ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

Effective Log File Mining by Pre-processing, User and Session Identification Algorithms

Abhishek Patil[1], Shweta Wani[2], Neha Patil[3], Shraddha Sonawane[4], Pooja Patil[5]

BE Student, Dept. of Computer Engineering, SSBT College of Engineering and Technology, North Maharashtra University, Jalgaon, Maharashtra, India[1,2,3,4,5]

ABSTRACT: The important part of any internet application is web log files. Web log file serves the purpose of directory in numerous facet of information mining for web users. Web users are not able to locate what they actually want within reasonable amount of time. There is a good style of logs to stock information regarding the search patterns of the users. There can be ample formats of convenience of logs, every internet application will develop format of its own logs. Generally, IP, date and time of the request, result for the request (with code), dealing size, protocol, request description, browser and software package utilized by the user area unit a number of the vital attributes of each request that get into the record of the log file. The user's activity search pattern is examined by the question log files. The users browsing behaviour is analysed using various methodologies. Completely different from existing ways, gift system avoids user's feedback to the browser and doesn't collect the data which can manufacture privacy problems, e.g. users browsing history, bookmarks, and so on. As a result of the data recorded by server in access logs are utilized to judge the page interest, a referrer-based knowledge pre-processing methodology is administrated to enhance the dependability of the access knowledge and extract the required info for interest estimation.

KEYWORDS: Data Pre-processing, Session Identification, UserIdentification, Web Usage Mining.

I. INTRODUCTION

The World Wide Web is a dense network of interconnected computers which hosts tons of information. At present, Google is indexing more than 30 trillion Web pages. The rapid expansion of the Web has provided a great opportunity to study user and system behavior by exploring Web access logs. In Web Mining [4], data can be collected at the server side, client side, proxy servers, or obtained from an organization's database(which contains business data or consolidated Web data). There are many kinds of data that can be used in Web Mining depending upon structure, usage and content. According to data analysis objective, web mining can be divided into three different types, which are web usage mining, web content mining and web structure mining, which is shown in the following Figure

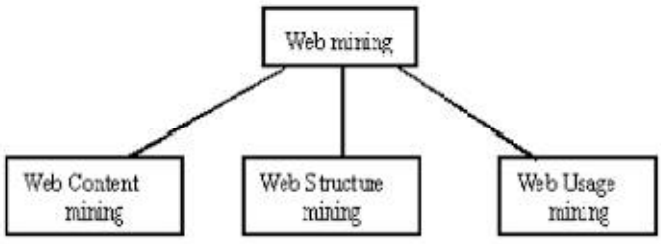

Fig.1. Web Mining Classification

Web content mining involves exploring of information resources available online, and involves mining of web data content. Web structure mining analyses the hyperlink and tree-like structure of a website. Web usage mining involves analyzing the web log files. Users show different interests when looking for internet. Some users might be looking at only documentary data, whereas some others might be engaged in multimedia data. Web usage mining (WUM) involves the automatic detection of user access patterns from one or more web servers. Organizations rely on internet

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

for their business work which often generate and collect bulk size data in their daily practices. Most of this information is generated automatically by web servers and collected in server access logs. The companies can establish better customer manager relationship by giving them exactly what they require. Companies can understand the requirements and serve them accordingly. They can also increase profitability and productivity based on the profiles generated [1]. Web mining is broadly divided into web usage mining, web content mining and web structure mining.

In the web usage mining process, the techniques of data mining are applied so as to discover the trends and the patterns in the browsing nature of the visitors of the web site. There is extraction of the navigation patterns as the browsing patterns could be traced and the structure of the web site can be designed accordingly [3]. When it is talked about the browsing nature of the user it deals with frequent access of the web site or the duration of using the web site. This information can be extracted from the log file. Only these log files record the session information about the web pages. The process of WUM [6] is divided into three phases: data preprocessing, pattern discovery, and pattern analysis.

