A Comparative Analysis of Web Usage Mining Techniques

Paridhi Nigam
PG Scholar
Dept Of CSE,
Shri Vaishnav Institute of
Technology & Science, Indore, India

Rajesh K. Chakrawarti
Reader
Dept Of CSE,
Shri Vaishnav Institute of
Technology & Science,
Indore, India

ABSTRACT

Web usage mining is the application of data mining techniques and is used to extract the important data which are present in the web. Nowadays web log mining is a very popular and computationally expensive task. Preprocessing, pattern discovery, and pattern analysis are the major task of web usage mining. In this paper we are presenting an overview of existing algorithms used in pattern discovery phase for mining the frequent item set by designing comparative analysis table i.e. Apriori, K-Apriori, FP growth which are used in pattern discovery phase.

General Terms

Web usage mining, Frequent item set mining.

Keywords

Web mining; Web log mining; Apriori; K-Apriori; FP growth;

1. INTRODUCTION

Data mining is the process of extraction of large amount of hidden information which are present in the database. There are various techniques to mine data from database that are clustering, classification and association rule. There are various fields in which data mining is used like e-commerce, health & science, web mining is also the application of data mining. Web mining is used to identify the hidden information or data from World Wide Web. In other words finding out the useful patterns from the unstructured data is known as web mining. It is also used to understand behavior of customer [1]. The web mining is divided into three main categories.

1.1 Web content mining

Web content mining is the process of gathering useful information from Web content [7]. Web content consists of several types of unstructured data like text, images, audio or video data, records such as lists or tables and structured hyperlinks.

1.2 Web structure mining

Web structure mining is one of three categories of web mining for data, is a tool used to identify the relationship between Web pages that are connected by information or direct link connections. Web structure mining is the collection of methods which are used for mining or analyzing the structure of a website or hierarchy or the links of a website. In web structure mining the structure of a web is mined on the basis of hyperlinks and intra-links.

1.3 Web usage mining

The web usage mining related to the application of data mining tools and technique. The web usage mining is used to discover usage patterns from web data in order to understand the user's need for navigating on the web. Web usage mining is used to discover the navigation patterns from web data, predicts the behavior of user while the user interacts with the web and also it helps to improve large collection of resources.

2. BACKGROUND

Internet has become the most significant medium for sharing and gathering of the information. As we know that the use of internet is increasing day by day. In recent years the size of the database has increased rapidly. There are so many private and public organization which produces large amount of data day by day for example customer care, financial forecast, marketing policies, even medical diagnosis and many other applications. It was very difficult to extract the important information from web. So Web usage mining used to extract the information from web servers on the basis of usage patterns. Web usage mining is use to find out the hidden patterns of web data. Web usage mining techniques is also used to discover the patterns of web data. This has led to a growing interest in the development of tools capable in the automatic extraction of knowledge from data.

3. TECHNIQUES

The web usage mining consists of the major steps. These steps are as follows [8].

3.1 Preprocessing

In the web usage mining, preprocessing phase is used to improve the quality of data. This can be achieving by extracting the data from web data set to remove the noisy data and then data is preprocessed. The data collection should be done before the preprocessing phase Preprocessed files consists of information such as who accessed the page, what page are accessed and how long the user accessed that page, which user accessed which page, access time, access date, access duration etc. The data preprocessing consist of another sub steps.

3.1.1 *Data cleaning*

In data cleaning phase the unusable data is eliminated from the given dataset. The data cleaning is usually site specific which includes removing of file extensions like gif, .pdf, jpg etc in target URL.

3.1.2 *User Identification*

After cleaning the HTTP log files , the next step in data preprocessing phase is the user identification. There are different methods for this, first method is converting the IP address to the domain name & second method is, ID is assigns randomly to the web browser by the web server while it connects to the web site. Third is, If the IP address of one log entry is same, but if there is a change in browser software and operating system, then the combination of IP address, browser and O.S represents a different user [9].

