Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 3, March -2015

SURVEY ON WEB USAGE MINING AND PRE-FETCHING

Devendra Parmar

CSE, Parul Institute of engineering and Technology

Abstract *- Web Mining plays a vital role in research area in the field of data mining. Hence, in this study some algorithms are presented which can be used according to one's requirement. Apriori and FP tree are the most popular algorithm used for the purpose of mining data. Apriori is not very effective in many cases because candidate key generation is costly. FP Tree is also used but it generates explosive quantity lacks of good candidate generation method. In this paper one way is to use Pre-fetching with web usage mining that makes the performance of proxy server improved in terms of response.*

I. INTRODUCTION

Data Mining is a process of identifying valid, useful, novel, understandable pattern in the data. Web mining is a process of mining the data from the large amount of data that is on web pages. It helps users to understand the nature and behavior of the users that are accessing the web. Web mining is a cross point of database, information retrieval and artificial intelligence. There are three sub types of web mining.

1) Web Usage Mining
2) Web Content Mining
3) Web Structure Mining

Web usage mining refers to the mining for the usage access and log information for the user. Web content mining refers for the content that belongs to web pages, which is mined in this web content mining. Web structure mining is mining for the web structure; this web structure is given by the hyperlinks to the web pages that are on a particular web page. In web structure mining to page ranking algorithm is widely used that provides page ranking for calculation of hyperlinks that are for a particular page. In this text mining is also includes because of availability of large amount of text data on web.

The Web has growing rapidly from a simple information-sharing mechanism offering only static text and images to a rich collection of dynamic and interactive services. The explosive growth of the web has compulsory a heavy demand on networking resources and web servers. Users often experience long and unpredictable delays when retrieving web pages from remote sites. Therefore, a clear solution to improve the class of web services would be the increase of bandwidth, but such type of solution increases the cost of system. However, higher bandwidth can resolve for the time being the problems since it would no difficulty the user to create more and more resource-greedy applications for the network. In many research, researcher proved and suggest that a cache-based approach is very helpful to improve the performance of the web for lower cost. We notice that a single user again and again requests the same web object many times during a small interval of time and web object access is non-uniform over all web servers. Additional, different users many times request the same web object. If we can store commonly requested objects closer to web clients, users should see lower latency when browsing. Web caches are the systems that keep copies of frequently accessed objects close to clients. The development of web caching has spurned new research in many areas.

Web page access prediction gained its importance from the ever increasing number of e-commerce and e-businesses. It involves personalizing, marketing, recommendations, helps in improving the web site structure and also guides web users in navigating through hyperlinks for accessing the information they need. In the proposed work, pre-fetching is done on the basic of proxy logs as it is the key requirement to make available user with best recommendations. Web log data is preprocessing for pattern discovery. We integrate the following two techniques together i.e. clustering and association using frequency support pruning, it achieves complete logs, better accuracy, less state space complexity and less number of rules. The predicted pages are pre-fetched and keep it in proxy server cache which reduce the accessing time of that page and increases the web proxy server performance.

II. WEB USAGE MINING

Web usage mining is the type of web mining activity that involves the automatic discovery of user access patterns from one or more web servers. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by web servers and collected in server access logs. Other sources of user information include referrer logs, which contain information about the referring pages for each page reference, and user registration or survey data gathered via tools such

as CGI scripts.

Day by day amount of data on web is increasing. For the purpose of getting particular data it is necessary to reduce the amount of that data. In web usage mining, apriori and FP Growth algorithms are used for pattern discovery that are giving meaningful pattern as a result. This algorithms works better in their environments, apriori needs candidate key for the pattern discovery and FP Growth has lack of quality of good candidate generation method.

An improved algorithm is much better than both of the above algorithms. This algorithms provides access pattern of the users for web means what web page is accessed and how much time that is accessed. According to this data one can restructure or redesign the websites for purpose of large access. In web mining another topics are for content mining and structure mining. Web content mining and web structure mining are relates with content of the web page and hyperlinks respectively. Web structure mining uses page ranking algorithm for ranking pages but is far away from my work. Web content mining focuses on text mining more because of web page contains mostly text documents.

Frequent patterns that are mined by algorithms can show the frequent occurrences of a particular task or in web mining frequent access for particular page or other. By finding this relation one can easily predict the importance for any content, file, document, webpage or link. This will help for kind of analysis for web mining.

