Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 9, September 2016, pg.16 – 22

An Efficient Preprocessing Method to Detect User Access Patterns from Weblogs

V.Pushpa[1], V. Vidyapriya[2]

[1]Research Scholar, Quaid-E-Millath Govt College for women (Autonomous), Anna Salai, Chennai-02
[2]Associate Professor, Quaid-E-Millath Govt College for women (Autonomous), Anna Salai, Chennai-02
[1] pushpapriyaa16@gmail.com, [2] vidyapriya2128@gmail.com

Abstract* ---The World Wide Web contains immense information, which is rapidly growing. It has the complex task to discover the exact data from web pages. Web mining technique is helpful for mining the information from the web pages, which is known as application techniques of data mining. It is used to determine the exotic patterns from the world wide web, in which there are three categories, namely web content, web structure and web usage mining. This paper concentrates on data preprocessing of web usage mining, which is performed to eliminate the inconsistence, irrelevant data and extract the interesting knowledge patterns from the weblogs, which assists to know about the user behaviour. The main objectives of this paper is to clean the server logs and acquiring the user session identification through a weblog explorer tool which are used to recognize the user activity and the pre-processed weblogs are used in the further stages of web usage mining.

Keywords--- web usage mining, data pre-processing, weblogs, data cleaning, user identification, session identification, web log explorer tool.

I. INTRODUCTION

The World Wide Web is a vast collection of web document data such as images, text document and multimedia yet it is a challenging task to retrieve the relevant data when the users are required. Every interaction of the user with the web pages is logged in a single record as a text file is known as web log file. The weblogs consist of abundant information, which include incomplete and irrelevant data too. Web mining is one of the application techniques of data mining, which is used to eliminate tedious patterns and mining interesting pattern from web pages. It includes three phases, namely web content mining, web structure mining and web usage mining. Web content mining is used for mining the text, graphs, and pictures from the millions of web pages and find out the relevant content to the search query which also known as text mining. Web structure mining helps to identify the association between web pages with the structural module of hyperlink connections, which is the structure data is discoverable by the providing of the web structure schema via database techniques for web pages [1] [2]. Web usage mining involves the analysis and automatic discovery of user access patterns from web log files, which apprehends the identity of web users based on the browser behaviour of the specific websites. This usage data is valuable to the business using online marketing and E-business.

II. WEB USAGE MINING

This paper discusses about web usage mining, which is the one of the phases of web mining. It primarily deals with the identifying of user behavioural patterns from web log files that is merely known as web log mining. Usually web usage mining has four methods such as Data collection, Data preprocessing, Pattern discovery**,** Pattern analysis.

A. Data Collection

The web usage mining uses the primary data source that holds the browser behaviour of user log data, which is gathered from different data sources like web server, client server and proxy server etc.

B. Data Preprocessing

Data preprocessing is the primary stage of web usage mining, which is used to remove the noisy, irrelevant, and inconsistent data from web logs. The data preprocessing techniques are data cleaning, user identification, session identification and path completion [1] [3].

C. Pattern Discovery

Pattern discovery is the major key component of web usage mining, which is used to detect the stimulating patterns or knowledge of the web user behaviour from the preprocessed weblogs. Pattern discovery uses different kinds of machine learning techniques such as association, sequential pattern, clustering, classification and so on. [2]

D. Pattern Analysis

Pattern analysis is the finishing phase of web usage mining, which is used to eradicate the uninteresting patterns and extract the useful patterns from the mined information of pattern discovery phase. The pattern analysis has some common methods of query mechanism such as SQL (structured Query Language) and OLAP operation [2].

III. WEB LOG FILES

A web log is the main source to be used in the web usage mining process. When the user requests the particular web page on any website that entry will be created as a record on the server automatically, which is maintaining a history of user browser behaviour [3]. Typically, the web log contains some fields such as IP address, date & time, request method, status, bytes, referrer, user agent etc.

A. Different Data Sources

Web logs are collected from different data sources like web server, proxy server, and client server.

1) Web server logs:

The log files are mostly collected from web server, which is containing the user log records. Server log maintains the relationship between the web server and the users. Generally, a server log file has four types, namely access log, error log, referrer log and agent log [2] [3].

- Access logs : access log files stores all incoming requests of the user access
- Error log: error log is used to store all the failed http error files
- Referrer logs: the referrer logs maintain the records, how the user has linked to each site and each page.
- Agent logs: it indicates information about browser, operating system that is used by the user.

