International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 2 (2018) pp. 1227-1232

An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining

Jayanti Mehra[1]
Research Scholar, Department of computer Application,
Maulana Azad National Institute of Technology (MANIT), Bhopal,
Madhya Pradesh, India.

Dr. R S Thakur[2]
Assistant Professor, Department of computer Application,
,Maulana Azad National Institute of Technology (MANIT), Bhopal,
Madhya Pradesh, India.

Abstract

World Wide Web is rising rapidly and enormous amount of information is produced due to user's communications with web sites. To exploit this information, recognizing usage pattern of users is very significant. Web Usage Mining is the application of data mining techniques to find out the useful, hidden information about the users and interesting patterns from data extracted from Web Log files. It supports to know frequently accessed pages, suppose user navigation, progress web site structure etc. In command to relate Web Usage Mining, variety of actions is executed. This paper describes the procedure of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has furthermore existing several approaches for example statistical analysis; clustering, association rules and sequential pattern are individual used to determine patterns in web usage mining In this paper describes data preprocessing technique of web usage mining, after completion of data preprocessing, any kind of irrelevant information can be sort out. We have also proposed an algorithm and its implementation for web log preprocessing in web usage mining. Every page has been allocated with an individual token. According to this token and frequency, data mining technique (Classification, Association Rules, and Clustering) can be applied. In this article we can simply discover the highest and lowest value according to page access frequency.

Keywords: Web Usage Mining, Web Log Data, Preprocessing, Frequency**.**

INTRODUCTION

The World Wide Web is a source of web pages that gives the lot of information to the internet users. For internet users the information presented on web has develop into a essential source. for the reason that , there is rising expansion and complication of websites accessible on internet, the amount of web is huge. A web site is the connection the consumer to company. The companies can revise visitor's performance during web investigation, and discover the patterns. Web mining is generally distinct as find outing and study of helpful information commencing the World Wide Web. Web mining split into three parts: Web Contents Mining, Web structure mining and Web Usage Mining. Web Contents Mining represents the taking out of helpful information and web knowledge from web resources or web contents such as text, image, audio, video, and structured records. Web Usage Mining can be as the find and analysis of access patterns of user, throughout the mining of log files. The output of the WUM can be used in web personalization, recovering the system performance, site alteration, usage description etc. Web log file is a server log file which is a fundamental data sources in Web usage mining, in which it include - access logs of the web server. The significant step in the WUM is Data Preprocessing segment. It includes of data cleaning, session identification, user identification, path completion. Data preprocessing is used to clean the inappropriate data from log file so it can be give to the pattern discovery to recognize the user pattern. [1]

Preprocessing is the initial step need after that step we include to execute data mining methods like association rules, classification and clustering for recovering interesting information [4]. The data preprocessing contains data cleaning, user identification, session identification and path completion [3].

Preprocessing is proficient then it can be simply search frequent pattern or interesting rule between web data with limited quantity of time. In this paper we have analyze the web log data and discriminate its attributes. After fetch the data we need to pre-process of that data and provide an algorithm for data preprocessing.

