International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol. 3, Issue. 5, Sep - Oct. 2013 pp-2816-2819 ISSN: 2249-6645

A Novel Method for Data Cleaning and User- Session Identification for Web Mining

Vijay Kumar Padala[1], Sayeed Yasin[2], Durga Bhavani Alanka[3]
[1]*M.Tech, Nimra College of Engineering & Technology, Vijayawada, A.P., India.*
[2]*Asst. Professor, Dept.of CSE, Nimra College of Engineering & Technology, Vijayawada, A.P., India.*
[3]*Assoc. Professor, Dept. of CSE, Usharama College of Engineering, A.P., India.*

ABSTRACT: *The World Wide Web(WWW) is serving as a huge widely distributed global information service center for technical information, news, advertisement, e-commerce and other information service. This makes information retrieval process very difficult. Most users may not have good knowledge of the structure of the information network, and may easily get bored by taking many access hops and losing their patience when waiting for the information. These challenges will have been solved efficiently by using Web mining, which is the application of data mining technologies. Web mining employs the technique of data mining process into the documents on the WWW. The overall process of web mining includes extraction of information from the WWW through the conventional practices of the data mining and putting the same into the website features. The task of web mining is to discover and extract interesting knowledge/patterns from Web is classified into three types as Web Structure Mining that focuses on hyperlink structure, Web Contents Mining that focuses on page contents as well as Web Usage Mining that focuses on Web logs. In this paper, we are concerned about Web Usage Mining (WUM), also called as Web log mining. We propose algorithms for cleaning a web log file, user and session identification.*

Keywords: *Log file, Session, Web Mining, WWW.*

I. INTRODUCTION

Web mining [1] that discovers and extracts interesting knowledge or patterns from Web is classified into three types as Web Structure Mining that focuses on hyperlink structure, Web Contents Mining that focuses on page contents as well as Web Usage Mining that focuses on Web logs. In this paper, we are concerned about Web Usage Mining (WUM), also called as Web log mining. In the web usage mining process, the techniques of data mining process are applied so as to discover the trends and the patterns in the browsing nature of the visitors of the website. There is an extraction of the navigation patterns as the browsing patterns could be traced and the structure of the websites can be designed accordingly. When it is talked about the browsing nature of the users, it deals with frequent access of the web site or the duration of using the web site. This information can be extracted from using a log file. Only these log files record the session information about the web pages [2]. The figure 1 shows the step wise procedure for web usage mining process.

Log files are files that list the actions that have been occurred on web sites. Such log files reside in the web server. Computers that deliver the web pages are called as "web servers". The Web server stores all of the files necessary to display the Web pages on the user's computer. All the individual web pages combines together to form the completeness of the Web site. Images or graphic files and any scripts that make dynamic elements of the site function. The browser requests the data from a Web server, and using HTTP, the server delivers the data back to the browser that had requested the web page. The browser in turn converts (formats) the files into a user viewable page. This gets displayed in the browser itself. In the same way the server can send the files to many user computers at the same time, allowing multiple clients to view the same page simultaneously.