2)Proxy server logs:

The proxy server log files are stored in the proxy server, which is used to handle the user request page when the main server is unable to respond to the user access [4].

3)Client server logs:

The client server log is also called as browser log, which consist of client browser itself. Client server logs helps to handle the web page caching and session reconstruction problems with the use of HTTP cookies [4].

B. Web Log File Format

There are different types of weblog format available in the various web servers such as

- w3c extended format – w3c is defined by the world wide web consortium, which is an access log for web server. It contains data about each access request.

- NCSA common format – this is regularizes text file format used by the server, which format is fixed ASCII text so the user unable to customize it.
- IIS log format- this is also having a fixed ASCII text format. But it contains more information than the NCSA log format.

These three are common web log file format on the web server [1][3].

IV. DATA PREPROCESSING

The weblogs usually contains the adequate information about the click stream data of user requests it may be incomplete, noisy and unstructured data. Data preprocessing is the one of the phases of web usage mining, which is used to remove the inconsistent, irrelevant and redundant data. The main intention of data preprocessing is to transform raw logs into the cleaned logs that can be given for further processing of web usage mining [5] [6].

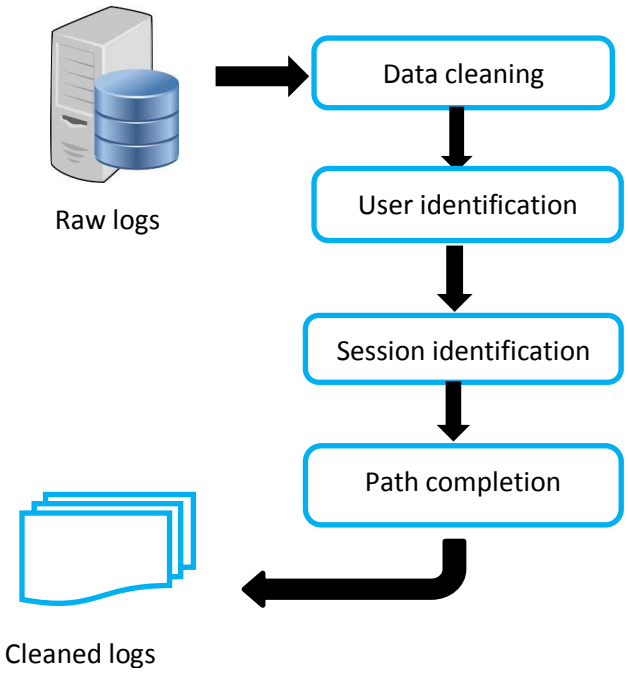

Fig 1. Data preprocessing steps

The data preprocessing has some techniques such as data cleaning, user identification, session identification and path completion, which can be used to clean the weblogs effectively. The above figure shows steps of data preprocessing.

A. Data Cleaning

Data cleaning is the essential step of data preprocessing which is used to remove the unwanted information from the web logs. The log files obtain very large data and it takes 80% process work in the data cleaning step, then only the further processes of pattern discovery are performed effectively. During data cleaning, we have to remove the irrelevant information, which is followed below:

- **Remove Image File:** The log file entries contain different extension format. But we need only relevant extension data other than the extension files of image file, graphics or multimedia and style sheet pages will be eliminated which is with the extension .gif, .jpg, .jpeg, .css to be removed from log files [8].
- **Remove Failed Status Code:** The log entries with failed status code which is to be removed from the weblogs. There are different failed status codes available it may be like this,
  400 - Invalid request page
  403 - Forbidden page
  404 - Page not found

206 - Partial content
304 –Not modified
412 - Condition failed

These are the some of the failed status code, but we need only the log entries with success that is 200 status codes and rest of others will be removed from the weblogs.

- **Remove Spider/ Robots Files:** The weblogs also entries with automated search engine such as robots, spider and crawler files removed from web logs. The robot file extension with robots.txt that is has to be removed.

B. User Identification

After the data cleaning process is over then the next step is user identification, which is used to identify which pages are accessed by whom. Each user obtains an individual IP address, which could be representing different user. The main task of user identification is to find the unique user to use of IP address, user agent and referrer URL. There are some following methods, which are used to identify the unique user [7].

- If an IP address is new then it represents a new user.
- If an IP address is same but agent logs (browser and operating system) are different then it represents distinguish user.
- If an IP address and agent log is same then check browsing path using referrer log if it is a mismatch then it represents a one more user in same IP address.

