International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 4 Issue 3, March 2015

A Survey on Frequent Web Page Mining with Improving Data Quality of Log Cleaner

Ms Shashi Sahu[1], Asst Prof. Leena Sahu[2]

Abstract— ***Data cleaning is the process of detecting and correcting the irrelevant, incomplete data from the datasets and log files and then replacing modifying this dirty data. Data cleaning is one of the major techniques used in the Data Preprocessing and Web Usage Mining. Data cleaning is very useful in the fields like banking, insurance, retailing, etc. There is lots of work on data cleaning of web server logs irrelevant items and useless data can not completely removed and Overlapped data causes difficulty during page ranking. Studied in previous paper there are many techniques of web log mining. They are Two-level clustering method, Effective and scalable technique. This paper presents an overview of web usage mining, its techniques and also provides a summary of LogCleaner that can filter out plenty of some irrelevant, inconsistent data based on the common of their URLs and improve the data quality and efficiency of Web Log.***

Index Terms— *Web Usage Mining (WUM); Data cleaning; Enterprise proxy logs; Web Page Mining; Preprocessing.*

I. INTRODUCTION

Data mining is the computational process of discovering patterns in large amount data sets involving methods at the intersection of artificial intelligence, machine learning of Data System. The World Wide Web is now a huge database with this growth there arises a need for analyzing the data. The process of discovery and analysis of Web is called Web mining. Web mining is the application of data mining techniques to discover patterns from the Web. Web mining can be divided into three different types 1) Web Structure Mining 2) Web Content Mining 3) Web Usage Mining. Web structure mining is the process of discovering the connection between web pages. Web content mining includes mining, extraction and integration of useful data and knowledge of Web page content. Web Usage Mining is a technique of extracting useful information from the Web Log, e.g. the pattern in which a user goes through different Web pages. EPLogCleaner that can filter out plenty of ir relevant items based on the common prefix

of their URLs of data cleaning methods. Mining enterprise proxy log plays an important role for enterprise manager and employer which makes it difficult to find the "right" or "interesting" information [1]. Web Log are generally noisy and ambiguous. Web applications are increasing at an enormous speed and its users, are increasing at exponential speed.

There are lots of work on data cleaning of web server logs irrelevant items and useless data can not completely removed. When multiple data sources need to be integrated, data quality problems are present in single data collections, such as files and databases. Shaa H. et al [1] has proposed EPLogCleaner Method . EPLogCleaner that can filter out plenty of irrelevant items based on the common prefix of their URLs. This method is improving data quality by removing the irrelevant items. Hussain T.et al [2] has described of data characteristics reveals the importance and difficulty of data cleaning in web mining.

The remainder of this paper is organized as follows: Section 2 describes the Web Usage Mining. Section 3 describes a literature review of some papers. Section 4 describes the problem statement of the research fields. Section 5 includes Methodology and at last we tend to conclude this paper in section 6.

II. WEB USAGE MINING

Web Usage Mining could be a technique of extracting useful information from the web log, e.g. the pattern in which a user goes through different Web Pages. Using usage mining a designer can work on improving the web site or to provide a personalized service. Web Usage Mining consists of three steps [6].

1) Data Preprocessing
2) Pattern Discovery
3) Pattern Analysis

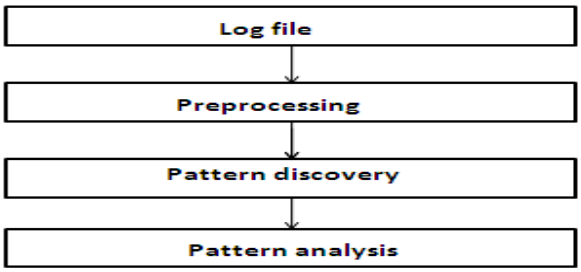

Figure1 Process of Web Log Mining

1. Data Preprocessing

The preprocessing of web logs is complex and time consuming and it is done using the following steps. The main task of data

Manuscript received March, 2015.

Ms. Shashi Sahu, *Computer Science & Engineering, RCET Bhilai (Chhattisgarh) India.*

Asst Prof Ms. Leena Sahu,, *Computer Science & Engineering, RCET Bhilai,(Chhattisgarh) India.*

ISSN: 2278 – 1323

preprocessing is to select standardized data from the original log files, prepared for user navigation pattern discovery algorithm [5].

1)Data Cleaning
2)Page view Identification
3)Path Completion
4)Formatting

1.1 Data Cleaning

Data cleansing is that the method of removing irrelevant logs from log entries. Since HTTP is a connectionless protocol, when a user browse a web page in several log entire graphics and scripts are downloaded along with the HTML file. Data cleaning involves:-

a) Removal of Global and local Noise
b) Removal of images, video etc.
c) Removal of records that failed HTTP status code
d) Robots cleaning
e) Web noise can be normally categorized into two groups depending on their granularities.
f) Global Noise are corresponds to the unnecessary objects with huge granularities, which are no smaller than individual pages.
g) Local (Intra-Page) Noise are corresponds to the irrelevant items inside a Web page.

