Improved Apriori Algorithm based on Query Planning for Business Analysis

Achu Thomas Philip
Department of Computer Science and Engineering
Mount Zion College of Engineering
Kerala, India

Dr. Smita C Thomas
Department of Computer Science and Engineering
Mount Zion College of Engineering
Kadammanitta, Kerala, India

***Abstract*—with the arrival of Data Mining and Big Data, according to how ably develop massive data (information) has become a hot study in the field of information technology. The intention of the paper is to make use of business data on transactional orders to focus on descriptive study on patterns, items which are bought together and units that are highly endorsed to facilitate reordering and maintaining sufficient data. In Multinational companies faces a problem to retrieve the company profits /loss/stocks in a single query with in a secured environment i.e. related data can be retrieved from the multiple tables that are located in the distributed databases using a single query. It could be done by evaluating the feasible data in such way that persistent item set can be discovered and can be consider to specifying an association rule. An apriori algorithm induce a combination by iteration methods that are using imitated database scanning process, pairing one itemset with another itemset and then recording the number of instances of the combination with the minimum limit of confidence values and support. After examine, it is found out that the traditional Apriori algorithms have two major bottlenecks: scanning the database frequently; generating a large number of candidate sets. In this Paper contains the processing flow of classic Apriori Algorithm, describes the main improvement ideas of improved Apriori Algorithm (DAA) contains Reduces the number of transactions scanning in iteration, we can merge two or more database or two or more table's scans to mine the frequent itemsets and dynamically add new itemset to the dataset. The experimental results reveal that the efficiency of improved Apriori Algorithm (DAA) is much higher than that of the traditional Apriori algorithms.**

Keywords— Data mining; Massive Data; Rule Mining; Apriori Algorithm; Improved Apriori Algorithm;

I. INTRODUCTION

Searching for aurous in a mine of information (data) data mining can help; Data mining for gold is scooping through earth and rock for the valuable data. Data mining is sorting through huge datasets to evaluate valuable information. The action of data mining involves using statistical methods to identify patterns in data to help answer business questions, predict future trends and behavior. Data mining techniques (method) are used in business fields like marketing, risk management, fraud detection, cyber security, medical diagnosis, mathematics and research disciplines like cybernetics and genetics. Big data is data that include greater variety, arriving in increasing volumes (size) and with more velocity. Commonly, big data is larger, more composite data sets, especially from new data sources. These data sets are so commodious that conventional data processing software just can't manage them. Mining association rule is treated an essential research (study) method in the field of data mining utilized to access useful knowledge and characterize the association between different valuable data. Association rules which facilitate in marketing, advertising, inventory control, and fault prediction are based on the discovered frequent set of items. Association rule analysis is the most efficient branch, especially worthy of in-depth research. In recent years, although some research results have been accomplished, it still has assured research value.

However, although the association rule mining algorithm has made some achievements(performance), the amount(measure) of data is increasing, which leads to some shortcomings (Scanning the database frequently, Generating large number of candidate sets) and weakness of the existing association rule mining algorithm, which seriously affects the aspect of association rule mining. Because the amount of data in the database is expanding every day, and each record has 100s of attributes. Plus for the reason that the Partition of continuous attributes, the data of itemset is very extensive. The classical Apriori algorithm returns the whole database as the solution space, and needs to scan the database frequently, which makes the efficiency of the whole algorithm decrease. For That Reason, how to improve the Apriori algorithm to meet the real needs of data mining in the era of big data and improve the efficiency of the algorithm has become the prime concern of researchers.

In this paper, we introduce an improved Apriori Algorithm called DAA (Dynamic Apriori Algorithm). The main feature of this algorithm is we have to add itemset dynamically. That means we can create any type of dataset (Itemset) of our own interest. After scanning the database for the first time, the new database is generated. The judgment data set is added, and the irrelevant and meaningless candidate item set is removed at the same time, which reduces the time consumption cost.The experiment (analysis) compares the time consumed of traditional Apriori algorithm. The time consuming in improved Apriori in each group of transactions is less than compared to traditional Apriori, and the difference increases more and more as the number of transactions increases. Apriori algorithm that has been modified with combination reduction and iteration limitation techniques has proven to be more efficient in terms of time than the performance of unmodified algorithms in generating association rules.