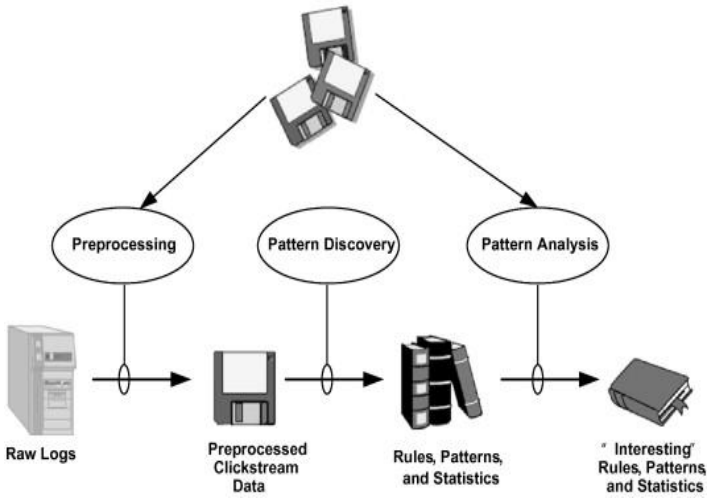

Figure 1: Web usage Mining Process

II. RELATED WORK

Data cleaning is a relatively new research field. The process is computationally expensive on very large data sets and thus it was almost impossible to do with using old technology. The new faster computers allow performing the data cleaning process in acceptable time on large amounts of data. There are many issues in the data cleaning area that researchers are attempting to tackle. They consist of dealing with missing data, erroneous data, determining record usability, etc. Some related research addresses these issues of data quality [3][4]. In [5], the authors introduced a new framework to separate human user and search engine access intelligently with less time span. And also Data Cleaning, User Identification, Session identification are designed correctly. The framework reduces the error rate and also improves significant learning performance of the algorithm.

In [6], the authors introduced a new data preprocessing technique to prune noisy data, irrelevant data, reduce data size and to apply pattern discovery techniques. This paper mainly focuses on data extraction and the data cleaning algorithms. Data cleaning algorithm eliminates unnecessary or inconsistent items in the analyzed data. In. [7], the authors described that the quality of data is an important issue in data mining, and 80 percent of mining efforts spend to improve the quality of data. The data quality depends on accuracy, consistency, completeness, timelines, believability, interpretability and accessibility. In [8], the authors presented a complete preprocessing technique such as data cleaning algorithm, filtering algorithm, user and session identification is performed. They proposed a new hierarchical session identification algorithm that generates the hierarchy of sessions. In [9], the authors classify data quality problems that can be addressed by data cleaning routines and provides an overview of the main solution approaches. In [10], the authors consider the cleansing of data errors in structure and content as an important aspect for data warehouse integration. In [11], the authors suggested that the quality of data is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements. The assessment of data quality dimensions should consider the degree to which data satisfy the user's needs.

III. PROPOSED WORK

A. Cleaning a Web Log File

Data ware house is the only viable solution that can bring that dream into reality. The enhancement of future endeavors to make decisions depends on the availability of correct information that is based on the quality of data underlying. The quality data can only be produced by cleaning data prior to loading into the data ware house since the data collected from different sources will be dirty. Once the data have been cleaned it will produce accurate results when the data mining query is applied to it. So correctness of the data is essential for well-formed and reliable decision making.

Data preprocessing is an important steps to filter and organize only appropriate information before applying any web mining procedure. Preprocessing reduce log file size and also increase quality of available data. The purpose of data preprocessing is to improve data quality and increase data mining accuracy. Preprocessing consists of: data cleansing, user identification, session identification. In this paper the main task is to clean the raw web log files and insert the processed data into a relational database. In this step remove noisy as well as unnecessary data. Remove log entry nodes contain file extension like jpg, gif means remove request such as multimedia files, image, page style file.

Data Cleaning Algorithm
Input: Web Server Log File
Output: Log Database
Step1: Read LogRecord from Web Server Log File
Step2: If((LogRecord.url-stem(gif,jpegjpg,cssjs)) AND
(LogRecord.mehod='GET') AND
(LogRecord.Sc-status<>(301,404,500)AND
(LogRecord.Useragent<>Crawler,Spider,Robot))
then Insert LogRecord in to LogDatabase.
End of If condition.
Step 3: Repeat the above two steps until eof (Web Server
Log File)
Step 4: Sop the process.

B. User Identification

The user identification step identify individual user by using their IP address. If there is new IP address, there is new user. If IP address is same but browser version or operating system is different then it represents a different user. The outcome of this algorithm1 is Unique Users Database gives information about total number of individual users, users IP address, browser used and user agent.

