Enhancing the Performance of BitApriori Algorithm in Data Mining using an Effective Data Structure

Abstract

Exploration frequent item sets within transaction databases, time-series databases, and far different form of databases has been analyzed popularly within information exploration investigation. An exploration frequent merchandise unit is one of the nearly all investigated investigated areas within information exploration. Association rules will be the principal way of information exploration. Apriori algorithm is really an established algorithm associated with connection tip exploration. A lot of algorithms regarding exploration connection guidelines as well as the mutations are usually planned in basis associated with Apriori algorithm; One of these algorithms is the BitApriori algorithm. However, this algorithm reduces the time counting amount of support, but when the Database is big, BitApriori may be faced with a shortage of memory; Recently, an algorithm provided called Enhanced BitApriori that by replacing some effective techniques on binary string, improved BitApriori algorithm has somewhat. But this method can still be improved by eliminating infrequent item sets. Therefore, in this study, a method was developed using factors Set Size and Set Size Frequency minimum number of candidates is presented with greater efficiency. The results showed that the proposed method is more efficient than the Enhanced BitApriori algorithm.

Key Words

Data Mining, Apriori Algorithm, Frequent Item sets, Association Rule Mining, Big Data
I. Introduction

Data mining can be a significant analysis area nowadays of which targets understanding discovery throughout data source. It's wherever data in the database are generally mined in order that educational data might be created along with used successfully along with correctly by means of human beings. Its aim is usually prediction along with outline [1]. These days, while using the swift growth of IT, specifically the internet service-based software, service-oriented buildings along with cloud-computing, continuously growing data are generally incorporated to come up with helpful information. Many approaches are used for data mining. Association rules mining (ARM) is among the most useful approaches. The challenges linked to ARM, especially for parallel along with ARM data mining, consist of minimizing I/O, increasing running velocity along with cutting down connection cost [3]. A major worry throughout ARM right now would be to still improve algorithm functionality.

Association Rule mining consists of a pair of treatments [1]: 1st, picking out the typical merchandise emerge the particular repository employing a minimum assistance and also building the particular relationship principle on the typical merchandise collection along with specific assurance. The idea relates to the particular relationship associated with products wherein for any happening of any, there may be the happening associated with B. This mining is a lot more relevant already in the market holder analysis [2]. In which software is useful towards the shoppers that will acquire selected products. In which for any merchandise them to obtained, precisely what could be the probable item/s as well as the particular obtained merchandise. Apriori algorithm could be the most widely used relationship principle mining algorithm [9]. Nonetheless, various restrictions are discovered in this particular technique [7] such as:

- Numerous iterations regarding data are needed with regard to mining data
- Normally builds items which tend to be unimportant
- Troubles locating uncommon situations
- With these disadvantages numerous is effective have been observed to raise the effectiveness regarding Apriori algorithm.

The Apriori-based algorithms discover regular product packages based upon the iterative bottom-up procedure for create choice item sets. Since the 1st estimate regarding affiliation policies mining by R. Agrawal [10], a lot of experiments have been carried out to create regular item sets mining scalable and efficient. Although you can still find a few inadequacies that will Apriori centered algorithms lived with, as well as: lots of verification from the deal data source any time trying to find regular item sets, great deal of choice item sets generated needlessly and so on.

This paper puts forward Enhancing the Performance of BitApriori algorithm using matrix data structure with simply counting rows and columns and transaction reduction strategies using top down approach for finding out largest regular itemset to smallest regular itemset. In this way, it can greatly reduce complexity and increases the efficiency of improved algorithm.
The remaining section of this paper is organized as follows: Section II contains review on Association Rule Mining and Apriori Algorithm. In section III contains Limitations of Apriori Algorithm. Section IV summarizes the related works. Section V; elaborate the proposed work for extracting regular itemset and experimental results. Section VI contains conclusion.