II. RULE MINING

A. Related Definitions

- *Itemset*: Group of items that occur together
- *Association Rule*: Possibility that particular items are purchased together.
- *Support*: Sup(X) of an itemset X is the ratio of transactions in which an itemset develop to the total number of transactions.
- *Confidence*: Confidence in a rule is evaluated by dividing the probability of the items take place by the probability of the occurrence of the antecedent means if Y (antecedent) is present, what is the chance that X (Consequent) will also be present.
- *Transaction Database*: It stores the transaction data. Transaction data may also be stored in some other form than (m x n) database.

B. Mining process

There are some essential rules between the values of some variables in the data. It finds the interesting relationships between variables in large databases. It aims to use some Interesting measures to identify powerful rules found in databases. The task of mining association rules can be divided into two independent steps, mining frequent Itemsets: - Apply the minimum support threshold to find all frequent itemsets in the database. Rules generation: - Minimum confidence constraint is applied to these frequent itemsets to form rules.

III. TRADITIONAL APRIORI ALGORITHM AND ITS IMPROVEMENTS

A. Related Works

According to Adie Wahyudi Oktavia Gama and Ni Made Widnyani et.al [1], This Apriori algorithm uses knowledge from an itemset previously formed with frequent occurrence frequencies to form the next itemset. Modification techniques are needed to optimize the performance of a priori algorithms so as to get frequent itemset and to form association rules in a short time Apriori algorithm is a method of finding (discovering) frequent item sets from candidate sets. Apriori algorithm is an algorithm for mining or discovering frequent item sets. It is based on preceding knowledge to mine frequent item sets, iterative approach of layer by layer search is used.
According to Hongqin Wang and Lina Yuan et.al [2], the traditional mining algorithm of association rules for Apriori is analysed. It has low efficiency and poor expansibility, because it scans the database many times and produces a large number of redundant frequent itemsets when dealing with big data. According to Prof. Dr. K. Rajeswariet.al [3], Association rule mining will be an important aspect as data in the world is increasing day by day. Discusses about improved Apriori algorithm, new techniques used, how they are more efficient as compared to traditional Apriori algorithm. Improved Algorithm for Weighted Apriori, in traditional Apriori meaningless frequent itemset exist that increases the database scans and requires lots of storage space. According to Mohit Ohri and Komal Thakur et.al [4], indicates the limitation of the original Apriori algorithm of wasting time for scanning the whole database searching on the frequent itemsets. It presents an improvement on Apriori by reducing that wasted time depending on scanning only some transactions. According to Xiuli Yuan et.al [5], Apriori technique, mining frequent item sets and interesting associations in transaction database, it is composed of two sub-problems: 1. Discover frequent item set according to some pre-defined threshold. 2. Generate association rules satisfying the confidential constraint. The main contributions of this study are: Proposing a new searching strategy for accelerating the searching process and reducing the storage cost by using compacted vector structure.

B. Existing System

Apriori algorithm is an algorithm for mining or discovering frequent item sets. It is based on preceding knowledge to mine frequent item sets, iterative approach of layer by layer search is used. At first, according to the minimum support degree, the frequent 1 item set is registered as L_1, and then it is pushed forward layer by layer. The frequent 2-item set L_2 is generated (activated) by frequent 1-item set L1. Finally, frequent k+1 item set L_{k+1} is created by frequent k-item set L_k, till frequency can no longer be found Item set L_k. Apriori algorithm belongs to join step and pruning step. In order to increase (improve) the performance of generating frequent item sets, the apriori algorithm states that if an item set is frequent, then all its non-empty subsets are also frequent. Apriori algorithm is a method of finding (discovering) frequent item sets from candidate sets.