3.1.3 Session Identification

Client session will be alive until he is proactive user with the particular website is known as the session of that particular client. The following are the steps we can use for identifying the session.There will be a new session, for every new clientThere will be a new session for user, if the page is associate by the user is null in a user session.After every 30 minutes new session is started automatically [4].

3.2 Pattern discovery tools

In this phase the activities of the users on the web are discovered. The frequent patterns discovery phase needs the web pages which are visited by the user. In the pattern discovery, sequences of the pages are irrelevant. Also the identical pages are ignored, and the pages are arranged in a predefined order. Frequent itemset mining, clustering, statistical analysis, classification and sequential analysis are the techniques which involved in pattern discovery phase.

3.2.1 Frequent itemset mining

This method is used to discover group of pages which are frequently accessed together with support exceeding a threshold. Frequent itemset mining also used to find information like Set of pages repeatedly accessed together by web users - The next page that will be fetched - Frequently accessed paths by web users.

3.2.2 Clustering

It is a process of grouping together a set of items having similar features. Two types of clusters can be found in web usage mining, user clusters and page clusters. User clusters will discover users having same browsing patterns whereas page clusters will discover pages possessing similar content. The techniques of finding user clusters and page clusters are called usage based clustering and content based clustering respectively.

3.2.3 Classification

Classification is the job of mapping a data item into one of the number of predefined classes or labels. In the Web Usage mining, one is interested in generating a user profile belonging to a particular class or category [3].

3.3 Pattern Analysis Tool

In this phase the patterns extracted from the patterns discovery phase are preprocessed to get most frequent pattern [2]. Pattern analysis makes the predictions of new data which are coming from the same source. Data are present in many forms like images, texts, audio or video data.

4. FRAMEWORK

To find bonafide information is very easy but extracting knowledge from structural information is difficult. Before applying mining technique, data cleaning and preparing the data is important. Elimination of unused data from web log mining such as image files, failed request entries generated by search engine is the basic challenge of data preprocessing. Relational database is ready for applying data mining technique after completion of data pre-processing. The architectural design for web usage mining is divided into two sections. Section one involves data pre-processing, session identification, and integration of relevant data and also it is responsible for converting the server access log data into appropriate user session format. Section two includes pattern matching techniques and basic data mining technique [6]. The architecture of web usage mining technique is given below.

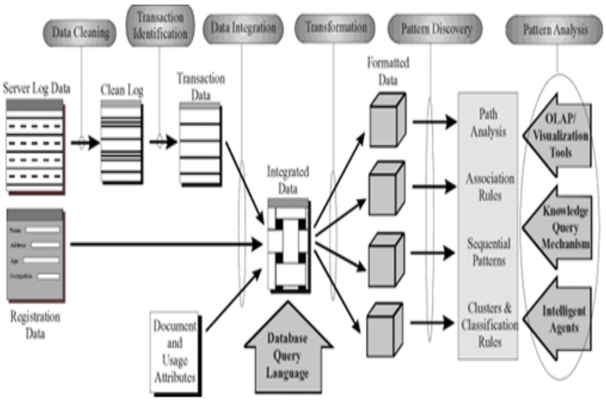

Fig 1: Architecture of Web Usage Mining

In web usage mining cleaning of data is the first step. The web mining techniques is partitions the log entries into logical groups called cluster but this can be achieved after the data cleaning task. There are two ways for cleaning web log files, the first way is thruogh a individual session of web pages & second way is through the set of numerous session and each session consist of a individual web page refrence. Partition of large session into so many small session or several small chunk of session is combining into a single one is the basic task of session identification [7].

The following are the table which shows the comparison between the algorithms used in pattern discovery phase for mining the frequent item set.

