A Novel Association Rule Mining in Large Databases

M. Manasa Rao

Department of CSE, MVGR, JNTUK.

N.Sushma Rani

Department of CSE, Asst.,professor.

Abstract

One of the core topics of data mining is mining association rules in large databases. The correct and appropriate decision made by decision makers is the advantage in discovering these associations. The key process in association rule mining is discovering frequent item sets. Main challenges in developing association rules mining algorithms are the large number of rules generated that makes the algorithms inefficient and makes it complicated for end users to comprehend the generated rules. It is because of the many traditional association rule mining approaches adopt an iterative technique to discover association rule, which requires many calculations and a difficult transaction process. Furthermore, the existing mining Due to high and repeated disk access overhead the existing algorithms cannot perform efficiently. By keeping this thing in mind, in this paper we present a novel association rule mining approach that can efficiently find the association rules in large databases. By using the conventional Apriority approach with added features to improve data mining performance has been derived in the proposed approach. We have performed many experiments and differentiated the performance of our algorithm with existing algorithms found in the literature. Experimental results show that our approach can quickly and easily discover the frequent item sets and effectively mine potential association rules .

1.Introduction

The amount of information or data being stored in database has increased from the earlier years. The mining for latent knowledge become essential to support decision making has become an exponentially growth in size of on hand databases, . The key step in the knowledge discovery process is Data mining. Data mining tasks are mainly divided in two types: Predictive and Descriptive. The main task of the predictive tasks is to predict the value of a particular attribute based on the values of other attributes, and the main motto of descriptive ones, is to extract earlier unknown and needed information such as significant structures, anomalies, changes, associations and patterns etc.,, from large databases. There are many techniques satisfying these two objectives of data mining. Some of these can be divided into the following types: They are analysis, sequential pattern discovery, association rule mining, classification and clustering. In this year the development of data mining systems has received a great deal of attention. In a wide variety of business environments it plays a key enabling for competitive businesses role. It has been extensively applied to a wide variety of applications like manufacturing Ecommerce, healthcare and sales analysis etc. A number of studies have been made on efficient data mining methods and the relevant applications. For knowledge discovery and generating the rules by applying our developed approach on real and synthetic databases we considered association rules in this study.

One of the techniques involved in Mining Associations is the process mentioned in the above data and it might be the most studied ones among the data mining problems. Discovering association rules is at the heart of data mining. Mining for association rules among items in large database of sales transactions has been recognized as an important area of database research. To uncover unknown relationships these rules can be effectively used ,by producing the results that can provide a basis for forecasting and decision making among sales of different products . Today, research work on association rules is motivated by an extensive range of the main problem addressed by association rule mining was to find a correlation From the analysis of a large set of supermarket data from the application areas such as telecommunications, healthcare, manufacturing and banking. From the text databases It is also used for building statistical thesaurus and also used for finding web access patterns from web log files , and Huge sized image databases can also be used in discovering associated images .

Many numbers of association rule mining algorithms have been discovered in the last few years, they can be divided into two types: (a) Generation of candidate/test approach such as Apriority Growth in pattern approach. The first category studies is the development of an Apriori based is the milestone in level wise mining method for associations, Apriori like association mining algorithms which has sparked the development of various kinds . Among these, the Apriori algorithm has been very influential. Many scholars have improved and optimized the Apriori algorithm and have presented new Apriori like algorithms, during its inception. The Apriori like algorithms adopt an iterative method to discover frequent item sets.

There are some drawbacks in the existing mining algorithms: Firstly, they are mostly designed in forms of several passes so that the whole database needs to be read from disks several times for each user's query under the constraint that the whole database is too large to be stored in memory. In considering the large overhead of reading the large database even though only partial items are interested in fact are very inefficient. As a result, they cannot perform efficiently in terms of responding the user's query quickly. Secondly, the algorithms generate an extremely large number of association rules, often in thousands or even millions in many cases. Further, the association rules are sometimes very large. To comprehend or validate such large number of complex association rules is nearly impossible for the end users thereby limiting the usefulness of the data mining results. Thirdly, For the constraints such as support and confidence such that an appropriate number of association rules are discovered and no guiding information is provided for users to choose suitable . Consequently, to get suitable number of rules the users must use Try and error approach. This is very time consuming and inefficient. Therefore, Developing fast and efficient algorithms that can handle large volumes of data are one of the main challenges in mining association rules.