Fig.1. Association Rule Base Apriori Algorithm

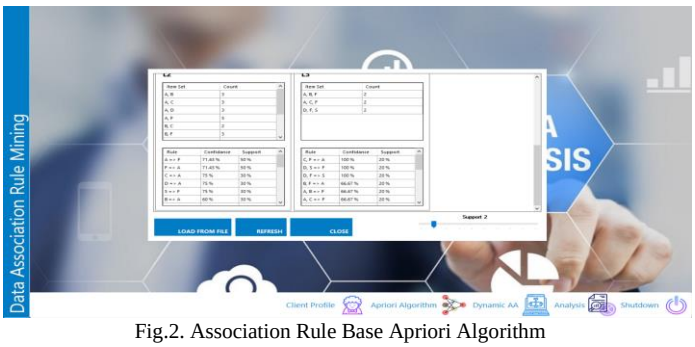

Fig.2. Association Rule Base Apriori Algorithm

The algorithm shows the (k+1) item set according to the k-item set using a procedure called layer-by-layer iteration of candidate generation tests.

STEP 1:- Apriori algorithm is a hierarchical iterative algorithm first, a set of frequent 1-item sets is found, which is denoted as L1, then L1 gets L2, L2 gets L3, and so on, until the frequent k-item set cannot be found [3]. Apriori algorithm mining produces all frequent items with no less than minimum support of minsup.
STEP 2:- Data are sorted in a transactional manner, the association is that the data are organized in the form of {Id, Item}, i.e. {trans. number, Item set}.
STEP 3:- Pruning method was adopted, make use of the property of frequent item set to optimize the search, because this optimization is the first in the algorithm, is called Apriori optimization. Apriori optimization is basically realized by
Pruning the candidate frequent item sets.
STEP 4:- Mining association rules relevant to transaction database.

C. Proposed System(Architecture)

In the existing data mining method (technology), there is some weakness in association rule mining methods. To decrease the impact of existing problems in Apriori algorithm and modifies the effectiveness of Apriori, many scholars have conducted a lot of research based on it and proposed some improved algorithms. Association rule mining algorithm has built (made) some achievements; the amount of data is increasing. It causes some shortcomings and deficiencies of the existing association rule mining algorithm, which critically affects aspects of association rule mining. Proposes a Software application based on market base analysis.