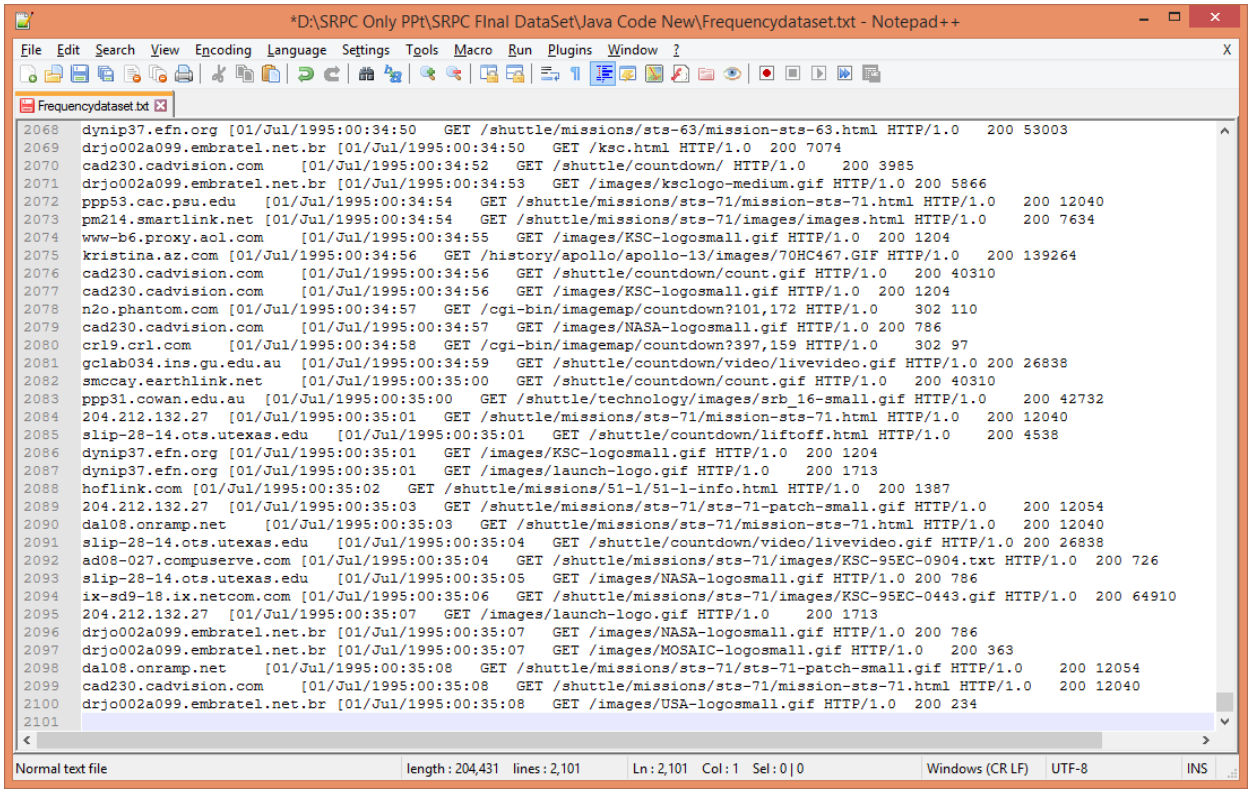


Figure 1. Sample of web log data

A. Data Preprocessing

The data gathered from the web log file is incomplete, noisy and not appropriate for mining initially. Pre-processing is required to exchange the data into relevant form for pattern finding. We start in on Pre-processing stage by data removal then data cleaning and data filtering because the source of web logs data causes are combined with inappropriate information. Data preprocessing acting an main function in Web usage mining. It is used to sift and systematize just suitable information by using web mining algorithms scheduled the web server logs. The innovative server logs are cleaned, formatted, and then grouped addicted to significant sessions earlier than organism used by WUM. This stage holds three sub steps: Data Cleaning, User Identification, and Session Identification [3]

B. Pattern Discovery

Pattern finding illustrates upon methods and algorithms expanded from numerous fields for example statistics, data mining, machine learning and pattern recognition though, it is not the objective of this paper to explain every the obtainable algorithms and techniques received from these fields.[3] This part explains the varieties of mining performance that have been functional to the Web field. Methods build up from other areas ought to obtain into deliberation the diverse types of data generalizations and previous knowledge obtainable for Web Mining. For illustration, in association rule finding, the idea of a deal for market-basket analysis does not get into deliberation the direct in which pieces are chooses. Yet, in Web Usage Mining, a server session is an efficient series of pages demanded by a user. Also appropriate to the complexity in recognizing unique sessions, extra previous knowledge is compulsory [3][8]

C. Pattern Analysis

Pattern analysis is the final action in the usually Web Usage mining procedure as explained in Figure 1. The reason at the back pattern analysis is to sort out unexciting convention or patterns commencing the set create in the pattern finding stage. The accurate analysis methodology is typically ruled by the function for which Web mining is completed. As a rule common form of pattern analysis consists of a knowledge query method. A further process is to load usage data into a data cube to facilitate perform OLAP operations. Visualization techniques, such as graphing patterns or allocating colours to diverse values, can frequently highlight on the whole patterns or developments in the data. Content and structure information be able of be used to sort out patterns enclosing pages of a usage category. Content type or pages that match a definite hyperlink organization.