By an apriori based approach specifically designed for the optimization in very large transactional databases can be used to attack the association rule mining. The developed mining approach called Feature Based Association Rule Mining Algorithm (FARMA).

The remaining paper is designed as follows. In section 2 , We introduce the theoretical properties and formal definition of association rule. Section 3, presents some related work in the. . In Section 4, the proposed method is described in details. With various Apriori algorithms and other algorithms found in literature Section 5 presents comparisons and experiments results of our proposed approach .Finally, a conclusion and further work is given in section 1

2. Association Rules Mining

Association is the discovery of association relationships or correlations among a set of items. In section[5] this problem was introduced. A set of binary attributes Let I = {i1, i2, .im} ,called items. Let D a set of transactions and each transaction T is a set of items such that T⊆I. Let X be a set of items. A transaction T is said to contain X if and only if X⊆T. An implication of the form X⇒Y is an association rule, where X⊂ I, Y⊂ I, and X∩Y= ∅. Furthermore, the rule X⇒Y is said to hold in the transaction set D with confidence c if there are c% of the transaction set D containing X also containing Y. If there are s% of transactions in D containing X∪Y Then rule X⇒Y is said to have support s in the transaction set D . An example of an association rule is: "35% of transactions that contain bread also contain milk; 5% of all transactions contain both items". The confidence of the rule here is called 35% and 5% the support of the rule. The selection of association rule is based on support and confidence. The strength of the implication rules is , The ratio of the number of transactions that contain X U Y to the number of transactions that contain X is called the confidence factor for an association rule; whereas the support factor indicates the frequencies of the occurring patterns in the rule. i.e., the support for an association rule is the percentage of transactions in the database that contain X U Y. The problem of mining association rules involves the generation of all association rules among all items in the given database D that have support and confidence greater than or equal to the user specified minimum support and minimum confidence. Smaller support and large confidence values are typically used. since the database is large we should satisfy both minimum support and minimum confidence are said to be as Strong rules .Only those frequently purchased items , Users concern about usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful.

Discovery of frequent item sets and the generation of association rules In a given dataset D the discovery of association rules is typically done in two steps [5, 16]: . Item sets are nothing but finding the each set of items in the first case. The rate of co-occurrence of these items is above the minimum support we say the item sets are called as either large or frequent item sets .Finding all sets of items (or item sets) whose transaction support is above the minimum support is The other definition .And very easy to Find large item sets but it is very expensive . The naive approach would be to Counting all item sets that appear in any transaction is the main motto of Naive approach. Suppose one of the large item sets is Lk, Lk = {I1, I2, …, Ik}, association rules with this item sets are generated in the following way: the first rule is {I1, I2, … , Ik1} ⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. By deleting the antecedent in the last items and the other rule is generated by inserting it to the consequent , To determine the interestingness of them further the confidences of the new rules are checked . Until the antecedent becomes empty those process is iterated. The size of an item set represents the number of items in that set. The item set is called as the k item set when the item set is equal to K. For finding the association rules in the second step the frequent item sets that are generated in the first step .Or directly by performing the second step we can get the item sets very easily.

Generating the association rules is straightforward way when all large item sets are found once .In the first step the processing time is dominated. This made data mining popular in research fields. Based on this fact this paper mostly concentrates on the first step.

Generally, an association rules mining algorithm contains the following:

- The set of candidate k item sets is generated by 1-extensions of the large (k 1) item sets generated in the previous iteration.
- Supports for the candidate k items sets are generated by a pass over the database.
- Item sets that do not have the minimum support are discarded and the remaining item sets are called large k item sets.

This process is repeated until no larger item sets are found.

