INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue2)
Available online at www.ijariit.com

A Novel Technique for Path Completion in Web Usage Mining

Padmapriya .R
Rathnavel Subramaniam College of Arts & Science, Sulur, Coimbatore - 402, TN, India
padmapriyaramesh86@gmail.com

D. Maheswari
Rathnavel Subramaniam College of Arts & Science, Sulur, Coimbatore - 402, TN, India
maheswari@rvsgroup.com

Abstract: World Wide Web is a huge repository of web pages and links. The Web mining field encompasses a wide array of issues, primarily aimed at deriving actionable knowledge from the Web, and includes researchers from information retrieval, database technologies, and artificial intelligence. The growth of web is tremendous as approximately one million pages are added daily. Users' accesses are recorded in web logs. Most data used for mining is collected from Web servers, clients, proxy servers, or server databases, all of which generate noisy data. Because Web mining is sensitive to noise, data cleaning methods are necessary. Web usage mining consists of three phases preprocessing, pattern discovery and pattern analysis. Web log data is usually noisy and ambiguous and data preprocessing system for web usage mining is an important process. A data preprocessing includes data cleaning, user identification, session identification and path completion. The inexact data in web access log are mainly caused by local caching and proxy servers which are used to improve performance and minimize network traffic. The proposed method uses path completion algorithm to preprocess the data. The proposed path completion algorithm efficiently appends the lost information and improves the consistency of access data for further web usage mining calculations.

Keywords: Web Mining, Data Preprocessing, Path Completion Algorithm, User Session Identification.

I.INTRODUCTION

Web mining is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces user's visiting behaviors and extracts their interests using patterns. Web mining method is one of the effective methods used in organizations to search the output from a large amount of the surface and WebPages from the hidden one. Many web mining algorithms are available to retrieve the WebPages. Web Usage Mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on. Web usage mining provides the support for the website design, providing personalization server and other business making a decision, etc.

The main parts of web usage mining are Data Preprocessing, Knowledge Extraction, and analysis of results. Data preprocessing includes data cleaning, user identification, session identification and path completion. Web Usage Mining consists of three main steps: data preprocessing, knowledge extraction, and results in analysis. Raw data is highly susceptible to noise, missing values. The quality of data affects the data mining results. In order to improve the data quality, that the data is preprocessed. Usage of data preprocessing deals with the preparation and transformation of the dataset. Data mining is one of the challenging to discover the large database. Data mining through these data preprocessing is increased and importance in industry.

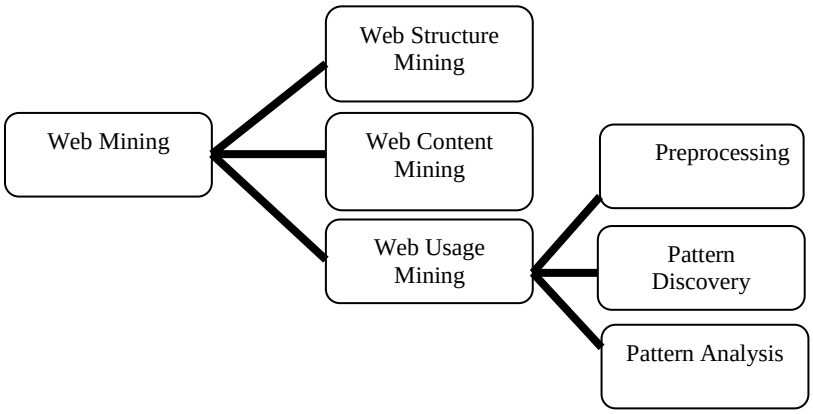

Figure 1: Web Usage Mining

Web usage mining is the main research area in Web mining focused on learning about Web users and their interactions with Web sites. The motive of mining is to find users' access models automatically and quickly from the vast Web log data, such as frequent access paths, frequent access page groups and user clustering. The proposed method uses path completion algorithm to preprocess the data. Then collect the data's from our college website and it is preprocessed based on the proposed method. The proposed path completion algorithm efficiently appends the lost information and improves the consistency of access data for further web usage mining calculations.

This remaining paper describes Literature Survey in Section II, Preprocessing methodology is discussed in Section III, Experiments and achieved results in Section IV. Finally, Conclusion of this work is given in Section V.

