To Enchanced & Optimize The A priori Algorithm Using Tokenization Based Association Rule Mining

Komal Thakur¹, Vinay Chopra²

¹²Department of Computer Science and Engineering, DAVIET Jalandhar

Abstract — In the term of Data Mining, Association Rule mining has remarkable role. Association rule mining is a popular mining technique that identifies interesting correlations between database attributes. This research paper gives the detailed introduction to proposed tokenization approach based on the apriori algorithm of association rule mining.

Keywords- association rules, frequent item sets, tokens.

I. INTRODUCTION

A. Data Mining- Data mining is a process that discovers the knowledge or hidden patterns from the large databases. Data mining is also known as the core processes of Knowledge Discovery in Databases (KDD). The KDD process is commonly defined as the following stages:

![Figure 1. KDD Process](image)

Data mining involves six main classes of tasks:
1. Anomaly detection - The identification of unusual records or data errors that require further investigation.
2. Association rule learning (Dependency modeling) - This searches the relationships between the variables.
3. Clustering - The clustering technique defines the classes and put objects in each class. Clustering is a DM technique which makes useful cluster of objects having similar characteristics using automatic technique.
4. Classification - This is used to classify each item in a set of data into one of pre defined set of classes or groups. Classification method uses mathematical techniques such as decision trees, linear programming, and neural network.
5. Regression - This attempts to find a function which models the data with the least errors.
6. Summarization- This provides a more compact representation of the dataset including visualization and report generation.

B. Association Rule Mining - Association Rule Mining are the data mining function that just discovers the probability of co-occurrence of items in a collection or dataset. As in the example, here are the 5 transactions in which in transaction has some items and in the next table, there is the probability of occurrence items i.e. in the T1 transaction there are only two items- (Bread, Milk). The (Bread, Milk) will be written as 1 and all remaining will written as 0.

![Figure 2. Association Rule Mining](image)
C. ALGORITHM USED FOR ARM- The two important Algorithm used for association rule mining are apriori algorithm and FP-Tree algorithm. The apriori algorithm is used for finding patterns called frequent item sets. A frequent item set is a set of items appearing together in a number of database records meeting a user specified threshold. The FP- Tree algorithm is to partition the original database to smaller sub-databases by some partition cells, and then to mine item sets in these sub databases. The FP-Tree construction takes exactly two scans of the transaction database. The first scan collects the set of frequent items, and second scan constructs the FP- Tree.

II. LITERATURE REVIEW

[1] Sotiris Kotsiantis, Dimitris Kanellopoulos, “Association Rules Mining: A Recent Overview” GESTS International Transactions on Computer Science and Engineering, vol.32 (1), pp.71-82, 2006. In this paper, they provide the preliminaries of basic concepts about association rule mining and survey the list of existing association rule mining techniques. They also describes the methods that had proposed for increasing the efficiency of association rules algorithms.


The authors improved the performance of mining in this paper. They use Sampling Technique to convert text document in to the appropriate format. This format contains data in the form of word and topic of word. This format take as an input in FP-Growth algorithm for giving support value and get association rules of that transaction data, and after getting association rules apply clustering process and then get clusters for that association rules.


In this paper, they proposed a novel frequent-pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree based mining method. Their performance study shows that the FP-growth method is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent-pattern mining methods.


In this paper, based on the original Apriori algorithm, an improved algorithm IAA was proposed by the authors. IAA adopts a new count-based method to prunecandidate itemsets and uses generation record to reduce totaldata scan amount.


On the basis of the association rule mining and Apriori algorithm, the authors proposed an improved algorithm based on the AntColony Optimization algorithm. They optimize the result generated by Apriori algorithm using Ant colony optimization algorithm. The algorithm improved result produces by Apriorialgorithm.


In this paper, the author surveys the most recent existing association rule mining techniques using Apriori algorithm. The conventional algorithm of association rules discovery proceeds in two steps. All frequent item sets are found in the first step. The frequent item set is the item that is included in at least minimum support transactions. The association rules with the confidence at least minimum confident are generated in the second step.


The authors proposed an algorithm for data mining called Ant-Miner (ant-colony-based data miner). The goal of AntMiner is to extract classification rules from data. They compare the performance of Ant-Miner with CN2, a well-known data mining algorithm for classification, in six public domain data sets. The results provide evidence that:

I. Ant-Miner is competitive with CN2 with respect to predictive accuracy

II. The rule lists discovered by Ant-Miner are considerably simpler (smaller) than those discovered by CN2.

In this paper, the author improves the performance of the conventional Apriori algorithm that mines the association rules. The approach is to attain the desired improvement is to create a more efficient new algorithm out of the conventional one by adding the encoding and decoding mechanisms to the latter in order to demonstrate the importance of the efficient decoding to high data mining performance and from various experiments it is proved that the logarithmic decoding method is the most efficient among the all methods it can speed up all the required processes.


The author proposes a new sequential covering strategy for ACO classification algorithms to mitigate the problem of rule interaction, where the order of the rules is implicitly encoded as pheromone values and the search is guided by the quality of a candidate list of rules.


In this paper, the author discussed the requirements and challenges of data mining. The author also describes about the major data mining techniques such as statistics, artificial intelligence, decision tree approach, genetic algorithms and visualization.


The author proposed the methodology to the classical Traveling Salesman Problem (TSP), and report simulation results. They also discuss parameter selection and the early setups of the model, and compare it with tabu search and simulated annealing using TSP. And also discussed the salient characteristics of global data structure revision, distributed communication and probabilistic transitions of the AS.


In this paper, the author designed a new technique which mines out all the frequent item sets without the generation of the conditional FP trees. Unlike FP tree it scans the database only once which reduces the time efficiency of the algorithm. It also finds out the frequency of the frequent item sets to find out the desired association rules.

III. PROPOSED METHODOLOGY

Proposed technique uses tokens i.e. those item-sets that co-occurrence with representative item can be identified quickly and directly using this simple and quickest token based method. This will avoid redundant operations of item-sets generation and many frequent items having the same supports as representative item, so the cost of support count is reduced hence the efficiency is improved. The proposed algorithm uses the following steps:
1. Scanning the database and converting it into vertical data format.
2. Generating Trans_tokenSet from vertical data format and also maintaining a list for no of iterations.
3. Finding the Frequent 1 itemset from the Trans_tokenSet i.e. the length of Trans_tokenSet of the item sets.
4. Sorting the itemset according to the ascending order of the Trans_tokenSet by its minimum support.
5. Gather these items as Keysets from the Trans_tokenSet.
6. Generate the Bit Table for each key that is available in Items keyset
7. Generate Subsume for each item in Items keyset
8. for each item in Items

   If item. Subsume<> " "
   If item. Support == min_sup
   FindItemsetsEqualsMinSup(item, item. Support)
   Else
   FindItemsetsGreaterThanMinSup (item, item. Support)
   End If
   Else
   If item. Support>min_sup
      AND Item_Sequence<Item.Length Then
      FindItemsetSubsumeNone(item)
   Endif
Endif

@IJAERD-2016, All rights Reserved 239
Figure 3. Flowchart of proposed algorithm

IV. EXPERIMENTAL SETUP

This algorithm is implemented in the Java Language because it contains the data set mining to develop the applications in the data mining and tools. The snapshot of frequent item sets mining is as follows which are acquired during implementation.

Figure 4. The screenshot for selecting the algorithm and the parameters for finding the minimum support and confidence for improved Apriori result
Figure 5. The screenshot for choosing the dataset.

Figure 6. The screenshot for the final output describing the parameters such as memory usage, time consumed and frequent itemset count are generated.
V. REFERENCES