C. Session Identification

Session identification is the one of the stages of data preprocessing which is used to identify the user session that is the sequence of web pages are accessed by a single user in the particular time. The default session time out is 30 minutes, but when the user requests for page, the user session is accessed until time out. In case of the same user, spend time more than 30 minutes on the same web page then the next session is started.

There are some condition which can be followed during session identification that are,

- If the IP address is new then the session is considered, has a new session.
- If the time exceeds 30 minutes, then the new session is started.

D. Path Completion

Path completion is the final step of preprocessing which are acquiring the entire path of the user access. The user page request is could not recorded in the log entries during the caching problems, POST method and during the use of "forward" or "back" button of a browser. If the user page request is not directly linked to the last page requested, then check the recent history, in case entries are not available in recent history then have to check the referrer URL in which the page request closest to the unknown page request that source is filled in this path and the pages is added to the user session.

V. TOOLS FOR PREPROCESSING

There are many log file analyser tools are available on the internet like log miner, deep log analyser, awstats, web log expert, logstash, log cruncher, weblizer and so on. But these types of tools provide only statistical analysis of log files and it is unable to perform the data preprocessing method of data cleaning. In this paper, use weblog explorer tool, which is used to analyse web server log and generate reports. It is the flexible system of filters, which give information about visitor who accessed the specific web page and this tool is used to remove the irrelative entries and generate useful log report. So the preprocessing steps of web usage mining will be done effectively.

VI. EXPERIMENTS AND RESULTS

In this experiment the dataset for log files are collected from "http//www.hoonlir.com/log/ access.log" which is dated from 9.1.2016 to 15.1.2016. There are totally 4193 access log entries recorded in the access log, which contains a page files, image files, error files, other files. The collected raw log file is below here,

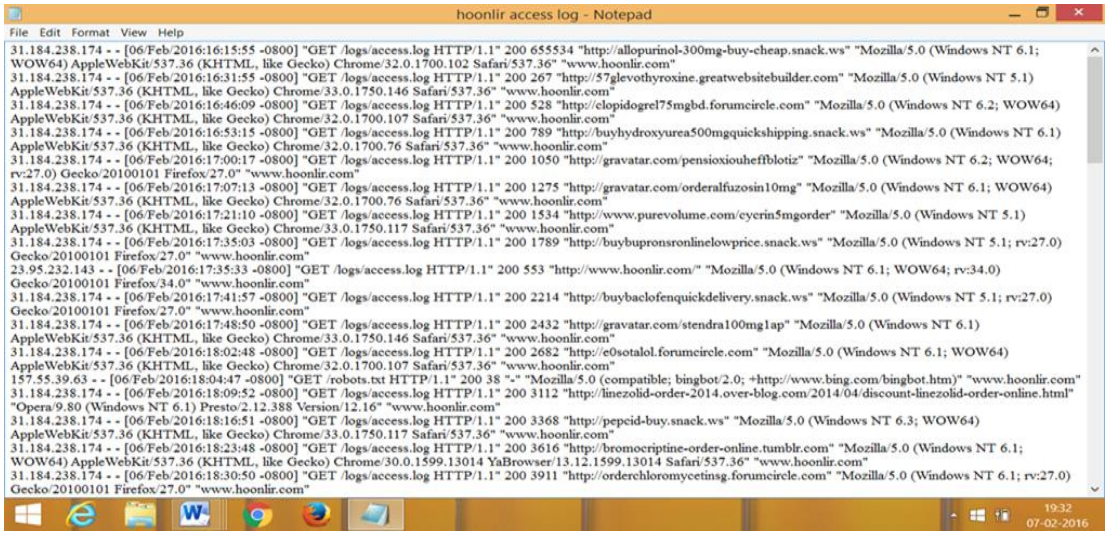

Fig 2 . Raw web log files

This weblog contains below type of irrelative file extension, and error code such as

- Image files - .jpg, .jpeg, .png
- Error code - 404, 206, 304,
- Other files - .js, .css,.log,

These unwanted log files will be eliminated from weblogs and only capture the useful files through web log explorer from which the required options will be chosen in the tool. The below figure shows that the log files which displays the respective fields such as IP address, date and time, status method, the user agent.