RELATED WORK

DATA PREPROCESSING

The information existing in the web is diverse and unstructured. Consequently, the preprocessing segment is a requirement for find out patterns. The objective of

preprocessing is to change the raw click stream data into a set of user profiles. Data preprocessing presents a number of exceptional challenges which led to a diversity of algorithms and heuristic techniques for preprocessing step such as integration and cleaning, user and session identification etc. A variety of research works are approved in this preprocessing part for combination sessions and transactions, which is used to determine user behavior patterns. [5][9]

A. Data Collection

In this article, the data source which is in IIS file format, for the finding hidden information of visitor is collected by NASA-HTTP. The log files .We use the part of the logs during the period of August 1995.For session identification, set the maximum elapsed time to 30 min, which is used in many commercial applications.

The raw data for mining purpose is collected from NASA website. It contains approximately 1727 records in Common log file format. The sample log file used for the task was in raw log format. Size of the file before cleaning was 164 KB with 1727 entries. We can see this in Fig 1.

Data cleaning Log data is stored in database for supplementary processing of data by way of queries and program .Data file acquired was very enormous and it obtains approximately 80% of total time to mine the data. [7] In data cleaning process, the unnecessary information is removed from the log database. The data cleaning obtains the following steps:

Step1: Elimination of the entries having image files, graphic or multimedia files. The records which are accessing file with extension gif,jpg, jpeg etc. are to be removed.[6]After performing this step around 392 records left.

Step 2: The elimination of entries with failed status code. A variety of status codes for HTTP 1.1 in this step the entries having status code of 200 will be retained, rest are removed.

Step3: Removal of records with bytes transferred field zero. The records having entries zero in the byte transferred field specifies that the requested page is not opened, and is to be removed. After performing the above two steps the number of records left are 380. [1]

C. An algorithm for data cleaning

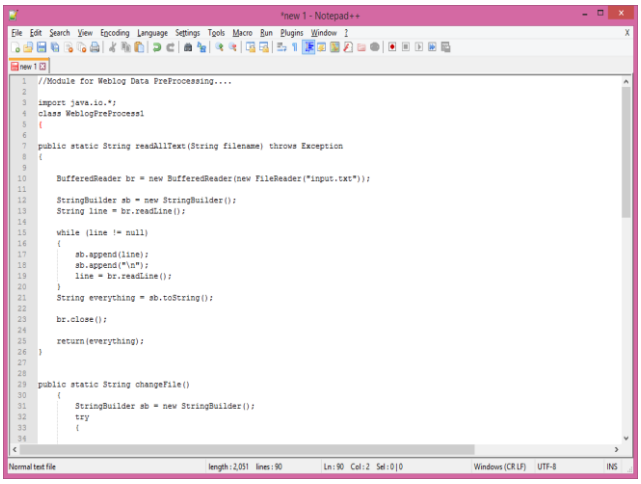
1) Start
2) Scan the Log Record in log file
3) For every record in log file
4) Read all the fields (Referrer, Methods, Status etc)
5) If status code = Success
6) then
7) Take IP Address and URL
8) If(Suffix of URL=*.txt, *.mpg,*.gif , *.css, *.jpg)
9) then
10) Remove suffix from URL
11) Otherwise
12) Save records
13) End if
14) Fetch the next record
15) End if
16) Stop

METHODOLOGY

We implemented the algorithm in java programming language. To clean the web log data, read the web log file and calculate all the record. The method is so as to, we read character by character from the file and evaluate the character from ASCII value of space and enter key and count up all the record from web log file. We can see this in the figure 2

Web log File Cleaning: In this action, the irrelevant log entries are deleted from the log file. This can be completed by examination in the request field of the log file, the suffix of the website URL requested by the user. These suffixes notify us the authentic format or extension of the web files requested by user. Contained by the log file, we will receive only those files which have extensions like .html, .asp, .aspx, .php. So we can also delete every log entries taking extensions like .gif, .jpeg, .flv, .mp3, .mp4, etc. We can also delete log entries with empty URL or having request methods other than GET and POST. We can also delete all those log entries with HTTP.[6]

Status code other than 200 .At the end the cleaned log file is organized for the next steps.