3. Previous Work

An algorithm called AIS was proposed for mining association rules and it is introduced first in for the problem of discovering association rules. So many algorithms for rule mining have been proposed for the last fifteen years. The algorithms mostly follow the representative approach mainly Apriori algorithm by Augural et al. . Using parallel computing various researches were done to improve the performance and scalability of Apriori. To improve the speed of finding large item sets with tree data structures, map and hash tables. To form a basis for our algorithm we have a review of some of the related work.

3.1 AIS Algorithm

The first algorithm proposed for mining association rules [5] is the AIS algorithm. There are two stages in this algorithm. The generation of the frequent item sets can be done in the first stage. To detect the frequent item sets the algorithm uses candidate generation. The generation of the confident and frequent association rules is followed in the second phase. AIS algorithm is that it makes multiple passes over the database is the main drawback. In order to turn out to be small, it generates and counts too many candidate item sets and due to this we requires more space and wastes much effort which is going to be useless.

3.2. Apriori Algorithms

The Apriori algorithm from [16] is based on the Apriori principle, which says that the item set X' containing item set X is never large if item set X is not large. Based on this principle, the Apriori algorithm generates a set of candidate large item sets whose lengths are (k+1) from the large k item sets (for k≥1) and eliminates those candidates, which contain not large subset. Then, for the rest candidates, only those with support over minimum support threshold are taken to be large (k+1) item sets. The Apriori generate item sets by using only the large item sets found in the previous pass, without considering the transactions.

It is a variation of the Apriori Tid algorithm of the Apriori algorithm. Before the pass begins the Apriori Tid algorithm also determines the candidate item sets. The Apriori Tid algorithm does not use the database for counting support after the first pass is the main difference from the Apriori algorithm Instead. Identifier TID is used for counting the large k item set in the transaction. The large item sets that would have been generated at each pass may be huge if we perform this scheme for counting large item sets in the downside manner. Another algorithm, called Apriori Hybrid is another algorithm, is introduced. It is to run the Apriori algorithm initially, and then switch to the Apriori Tid algorithm when the generated database, i.e. large k item set in the transaction with identifier TID, would fit in the memory is the basic idea of the Apriori Hybrid algorithm.

3.3. DHCP Algorithm

For the candidate set generation the DHP (Direct Hashing and Pruning) algorithm [12] is an effective hash based algorithm. By filtering any k item set out of the hash table it reduces the size of candidate set if the hash entry does not have minimum support. Regarding the support of each item set the hash table structure contains the whole information. The algorithm DHP algorithm contains of three steps. Getting a set of large 1-itemsets and constructing a hash table for 2 item sets is the first step. The generation of the set of candidate item sets Ck is done in the second step.Inorder to perform the other third step the same second step is repeated in determining whether to include a particular item set into the candidate item sets except it does not use the hash table. When the number of hash buckets with a support count greater than or equal to the minimum transaction support required is less than a predefined threshold it should be for later iterations further.

3.4. Partition Algorithm

For logically partition the database D into n partitions it requires just two database scans which is used to mine large item sets in the Partition algorithm. This algorithm mainly consists of two stages. The algorithm subdivides the database into n no overlapping partitions in the first stage, which can fit into main memory. During each iteration only one partition is considered algorithm iterates n times. These algorithm performs the counting the actual support of each global candidate item sets and generating the global large item sets is performed in the second phase.

3.5. AIS Algorithm

Apriori by using a novel data structure which is the frequent pattern tree, or FP tree improved by the frequent pattern growth (FP growth) algorithm .This algorithm stores the frequent patterns information. By using only two passes over the database this algorithm adopts a divide and conquers strategy and a frequent pattern tree to mine the frequent patterns and without candidate generation .And this is going to be a rapid development over Apriori.

4. Proposed Aproach

In order to reduce the time execution of the algorithm the developed approach adopts the philosophy of Apriori approach with some modifications. During the processing the generating the feature of items used and second, the weight for each candidate item set is calculated.