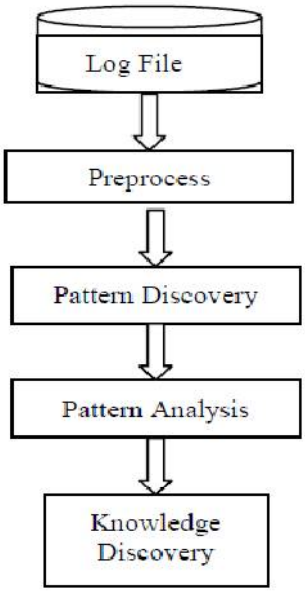

Fig.2. Web log mining structure

II. BACKGROUND

Important step in knowledge discovery in databases is to build a data set on which data mining tasks can be performed. In Web Mining, data can be accumulated at various places like, server side, client-side, proxy servers, or can be obtained from an organization's database. Each type of data collected differs not only in location from where it was collected, but also the kind of data, particular population for which the data is collected and its implementation method[2]. There are many kinds of data that can be used in Web Mining. This includes:

- Content

The real data in the Web pages, i.e. the data the Web page was designed to convey to the users. This usually consists of, but is not limited.

- Structure

Data which describes the content of organization. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. With the HTML tag at the root, this resembles a tree like structure. The primary kind of inter-page structure information is how pages are inter-linked.

- Usage

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Usage: Data that gives information about usage pattern of a web page, which is extracted by parameters such as IP addresses, date and time of access, time zone, request methods, HTTP protocol version, status code, bytes exchanged, etc.

- User Profile

Profile: Data that provides demographic information about users of the Web site. This includes registration data and customer profile information

Among the various kinds of data available, a particular data can be used to extract information. Certain techniques can be applied on given data set to get the required information. Log file mining gives information about different users by analysing log files generated for those users. Numerous approaches have been proposed to analyse web log files.

III. PROPOSED ALGORITHM

The commonly used technique to augment the users search experience is the utilization of the knowledge contained within past queries in the log files. Log files are files that list the actions that have been occurred. These log files reside in the web server. A log files typically, contains information about users, issued queries, clicked results, etc. From this information, knowledge can be extracted to improve the quality in terms of effectiveness and efficiency of their system. The technique used for analysing the search pattern of user is a part of web usage mining.

Web Log Information

A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. Most logs use the format of the common log file (CLF). Each entry in the log _le consists of a sequence of fields relating to a single HTTP transaction with the various fields separated by a space. The following is a fragment from the server logs [5] for proposed system.

64.242.88.10 - - [07/Mar/2004:17:01:53 -0800] "GET /razor.html HTTP/1.1" 200 2869

This reflects the information as follows:

- Remote IP address or domain name ("64.242.88.10"): An IP address is a 32-bit host address defined by the Internet Protocol.
- Authuser: Username ('-') and password ('-') if the server requires user authentication.
- Entering date and time ("07/Mar/2004:17:01:53") and time zone (0800).
- Modes of request: GET, POST or HEAD method of (CGI) (Common Gateway Interface).
- HTTP Protocol (HTTP/1.1)
- HTTP status code: The HTTP status code returned to the client, e.g., 200 is ok and 404 if page not found.
- Bytes sent (2869): This is the size of object returned to the client, measured in bytes.

A. DATA CLEANING

Data cleaning means eliminate the irrelevant information from the original Web log file. Usually, this process removes requests concerning non-analysed resources such as images, multimedia files, and page style files. For example, requests for graphical page content like images or videos and requests for any other file which might be included into a web page or even navigation sessions performed by robots, crawlers and web spiders. By filtering out useless data, the log file size can be reduced to use less storage space and to facilitate upcoming tasks. For example, filtering out image requests in a web log file reduces the log file size to less than 50% of its original size. Thus, data cleaning includes the elimination of irrelevant entries like:

- Requests for image files associated with requests for particular pages; an users request to view a particular page often results in several log entries because that page includes other graphics, while we are only interested in what the users explicitly request, which are usually text files.
- Entries with unsuccessful HTTP status codes; HTTP status codes are used to indicate the success or failure of a requested event, only successful entries with status codes with value 2XX are considered to be successful.
- Entries with request methods except GET and POST.

The proposed algorithm for data cleaning is given below [8].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

Data Cleaning Algorithm
Input:Web Server Log File
Output:WebLog Database
Step1: Read LogFileRecord from Web Server Log File
Step2: If((LogFileRecord.url-stem(gif,jpeg,jpg,css,js)) AND (LogFileRecord.method='GET' AND 'POST') AND (LogFileRecord.Sc-status<>(301,404,500) AND
(LogFileRecord.Useragent<>Spider,Robot))
then Insert Log Record in to Web Log Database.
End of If condition.
Step 3: Repeat the above two steps until EOF (Web Log File)
Step 4: Stop the process.