Log data have different types of attribute like IP Address, User Name, Timestamp, access Request, Status code, Byte Transferred, Referrer, User Agent etc.Once studying every these diverse types of attributes we experiment on the access request

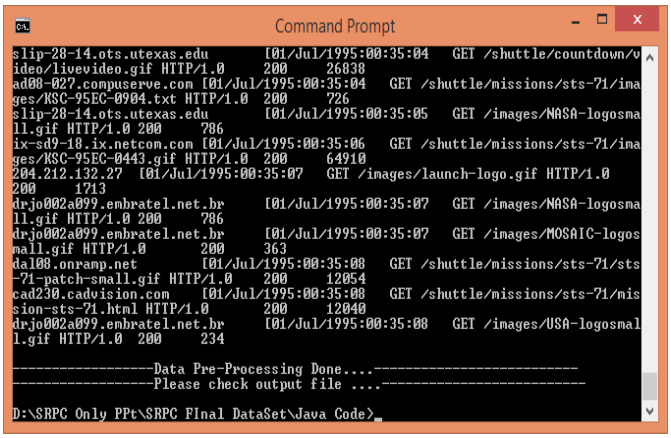
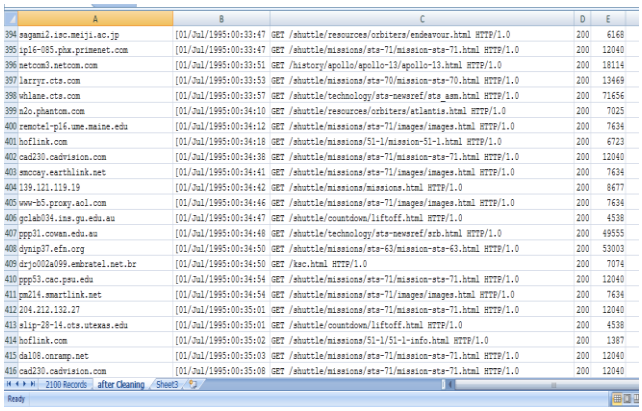


Figure 3. Result after cleaning process

This output already stored in text file we can also store above output in excel sheet below screen shot is output within the excel sheet.



Result after data cleaning in excel sheet

Comparison in Size Before and After Cleaning Size (KB)

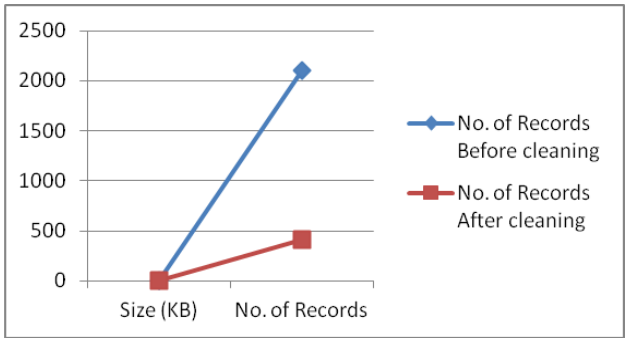


Figure 4. Comparison graph of Data cleaning

Find out the most popular page from web log file

In this section we can find out most popular Page. From this section firstly we can apply data preprocessing technique. Then we counted the frequency of the access page, the term Frequency of page means that the numbers of visibility of that page in web log file. This method we implement in java language which read the string line by line and checked it with other string. If the string gets matched again and again then we increment its counting by one each time and this counting only shows its frequency. This process is repeated until we reach the end of file.[4]

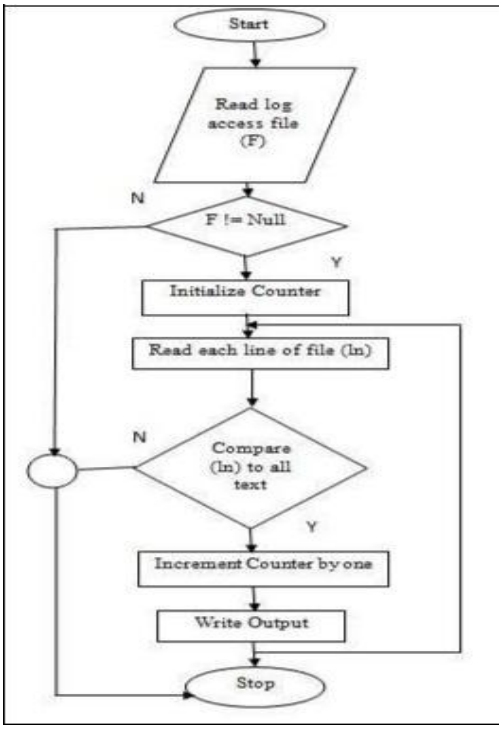


Figure 5. Flow diagram for frequency of page

	Size (KB)	No. of Records
No. of Records Before cleaning	200 KB	2100
No. of Records After cleaning	41 KB	416

