ISSN 2250-3153

Clustering of Web Log Data to Analyze User Navigation Patterns

Aditi Shrivastava

Shri Ram Institute of Technology
Jabalpur (M.P), India

Abstract- As we know the amount of data available online is increasing day by day, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. Application of web usage data can be used to better understand web usage, and apply this specific knowledge to better serve users. Web usage mining is the base for navigation pattern mining and approach of clustering is used to perform that Mining, usage mining deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications. A web navigation behavior is helpful in understanding what information of online users demand. In our study we extract the common pattern and do clustering, following that, the analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps, web based personalization system and improving searching accuracy when seeking. The experimental results shows the clusters of navigation patterns of user and also the approach can improve the quality of clustering for user navigation pattern in web usage mining systems. These results can be used for predicting user's intuition in the large web sites.

Index Terms- web usage mining, pre-processing, pattern discovery, pattern analysis, navigation pattern mining, clustering

I. INTRODUCTION

The web is vast, varied and dynamic and thus raises the scalability, multimedia data and temporal issues respectively. The expansion of the web has resulted in a large quantity of data that is now freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users effectively and efficiently. Therefore, the application of data mining techniques on the web is now on the focus of an increasing number of researchers. The Web Mining is the set of techniques of Data Mining applied to extract useful knowledge and implicit information from Web data. As more organizations rely on the Internet to conduct daily business, the study of Web mining techniques to discover useful knowledge has become increasingly important. However, with the magnitude and diversity of available information from the Internet, it is not insignificant to locate the relevant information to satisfy the requirements of people with different backgrounds. Web mining enables one to discover web pages, text documents, multimedia files, images and other types of resources from web generally; data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

A. Web Mining

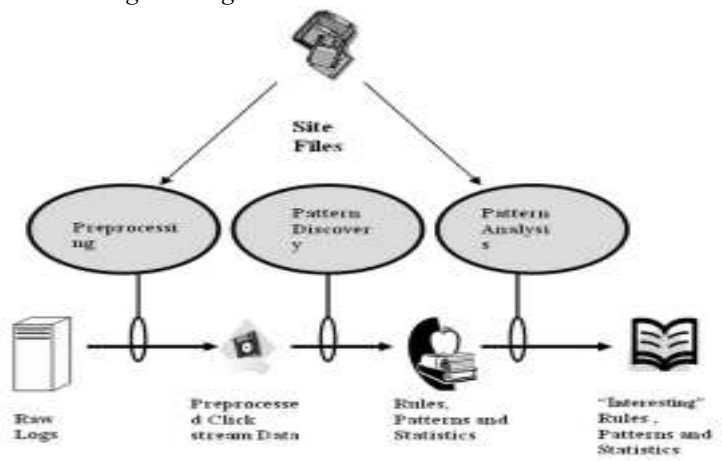
Web mining can be categorized into three different classes based on which part of the Web is to be mined. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining. Web Usage Mining (WUM) is the process of extracting knowledge from Web user's access data by exploiting Data Mining technologies. It can be used for different purposes such as personalization, system improvement and site modification. Web usage mining tries to make sense of the data generated by the web surfer's session or behaviors. While the web content and structure mining utilize the real or primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web.

B. Web Usage Mining

In our study we will give emphasize on Web usage mining. As Now a day the web is not the place where only transaction has occurred. Millions of visitors interact with the web in daily life which generates an enormous amount of data. Web usage mining helps to know information about users' behaviors and their usage patterns, can lead to interesting results that go over. Web Usage Mining (WUM) is the automatic discovery of user access pattern from web servers. Organizations collect large volumes of data in their daily operations, generated automatically by web servers and collected in server access logs. It can also provide information on how to restructure a website to service effectively. The Web Usage mining 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, bookmark data, mouse clicks and scrolls, and any other data as the results of interactions. Web Usage Mining is the discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities. Web usage mining focuses on techniques that could predict user behavior while the user interacts with the web. As mentioned before the mined data in this category are the secondary data on the web as the result of interaction. These data could range very widely but generally it is classified into usage data that resides in the web client, proxy server and servers. The aim of understanding the navigation preferences of the visitors is to enhance the quality of

electronic commerce services ecommerce, to personalize the Web portals or to improve the Web structure.

C. Web Usage Mining Architecture



1) Pre-processing

Pre-processing "consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery". This step can break into at least four sub steps: Data Cleaning, User Identification, Session Identification and Formatting. Unneeded data will be deleted from raw data in web log files in the data cleaning step. At least two log file formats exists: Common Log File format (CLF) and Extended Log File

2) Pattern Discovery

After data preparation phase, the pattern discovery method should be applied. This phase consists of different techniques derived from various fields such as statistics, machine learning, data mining, pattern recognition, etc. applied to the Web domain and to the available data. The task for discovering the patterns offer some techniques as statistical analysis, association rules, sequential pattern analysis, clustering and so on. Here we will briefly describe some techniques to discover patterns from processed data.

