scope of Web Log mining is in Web Personalization and to improve the overall performance of future accesses. In today's era of advancements it can also be used in e-commerce, digital libraries etc, using techniques of data mining at group level instead at individual level for high accuracy.

References

[1] R. Kosala, and H. Blockeel, "Web Mining Research: A Survey" , SIGKDD Explorations, 2(1):1-15, 2000.

[2] O.R. Zaıane, "Building Virtual Web Views", Data and Knowledge Engineering , 39:143–163, 2001.

[3] G.S. Linoff, and M.J.A. Berry, Mining the Web, John Wiley and Sons, first edition, 2001.

[4] Magdalini Eirinaki, and Michalis Vazirgiannis, "Web Mining for Web Personalization", PKDD, 2005.

[5] Internet: Web Log files overview: http://www.si.umich.edu/Classes/540/Readings/ServerLogFileAnalysis.htm

[6] Zhang Huiying, and Laing Wei, "An Intelligent Algorithm of Data Pre-processing in Web Usage Mining ", Proceedings of the 5th world Congress on Intelligent Control and Automation, June15-19, 2004 , Hangzhou, P.R.China.

[7] Doru Tanasa et.al., "Advanced data preprocessing for inter sites Web Usage mining", IEEEE computer society, 2004.

[8] M. Eirinaki, and M. Vazirgiannis, "Web mining for web personalization", ACM Trans. Inter. Tech., vol. 3, no. 1, pp. 1-27, 2003.

[9] J. Punin, M. Krishnamoorthy, and M. Zaki, "Web usage mining: Languages and algorithms", in Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2001.

[10] P. Batista, M. ario, and J. Silva, "Mining web access logs of an on-line newspaper", NetLab, Lund University Libraries, Sweden April 2002.

[11] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs", in PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, London, UK: Springer-Verlag, 2000, pp. 396-407 .

[12] Internet: Typical softwares : http://www.kdnuggets.com/software.html

[13] Suneetha, K. R. and D. R. Krishnamoorthi , "Identifying User Behavior by Analyzing Web Server Access Log File", in IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[14] M. H. A. Wahab, M. N. H. Mohd, et al. , " Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology, 2008.

[15] G. Stermsek, M. Strembeck, et al. , " A User Profile Derivation Approach based on Log-File Analysis", IKE 2007, pp. 258-264.

[16] G. Castellano, F. Mesto, et al. , " Web User Profiling Using Fuzzy Clustering", WILF 2007, pp. 94–101.

[17] J Vellingiri, and S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology, Volume 11, Issue 4, Version 1.0, March 2011.

Vikas Verma is currently working as a Lecturer in M. M. Institute of Computer Technology & Business Management, M. M. University, Mullana, Haryana. He received his MCA from Punjabi University, Patiala, Punjab, India, in 2003. He is pursuing Ph.D. from School of Mathematics and Computer Applications, Thapar University, Patiala, Punjab, India. He is in teaching since 2003. He has published 08 research papers in International/National Journals and Conferences. His research interests are Web Mining, Knowledge Discovery, Information Systems, Data base management systems.

Dr. A.K. Verma is currently working as an Associate Professor in the department of Computer Science and Engineering at Thapar Institute of Engineering & Technology (Deemed University), Patiala. He received his B.S., M.S. and Ph.D. in 1991, 2001 and 2008 respectively, majoring in Computer science and engineering. He has worked as Lecturer at M.M.M. Engg. College, Gorakhpur from 1991 to 1996. He joined Thapar Institute of Engineering & Technology in 1996 as a Systems Analyst in the Computer Centre and is presently associated with the same Institute. He has been a visiting faculty to many institutions. He has published over 35 papers in referred journals and conferences (India and Abroad). He is a MISCI(Turkey), LMCSI (Mumbai), GMAIMA (New Delhi). He is a certified software quality auditor by MoCIT, Govt. of India. His research interests include wireless networks, routing algorithms and securing ad hoc networks.

Dr. S.S. Bhatia, is currently working as Professor and Head, School of Mathematics and Computer Applications, Thapar University, Patiala, Punjab, India. He received his M.Sc, M.Phil and Ph.D in 1985, 1986 and 1994 respectively, majoring Mathematics. He is associated with Thapar University for the last 24 years. He has vast experience of teaching at UG and PG level to Science and Engineering students at Thapar University. He has published over 55 research papers in Journals, International and National Conferences in the areas of Functional Analysis, Reliability Analysis and Image processing. He has guided 2 Ph.D. and 7 M.Phil scholars. He has accomplished 2 UGC Major Research Projects. He's a Life member of Punjab Academy of Sciences, Indian Society of Industrial and Applied Mathematics and Indian Society of Technical Education.