ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 4, April 2013

Web usage mining: Review on preprocessing of web log file

Sunita sharma
M.Tech., CSE Deptt.
Hindu College of Engg.
Sonepat, Haryana

Ashu bansal
A.P., CSE Deptt.
Hindu College of Engg.
Sonepat, Haryana

Abstract: Web log data is one of the major source which contain all the information regarding the users visited links browsing pattern, time spent on a particular page. A web server log file contain information about user name, ip address, date time, byte transferred, access request. A web log file range is 1kb to 100kb. Many interesting pattern available in web log data. But it is very complex process to extract the interesting pattern without preprocessed phase. Preprocessing step is used to give a reliable input for data mining task. In this paper we present the data preprocessing method for improving the efficiency & ease of mining process.

Keywords: Data preprocessing, Data cleaning ,web log file, web usage mining session identification, user identification.

I. INTRODUCTION

With the techniques of web log data mining we can find the law of of user accesing web pages and better the performance and organizational structure of website ,there by improving the quality and efficiency of user seeking information. several data mining methods are used to discover the hidden information in the web.web log data can provide useful information that helps a website engineer provide in enhancing the website structure in a way that will make the Website usage easier and faster. Web log file formats are usually designed for debugging purposes, therefore, web accesses are recorded in the order they come. Due to the stateless nature of the HTTP (i.e., each request is handled in a separate connection), web log records for a single user do not necessarily appear contiguously since they could be interleaved with records from other users. Thus, for each page component—such as an image, a cascading style sheet file, an HTML file, scripting file, or a Java script a separate record is recorded in the web log file. For web usage mining purposes, the only interesting elements are extract are from web log fille.

II. WEB USAGE MINING

The amount of data kept in computer files and data bases is growing at a phenomenal rate. At the same time users of these data are expecting more sophisticated information from them A marketing manager is no longer satisfied with the simple listing of marketing contacts but wants detailed information about customers' past purchases as well as prediction of future purchases. Simple structured / query language queries are not adequate to support increased demands for information. Data mining steps is to solve these needs. Data mining is defined as finding hidden information in a database alternatively it

has been exploraroty data analysis, data driven discovery and deductive learning.
web usage mining is the application of data mining techniques to discover usage pattern from web data. Web usage mining include three main steps : data preprocessing, pattern discovery and pattern analysis. In preprocessing phase data need to be cleaned and analyzed for pattern discovery phase. In preprocessing phase we get rid off irrelevant data from the web log file. The purpose of data preprocessing is to improve the data quality and increase mining accuracy. preprocessing converts the raw data into the data abstraction necessary for pattern discovery.
The following are the some web usage mining task:
(a) **Data preprocessing**: this step is used to extract useful information from web server log file. server log is examined to remove irrelevant items.
(b) **Pattern Discovery phase**: this phase is key component of the web usage mining. It converges the techniques from several research areas, such as data mining, machine learning and pattern recognition applied to the available data.
(c) **Pattern analysis**: This is the last phase of web usage mining process. This process involve the user evaluating each of pattern identified in the pattern discovery phase.
Web usage mining analyzes results of user interactions with a Web server, including Web logs or other database transactions for a web site or a group of related sites. It introduces privacy concerns and is currently the topic of extensive debate.It includes the data from the web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions cookies, user queries, book mark data, mouse click and scrolls, and any other data as the result of interactions. It aims at discovering general patterns in Web Access logs. In order to discover usage patterns from the available data, it is necessary to perform: Pre processing, Pattern Discovery, and Pattern Analysis. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web server logs, different web sites can help understand the user behavior and the web structure, thereby improving the data quality.