By storing the decimal equivalent of the location of the item in the transaction the feature array data structure is built. The transaction database is transformed into the feature matrix is the other definition. Reorganizing and transforming a large database into manageable structure to fulfill two objectives is called as Transforming. In the first objective (a) The number of I/O accesses in data mining have been reduced, and (b) The mining process is made speeding up. For the transforming technique there is one mandatory requirement, within the whole life cycle of data mining, that the transaction database should be read only once. The size of the database to be accessed can be reduced greatly by storing the appearing feature of each interested item as a compressed vector separately .The weight for each candidate item set Ck is calculated, through our developed approach first scans the array data structure and second the items contained in Ck are accessed and by summing the decimal equivalent of each item in the transaction the weight

$$\text{Leverage}(X \longrightarrow Y)=P(X \text{ and } Y)-(P(X)P(Y))$$

is obtained.

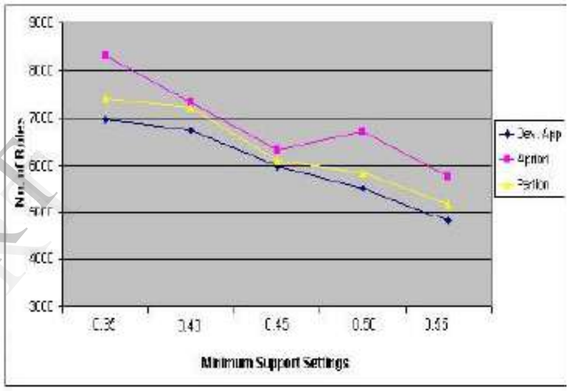
Similar process is done for calculating the support value for each item is calculated is done in similar way. To calculate the support value for each candidate item set Ck, the array of data structure was scanned by the developed approach and the items contained in Ck are accessed for and the value of support is obtained by counting the number of decimal equivalent appeared in the transaction. The process from the beginning is repeated; If a certain number of generations have not passed otherwise generate the large item sets by doing the union of all

Lk. The rules can be discovered in a straight forward manner Once the large item sets and their supports are determined as follows: The ratio of support (l) / support (a) is computed if I is a large item set, then for every subset a of I. According to that rule a ⇒ (1a) is an output if the ratio is at least equal to the user specified minimum confidence. At least N item sets are discovered with the user specified minimum support level, or until the user specified minimum confidence is reached multiple iterations of the discovery algorithm are executed until. To filter the found item sets and to determine the interestingness of the rule this algorithm uses Leverage measure introduced by Piatetsky [20].

Leverage measures the difference of X and Y appearing together in the data set and what would be expected if X and Y where statistically dependent .An implicit frequency constraint incorporates. Example for setting a minimum. Using minimum leverage thresholds at the same time leverage thresholds to 0.01% (corresponds to 10 occurrence in a data set with 100,000 transactions) by one first can use an algorithm to find all item sets with minimum support of 0.01% and then filter the found item sets using the leverage constraint. By reducing the generation of candidate's item sets and thus we reduce the memory requirements to store a huge number of useless candidates is done by using Leverage measure. And its going to be one of the main contributions of this paper.

TABLE I. EXPERIMENTAL RESULTS

	Data set 1	*Data set 2*	*Data set 3*
no. of transactions	120,000	400,000	750,000
No. of items	420	557	682
Max items/ transaction	18	26	38
Min items/ transaction	5	7	11
Support %	0.35%, 0.40 %, 0.45%, 0.50%, 0.55%	0.35%, 0.40 %, 0.45%, 0.50%, 0.55%	0.35%, 0.40 %, 0.45%, 0.50%, 0.55%
Confidence %	0.40%, 0.47 %, 0.55 % 0.60%, 0.65%	0.40%, 0.47 %, 0.55 % 0.60%, 0.65%	0.40%, 0.47 %, 0.55 % 0.60%, 0.65%
Avg # rules / FARMA	6262	6823	8118
Avg # rules / Apriori	7005	7597	9011
Avg # rules /Partition	6618	7107	8391