The outcome of this algorithm is the Log Database consisting relevant set of records with the entries as user IP address, date, time, and URL details etc. By filtering out the irrelevant entries the size of web log files reduces to more than 50% of its original size.

B. USER IDENTIFICATION
User identification means identifying each user accessing Web site, where goal is to mine every user's access characteristic. This paper works with the assumptions, that each user has unique IP address and each IP address represents one user. But user identification is greatly complicated due to the existence of local caches, corporate firewalls and proxy servers. Browser cache does not allow the user request being recorded into the log file, in the same way existence of proxy servers makes it difficult to record the IP of the originating device. In order to overcome these problems some rule is proposed for user identification. If there is a new IP address, then there is a new user.
This rule is used in the proposed algorithm to identify users mentioned below [8].

User Identification Algorithm
Input: Log Database
Output: Unique Users Database
Step 1: Initialize IPList=0; UsersList=0; No-of-users=0;
Step 2: Read Record from LogDatabase.
Step 3: If Record.IP address is not in IPList
then add new Record.IPaddress into IPList
addRecord.Browser in to BrowserList
addRecord.OS in to OSList
increment count of No-of-users
insert new user in to UserList.
Else
If Record.IP address is present in IPList
then
increment count of No-of-users
insert as new user in to UserList.
End of If
End of If
Step 4: Repeat the above steps 2 to 3 until
eof (Log Database)
Step 5: Stop the process.

The outcomes of this algorithm are the unique users which gives information about total number of individual users, users IP address, user agent, browser and operating system used.

C. SESSION IDENTIFICATION
After user identification, session identification is done on the basis of the pages accessed by each user. The goal of session identification is to find each users access pattern and frequently accessed path. The simplest method is using a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

session timeout, where if the time between page requests exceeds a certain threshold, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout. A session timeout suitable for a particular website can be fed into this algorithm and sessions can be identified. Following are the rules for session identification [8]:

1. If there is a new user, then there is a new session.
2. If the time between page requests exceeds a certain limit (session timeout), it is assumed that the user is starting a new session.

Session Identification Algorithm

Input: Log Database
Output: Session Database
Step1: Initialize SessionList=0,UserList=0,No-of-Sessions=0
Step 2: Read Log Record from Log Database
Step 3: If LogRecord.time-taken>session_timeout OR
LogRecord.UserID not in UserList)
then
Increment No-of-Sessions
Get URL address of corresponding Session and
Insert in to SessionList
End of If
Step 4: Repeat the above steps 2 and 3 till eof (Log Database)
Step 5: End of process.

This algorithm gives user IP address along with page accesses performed by individual users during a visit in a web site. The Session Database gives details about total number of sessions, Session key with start time and end time details.

IV. SIMULATION RESULTS

In this work a server log file of size 170 KB is analysed. Several experiments on log files have been conducted. Through these experiments it is shown that the proposed pre-processing methodology reduces the size of the initial log files is reduced by eliminating unnecessary requests and also increases quality through better structuring of the web data. Following figure shows the result of pre-processing log file.

Web Server Log File:	Access_log
Duration:	1-7 Days
Original Size:	0.174449 in MB
Reduced Size After Preprocessing:	0.143551 in MB
Percentage in Reduction:	17.711

Fig.3. Result after pre-processing log file

Fig.4 represents total number of unique users identified, and number of sessions created by the users on particular date is shown in Fig. 5.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

