Minimizing the Repeated Database Scan Using an Efficient Frequent Pattern Mining Algorithm in Web Usage Mining

Devinder Kaur¹, Ravneet Kaur²
Student of Master Technology¹, Assistant Professor²
Department of Computer Science and Engineering¹,²
Sri Guru Granth Sahib World University
Fatehgarh Sahib, Punjab, India

Abstract—Data Mining is the process of discovery of new patterns and knowledge from large dataset. Web mining is the application of data mining techniques to extract and mine useful knowledge and interesting patterns from World Wide Web. Web data including web documents, hyperlinks between documents, usage logs of web sites. The web usage data captures the identity and origin of the web user along their surfing behaviour at a website. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behaviour of the users. Mining frequent patterns from web log data can help to optimise the structure of a web site and improve the performance of web servers. In this paper First we investigate process used for maximal forward reference. And then a new approach is proposed to modify the process of find maximal forward reference through backward scan algorithm. After use proposed modified algorithm time and space complexity will reduce. The backward scan algorithm is proposed for frequent pattern mining in web usage mining. It scan the web log database from backward level.

Index Terms—Data mining, World Wide Web, traversal pattern.

1. INTRODUCTION

Data mining is the process of discovering interesting knowledge from large amount of data. Data mining refers to discover knowledge in huge amounts of data. Knowledge discovery in database is the non trivial process of identifying valid, potentially useful and ultimately understandable patterns in data [1]. The web mining is a combination of the two singular areas of in progress one is the data mining and second one is world wide web (WWW). It can be able to be mostly defined as the finding and discovering the useful information from WWW. Web mining is the make use of data mining discover and mine information from Web documents and services. Web Data Mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data. Web mining has three types web content mining, web structure mining and web usage mining. It is normally expected that either the hyperlink structure of the web or the web log data or both have been used in the mining process [2].

<table>
<thead>
<tr>
<th>Type</th>
<th>Structure</th>
<th>Forms</th>
<th>Object</th>
<th>Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>Accessing</td>
<td>Click</td>
<td>Behaviour</td>
<td>Logs</td>
</tr>
<tr>
<td>Structured</td>
<td>Pages</td>
<td>Text</td>
<td>Index</td>
<td>Pages</td>
</tr>
<tr>
<td>Content</td>
<td>Map</td>
<td>Hyperlinks</td>
<td>Map</td>
<td>Hyperlinks</td>
</tr>
</tbody>
</table>

2. LITERATURE REVIEW

Hemant Kumar Singh and Brijendra Singh [3] introduced about the web content mining, web structure mining and web usages mining. The aim of this paper is to provide past, current evaluation and update in each of the three different types of web mining i.e. web content mining, web structure mining and web usages mining. It also represents the comparisons and summary of various methods of web data mining with applications. Jaideep Srivastava et al. [4]. Describe the Web usage mining phases, namely preprocessing, pattern discovery, and pattern analysis and its application. It provides a detailed taxonomy of the work in this area, including research as well as commercial offerings and practice communities. Pawel Weichbroth et al. [5] describe problem of mining
access patterns from Web logs efficiently. A novel data structure, called Web access pattern tree, or WAP-tree in short, is developed for efficient mining of access patterns from pieces of logs. The Web access pattern tree stores highly compressed, critical information for access pattern mining and facilitates the development of novel algorithms for mining access patterns in large set of log pieces. Renata Ivancsy and Istvan Vajk [6] presented discovering frequent patterns in Web log data is to obtain information about the navigational behaviour of the users. The different patterns in Web log mining are page sets, page sequences and page graphs. This can be used for advertising purposes, for creating dynamic user profiles etc.

Yan Li et al. [7] the proposed path completion algorithm efficiently appends the lost information and improves the reliability of access data for further Web usage mining calculations. Ming-Syan et al. [8] proposed a new data mining algorithm that involves mining path traversal patterns in a distributed information-providing environment where documents or objects are linked together to facilitate interactive access. Their solution procedure consists of two steps. First, derive an algorithm to convert the original sequence of log data into a set of maximal forward references. Second derive another algorithm to determine the frequent traversal patterns i.e. large reference sequences from the maximal forward references obtained. Jianhan Zhu [9] applied the Markov chains to model user navigational behaviour. They proposed a method for constructing a Markov model of a web site based on past visitor behaviour. Then the Markov model is used to make link predictions that assist new users to navigate the Web site. WANG Tong [10] offers an improved algorithm based on the original AprioriAll algorithm. The new algorithm adds the property of the User-id during the every step of producing the candidate set and every step of scanning the database by which to decide whether an item in the candidate set should be put into the large set which will be used to produce next candidate set. Hengshan Wang [11] introduced two prevalent data mining algorithms – Fpgrowth and PrefixSpan into WUM. Maximum Forward Path (MFP) is also used in the web usage mining model during sequential pattern mining along with PrefixSpan so as to reduce the interference of “false visit” caused by browser cache and raise the of mining frequent traversal paths. Sandeep Singh Rawat [12] proposed a custom-built apriori algorithm which is based on the old Apriori algorithm, to find the effective pattern analysis. Ankit R Kharwar et al. [13] describe implements the high level process of Web Usage Mining using basic Association Rules algorithm call Apriori Algorithm. It presents finding association Rule from server log which are useful in many application like cache for web page, Marketing, Targeted and Advertising etc. Mr. Rahul Mishra and Ms. Abha Choubey [14] proposed the FP growth algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the user’s interest.