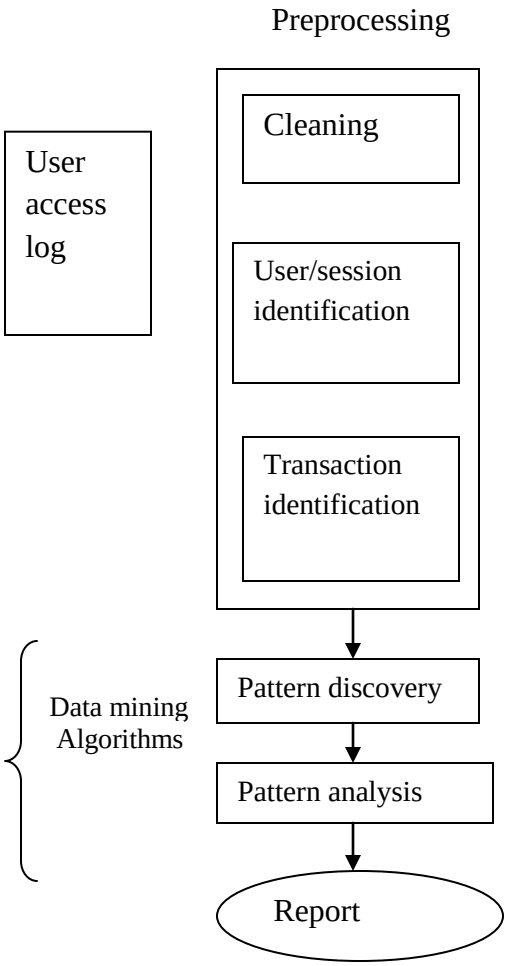

Figure 1. web usage mining system

Web mining is the use of data mining techniques to automatically discover and extract knowledge from web document .

web mining is the information service centre fornews, e-commerce, and advertisement, government, education, financial management, education, etc.

III. THE PROCESS OF DATA PREPROCESSING IN WEB LOG MINING

The purpose of preprocessing is to convert the Web log into reliable, complete and accurate data sources, to meet the needs of Web data mining algorithm implementation process needs. However, as Web log file format is different from the traditional sense of database or data warehouse data which has a good data structure, which is semi-structured, together with the existence of a variety of reasons lead to the data in the logs incomplete. All this has made the work of pretreatment face many technical problems. And data pre-processing has become the most difficult task in the Web data mining.

III. WEB LOG DATA PREPROCESSING PROCESS

Data preprocessing phase include three basics steps:

(a) Data cleaning: data cleaning focus on the getting rid of irrelevant & unimportant data from the web log server. A log file can provide useful information that helps a website engineer in enhancing the website structure in a way that will make the website usage easier and faster in future. This step consists of removing useless requests from the log files. Usually,this process removes requests concerning *nonanalyzed resources* such as images and multimedia files. Data cleaning also identifies *Web robots* and removes their requests.For Web portals and popular Web sites, log file size is measured in gigabytes per hour. For example, as of September 2000, Yahoo, then the most popular Web site according to Nielsen, had collected 48 Gbytes of log data for one hour, state Shahabi and Kashani. Manipulating such large files is complicated even with topnotch hardware tools. By filtering out useless data, we can reduce the log file size to use less storage space and to facilitate upcoming tasks. For example, by filtering out image requests, we reduced INRIA's Web server log files to approximately 40 to 50 percent of their original size.

(b)User Identification: user identification is identify each user from the weblog who access website.user identification group together the record for the same user from log records which are recorded in a sequential manner as they are coming from different user.

(c) Session identification: the task of session identification is to identify the sequences.session is the sequence of consective pages requested when the same user accessed a site.

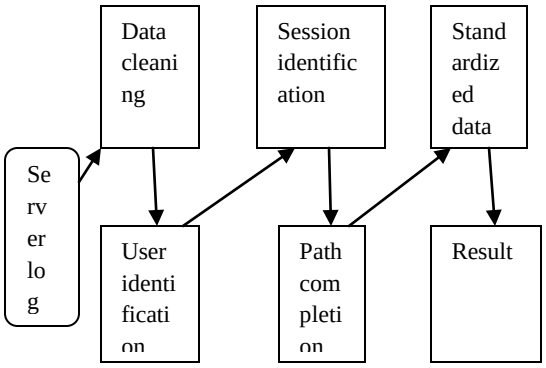

Figure 2: The process of data preprocessing

Data cleaning is mainly to delete items in the Web server logs without relation to data mining algorithm and to merge some records and to handle the records gracefully when the error page occurs.