II. REVIEW ON ASSOCIATION RULE MINING AND APRIORI ALGORITHM

Association Mining is just about the most critical info mining's functionalities which is typically the most popular process has become analyzed by simply analysts. Removing association regulations may be the key involving info mining [3]. It can be mining intended for association regulations throughout data source involving product sales deals involving things that is very important area on the study throughout dataset [4]. The advantages of these types of regulations are usually detecting unidentified romantic relationships, producing results which could conduct time frame intended for choice producing and also prediction [3]. The actual development involving association regulations are actually divided straight into a couple of levels [5]: discovery the particular frequent item sets and also age group involving association regulations. Inside the very first stage, each pair of things is known as item set, when they transpired with each other in excess of the particular lowest service threshold [5], this kind of item set is known as frequent item set. Acquiring frequent item sets put in at home however costly and this stage is actually much more crucial as compared to second stage. In the subsequent step, it could possibly create quite a few policies in one item set like kind, if item set \( I_1, I_2, I_3 \), it's policies are \( I_1 \_ I_2, I_3, I_2 \_ I_1, I_3, I_3 \_ I_1, I_2, I_1, I_2 \_ I_3, I_1, I_3 \_ I_1, I_2, I_3 \_ I_1 \), variety of individuals policies will be \( n^2 - 1 \) exactly where \( n = \) variety of items. In order to validate the association rule (e. grams. \( X \_ Y \)), exactly where \( X \) and \( Y \) are items, dependent on confidence threshold that establish the particular percentage in the deals that includes \( X \) and \( Y \) for the deals \( A\% \) that includes Times, because of this \( A\% \) in the deals that includes \( X \) likewise include \( Y \). minimum support and confidence will be explained from the end user that shows concern in the policies. And so the support and confidence thresholds must be used for all your policies in order to prune the rules that the idea beliefs under thresholds beliefs. The issue that may be resolved straight into connection mining will be locating the relationship involving different things via substantial group of deals effectiveness [3]. The study involving association rules will be inspired through much more programs for example telecommunication, banking, health and making, etc. [2].

The apriori algorithm is standard and classical algorithm for mining regular/frequent itemsets (if an itemset satisfies minimum threshold i.e., min_support, it is called regular itemset. The set of regular k-itemsets is commonly denoted by \( L\_k \) ) brought by R. Agarwal and R. Shrikant in 1994, that leads to generate association rules called association rule mining. It uses bottom-up and iterative approach known as breadth first search (level wise search where k-itemset used to discover k+1 itemset). It generates all regular itemsets (a set of items is referred
as an itemset, n-itemset consist of n items) and apriori property is introduced to reduce the search space [1].

The actual operating of Apriori algorithm is rather is dependent upon this Apriori residence which usually expresses that” Most nonempty subsets of your repeated item sets need to be frequent”. What's more, it referred to this anti-monotonic residence which usually claims should the technique cannot move this lowest assist check; all the supersets will probably don't move this check. Whenever the main one fixed will be irregular subsequently all the supersets will also be repeated and vice versa. This specific residence can be used to prune this irregular customer element. At the start, this number of repeated 1-itemsets can be found. The actual number of made up of one particular merchandise, which usually fulfill the assist threshold, will be denoted by means of L. Within every future move, all of us start out with a seedling number of itemsets discovered for being large in the previous move. This specific seedling fixed can be used pertaining to generating completely new probably large itemsets, referred to as customer itemsets, and matter the particular assist pertaining to these kinds of customer itemsets during the move in the information. By the end on the move, all of us determine which usually on the customer itemsets have been large (frequent), plus they end up being the seedling for the future move. For that reason, L can be used to uncover, this number of repeated 2-itemsets, which can be employed to uncover L, etc., until eventually get rid of repeated k-itemsets is found.

It is undoubtedly which Apriori criteria productively finds your regular components on the database. Yet because the dimensionality with the database rises together with how many items after that:

- A lot more look for area should be applied along with I/O cost increase.
- Volume of repository scan will be elevated therefore choice age group increase ends in increase in computational cost.

Thus many variations happen to be takes place within the Apriori criteria to minimize the above mentioned restrictions develops caused by improve bigger involving repository. Most of these later planned algorithms follow identical repository scan stage by means of stage as in Apriori criteria, while the ways of customer creation and also trimming, assistance keeping track of and also customer manifestation may differ. The actual algorithms improve the Apriori algorithms by means of:

- Minimize travels connected with deal data bank verification
- Reduce in size quantity of candidates
- Help support counting connected with individuals

### III. Limitations of Association Rule Mining and Apriori Algorithm

Association rule mining is an intriguing topic connected with analysis in the field of info mining. On the other hand organization tip mining continues to be in a very phase connected with query as well as development. You can still find some essential conditions that have to be studied pertaining to figuring out useful organization principles. There are several complications pertaining to Association rule mining since beneath:
More and also scalable techniques for Association Rule mining must be designed. 
Individual check out and also on the internet exploration strategies must be designed. 
Approaches for exploration connection principles inside multi-databases must be explored. 
Effective processes for World Wide Web Usage Exploration must be designed.