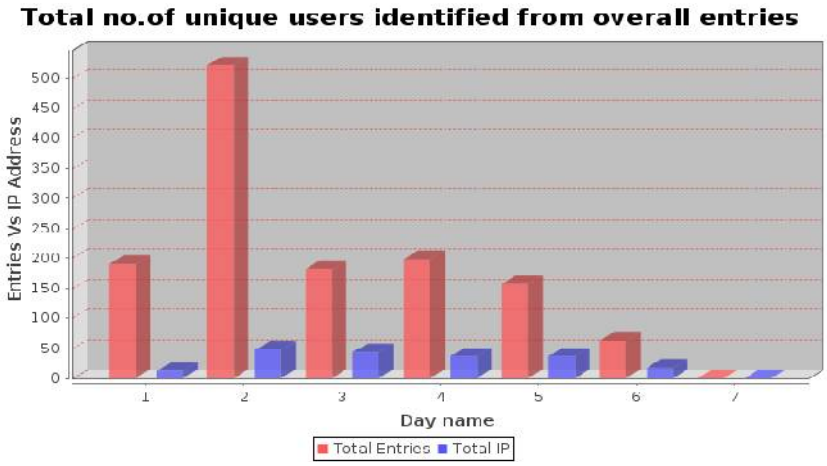

Fig.4. Unique users identified

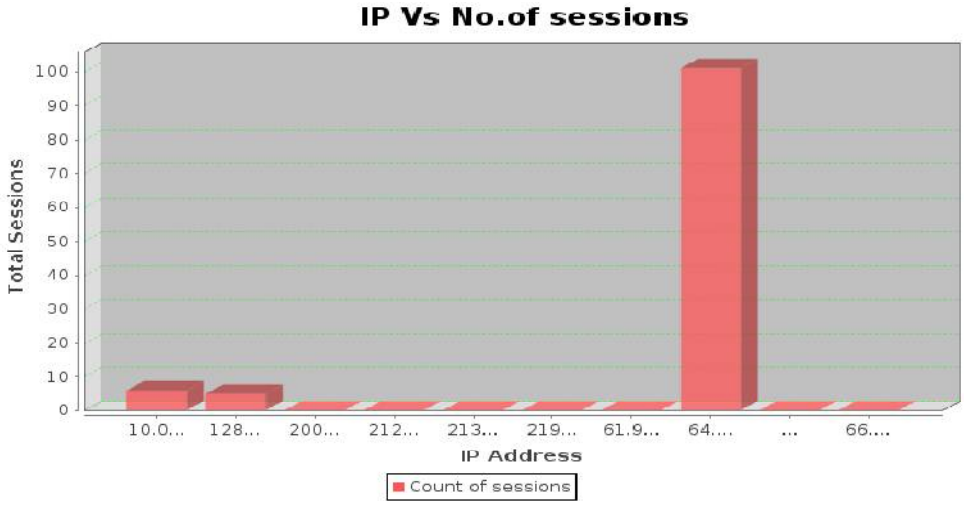

Fig.5. Sessions identified per user

V. CONCLUSION AND FUTURE WORK

As a result of lots of research and development, quality of web related services has improved significantly. But because of unprecedented growth of the Web and abundance of information, Web Users are not able to locate what they actually want within a reasonable amount time. This work has proposed one efficient tool towards an intelligent web. The proposed tool provides a good model for accessing the information related to particular log from the log files. The tool can be used in different type of mining areas of data mining applications. But due to the fast development in the technology and explosion in the number of users, Web Mining area still gives lots of research opportunities. Mining log files will provide the desired information. This information can be out to use by companies, organizations, businesses and individuals.

Pre-processing is an important task before beginning data mining. Requests which do not lead to any meaningful pattern discovery were cleaned. Only valid and relevant requests were regrouped in to user sessions and finally the results were saved in a database. Data warehousing and mining provides the OLAP tools using which multi-dimensional data can be analysed at various levels which further helps in effective data mining. Moreover, data mining functions such as classification, prediction, association, and clustering along with OLAP operations can be used for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

interactive knowledge mining at varied levels of abstraction. The pre-processed data can be stored in a snowflake schema for retrieval and analysis.

REFERENCES

[1]Chandana S. Khatavkar (2015), A Hybrid Approach For Clustering Weblog(2015). International Journal of Advanced Research In Computer Science And Software Engineering,
Volume 5,Issue3.
[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In Proc. ACM KDD, 1994.
[3] Mehak(2013), Web Usage Mining: An Analysis, Journal of Emerging Technologies in Web Intelligence, Vol. 5, No. 3
[4] R. Kosala, H. Blockeel. Web Mining Research: A Survey, In SIGKDD Explorations, ACM press, 2(1): 2000, pp.1-15.
[5] W.W.W Consortium the CommonLogFileformat http://www.w3.org/Daemon/User/Config/Logging.html#common-Log file-format, (1995)
[6] R. Srikant, R. Agrawal. Mining sequential patterns: Generalizations and performance improvements, In 5th International Conference Extending Database Technology, Avignon, France, March1996, pp. 13-17
[7]Zaïane, O.R. Xin and M. Han. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," In Proceedings of Advances in Digital Libraries Conference(1998), pp. 19-29.
[8]Suneetha K.R, Dr. R. Krishnamoorthi, Data Preprocessing and Easy Access Retrieval of Data through Data Ware House, World Congress on Engineering and Computer Science 2009 Vol I WCECS 2009, October 20-22, 2009, San Francisco, USA.