Then processed data is imported into relational database for further data processing.

IV. CONCLUSION

Web log data is a collection of huge information. Many interesting patterns available in the web log data. But it is very complicated to extract the interesting patterns without preprocessing phase. Preprocessing phase helps to clean the records and discover the interesting user patterns and session construction. But understanding user's interest and their relationship in navigation is more important. For this along with statistical analysis data mining techniques is to be applied in web log data. Data preprocessing treatment system for web usage mining has been analyzed and implemented for log data.

Data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the failed HTTP status code and finally robots cleaning. Different from other implementations records are cleaned effectively by removing local and global noise and robot entries.

REFERENCES

[1] Ling Zheng, Hui Gui and Feng Li, "Optimized Data Preprocessing Technology for WebLog Mining", International Conference On Computer Design And Applications (ICCDA 2010) ,volume 1, pp. 19-22, 2010.

[2] Natheer Khasawneh,Chien-Chung Chan, "Active User-Based and Ontology-Based Web Log Data Preprocessing for Web usage mining", Proceeding of the 2006 IEEE/WIC/ACM International Conference,2006.

[3] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood, "WebUsageMining: A Survey on Preprocessing of Web Log File" , 2010,

[4]Sumian peng, qingqing cheng,"Reasearch on data preprocessing in web log mining", The 1st International Conference on Information Science and Engineering (ICISE2009) , 978-0-7695-3887-7/09/$26.00 ©2009 IEEE , pp. 942- 945 , 2009.

[5] P.Nithya, Dr.P.Sumathi ," Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots" ,2012 National Conference on Computing and Communication Systems (NCCCS), 978-1-4673-1953-9/12/$31.00 ©2012 IEEE, 2012.

[6] Brijendra Singh, Hemant Kumar Singh, "web data mining research: A survey" , 978-1-4244-5967-4/10/$26.00 ©2010 IEEE , 2010.

[7] P. Sampatht, C. Ramesht, T. Kalaiyarasit, S. Sumaiya Banut, G. Arul Selvan2, "An Efficient Weighted Rule Mining for Web Logs" , IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012.

[8] R.Manjusha , Dr.R.Ramachandran "Web Mining Framework for Security in E-commerce" IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011 .MIT, Anna University, Chennai. June 3-5, 2011.

[9] Sanjay Kumar Malik , SAM Rizvi "Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation " ,

2011 International Conference on Computational Intelligence and Communication Systems.

[10] Qingtian Han, Xiaoyan Gao, Wenguo Wu , "Study on Web Mining Algorithm Based on Usage Mining " , 978-1-4244-3291-2/08/$25.00©2008 IEEE.

[11] Yonghe Niu, Tong Zheng, Jiyang Chen, Randy Goebel , "Visualizing Structure and Navigation forWeb Mining Applications ", Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03) 0-7695-1932-6/03 $17.00 © 2003 IEEE.

[12] Hu0yuki KAWANO , "Web Archiving Strategies by using Web Mining Techniques ".

[13] I-Hsien Ting, "Web Mining Techniques for On-line Social Networks Analysis" 978-1-4244-1672-1/08/$25.00 ©2008 IEEE.

AUTHORS

Sunita Sharma was born in Haryana, India in1989. He received his Bachelor of Engineering Degree in C.S.E from Kurukshetra University, kururkshetra and Master of Engineering in Computer Science from DECRUST murthal, India in 2011 and 2013 respectively. Presently she is pusuing her M.tech from DECRUST murthal(final year). His research interests includes web usage mining, data mining.

Ashu Bansal was born in Haryana, India in 1982. He received his Bachelor of Engineering Degree in I.T. from M.D. University Rohtak and Master of Engineering in Computer Science from PEC Chandigarh, India in 2004 and 2007 respectively.Presently, he is working as a Faculty of the Department of CSE/IT at Hindu College of Engineering, Sonepat ,Haryana, India. He has Teaching experience of around 6 years. His research interests includes Artificial Intelligence, Logic and Reliability.