User Identification Algorithm
Input: Log Database
Output: Unique Users Database

Step1: Initialize
IPList=0; UsersList=0; BrowserList=0;
OSList=0; No-of-users=0;

Step2: Read Record from LogDatabase
Step3: If Record.IP address in not in IPList
then add new Record.IPaddress in to IPList
add Record.Browser in to BrowserList
add Record.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 OR
Record.Browser not in BrowserList OR
Record.OS not in OSList
then
increment count of No-of-users
insert as new user in to UserList.
End of If
End of If
Step 5: Repeat the above steps 2 to 3
until eof (Log Database)
Step 6: Stop the process.

C. Session Identification

Each user spends total time in each web page of a web site. Session means time duration spent in web pages of the web site. A referrer-based method is used for identifying the sessions. If IP address, browsers and operating systems are same, then the referrer information should be taken. The cs_referer is checked, and a new user session is identified if the URL in the Refer URI – field is a large interval usually more than 30 minutes between the accessing time of this record.

Session Identification Algorithm
Input: Log Database
Output: Session Database

Step1: Initialize
SessionList=0
UserList=0
No-of-Sessions=0
Step2: Read LogRecord from Log Database
Step3: If (LogRecord.Refer='-') OR
LogRecord.time-taken>30min 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
Step4: Repeat the above steps 2 and 3 till eof (Log
Database)
Step5: End of process.

IV. CONCLUSION

Data preprocessing is an important steps to filter and organize only appropriate information before applying any web mining procedure. Preprocessing reduce log file size and also increase quality of available data. The purpose of data preprocessing is to improve data quality and increase data mining accuracy. Preprocessing consists of: data cleansing, user identification, session identification. In this paper main task is to clean the raw web log files and insert the processed data into a relational database. By cleaning the data, we can create a new database according to our application which includes the information about user identification, session identification. In some web sites the user identification is made by getting the user's profile and allows them to access the web site by using a user name and password. In this kind of access the user is being identified uniquely so that the revisit of the user can also be identified quickly. Next session identification. Session is the time duration spent in the web page of a web site. This done by using the time stamp details of the web pages of a site. The total time used by each of the user of each web page. This can also be done by noting down the user identification number those who have visited the web page and had traversed through the links of the web page.

REFERENCES

[1] R. Kosala, H. Blockeel. "Web Mining Research: A Survey," In SIGKDD Explorations, ACM press, 2(1): 2000, pp.1-15.
[2] R. Cooley, B. Mobasher, and J. Srivastava,(1999) "Data Preparation for Mining World Wide Web Browsing Patterns," KNOWLEDGE AND INFORMATION SYSTEMS, vol. 1.
[3] Ballou, D., Tayi, K.,"Methodology for Allocating Resources for Data Quality Enhancement", CACM, pp. 320-329, 1989.
[4] Redman, T.,"Data Quality for the Information Age", Artech House, 1996.
[5] Maheswara Rao.V.V.R, Valli Kumari.V,"An Enhanced Pre- Processing Research Framework for Web Log Data Using a Learning Algorithm", Computer Science and Information Technology, pp. 01–15, 2011.
[6] Aye.T.T,"Web Log Cleaning for Mining of Web Usage Patterns", International Conference on computer Researchand Development, IEEE, pp. 490-494, 2011.
[7] Marquardt.C, K. Becker, D. Ruiz,"A Preprocessing Tool for Web usage mining in the Distance Education Domain", InProceedings of the International Database Engineering and Application Symposium (IDEAS' 04), pp.78-87, 2004.
[8] Tasawar Hussain, Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level ofWUM Based on Swarm Intelligence", IEEE Conference on Emerging Technologies, pp.21-26, 2010.
[9] E. Rahm, H. Hai Do,"Data Cleaning: Problems and Current Approaches", University of Leipzig, Germany, [Online] Available: http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/ dw/paper/data_cleaning.pdf
[10] Raman, V.; Hellerstein, J.M.: Potter's Wheel: An Interactive Framework for Data Cleaning. Working Paper, 1999. http:// www.cs.berkeley.edu/~rshankar/papers/pwheel.pdf.
[11] Cinzia Cappiello, Chiara Francalanci, Barbara Pernici. "A Self-monitoring System to Satisfy Data Quality Requirements", OTM Conferences, pp. 1535-1552, 2005.