3. WEB MINING CHALLENGES

Today the World Wide Web is popular and interactive medium to distribute information. The web is huge, diverse, dynamic and unstructured nature of web data, web data research encountered lot of challenges for web mining. Information user could encounter following challenges when interacting with web.

1. Finding Relevant Information:- People either browse or use the search service when they want to find specific information on the web. Today’s search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.

2. Creating new knowledge out of the information available on the web:- This problem is basically sub problem of the above problem. Above problem is query triggered process (retrieval oriented) but this problem is data triggered process that presumes that already have collection of web data and extract potentially useful knowledge out of it.

3. Personalization of information:- When people interact with the web they differ in the contents and presentations they prefer.

4. Learning about Consumers or individual users:- This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem related to web site design and management and marketing.

4. BASIC PROBLEM OF FREQUENT ITEMSET

Association Rules find all sets of items that have support greater than the minimum support and then using the large item sets to generate the desired rules that have confidence greater than the minimum confidence. An algorithm for finding association rule named as AIS was proposed by R.S. Aggarwal in 1993. There are several algorithms for frequent itemset. All the variations make for apriori algorithms have some advantages and disadvantages.
### Table 1: Algorithms, Storage Structure, Advantages, and Disadvantages

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Storage Structure</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apriori</td>
<td>Array based</td>
<td>Any subset of frequent item set is also frequent item set.</td>
<td>Multiple scans have to be done on database. Its time and memory complexity is very large.</td>
</tr>
<tr>
<td>Apriori Tid</td>
<td>Array based</td>
<td>Number of entries may be smaller than the number of transactions in the database.</td>
<td>The time and space complexity is also very large.</td>
</tr>
<tr>
<td>FP tree</td>
<td>Tree based</td>
<td>Scans the database only twice. It has interactive rule mining.</td>
<td>It seems to be difficult in incremental. It require less memory and more execution time.</td>
</tr>
<tr>
<td>Custom built Apriori</td>
<td>Array based</td>
<td>Based on old apriori algorithm. It is efficient and effective pattern analysis.</td>
<td>It requires more memory and more execution time.</td>
</tr>
</tbody>
</table>

### 5. PROPOSED APPROACH

The proposed algorithm has based on frequent pattern mining using web log data. The basic objective of the algorithm to obtain the maximum traversing path of the users from web log database. It scan the web log database from the backward. This algorithm will makes the pattern mining process effective.

#### Algorithm : Backward Scan.

**Input** : User traversal pattern.

**Output** : Maximal forward reference.

Step 1 : Input data set and min threshold value.

Step 2 : Calculate the length of longest itemset in dataset.

Step 3 : Repeat from longest itemset length.
   a. Generate candidate of k level.
   b. Calculate the count of candidate
   c. If count of any candidate is greater than min threshold value.
      Then print result.
      Else repeat step 3.

Step 4 : Exit.

### Illustrative Example

The proposed algorithm consists of various steps which will explain with the user traversal pattern. The web has like tree structure with pages being represent as nodes which denoted by alphabets and hyperlinks represented by arrows.

![User Traversal Pattern](image)

**Fig 1 User Traversal Pattern**

<table>
<thead>
<tr>
<th>TID</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>ABE</td>
</tr>
<tr>
<td>200</td>
<td>ADH</td>
</tr>
<tr>
<td>300</td>
<td>ACG</td>
</tr>
<tr>
<td>400</td>
<td>ACG</td>
</tr>
<tr>
<td>500</td>
<td>AD</td>
</tr>
<tr>
<td>600</td>
<td>ACG</td>
</tr>
</tbody>
</table>

**Table 2 web log database**

The above database is generate from the user traversal pattern graph to obtain the maximal forward reference.

**Step1** : In the first step generate candidates from graph with the transaction id and number of nodes denoted by alphabet during access the transaction.

**Step 2** : In the next step generate the length of longest candidate level from the graph.
Step 3: After generation of candidate level then count the number of occurrence of each candidate from the web log database. The candidate count satisfy the minimum threshold value.

<table>
<thead>
<tr>
<th>length</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABE</td>
<td>1</td>
</tr>
<tr>
<td>ACF</td>
<td>0</td>
</tr>
<tr>
<td>ACG</td>
<td>3</td>
</tr>
<tr>
<td>ADH</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Candidate length

The above shows count of each candidate. The maximum candidate count is 3 which has maximum traverse path of the user. Otherwise generate level 2 candidate from the graph.

The proposed algorithm will improve the efficiency of traditional Apriori algorithms. The backward scan algorithm is modified approach for frequent pattern mining in web usage mining. It will efficient from all previous algorithms. The time and space complexity will reduce used by this algorithm because it’s minimizing the repeated database scan for Frequent Pattern Mining in web usage mining. The maximal forward reference will easily obtain by use this algorithm.

6. CONCLUSION

In this paper first data mining and web mining categories have been discussed. Frequent mining algorithms also has been analyzed along their advantages and disadvantages. An algorithm has been proposed for frequent pattern mining in web usage mining which is efficient than traditional Apriori algorithms. The first part of algorithm, i.e. backward scan, firstly scan the web log database and obtain the longest candidate level length. After that count the occurrence of each candidate. Each candidate count satisfy the minimum threshold value and then obtain the maximum forward reference from candidate count length. The new approach requires minimum repeated database scan for frequent pattern mining in web usage mining. It will reduce the time and space execution.

REFERENCES

[14] Mr. Rahul Mishra and Ms. Abha Choubey, “Discovery of Frequent Patterns from Web Log Data by Using FP-Growth Algorithm for Web Usage Mining”, in proceeding of Internation Journal of Advanced