II. LITERATURE SURVEY

Analyzing the unique types of data that come from educational systems can help find the most effective structure of the e-learning courses, optimize the learning content, recommend the most suitable learning path based on student's behavior, or provide the more personalized environment. Munk and Martin (2011) focus only on the processes involved in the data preparation stage of web usage mining. Preprocessing is an important process which converts raw web log data into transactions. A new technique for identifying sessions is being proposed by Chitraa and Antony (2011) for extraction of user patterns. Log data is usually noisy and ambiguous and preprocessing is an important process for efficient mining process. Vellingiri and Chenthur Pandian (2011) focus on providing techniques for better data cleaning and transaction identification from the weblog. Web log data is one of the major sources which contain all the information regarding the users visited links, browsing patterns, time spent on a particular page or link and this information can be used in several applications like adaptive websites, modified services, customer summary, pre-fetching, generate attractive websites etc. Nithya and Sumathi (2012) continue the line of research on Web access log analysis is to analyze the patterns of website usage and the features of user's behavior.

Web Usage Mining (WUM) integrates the techniques of two popular research fields - Data Mining and the Internet. Wang et al (2015) introduce two prevalent data mining algorithms - FPgrowth and PrefixSpan into WUM and they are applied in a real business case. Web usage mining is the process of data mining techniques. Web usage mining consists following sections. 1) Pre-processing 2) Pattern discovery 3) Pattern Analysis. Mitharam (2012) describes the First phase in detail. The purpose of using web usage mining methods in the area of learning management systems is to reveal the knowledge hidden in the log files of their web and database servers. Munk and Martin (2011) help us to find the most effective structure of the e-learning courses, optimize the learning content, recommend the most suitable learning path based on student's behaviour, or provide the more personalized environment. Ketul and Patel (2012) discuss the process of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. Sisodia et al (2012) review the process of discovering useful patterns from the web server log file of an academic institute. Kotiyal et al (2013) focus on adopting an intelligent technique that can provide personalized web service for accessing related web pages more efficiently and effectively so that it can be determined which web pages are more likely to be accessed by the user in future. Web Usage Mining (WUM) is the application of data mining techniques to discover the knowledge hidden in the web log file, such as user access patterns from web data and for analyzing users' behavioral patterns. Pamutha et al (2012) focus on the preprocessing of the web log file methods that can be used for the task of session identification from web log file.

Weblogs take an important role to know about user behavior. Suguna and Sharmila (2013) discusses the basics of weblog preprocessing, existing preprocessing techniques, the proposed UILP algorithm, and performance of the proposed UILP algorithm with existing algorithms to identify user interest level. Dohare et al (2012) proposed a new reactive session reconstruction method. Securing e-commerce sites has become a necessity as they process critical and sensitive data to customers and organizations. Salama et al (2011) discuss how different web log files with different formats will be combined together in one unified format

using XML in order to track and extract more attacks. Web Mining is an area of Data Mining dealing with the extraction of interesting knowledge from the World Wide Web. Verma et al (2011) present a comprehensive survey of over 100 research papers dealing with Web Mining framework. Nowadays, user rely on the web for information, but the currently available search. A real problem before web master of a website is to match the user needs and keep their attention focused on his website. Bhushan and Rajender (2012) propose a web recommendation approach which recommends user a list of pages based upon user's historic pattern and a list of web pages which have not been visited yet.

III. RESEARCH METHODOLOGY

Preprocessing

Some databases are insufficient, inconsistent and including noise. The data preprocessing is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data preprocessing work, mainly include data cleaning, user identification, session identification and path completion. Data preprocessing techniques can improve the quality of the data, thereby help to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, because quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making.

Path completion

Another critical step in data preprocessing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. Using the local caching and proxy servers also produce the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time. It is especially important to web log files, in respect that the structure of web log files are not the same as the data in database or data warehouse. They are not structured and complete due to various causations. So it is especially necessary to pre-process web log files in web usage mining. Through data pre-processing, weblog can be transformed into another data structure, which is easy to be mined.

IV. EXPERIMENTAL RESULTS

The experiments are conducted in the proposed technique by using the log obtained from Google website during Nov. 1st, 2012 to Nov. 5th, 2012. The obtained record consists of 1000 records in the log file. Initially, the data cleaning process is carried out and 520 records are obtained. Remaining 480 records are eliminated from log file. Users are identified according to the IP address, operating systems, and browsers. Then the algorithm in the proposed method is used to carry out the user access path completion. The accuracy of the results depend on the size of web access log to some extent, usually the weblog of longer period results in a more accurate result. The access path of one user session before path completion is shown in Table 2 as an example. The corresponding result of path completion is given in Table 3. The result shows that the user access path can be efficiently acquired by using the algorithms proposed in this study if necessary referrer information is available in the web access log.