- Statistical Analysis such as frequency analysis, mean, median, etc.
- Clustering of users help to discover groups of users with similar navigation patterns (provide personalized Web content).
- Classification is the technique to map a data item into one of several predefined classes.
- Association Rules discover correlations among pages accessed together by a client.
- Sequential Patterns extract frequently occurring inter-session patterns such that the presence of a set of items s followed by another item in time order.
- Dependency Modelling determines if there are any significant dependencies among the variables in the Web.

3) Pattern Analysis

Pattern Analysis is the final stage of WUM (Web Usage Mining), which involves the validation and interpretation of the mined pattern.

Validation: To eliminate the irrelevant rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process.

Interpretation: The output of mining algorithms is mainly in mathematic form and not suitable for direct human interpretations

II. PROBLEM STATEMENT

The proposed system aims at presenting a useful information extraction system from web access log files of web servers and using them to achieve clusters from related pages in order to help web users in their web navigation. In our work we prepare a tool to help the website owners to keep the record of their visitors. This will be helpful as they will be able to know how many times the website is being visited and who all their visitors are. The information will be beneficial as they can customize the website and the contents according to the people visiting it. It will also be commercially able to benefit as it will help them plan their advertisements. Before mining, we need to gather the Web document together. Secondly, Web pages are semi-structured, in order for easy processing; documents should be extracted and represented into some format. The main objective of our work is basically clustering of common browsing patterns these records are in the form of pattern they follow to leaf through the site. In our work the common pattern will be the path the user navigates through the site.

III. DESIGN CONSIDERATION

Various steps are involved in identifying the clusters of user navigation pattern are shown below: first data is collected which is known as web logs then preprocessing techniques are applied to that web log so that we can get relevant information Generally, several pretreatment tasks need to be done before performing web mining algorithms on the Web server logs. Data pretreatment in a web usage mining model (Web-Log preprocessing) aims to reformat the original web logs to identify all web access sessions. The Web server usually registers all users' access activities of the website as Web server logs.

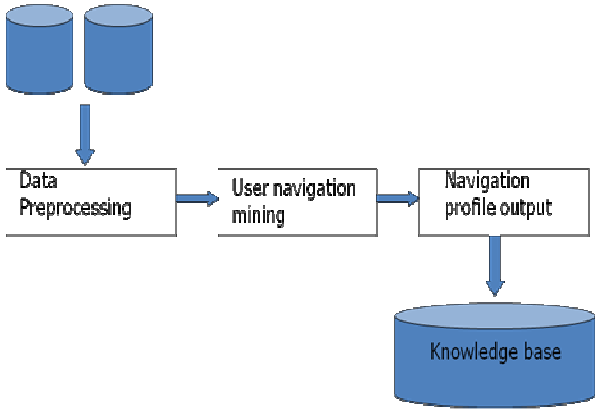


Fig. 1 Model of the System

A. Web Logs

Data sets consisting of web log records collected from own fashioned website. Web log is unprocessed text file which is

recorded from the Server. Web log consist of various attributes with the data values in the form of records

This log records restrain session id, login date, username user IP address and email id.

B. Data Preprocessing

Generally, several preprocessing tasks need to be done before performing web mining algorithms on the Web server logs. Data preprocessing a web usage mining model (Web-Log preprocessing) aims to reformat the original web logs to identify user's access sessions. This is done by following methods which was proposed in [1].

1. Data collection: This is done mostly by the web servers; however there exist methods, where client side data are collected as well.
2. Data cleaning: As in all knowledge discovery processes, in web usage mining can also be happen that such data is recorded in the log file that is not useful for the further process, or even misleading or faulty. These records have to be corrected or removed.
3. User identification: In this step the unique users are distinguished, and as a result, the different users are identified. This can be done in various ways like using IP addresses, cookies, direct authentication and so on.
4. Session identification: A session is understood as a sequence of activities performed by a user when he is navigating through a given site. To identify the sessions from the raw data is a complex step, because the server logs do not always contain all the information needed. There are Web server logs that do not contain enough information to reconstruct the user sessions; in this case (for example time-oriented or structure-oriented) heuristics can be used.

C. Navigation Pattern Mining

After the data pre-processing step, we perform navigation pattern mining on the derived user access session. As an important operation of navigation pattern mining, clustering aims to group sessions into clusters based on their common properties. User navigation patterns are described as the common browsing behaviors among a group of users. Since many users may have common interests up to a point during their Navigation, navigation patterns should capture the overlapping interests or the information needs of these users. In addition, navigation patterns should also be capable to distinguish among web pages based on their different significance to each pattern We extract the common pattern and do clustering while the user browses the site and apply it to knowledge base. As navigation output clusters of patterns are made which is done by observing common pattern. In our work common browsing behavior is the path in which user trace the site.

D. Knowledge Base

Following that, the analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps, web based personalization system and improving searching accuracy when seeking. To analyze the path the administrator can understand what pages the visitors like most or how long path they like to visit in a web site. The algorithm which we use is as follows.