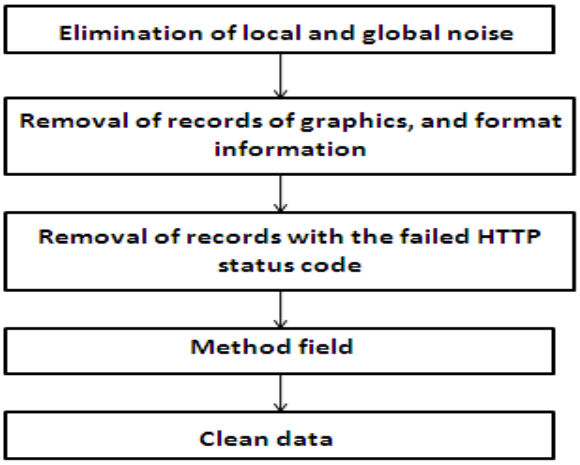

Figure 1 - Steps of Data cleansing

1.2 Page view Identification

Page view is a collection of web object. Page view identification is the process of identifying which page access files belong to a single page view. All the page views are assigned with a page view id.

2. Pattern Discovery

It is a method used in various fields such as data mining, pattern recognition, etc. pattern discovery involves finding a pattern in which the web user uses the web. There are various algorithms available to do this process such as the Association Rule for data mining.

3. Pattern Analysis

It is the last step in mining. It involves analyzing the pattern that is discovered in pattern discovery process. Useful and interesting pattern are kept and rest of the pattern, which are least useful, and interesting are removed.

III. LITERATURE REVIEW

The important task of web mining is web usage mining, which mines log records to discover user access patterns of web pages. This section provides a detailed discussion about several WUM techniques. The following section discusses the various works of several authors. Shaa H. et al [1] has proposed EPLogCleaner Method. EPLogCleaner that can filter out plenty of irrelevant items based on the common prefix of their URLs. Experimental results show that EPLogCleaner can improve data quality of enterprise proxy logs by further filtering out more than 30% URL requests comparing with traditional data cleaning methods.

Tyagi N.et al [3] provides an algorithmic approach to data preprocessing in web usage mining. They take requests for graphical page content, or the other file which can be induced into a web page, or navigation sessions performed by robots and web spiders into consideration.

Zheng L.et al [4] has proposed Optimized User Identification, Optimized Session Identification. The optimized data preprocessing technology is used to improvement of the technology betters the quality of data preprocessing results. The strategy based on the referred web page is adopted at the stage of user identification. Experiments have proved that advanced data preprocessing technology can enhance the quality of data preprocessing results.

Munk M.et al [5] has tried to assess the impact of reconstruction of the activities of a web visitor on the quantity and quality of the extracted rules which represent the web user behavior patterns. Experiment, find out to which criteria are necessary to realize this time-consuming data preparation and specifying the inevitable steps that are required for obtaining valid data from the log file.
Nithya P and Sumathi P [6] have proposed novel pre-processing technique This method describes removing local and global noise and web robots. This paper continues the line of research on Web access log analysis is to analyze the patterns of web site usage and the features of users' behavior. It is vital to preprocess the log data for efficient web usage mining process.

ISSN: 2278 – 1323

SUJATHAa V.et al [7] has planned Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data. In the initial stage PUCC focuses on separating the potential users in web log data, and in the second stage clustering process is used to cluster the potential users with similar interest and within the third stage the results of classification and clustering is used to predict the user future requests.

Theint Theint [8] has proposed data mining techniques to discover user access patterns from web log. This paper mainly focuses on data preprocessing stage of the first phase of web usage mining with activities like field extraction and data cleaning algorithms. Field extraction algorithm performs the process of separating fields from the single line of the log file.

Hui Lee C.et al [9]has proposed an efficient prediction model, two-level prediction model (TLPM), TLPM can decrease the size of the candidate set of web pages and increase the speed of predicting with adequate accuracy. The experiment results prove that TLPM can highly enhance the performance of prediction when the number of web pages is increasing.

Losarwar V.et al [10] has discussed the importance of data preprocessing methods and various steps involved in getting the required content effectively. A complete preprocessing technique is being proposed to preprocess the web log for extraction of user patterns. Data cleaning algorithm removes the irrelevant entries from web log.

Hussain H. et al [12] has surveyed the preprocessing techniques to identify the issues and how WUM preprocessing can be improved for pattern mining and analysis. Summarized the existing web log preprocessing techniques and concluded some results.