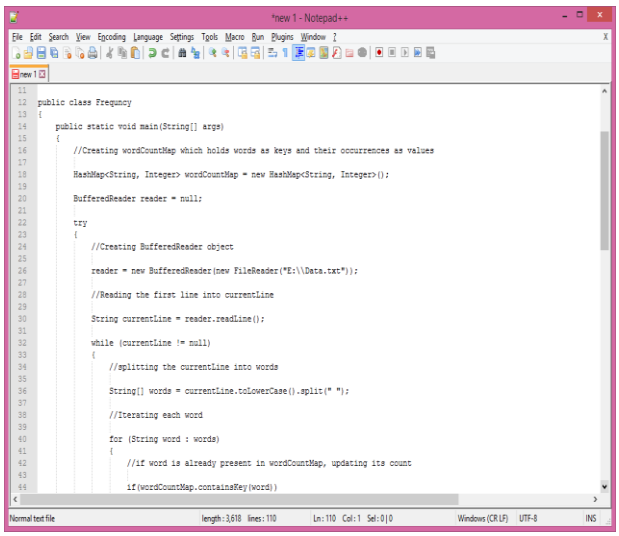


Figure 6. Java code for page access Frequency

In this we can find out most popular page after that

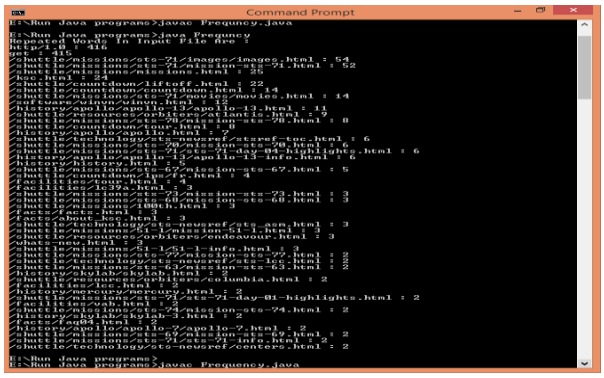
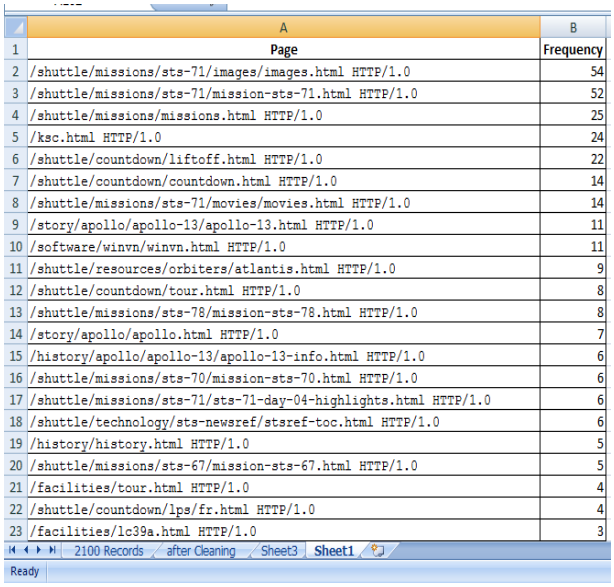