Table 1: The Access Path of One User Session

Date	Access page number	Referrer number
2012-9-2 13:45:05	16	
2012-9-2 13:45:07	17	16
2012-9-2 13:45:10	18	17
2012-9-2 13:45:13	19	20
2012-9-2 13:45:13	20	15
2012-9-2 13:45:15	11	25
2012-9-2 13:45:18	26	45
2012-9-2 13:45:22	26	22
2012-9-2 13:45:24	41	25
2012-9-2 13:45:28	25	26
2012-9-2 13:45:30	30	35

Table 2: The Path Completion Result

	User's information: 116.128.56.89 Google chrome
Page sequence	16-17-18-19-20-11-26-26-41-25-30
Combination	16-17-18-19-20-11-26-41-25-30
Path completion	16-17-18-17-18-19-20-25-26-30-35

The access path of one user session before path completion is shown in Table 1 as an example. The corresponding result of path completion is given in Table 2. The result shows that the user access path can be efficiently acquired by using the algorithms proposed in this study if necessary referrer information is available in the web access log.

CONCLUSION

An important task in data mining application is the creation of a suitable data set to which mining and algorithms can be applied. This is an important activity in web usage mining due to the various characteristic features of the click stream data. The data preparation process is the most time-consuming and intensive step in mining web usage data. This work has presented various details about data preprocessing activities that are necessary to perform Web Usage Mining. In every phase of the data preprocessing, we give some rules to design and implement them easily and efficiently. Proposed method is used to reduce the size of the log file but also increases the quality of the data available. The path completion process which is used to append lost pages and construction of transactions in preprocessing stage

REFERENCES

1. Munk, Michal, and Martin Drlík. "Impact of Different pre-processing tasks on effective identification of users' behavioral patterns in the web-based educational system." *Procedia Computer Science* 4 (2011): 1640-1649.
2. Chitraa, V., and Antony Selvadoss Thanamani. "A novel technique for session's identification in web usage mining preprocessing." *International Journal of Computer Applications* 34, no. 9 (2011): 23-27.
3. Nithya, P., and P. Sumathi. "Novel pre-processing technique for web log mining by
4. Removing global noise and web robots." In 2012 NATIONAL CONFERENCE ON COMPUTING AND COMMUNICATION SYSTEMS. 2012.
5. Vellingiri, J., and S. Chenthur Pandian. "A novel technique for web log mining with better data cleaning and transaction identification." Journal of Computer Science 7, no. 5 (2011): 683.
6. Wang, Hengshan, Cheng Yang, and Hua Zeng. "Design and Implementation of a Web Usage Mining Model Based On Upgrowth and Preflxspan."*Communications of the IIMA* 6, no. 2 (2015): 10.
7. Mitharam, Marathe Dagadu. "Preprocessing in Web Usage mining."International Journal of Scientific & Engineering Research 3, no. 2 (2012): 1.
8. Munk, Michal, and Martin Drlik. "Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining." In*Digital Information and Communication Technology and Its Applications*, pp. 60-74. Springer Berlin Heidelberg, 2011.
9. Patel, Ketul B., and A. R. Patel. "Process of Web Usage Mining to find Interesting Patterns from Web Usage Data." *International Journal of Computer Applications & Technology* 3, no. 1 (2012): 144-148.
10. Sisodia, Dilip Singh, and Shrish Verma. "Web usage pattern analysis through weblogs: A review." In *Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on*, pp. 49-53. IEEE, 2012.
11. Kotiyal, Bina, Ankit Kumar, Bhaskar Pant, R. H. Goudar, Shivali Chauhan, and Sonam Junee. "User behavior analysis in weblog through a comparative study of eclat and apriori." In *Intelligent Systems and Control (ISCO), 2013 7th International Conference on*, pp. 421-426. IEEE, 2013.
12. Pamutha, Thanakorn, Siriporn Chimphlee, Chom Kimpan, and Parinya Sanguansat. "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns." *International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol* 2 (2012).
13. Suguna, R., and D. Sharmila. "User interest level based preprocessing algorithms using web usage mining." *International Journal of Computer Science and Engineering* 5, no. 9 (2013): 815-822.
14. Dohare, Mahendra Pratap Singh, Premnarayan Arya, and Aruna Bajpai. "Novel Web Usage Mining for Web Mining Techniques." *International Journal of Emerging Technology and Advanced Engineering* 2, no. 1 (2012): 253-262.
15. Salama, Shaimaa Ezzat, Mohamed I. Marie, Laila M. El-Fangary, and Yehia K. Helmy. "Web Server Logs Preprocessing for Web Intrusion Detection."*Computer and Information Science* 4, no. 4 (2011): p123.
16. Verma, Vikas, A. K. Verma, and S. S. Bhatia. "Comprehensive survey of a framework for web personalization using web mining." *International Journal of Computer Applications* 35, no. 3 (2011): 23-28.
17. Bhushan, Ravi, and Rajender Nath. "Automatic recommendation of web pages for online users using web usage mining." In *Computing Sciences (ICCS), 2012 International Conference on*, pp. 371-374. IEEE, 2012.