Apriori algorithm afflicted with a number of a weakness notwithstanding being clear in addition to straightforward. The key limit is actually expensive throwing away of time to hold a massive amount of applicant pieces along with much typical itemsets, small minimal service as well as large itemsets. Consequently, it will look for several pieces by applicant itemsets, and yes it will probably check out data source often times continuously for locating applicant itemsets. Apriori will probably be minimal in addition to inefficiency as soon as recollection capacity is fixed along with many deals [2].

This paper will focused on proposing an approach to reduce the time spent for searching in database transactions for frequent itemsets.

IV. RELATED WORKS

Various improved algorithms have been proposed to vanquish the weaknesses of Apriori algorithm in various ways. Here presents some different approaches that face the common drawback.

To weed out obstructions of frequent itemsets mining in Apriori algorithm the authors Wanjun Yu, Xiaochun Wang and Fangyi Wang, Erkang Wang and Bowen Chen [6] has given a new algorithm named as Reduced Apriori Algorithm with Tag (RAAT). In Apriori algorithm, there is large number of candidate 2-itemsets and less tendency to determine support value. So, it takes lot of time to scan the database repeatedly and decreases the efficiency. To improve this, RAAT uses Apriori-gen operation to form candidate 2-itemsets which results in diminishing the pruning operation. RAAT also follows the concept of tag to increase the speed of support calculation. As a result, RAAT shortens the time and improves efficiency. The experimental results have shown that RAAT performs well when it is compared with Apriori algorithm in a number of times.

The authors Jaishree Singh, Hari Ram, Dr.J .S.Sodhi [7] have introduced a modified Apriori Algorithm called an improved Apriori Algorithm (IAA) to conquer the limitations of classical Apriori Algorithm. The classical Apriori algorithm scans the database many times. If database contains ample number of records, it takes huge time to scan the database which results in increasing I/O cost. The improved Apriori Algorithm reduces the scanning time by eliminating the transactions containing irrelevant records. It uses the concept of attribute named as Size_Of_Transaction (SOT) which contains the number of items exists in specific transaction. It also decreases the I/O cost. By comparing improved Apriori Algorithm with classical apriori algorithm, it was shown that improved

Apriori Algorithm is better on the basis of efficiency and optimization. This algorithm has certain drawback also as it has to deal with new database after every generation of frequent itemset. In future, we can divide the database among processors to remove this drawback.
Shuo Yang et al [8] proposed a theorem to improve the traditional Apriori Algorithm. The traditional Apriori Algorithm takes more time to scan the database in order to find out the frequent itemsets. This increases the complexity and decreases efficiency. The proposed algorithm decreases the database access on the basis of customer habits. It uses relative theorems to find frequent itemsets. For applying improved Apriori algorithm to Ecommerce, there will be a need to develop a shopping site because when customers visit the shopping site the system will automatically find out their next purchasing goods based on goods already available in their shopping basket.

So, it will save time and increases the efficiency and provides more benefit. The customers could easily generate association rules with the help of improved Apriori algorithm and suggests useful products to customers within a reasonable time. The tool used for developing the site is Macromedia Dreamweaver 8; database management tool is Microsoft SQL Server 2000 and Web server is Tomcat 6.0. According to experimental results, it was shown that improved Apriori Algorithm when compared with traditional Apriori algorithm is more efficient.

The authors Mrs. R. Sumithra and Dr (Mrs) [9], have developed a new method distributed apriori association rule to recover the limitations of classical apriori algorithm. In this paper, main focus is on eliciting the knowledge. The implementation of both algorithms is shown using concept of grid computing. Grid computing is a form of distributed computing that enables the developers to work together on a single task at same time. Grid computing has capability to increase the efficiency and decreases the cost of computing networks by optimizing the resources. It is best for large workloads. By comparing, it was shown that distributed apriori association rule on grid based environment is better than classical apriori algorithm. The future scope is that knowledge could be extracted in parallel to produce more optimized result.

Association rule mining is a major technique in the area of data mining. Apriori algorithm is one of the earliest algorithms of association rule mining. This section has presented a survey of most recent work that has been done in Association rule based mining using Apriori algorithm.

V. PROPOSED WORK AND EXPERIMENTAL RESULTS

The improved algorithm proposed in this paper works in two phases. In first phase required compressed data structure i.e, b_matrix is constructed and then this compressed data structure is used in second phase to generate regular itemsets. This algorithm employs top down approach to discover regular item sets from largest regular item set to smallest regular itemset.