Figure 7. Result for page access frequency



	A	B
1	Page	Frequency
2	/shuttle/missions/sts-71/images/images.html HTTP/1.0	54
3	/shuttle/missions/sts-71/mission-sts-71.html HTTP/1.0	52
4	/shuttle/missions/missions.html HTTP/1.0	25
5	/ksc.html HTTP/1.0	24
6	/shuttle/countdown/liftoff.html HTTP/1.0	22
7	/shuttle/countdown/countdown.html HTTP/1.0	14
8	/shuttle/missions/sts-71/movies/movies.html HTTP/1.0	14
9	/story/apollo/apollo-13/apollo-13.html HTTP/1.0	11
10	/software/winvn/winvn.html HTTP/1.0	11
11	/shuttle/resources/orbiters/atlantis.html HTTP/1.0	9
12	/shuttle/countdown/tour.html HTTP/1.0	8
13	/shuttle/missions/sts-78/mission-sts-78.html HTTP/1.0	8
14	/story/apollo/apollo.html HTTP/1.0	7
15	/history/apollo/apollo-13/apollo-13-info.html HTTP/1.0	6
16	/shuttle/missions/sts-70/mission-sts-70.html HTTP/1.0	6
17	/shuttle/missions/sts-71/sts-71-day-04-highlights.html HTTP/1.0	6
18	/shuttle/technology/sts-newsref/stsref-toc.html HTTP/1.0	6
19	/history/history.html HTTP/1.0	5
20	/shuttle/missions/sts-67/mission-sts-67.html HTTP/1.0	5
21	/facilities/tour.html HTTP/1.0	4
22	/shuttle/countdown/lps/fr.html HTTP/1.0	4
23	/facilities/lc39a.html HTTP/1.0	3

2100 Records | after Cleaning | Sheet3 | Sheet1
Ready

Figure 8. Frequency of an individual page

CONCLUSION AND FUTURE SCOPE

Web data preprocessing is a significant research way of in the field of Web Mining. Web log files are the greatest source to predict user's behavior. Web log file has useful information and it is also contains entries for unnecessary details like image access, failed entries etc. which are not needed to our mining process. Therefore, it becomes necessary to get divest of this irrelevant information. In this paper, the different phases of data pre-processing have been described. Algorithms for performing the data cleaning technique on server log have also been discussed. The proposed algorithm was successfully tested on the log files for data cleaning. The results which were found after the analysis were acceptable and included important information concerning the log files. The data cleaning approach demonstrated a quite salient reduction in the number of records and in the log files size and therefore enlarges the quality of the available data. Here we also counted the page access frequency and distinct different pages. So that most popular page and least popular page can be find out.

ACKNOWLEDGEMENT

We take opportunity to express our obligation and very thankful to all those who have helped us directly or indirectly to successful completion of this review paper.

REFERENCES

[1] Surbhi Anand and Rinkle Rani Aggarwal, "An Efficient Algorithm for Data Cleaning of Log File using File Extensions" *International Journal of Computer Applications*, 2012.

[2] Sheetal A.Raiyani,Shailendra, " Efficient Preprocessing technique using Web log mining," International Journal of Advancements in Research & Technology, Volume 1, Issue6,November-2012.

[3] V.Chitraa and Dr.Antony Selvadoss Thanamani,"A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011.

[4] S. Umamaheswari and S. K. Srivatsa," Algorithm for Tracing Visitors' On-Line Behaviors for Effective Web Usage Mining,International Journal of Computer Applications (0975 – 8887) Volume 87 – No.3, February 2014

[5] Sujith Jayaprakash and Balamurugan E.," Comprehensive Survey on Data Preprocessing Methods in Web Usage Minning", International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3170-3174

[6] Ketan D. Patel and Satyen M. Parikh," Preprocessing on Web Server Log Data for Web Usage Pattern Discovery "International Journal of Computer Applications (0975 – 8887) Volume 165 – No.10, May 2017.

[7] Arshi Shamsi, et. All," Web Usage Mining by Data Preprocessing", IJCST Vol. 3, Iss ue 1, Jan. - March 2012.

[8] Zhuang Like, Kou Zhongbao and Zhang Changshui, "Session identification based on time intervals in Web log mining," Journal of Tsinghua University (Science and Technology), 2005.

[9] Brijesh Bakariya and G.S.Thakur, "Preprocessing on Web Log Data in Web Usage Mining," International Conference on Intelligent Computing and Information System ICICIS, 2012

[10] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood, " Web Usage Mining: A Survey on Preprocessing of Web Log File," IEEE, 2010.

[11] T. Murata and K. Saito, "Extracting Users' Interests from Web Log Data," Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings, 2006.

[12] R. Cooley, B. Mobasher and J. Srivastava, "Data preparation for mining world wide web browsing patterns," Knowledge and Information System, 1999.

[13] Thi Thanh Sang Nguyen, Hai Yan Lu and Jie Lu, "Web-page Recommendation based on Web Usage and Domain Knowledge," IEEE, 2013.