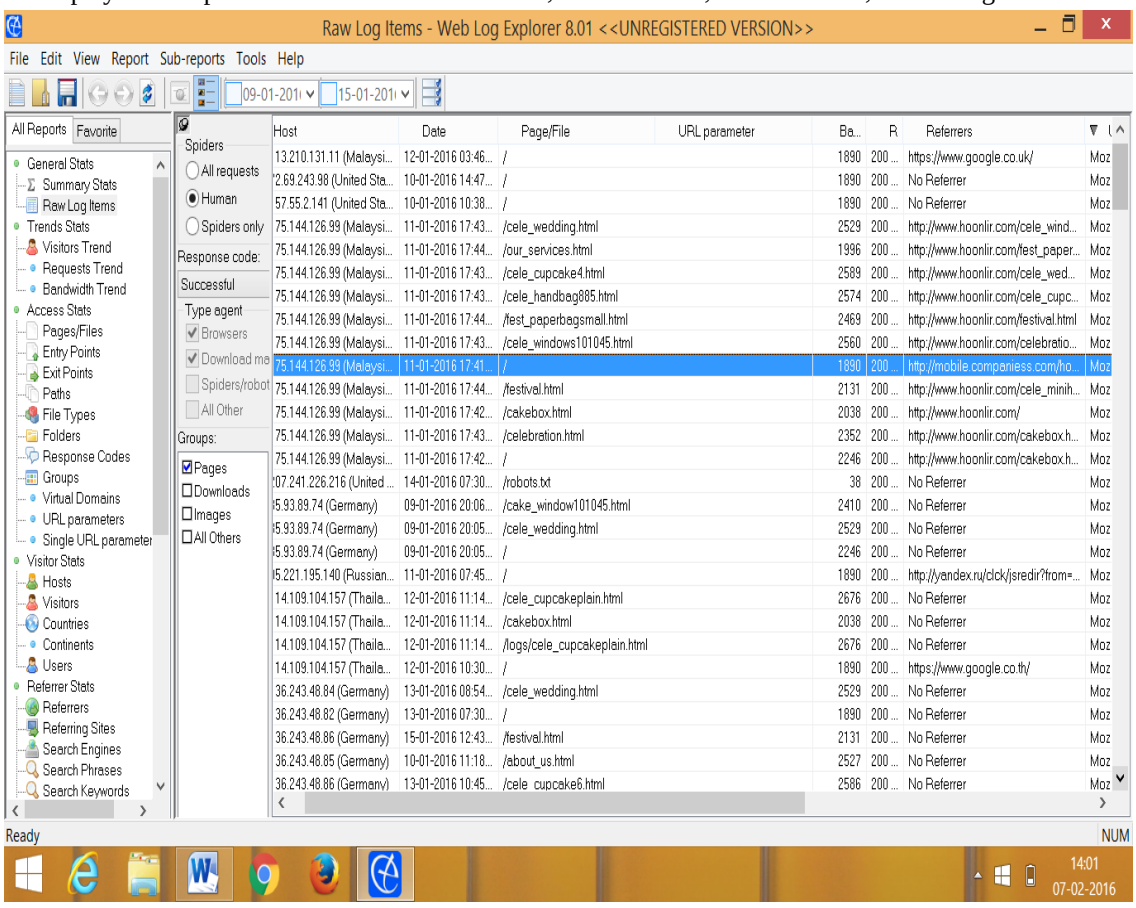

Fig 3. Weblog explorer tool- preprocessing process

In this Fig 3 shows that how the weblog explore tool taking the preprocessing step, in which there are various options available the above window shows the raw log items from that is to select the request option whether is

human or spider. In those logs, only human request is needed and to select the success response code which displays as the 200 status code entries. There are totally 4193 log entries are recorded in the web log file from 995 spider logs are presented. After removing the spider from the weblog, the number of entries becomes 3198. Then remove error code files such as 404, 206, 304, which has 799 entries in log files. Then to remove the image files and other files, which have totalled 2145 entries, are recorded in the web log files. Finally, the web log is cleaned effectively now the log files have only 254 useful page file entries. This is the useful exact pattern detected from the data preprocessing step. The below table summarizes the number of weblogs entries in the cleaning process

Raw logs	After removing spider files	After removing error files	After removing image files	After removing other files	Cleaned logs
4193	3198	2399	1414	254	254

Table1. Entries of log file

This table shows that the number of entries after removing the irrelative files where the entries are shows in graphical representation in the below Fig 4.

