ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 4 Issue 3; May-June-2017; Page No. 65-67

Survey on Web Usage Mining using Association Rule Mining

Shanthi S

Asst. Professor, Department of Commerce PA,

KG College of Arts & Science, Coimbatore-35.

shanthisenthilmurugan.s@gmail.com

Received 02 April 2017; Accepted 04 May 2017

ABSTRACT

Web mining is a data mining technique to extract information from web documents. Web usage mining is a type of web mining and widely used in e-commerce applications to understand the behavior of the consumers. It is used to mine the information from the server log files. Server log files are automatically created and maintained by a server consists of pages requested URL and user information. Dramatic increase in the number of online shopping websites has increased the research on understanding customers preferences. Hence, various algorithms are used in mining the information from the server log files and one of the most extensively used algorithm is the association rule algorithms. This survey paper is about the recent researches and developments in the field of web mining using association rule algorithm.

Keywords: Web mining, Web usage mining, Association rule minin

1. INTRODUCTION

The proliferated growth of web users in today's world has created a new path in the field of research. Web mining is nothing but an old wine in a new bottle. A lot of research works has been done using data mining techniques in Information retrieval, pattern mining and pattern analysis. Web mining also falls under the same category where the mining is done on Web usage, Web structure and Web content by retrieving the useful information from the log files. Recent researches are focused highly on Web mining because of its multi-disciplinary nature and due to the increase of e-commerce applications. Basically, Web mining tasks are classified in to three categories namely Content mining, Structure Mining and Usage mining. Association rule learning is used for discovering interesting relations between variables in large dataset. Many algorithms are used for generating association rules and they are Apriori, Eclat, FP-Growth, Context based association rule mining, node set based algorithms, etc. This survey papers is mainly focused on the Association rule mining in Web usage mining. The paper is organized as follows: Section 2 is about the Web Usage Mining, Section 3 is about the related research works, Section 4 is about the limitations of the existing work and section 5 is about the Scope of further work.

2. WEB USAGE MINING

Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the Web. The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (e-commerce), to personalize the Web portals [1] or to improve the Web structure and Web server performance

Web usage mining consists of three phases namely data processing, pattern discovery, and pattern analysis.

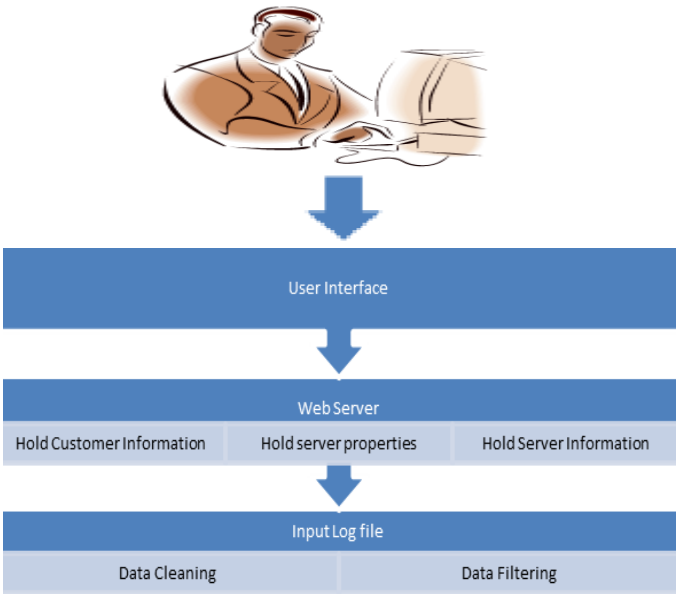

Figure 1 : Preliminary steps in Pre-processing

Corresponding author: Shanthi S |

Several algorithms are used for Data pre-processing, Knowledge extraction and analysis. Fig1. represents the preliminary steps followed in the pre-processing. Quality of the analysed report from the web data is totally dependent on the quality of data used in the mining process. When a user tries to access a web site using a URL from a web browser like IE, Mozilla, Google Chrome or Opera, the information related to the operation is recorded in a access log file stored in the web server which will contain IP address, Access date and time, URL of the page, Transfer protocol, Success of return code, Number of bytes transmitted. Webserver contains numerous log files and applying associate rule in the entire dataset is cumbersome. Hence, data pre-processing has to be carried out by cleaning the data. This can be achieved through filling the missing values, identifying the outliers, clearing the noisy data and correcting the inconsistent data. The next phase is the pattern discovery where various patterns can be discovered using Association, Clustering and Sequential Analysis. Last phase of the Web usage mining is the Pattern Analysis. The result of analysing the web log files in the server is to identify the frequently access web sites from the client side and maintain those information in the cache to increase the efficiency of the web pages.

Table 1 : Sample Log file information.

S. No	Fields
1	IP Address of the userm
2	Access Date and Time
3	URL of the page
4	Transfer Protocol
5	Success of return code
6	Request Method
7	Number of bytes transmitted

3. *RELATED WORKS*

Interesting research works have been done in discovering useful access patterns using the server log files.

Masseglia, Poncelet and Ciccheti [2] have proposed a very efficient algorithm for the “market basket” with the problem of web mining. Information from Web servers are analyzed to find the relationships such as: 60 % of clients who visited /jdk1.1.6/docs/ api/ package-java.io.html and /jdk.1.6/docs/api/ java.io.BufferedWriter.html in the same transaction, also accessed /jdk1.6/docs/ relnotes/ deprecated list. html within 20th September and the 30th October. They have described an efficient algorithm for finding all frequent user access patterns from one or more web servers. They have described an algorithm for finding all frequent access patterns from one or more web servers. The algorithm was based on a new prefix tree structure wich is very adequate.