Algorithm steps In first phase b_matrix is constructed for the given transactional database. rows in b_matrix represents each transaction and column represents items in transactional database. In b_matrix, each cell will contain values either 1 or 0 for showing the existence of items in transactional database. Entry value will be 1, if the item is present in the respective row else 0, if the item is absent in the row. With two more columns count and redundant transaction counter (TC). Here count column represents the size of row(the sum of total no of 1’s in that particular row) and remove those columns whose sum is not equal or greater than predefined min_support value and then update count column. If row is duplicated in database then it is represented by the value in the redundant transaction counter column and delete unnecessary...
duplicate transaction/row and if row is not duplicate then redundant transaction counter column is set to 1. Then rearrange the b_matrix in descending order based on count column. This is our required compressed data structure and here the phase 1 of our improved algorithm completes.

Now generate regular itemsets directly from b_matrix. Select first row from b_matrix and match its count value with next row count respectively. In the event the following line count will be a lot more or equal to the processing line count then perform AND function one of many series, in the event that consequence will be similar on the processing line product established composition then improve the count benefit connected with help connected with processing line product established by a single and also continue this action connected with matching and also AND function as a result of remaining portion of the series throughout b_matrix and then look at the worthiness connected with complete help. When it is greater or equal to predefined min_support count then remove that established and its subsets and also shift the crooks to regular assortment record. The same procedure will be repeated for rest of the transactions in b_matrix until all rows are not checked.

The gain of improved algorithm is that it lessen the no. of comparisons to mine largest regular item set for duplicate transactions and transactions having smaller item sets in size that is count value (since they do not have all the items of row under process) and another major advantage is once largest regular item set is discovered then its subsets are searched and moved into frequent array list. While searching for next largest regular item set it checks first, transaction under processing is previously present in frequent array list because of prior largest itemset and its subsets, if itemset is already in frequent array list, it avoids number of comparisons needed to calculate the support count of itemset. Hence decreases number of scans and time needed to extract the regular itemset.

Even as notice throughout Table I, which the difficult throughout recommended technique throughout every single benefit associated with minimum help will be less than the item from the Enhanced BitApriori, plus the change improves a lot more because benefit associated with minimum help lessens. Table I shows that the particular recommended technique reduce the difficult by 84.09% from your Enhanced BitApriori the place that the minimum help will be 0.02, and also by 56.02% throughout 0.10. Because the benefit associated with minimum help raise the pace will be decreased likewise. The typical associated with minimizing period pace from the recommended technique will be 68.39%.

<table>
<thead>
<tr>
<th>Min-Sup</th>
<th>Enhanced BitApriori</th>
<th>Proposed Technique</th>
<th>Time Minimizing Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>6.648</td>
<td>1.066</td>
<td>83.09%</td>
</tr>
<tr>
<td>0.04</td>
<td>1.865</td>
<td>0.432</td>
<td>76.25%</td>
</tr>
<tr>
<td>0.06</td>
<td>1.158</td>
<td>0.340</td>
<td>72.50%</td>
</tr>
<tr>
<td>0.08</td>
<td>0.434</td>
<td>0.189</td>
<td>52.07%</td>
</tr>
<tr>
<td>0.10</td>
<td>0.372</td>
<td>0.158</td>
<td>57.02%</td>
</tr>
</tbody>
</table>
VI. CONCLUSION & FUTURE WORKS

Association rule mining is a major technique in the area of data mining. Apriori algorithm is one of the earliest algorithms of association rule mining. This paper has presented a survey of most recent work that has been done in Association rule based mining using Apriori algorithm. After accomplishing study of algorithms, some sort of proposed method can be proposed as a result of minimizing time taken throughout purchases encoding regarding candidate itemsets by simply minimizing the amount of purchases to get scanned. Whenever k regarding k-item set boosts, your difference involving our proposed method along with the Enhanced BitApriori boosts from see of your energy taken, and each time the significance regarding bare minimum help boosts, your difference involving our improved Apriori along with the Superior BitApriori diminishes from see of your energy taken. Any time taken to get customer help rely throughout proposed method can be less than time taken in the Superior BitApriori. Ideas to locate other methods to get combining are inspire. Currently, further research in finding a faster way of pruning candidate keys is undergoing in finding the ideal starting size of combination size.

REFERENCES