Forsati R. et al [15] has proposed Effective and scalable technique to solve the web page recommendation problem. This technique use for distributed learning automata to learn the behavior of previous users' and cluster pages based on learned pattern.

IV. PROBLEM STATEMENT

Discuss the problem relating to Data cleaning of web log. Web log is generally noisy and ambiguous Web applications are increasing at an enormous speed and its users are increasing at exponential speed. Difficult to find the "right" or "interesting" information, There are a lot of work on data cleaning of web server logs irrelevant items and useless data can not completely removed. Difficulty in specifying the valid data from the log file with unlimited accesses to websites, web requests from multiple clients to multiple web servers.

Overlapped data cause difficulty during Page Ranking, When multiple data sources need to be integrated, data quality issues are present in single data collections, like files and databases, e.g., because of misspellings during data entry, missing information or alternative invalid data. The Standard Log file contains irrelevant inconsistent data. Difficulty of knowledge extraction during Web Log Mining.

V. METHODOLOGY

In previous papers, we studied the different techniques and method of data preprocessing are used for data cleaning of server log.

A. Two-level clustering method
B. Noise Detector as an effective technique
C. Novel pre-processing technique
D. Community Detection Technique
E. EPLogCleaner filtering method
F. Effective and scalable technique

A. Two-level clustering

The Two-level clustering method is improving the quality of data in the WUM process, which is the two-level clustering. Based on the results of two level clustering method on web log data, it can be concluded that this method can improve the quality of data web log.

- The first level clustering is done in the form of data frequently user access using non-hierarchical clustering method.

- The second level clustering is done by first changing the form of web log data into user access behavior patterns.

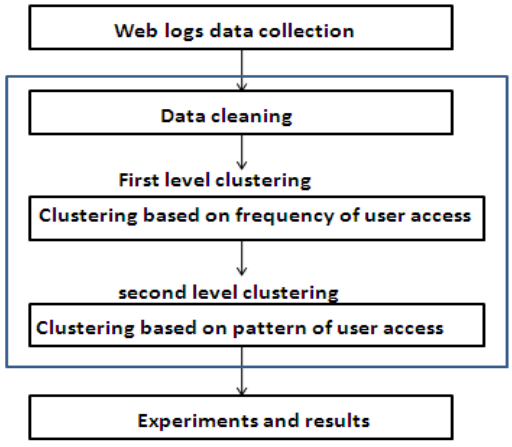

Figure 2 - Two level clustering proces

B. Noise Detector as an effective technique

This methodology is capable of work and eliminating interest noise. Noise Detector can detect the template with high accuracy using two pages only. However, it can be dilated to detect multiple templates per website, and also the challenge will go to reduce the quantity of pages to be checked [16].

C. *Novel pre-processing technique*

ISSN: 2278 – 1323

The Novel pre-processing technique is proposed by removing native and global noise and web robots. Preprocessing is a crucial step since the Web architecture is very complex [6].

D. Community Detection Technique

The Community Detection technique in advanced networks are groups of nodes that share probably a common proprieties or similar functions.

E. EPLogCleaner filtering method

EPLogCleaner that can filter out plenty of irrelevant, inconsistent data based on the common prefix of their URLs. Filtering method can improve data quality and Efficiency of web log. It uses different filter for removing the multimedia data and time, date, status code filtered by some new filtered technique of a log [1].

F. Effective and scalable technique

Effective and scalable technique to solve the web page recommendation problem. Use this technique for distributed learning automata to learn the behavior of previous users and cluster pages based on learning pattern. One in every of the difficult issues in recommendation systems is dealing with unvisited or new extra pages [15].

In our work, each item in the Web log contains the following six fields: client address (host name or IP address), request time, access method (GET, HEAD, POST and so on), accessed URL HTTP status code (200, 400 etc.), webpage size (only for successful request).

```
41.200.89.109 - - [12/Oct/2008:20:18:23 +0100]
"GET/citic2008/soumission.html HTTP/1.1" 200 23247
"http://www.univ-setif.dz/citic2008/index.html" "Mozilla/5.0
(Windows; U; Windows NT 5.1; fr; rv:1.9.0.3)
Gecko/2008092417 Firefox/3.0.3
```

Figure 3 Log Entry in Server log

- The name or IP address of the appealing machine.
- The name and the login HTTP of the user.
- The date and the hour of the request.
- Method used by request. (Get, Post etc)
- The URL of request.
- The used Protocol.
- The request status
- Size of the sent file
- The URL which referred the request

Table 1 Common log format

IP Address	Date & time	Me-thod	Request	Sta-tus	Size
66.249.69.xxx	04/Sep/2011:06:45:13	GET	/news.html	200	15319
valley.net	05/Sep/2011:19:08:48	GET	/info.html	200	15582
206.53.148.xxx	05/Sep/2011:19:08:50	GET	/media.jpg	200	1324
Evins.net	05/Sep/2011:19:20:20	POST	/button.gif	200	30462
114.79.16.xxx	06/Sep/2011:20:00:01	GET	/favicon.ico	200	3798

VI. CONCLUSION

There are many techniques proposed by totally different researchers for the web usage mining. This paper mentioned about various techniques like Two-level clustering method, Noise Detector as an efficient technique, Community Detection Technique, Effective and scalable technique and EPLogCleaner filtering method available for web usage mining.