Test Result of Mining Setting

5. Experimental Results

We have extensively studied our algorithm's performance by comparing it with the Apriori algorithm as well as partition algorithm and consider The superiority of it to evaluate the efficiency of the proposed method. Using Microsoft Visual Basic for Applications (VBA) and run on a 3.2 GHz Pentium 4 PC with 2 GB of RAM and 250GB Hard Disk running the XP operating system all the algorithms are implemented .The transaction database provided with Microsoft SQL Server 2000 The test database .The food mart transaction database whose experimental data are randomly sampled and the three data sets of 1200,000, 400,000, and 750,000 . The test database contain 682 , 557 and 420 items respectively; with the longest transaction record contains 38,26, 18 items respectively and the lowest transaction record contains 11,7and 5 items respectively. We studied the effect of different values of minimum support (Minimum support) which are set at 0.35%, 0.40%, 0.45%, 0.50%, 0.55% and different values of minimum confidence (Minimum confidence) which are set at 0.40%, 0.47%, 0.55%, 0.60%, 0.65% on the processing time for the algorithms. As compared to other competing algorithm We have observed considerable reduction in the number of association rules generated by our algorithm. Our experimental results are presented in Table 1.Based on the support and confident values used the reduction table is also dependent. By plotting the number of rules generated versus the possible settings of minimum support over dataset 2 In Figure 1 we compare different possible settings of minimum support (Minimum support). The smaller the value of minimum support the development of the number of rules generated in Figure1 show that , the more the number of rules generated.

5.1Test Result of Mining Setting

The better the solution quality obtained (smaller number of rules generated)with increasing value of minimum support to a reasonable value. In the figure the difference is more notable. The algorithm to be flexible enough over all the datasets when the value for Minimum support is between 0.4 and 0.6 and seems to give a reasonable compromise between the solutions qualities obtained.

CONCLUSION

The aim of this paper is to improve the performance of the conventional Apriori algorithm that mines association rules by presenting fast and scalable algorithm for discovering association rules in large databases. The approach to attain the desired improvement is to create a more efficient new algorithm out of the conventional one by adding new features to the Apriori approach. The proposed mining algorithm can efficiently discover the association rules between the data items in large databases. In particular, at most one scan of the whole database is needed during the run of the algorithm. Hence, the high repeated disk overhead incurred in other mining algorithms can be reduced significantly. We compared our algorithm to the previously proposed algorithms found in literature. The findings from different experiments have confirmed that our proposed approach is the most efficient among the others. It can speed up the data mining process significantly as demonstrated in the performance comparison. Furthermore, gives long maximal large item sets, which are better suited to the requirements of practical applications. We demonstrated the effectiveness of our algorithm using real and synthetic datasets. We developed a visualization module to provide users the useful information regarding the database to be mined and to help the user manage and understand the association rules. Future work includes: 1) Applying the proposed algorithm to more extensive empirical evaluation; 2) applying our developed approach to real data like retail sales transaction and medical transactions to confirm the experimental results in the real life domain; 3) Mining multidimensional association rules from relational databases and data warehouses (these rules involve more than one dimension or predicate, e.g. rules relating what a customer shopper buy as well as shopper's occupation); 4) Mining multilevel association rules from transaction databases (these rules involve items at different levels of abstraction

REFERENCES:

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2000.

[2] J. D. Holt and S. M. Chung, "Efficient Mining of Association Rules in Text Databases" CIKM'99, Kansas City, USA, pp. 234242,Nov. 1999.

[3] B. Mobasher, N. Jain, E.H. Han, and J. Srivastava, "Web Mining: Pattern Discovery from World Wide Web Transactions" Department of Computer Science, University of Minnesota, Technical Report TR96-050, (March, 1996).

[4] C. Ordonez, and E. Omiecinski, "Discovering Association Rules Based on Image Content" IEEE Advances in Digital Libraries (ADL'99), 1999.

[5] R. Augural, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases". In Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '93), pages 207216, Washington, USA, May 1993.