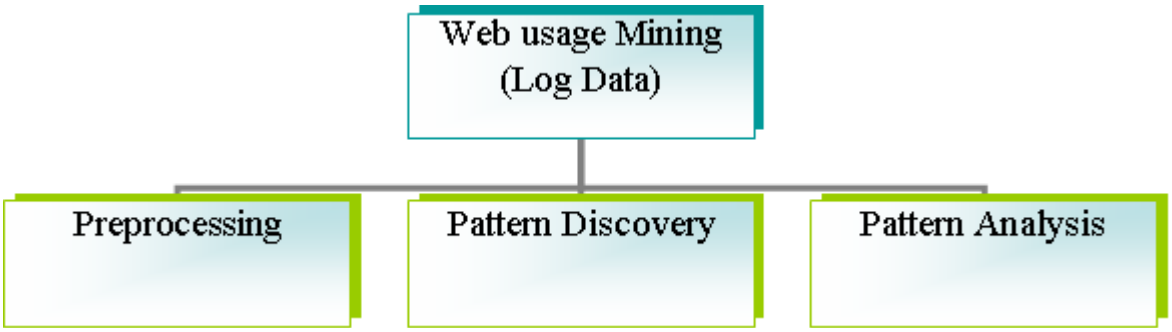

Figure 2.1 Web Usage Mining Steps

Web data being mined is by use of web log files that is web server log file, client side log file or proxy server log file. Preprocessing step in that data cleaning, user identification and session identification is done. Another steps are in data preprocessing are path completion and transaction. Transaction refers to grouping of set of operations which are atomic, logically identical and which are performed and recorded over certain period of time[6]. The next step is pattern discovery in that method of the above is applied for the pattern extraction like clustering, association rule mining, sequential pattern mining, classification or prediction. After this step the resultant output from this pattern discovery step is pattern that is analyzed in pattern analysis step. Various OLAP operations are performed on that. Paper discusses the importance of the web and the web data. For this web data provided methods and processes are necessary and important one for us.

This paper gives introduction about web mining and its sub types web usage mining, web structure mining, and web content mining.

It mainly discuss about web log files and information extraction from that log file. This pattern discovery is done in three steps pre-processing, pattern discovery, pattern analysis. Also discuss various techniques for data usage mining like association rule mining, sequential pattern mining, clustering and classification. At the end also shows the applications of the web usage mining. This paper also includes the basics of the web usage mining and steps in web usage mining like preprocessing, pattern discovery and pattern analysis.

Pattern discovery for web usage mining, includes topics for web log files, web usage mining, data preparation, pattern discovery, and pattern analysis. It also shows the need for the data preparation and its steps like data cleaning.

Study provides the information for user and session identification. It also gives algorithms for user identification and session identification. It also provides algorithm for data preparation.

Preprocessing step in web usage mining performed on the log file. The study says that the log file contains noisy data also. So it is vital step to remove that noisy data.

The study shows the importance of that preprocessing step of the information extraction steps. It also discusses about the user identification and session identification for the data. In this pre-processing data cleaning, user identification, session identification and data formatting and data summarization is done.

The study of that all three algorithms and compares the results of that three algorithms. Candidate key generation is costly in apriori algorithm if large number of pattern or long pattern exists. Here FP Growth algorithm having drawback that lack quality of good candidate generation method. The result of the comparison shows that apriori has drawback that candidate key generation is costly while large number of pattern exists. And the FP Growth algorithm has a drawback that lacks of good candidate generation method. The comparison is done based on different types of databases. The result shows that FP Growth works better than Apriori in terms of execution time.

III. COMPARISON

Algorithm	Parameter	Advantage/Disadvantage
Apriori	Use candidate key	Needs candidate generation that is costly
FP-Growth	Tree structure	No needs for candidate generation Better and fast than Apriori
Improved FP-Growth	Tree structure	fast and better than both of above uses URI from log for finding access pattern

IV. CONCLUSION

Frequent itemsets and web usage mining generally serve as building blocks for various patterns in many real-life applications. User perceived latency's from several sources such as bandwidth, speed, overhead, accessing the web page etc. Most of the existing algorithms find unconstrained frequent itemsets from traditional static transaction databases consisting of precise data. However, there are situations in which ones are uncertain about the contents of transactions. There are also situations in which users are only interested in some subsets of all the mined frequent itemsets. Furthermore, a flood of data can be easily produced in many situations.
If we use the prefetching with caching then the performance of cache is improved. Prefetching fetches objects that are likely to be accessed in the near future and store them in advance thus the response time of the user request is reduce.

REFERENCES

[1] Ashika Gupta, Rakhi Arora, Ranjana Sikarwar, neha Saxena "Web Usage Mining Using Improved Frequent Pattern Tree Algorithms" IEEE 2014
[2] Mr.Rajesh K. Ahir, Ms. Mital B. Ahir "Algorithms For Mining Frequent Patterns: A Comparative Study" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013
[3] Ankita Kusmakar, Sadhna Mishra "Web Usage Mining: A Survey On Pattern Extraction From Web Logs" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 9, September 2013
[4] Ramya C. Shreedhara K. S., And Kavitha G "Preprocessing: A Prerequisite For Discovering Patterns In Web Usage Mining Process" International Journal of Information and Electronics Engineering, Vol. 3, No. 2, March 2013
[5] Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi "Overview Of Web Content Mining Tools " IJES 2013
[6] Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod "Web Usage Mining: A Review on Process, Methods and Techniques" IEEE, 2012.
[7] Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul Islam Bhuiyan, Khandakar Entenam Unayes Ahmed "Pattern Discovery of Web Usage Mining" IEEE, 2009.
[8] Greeshma G., Vijayan, Jayasudha J. S. "A Survey On Web Pre-Fetching And Web Caching Techniques In A Mobile Environment" AIRCCJ, 2011.
[9] Sandhaya Gawade, Hitesh Gupta "Review of Algorithms for Web Pre-fetching and Caching" IJJARCCE-ISSN 2278 – 1021, Vol. 1, Issue 2, April 2012