This previous paper has attempted to give an overview of how weblog mining is done. Web log mining consists of data preprocessing, pattern discovery and analysis. The results of Web Log mining can be used for various applications such as web personalization, site recommendation, site improvement, etc. In this survey, we summarized the existing web log preprocessing techniques and concluded some results. EPLogCleaner can filter out more than 30% URL requests which cannot be filtered by traditional data cleaning methods for proxy logs and improve the design of threshold and the estimation method of precision rate in order to improve in future much more accurate and reliable.

REFERENCES

[1] Hongzhou Shaa,c, Tingwen Liub,c, Peng Qinb,c Yong Sunb,c, Qingyun Liub,c, "EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining", Information Technology and Quantitative Management , ITQM 2013 Procedia Computer Science 17 (2013).

[2] T. Hussain, S. Asghar, N. Masood, "Web Usage Mining: A Survey on Preprocessing of Web Log File", in: Proceedings of the 2010 InternationalConference on Information and Emerging Technologies (ICIET), IEEE, 2010, pp. 1–6.

[3] N. Tyagi, A. Solanki, S. Tyagi, "An Algorithmic Approach to Data Preprocessing in Web Usage Mining", International Journal of Information Technology and Knowledge Management 2 (2) (2010) 279–283.

ISSN: 2278 – 1323

[4] Ling Zheng Hui Gui. Feng Li, "Optimized Data Preprocessing Technology for Web Log Mining",International Conference On Computer Design And Applications (ICCDA 2010).

[5] Michal Munk, Jozef Kapustaa, Peter Šveca*, "Data Preprocessing Evaluation for Web Log Mining:Reconstruction of Activities of a Web Visitor" International Conference on Computational Science, ICCS 2011 Procedia Computer Science 1 (2012).

[6] P.Nithya, Dr.P.Sumathi , "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots. " National Conference on Computing and Communication Systems (NCCCS) IEEE 2012.

[7] V. SUJATHAa, PUNITHAVALLIb, a*, "Improved user navigation pattern Prediction technique from web log data."International Conference on Communication Technology and System Design 2011 Procedia Engineering 30 (2012) 92.

[8] Theint Theint Aye, "Web Log Cleaning for Mining of Web Usage Patterns."IEEE 2011.

[9] Chu-Hui Lee, Yu-lung Lo, Yu-Hsiang Fu, "A novel prediction model based on hierarchical characteristic of web site", Elsevier (2010), Expert Systems with Applications 38 (2011) 3422–3430.

[10] Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012.

[11] Rohit Agarwal,K. V. Arya, Shashi Shekhar, Rakesh Kumar, "An Efficient Weighted Algorithm for Web Information Retrieval System", IEEE(2011).

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

[13] 1Yuhefizar,2Budi Santosa,3I Ketut Eddy P.,4Yoon K. Suprapto, "Two Level Clustering Approach for Data Quality Improvement in Web Usage Mining", Journal of Theoretical and Applied Information Technology 20th April 2014.

[14] 1B.Uma Maheswari, 2 Dr. P.Sumathi, "A New Clustering and Preprocessing for Web Log Mining", IEEE (2014).

[15] Rana Forsati,Mohammad Reza Meybodi,Afsaneh Rahbar, "An Efficient Algorithm for Web Recommendation Systems", IEEE (2009).

[16] Derar Alassi, Reda Alhajj, "Effectiveness of templatedetection on noise reduction and website summarization", Elsevier (2012).

Ms. Shashi Sahu received the B.E. degree from Chhattisgarh Swami Vivekanand Technical University, Bhilai (C.G.) India in Computer Science & Engineering in the year 2013. She is currently pursuing M.Tech. Degree in Computer Science Engineering with specialization in Software Engineering from CSVTU Bhilai (C.G.), India. Her research area includes Data Mining.

Ms. Leena Sahu is currently Assistant professor in Department of Computer science & Engineering RCET, Bhilai (C.G.) India. She completed her B.E and M.Tech. in Computer Science and Engineering Branch. Her research area includes Data Mining, Computer Network etc. She has published many Research Papers in various reputed National & International Journals, Conferences, and Seminars.

ISSN: 2278 – 1323