1. Input: Set of web page P={ X_1 , X_2 ,…,…, X_n}
2. Output: Web page Clusters
3. Begin

4. For i=1 to m do
5. $X_i \rightarrow Pattern_i$
6. cluster=m;
7. while cluster > τ do
8. Begin
9. For i=1 to cluster do
10. Begin
11. Find Two Cluster With max similarity();
12. if Condition( ) then
13. Begin
14. CombineTwoCluster( );
15. cluster --;
16. End
17. End
18. End
19. End

IV. EXPERIMENTAL EVALUATION

Measuring the quality of the clustering in navigation patterns mining systems needs to characterize the quality of the results obtained. The experimental evaluation was conducted using Log file of our own adaptive website The only cleaning step performed on this data was the removal of references to auxiliary files (e.g., image files). No other cleaning or preprocessing has been performed in the first phase. The data is in the original log format.

Step wise results are shown below for records

Fig 1shows the different paths that user follows while navigate through the site with unique session id. Cluster patterns of similar path each time user login the site with the unique session id.

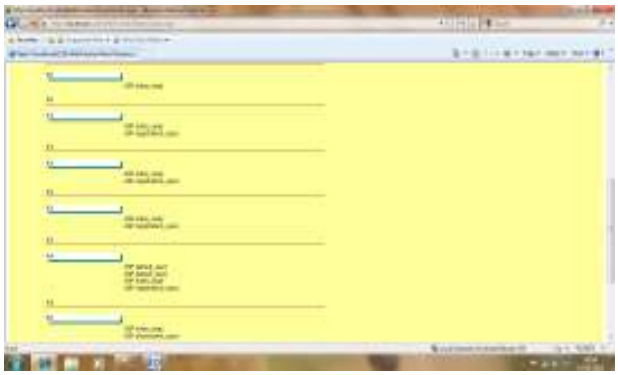


Fig 1: Navigation patterns

Fig 2 screen shows the output after we apply the clustering technique. Now the clusters of navigation patterns are made by which we get the number of users who follows similar type of pattern while browsing the site.

Fig 2: Cluster of Patterns

V. CONCLUSION

Web usage and data mining to find patterns is a growing area with the growth of Web-based applications Application of web usage data can be used to better understand web usage, and apply this specific knowledge to better serve users Web usage mining is the base for navigation pattern mining and approach of clustering is used to perform Navigation Pattern Mining. The web is a most important medium to conduct business and commerce. Therefore the design of web pages is very important for the system administrator and web designers. These features have great impact on the number of visitors. So the web analyzer has to analyze with the data of server log file for detecting pattern. In this paper we tried to give a clear understanding of the web usage mining and clustering of patterns formed while navigation. By this we can shorten the pages which are not in the user pattern, also we can record the information of the user, This will also help in evaluating address campaigns, restructuring and redesigning of website also it will remove the page which is in the category of low access or it can merge into the page which is frequent access Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

REFERENCES

[1] Pattern Discovery of Web Usage Mining Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul Islam Bhuiyan, Khandakar Entenam Unayes Ahmed Computer Science & Engineering Shahjalal University of Science & Technology Sylhet, Bangladesh 978-0-7695-3892-1/09 $26.00 © 2009 IEEE DOI 10.1109/ICCTD.2009.199

[2] International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 279-283 AN ALGORITHMIC APPROACH TO DATA PREPROCESSING IN WEB USAGE MINING

[3] R. Baraglia and F. Silvestri, "Dynamic personalization of web sites without user intervention," Communications of the ACM,vol. 50, pp.63-67, 2007.

[4] A New Clustering Approach based on Graph Partitioning for Navigation Patterns Mining978-1-4244-2175-6/08/$25.00 ©2008 IEEE

[5] Journal of Theoretical and Applied Information Technology© 2005 - 2008 JATIT. All rights reserved.www.jatit.org1125WEB USER NAVIGATION PATTERN MINING APPROACH BASED ONGRAPH PARTITIONING ALGORITHM

[6] Jalali, M., Mustapha, N., Sulaiman, N.B. and Mamat, A. (2008c) A web usage mining approach based on LCS algorithm in online predicting recommendation systems, 12th International Conference Information

[7] "An Online Recommender System for Large Web Sites,"Proceedings of the Web Intelligence, IEEE/WIC/ACM International Conference on (WI'04)-Volume 00, pp. 199-205, 2004.

[8] Expectation Maximization Clustering Algorithm for User Modeling in Web Usage Mining Systems European Journal of Scientific Research ISSN 1450-216X Vol.32 No.4 (2009), pp.467-476

[9] Classification and Clustering of Web Log Data to Analyze User Navigation Patterns Volume 1, No. 1, August 2010_Journal of Global Research in Computer Science

[10] Jaideep Srivastava , Robert Cooley, Mukund Deshpande, Pang-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDDExplorations, Volume 1, Issue 2- Pages 12-23.[2].

[11] Adel T. Rahmani and B. Hoda Helmi, EIN-WUM an AIS-based Algorithm for WebUsage Mining, Proceedings of GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA, ACM978-1-60558-130-9/08/07 (Pages 291-292)

[12] www.google.com

[13] www.wikipedia.com

[14] www.webmining.com

First Author – Aditi Shrivastava ,ME(C.S.E) IV SEM, Shriram Institute of Engineering and Technology, Jabalpur(M.P) India
Email – aditi.aditi25@gmail.com