Table 1. Comparison between frequent itemset mining algorithms

Parameters	Apriori	K-Apriori	FP growth
Required memory	Apriori algorithm used large memory, because it generates large candidate set [7].	K Apriori algorithm used less memory as compared to Apriori	FP growth algorithm required less memory
Number of visit to scan database	Database is scanned multiple times	Database is scanned less as compared to Apriori	In this only two times database is scanned.
Time complexity	It requires large time complexity	It requires small time complexity	Large execution time is required

Efficient	Less efficient	More Efficient then Apriori	It reduces the large number of candidate set so it is more efficient then k Apriori
Storage	Array	Matrix	FP tree
Technique	Uses Apriori property and join and prune property	Uses iterative approach. The data is partitioned into K clusters using multipass K-means algorithm [9].	Divide and conquer methodology is used.

5. BENEFITS

5.1 Apriori

5.1.1 This algorithm uses an iterative approach called level wise search. So the implementation is easy and simple [2].

5.1.2 It is an efficient algorithm for finding all frequent item sets

5.1.3 Uses large itemset property

5.1.4 Easily parallelized.

5.2 K-Apriori

5.2.1 It is more efficient than Apriori.

5.2.2 It requires less time complexity because large datasets are divided into clusters [3][7].

5.3 FP Growth

5.3.1 FP growth algorithm uses compact data structure.

5.3.2 Required less memory.

5.3.3 Reduces repeated scans of database.

6. LIMITATIONS

6.1 Apriori

6.1.1 Because of huge number of candidate set, it is costly [8].

6.1.2 It is tedious to repeatedly scan the database and check the large set of candidates by matching patterns.

6.1.3 It requires large time and space complexity because of multiple scans of database.

6.2 K-Apriori

6.2.1 In K-Apriori algorithm, implementation is easy.

6.2.2 Requires large space.

6.3 FP Growth

6.3.1 Requires large execution time [3]

7. FUTURE SCOPE

In web mining, web usage mining is the main area in research which identifies the web usage patterns of user's such as web access log, web structure, and web contents. Pre processing, pattern discovery, and pattern analysis is the three main steps of web usage mining. In this research paper we studied and compared the various web usage mining techniques having many advantages & disadvantages. So on the basis of time complexity and space complexity, we select an algorithm and will propose a novel algorithm for reducing the space complexity and time complexity.

8. ACKNOWLEDGEMENT

The best way I have gained while working in this research paper was to be efficient and clean in my work. Any work pays you good when planned & undertaken in an organized way. The best way to proceed is to have good ideas and always remembering that there exists many a ways to tackle a given problem. This work is the result of inspiration, support, guidance, cooperation that were extended to me.

I express deep sense of gratitude and indebtedness to my guide Prof. Rajesh Kumar Chakrawarti (Reader, Deparment of Computer Science and Engineering) for their valuable, affectionate guidance and constructive criticism at all stages of my research paper. This acknowledgement can never be complete without thanking to Mr. Anand Rajavat (Head of Department of Computer Science) for providing all facilities and resources needed for this research paper.

9. REFERENCES

[1] Suhasini Parvatikar, Bharti Joshi, "Analysis of User Behavior through Web Usage Mining", Department of Computing.

[2] R. Kousalya, V Sarvanan, "Improving efficiency of Web Usage Mining using K-Apriori and Fp-Growth Algorithm"

[3] Parth Suthar, Bhavesh Oza Department of Computing Science and engineering, "A Survey of Web Usage Mining Techniques", L.D College of Engineering, Ahmedabad, Gujarat, India.

[4] Li Chaofeng School of Management, "Research and Development of Data Preprocessing in Web Usage," South-Central University for Nationalities ,Wuhan 430074, P.R. China.

[5] Alexandros Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos, "Effective prediction of web user

accesses: A data mining approach." in Proc. Of the workshop WEBKDD, 2001.

[6] Sarita Dalmia, "Wed Mining:survey and Research,".

[7] B.Santhosh Kumar, K.V Rukmani Department of Computing Science, " Implementation of Web Usage Mining APRRIORI and FP Growth algorithm", C. S. I College of Engineering, Ketti- 643215. The Nilgiris.

[8] Jaideep Srivastava , Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data".

[9] Monika Dhandi, Rajesh Kumar Chakrawarti Department of Computer Science, " A comprehensive study of Web Usage Mining", SVITS, Indore.