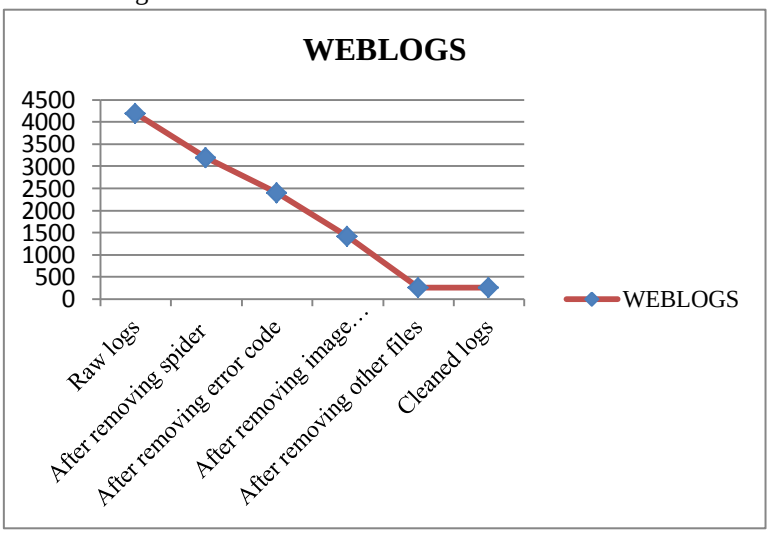

Fig 4. Number of weblog entries

The below Fig 5 shows that the result of final preprocessed weblog that contains the total raw log entries has 4193 after data cleaning the weblogs become 254. In cleaned logs individual has a unique IP address, user agent, and URL to represent as a unique user.

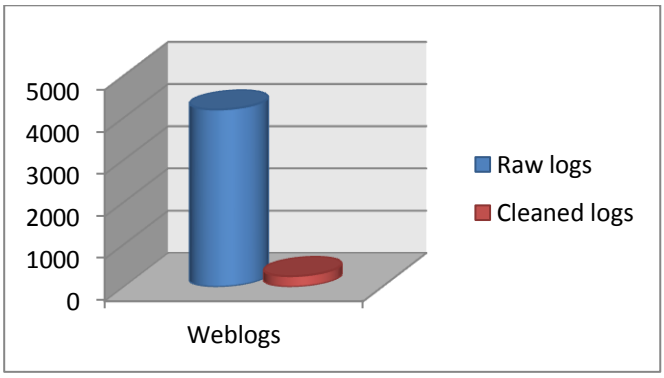

Fig 5. Result of preprocessed weblogs

VII. CONCLUSION

The World Wide Web is a vast collection of web document such as text, images, and multimedia data too, which is not an easy task to retrieve the exact data from the web. Web mining technique is one of the data mining techniques, which is used to mine the data from the web. This paper described the data preprocessing of web usage mining from which the weblogs are cleaned effectively using the web log explorer tool. This tool has the flexible system of filters, which gives information about visitors who accessed the specific web page. It is used to remove the irrelative entries and generate useful log report. Now these preprocessed weblogs can be used for further stages of pattern discovery and pattern analysis of web usage mining.

REFERENCES

[1]. Mugali, Chaitra L., AyeshaAzeema Maniyar, and Asst Prof Padma Dandannavar. "Pre-Processing and Analysis of Web Server Logs." (2014).

[2]. ERRITALI, Mohammed, and Hanane EZZIKOURI. "Pretreatment of web log files." *Journal of Information Sciences and Computing Technologies* 2.1 (2015): 108-121.

[3]. Chitraa, V., and A. Selvadoss Thanamani. "A novel technique for sessions identification in web usage mining preprocessing." *International Journal of Computer Applications* 34.9 (2011): 23-27.

[4]. Khosla, Mr Shivkumar, and Mrs Varunakshi Bhojane. "Capturing Web Log and Performing Preprocessing of the User's Accessing Distance Education System." *Int. J. Mod. Eng. Res.(IJMER)* 2.5 (2012): 3128-3130.

[5]. Sait, Abdul Rahaman Wahab, and Dr T. Meyappan. "Data Preprocessing and Transformation Technique to Generate Pattern from the Web Log." *International conference on Computer Science and Information Systems (ICSIS'2014) Oct*. 2014.

[6]. Chitraa, V., Dr Davamani, and Antony Selvdoss. "A survey on preprocessing methods for web usage data." *arXiv preprint arXiv:1004.1257* (2010).

[7]. Raiyani, Sheetal A. "Efficient Preprocessing technique using Web log mining." *International Journal of Advancements in Research & Technology* 1.6 (2012): 59-63.

[8]. Ramya, C., G. Kavitha, and Dr KS Shreedhara. "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process." *arXiv preprint arXiv:1105.0350* (2011).