Veeramalai, Jaisankar and Kannan [3] analyzed the patterns using different algorithms like apriori, hash tree and fuzzy. They have also proposed an enhanced apropri algorithm to give solution for crisp boundary problem with higher optimized efficiency while comparing to other algorithms. Web usage mining are the reconstruction of user sessions by using heuristics techniques and discovering useful patterns from these sessions by using pattern discovery techniques like association rule mining, Apriori. Proposed algorithm is an integrated system (Web Tool) for applying data mining techniques such as association rules or sequential patterns on access log files.

Mishra and Choubey [4] in their research have used FP-algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the users interest. The algorithm was tested on available log files on HTTP requests to the NASA Kennedy Space Center WWW server in Florida. The log was collected from 00:00:00 July 1, 1995 through 23:59:59 July 31, 1995, a total of 31 days. Now to extract the information such as requested files and most frequently accessed files, log file has been analysed.

The strings extracted from the log files are represented by unique index id. Similar indexing operation has been done for the gif files available in the web server log data and then applying the FP-growth algorithm to obtain various results such as most frequently visited pages, Top downloaded Pages from the web site, Top downloaded gif files and most frequently downloaded gif files from the web server log data. By using the concept of web usage mining we can easily find out the user’s interest and we can modify and make our web site more valuable and more easily accessible for the users.The main goal of the proposed system is to identify usage pattern from web log files.

Ravindra and Prateek proposed an efficient improved iterative FP Tree algorithm for generating frequest access patterns from the access paths of the users. The frequent access patterns are generated by backward tree traversals [5].

Dimitrijević and Bošnjak [6] proposed to alleviate the problem of web usage association rule over-generation by pruning the rules that contain directly linked pages out of the rule set. The experiments showed that interestingness measures can successfully be used to sort the discovered association

rules after the pruning method was applied. Most of the rules that ranked highly according to the interestingness measures proved to be truly valuable to a web master. Apriori algorithm is used to find the interestingness measures.

Veleti and Nagalakshmi [7] in their paper proposed an incremental algorithm (IPNAR) that mines positive and negative association rules in web usage data. The incremental based algorithm incrementally update web log association rules by utilizing the metadata of old database transactions as well as old mined rules, performs single scan over the dataset , and it overcomes the limitations of other mining methods. When comparing with the other existing algorithms, Incremental algorithm is highly efficient, reduced number of passes over the database, reduce the number of non-interesting negative rules and it will find all the association rules quickly .

4. LIMITATIONS OF THE EXISTING WORK

From the related research works, it is clear that most of the research carried out in the field of web mining concentrated on the Web usage mining and a very little attention is give in the area of web content and web structure mining. Apriori, FP-Growth, Market Basket algorithms are most commonly used algorithm in pattern discovery. Below graph shows the usage of algorithms in Web usage mining.

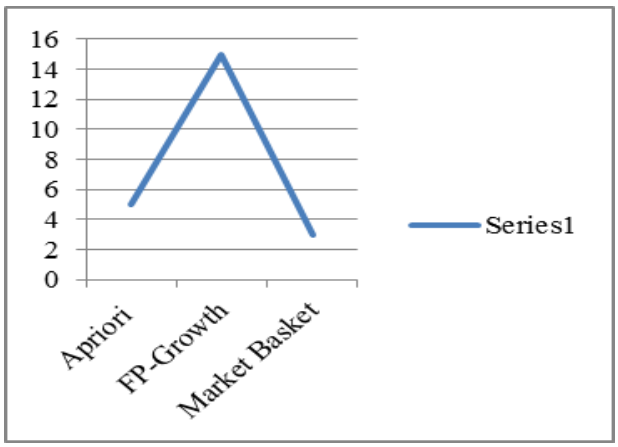

Figure 2:

More research works can be done using various algorithms to discover the patterns and the existing results can be enhanced.

5. CONCLUSION AND FUTURE WORK

This paper is just an overview on Web usage mining using association rule. Huge amount of information is stored in the web services. The information overhead leads to difficulty in finding relevant and useful knowledge, therefore web mining is a tool to discover and extract knowledge from the web. The massive growth in the field of E-Commerce attracts lot of web consumers, and hence understanding the behaviour of the consumer is one of the challenging task for companies. Improvised algorithms by incapacitating the existing and implemented algorithm in the field of web usage mining will aid the companies in understanding the consumer. In future, a detailed survey will be conducted on the Web mining using FP-Growth algorithm to understand the advancement and limitations in this field.

REFERENCES

1. Aggarwal and Philip S. Yu, “An Automated Sys – tem for Web Portal Personalization”, Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002
2. Masseglia F, Poncelet P and Cicchetti R, “An efficient algorithm for Web usage mining”, Networking and Information Systems Journal. Volume X, 2000
3. Veeramalai S, Jaisankar N and Kannan, “Efficient web log mining using Enhanced Apriori Algorithm with Hash Tree and Fuzzy”, International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 201, PP 3-4
4. Rahul M and Abha Choubey, “Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining”, International Journal of Advanced Research in Computer Science and Software Engineering Research Paper, Volume 2, Issue 9, September 2012, PP 2-3
5. Ravindra and Prateek, “Fast Processing of Web Usage Mining with Customized Web Log Pre-processing and modified Frequent Pattern Tree” , International Journal of Computer Science & Communication Networks,Vol 1(3), PP 277-279
6. Dimitrijević and Bošnjak, “Web Usage Association Rule Mining System”, Interdisciplinary Journal of Information, Knowledge, and Management, Volume 6, 2011
7. Veleti and Nagalakshmi, “Web Usage Mining: An Incremental Positive and Negative Association Rule Mining approach”, International Journal of Computer Science and Information Technologies, Vol. 2 (6) , 2011, PP 2862-2866