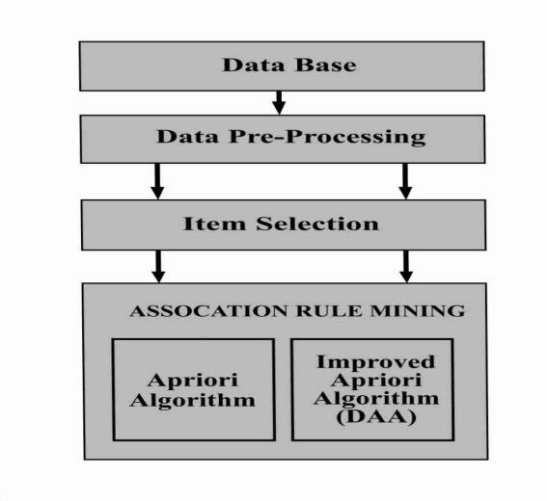

Fig.3. Proposed System Architecture

Data is a collection of a well-defined small unit of information. The database is a group of correlated data and a set of programs to access those data. Examples of DBMS (Data Base Management System) packages are dBASE, FoxPro, FoxBase, Oracle, Ms-Access etc. Data preprocessing is a data mining method which is used to convert the raw data in an efficient and useful format i.e. Data preprocessing is the process of transforming raw data into a recognizable format. It is also an important step in data mining as we can't work with raw data.
Association rule mining discovers interesting relationships and associations among large sets of data items. This rule shows how frequently an itemset exists (occurs) in a transaction. Market Based Analysis is a typical example. Market Based Analysis is one of the key method that used by large relations to reveal associations between items. It allows retailers to recognize relationships between the items that people buy together frequently. Association between the data: - There are some essential rules between the values of some variables in the data. It finds the interesting relationships between variables in large databases. It aim to use some Interesting measures to identify powerful rules found in databases. The task of mining association rules can be divided into two independent steps mining frequent Itemsets. Apply the minimum support threshold to find all frequent itemsets in the database, Rules generation: - Minimum confidence constraint is applied to these frequent itemsets to form rules. Data preprocessing is key factor in any data mining process as they s impact success rate of the project. Data is said to be unclean if it is attribute values, contain noise or outliers and duplicate or wrong data, missing attribute. Presence (existence) of any of these will degrade quality of the results. It is a data mining method that transforms raw data into an understandable format. Raw data (real world data) is always incomplete (insufficient) and that data cannot be sent through a model. So we need to preprocess data before sending through a model. In order to understand the working of architecture of improved apriori algorithm (DAA), it is essential to study the required blocks for architecture. The architecture of improved Apriori algorithm includes the business database, items selection and quantity selection, transactional database(Microsoft SQL Server dataset) two apriori algorithms such as Apriori algorithm and Dynamic apriori algorithm(we have to add itemset dynamically), Association rule mining(Frequent Itemset mining) and Result Analysis. Association rule adopt (uses) two criteria support and confidence to identify the relationships and rules are generated by analyzing data for frequent if/then pattern. Association rules are normally wants to satisfy a user specified minimum support and a user specified minimum confidence at the same (concurrent) time.

D. Improved Apriori Algorithm (DAA)

Step_1: Start
Step_2: Connect and authenticate database using Microsoft Server.
Step_3: Generate item set from different columns of multiple tables or database dynamically.
Step_4: Scan all the transactions.
Step_5: Initialize minimum support threshold value L1.
Step_6: Scan the database to check whether each item in transaction-library.
Step_7: Based on the minimum support threshold value, counting the occurrence of each item generate frequent 1 itemset (C1).
Step_8: Use new minimum support calculation by dividing the average number of transactions with total number of transactions.
Step_9: Then again candidate set generation is carried out and the 2-itemset which is generated known as C2.
Step_10: Again we will calculate the support of the 2- Itemset and we will prune C2 using minimum support and generate L2.

Step_11: The transactional database is scanned repeatedly, This is because every candidate of candidate set (C_k) generated after Join operation must be checked in all transactions of transactional database for the presence of candidate.
Step_12: At the end of the pass, determine which of the candidate item sets are actually large, and those become the seed for the next pass.
Step_13: Exit

Features of DAA includes

- Reduces the number of transactions scanning in iteration
- The Algorithm works on any type of databases with only one time scanning on databases.
- This method can merge two or more database or two or more table's scans to mine the frequent itemsets.
- In the similar item set and heterogeneous item set the algorithm work properly.
- The transactions which do not contain frequent itemsets are marked or removed
- It is a deep analysis method i.e. It will works in more than 4 layers in association rule after that it will give accurate results**.**
- Proposed system can run on any type of database.
- New itemset dynamically added to the dataset.

IV. RESULT ANALYSIS

The Association Rule Mining Apriori Algorithm has few drawbacks such as the iterations involved. We compared the performance of our advance algorithm with the Apriori Algorithm. The experimental platform is Intel Pentium® CPU B950 2.10GHz, 8GB RAM Windows 10 Operating System.

TABLE I. COMPARISON OF APRIORI AND IMPROVED APRIORI

k-ITEM SET	NO.OF TRANSACTION	APRIORI ALGORITHM (time (s))	DAA (IMPROVED APRIORI ALGORITHM) (time (s))
1	5	0.4	0.2
2	10	1.2	0.8
3	15	1.4	0.9
4	20	1.5	1
5	25	1.6	1
6	30	1..8	1.2

In Table I illustrate the comparison of traditional apriori and improved apriori based on the Number of transactions and time, here we take six groups of transactions applied to both algorithm.

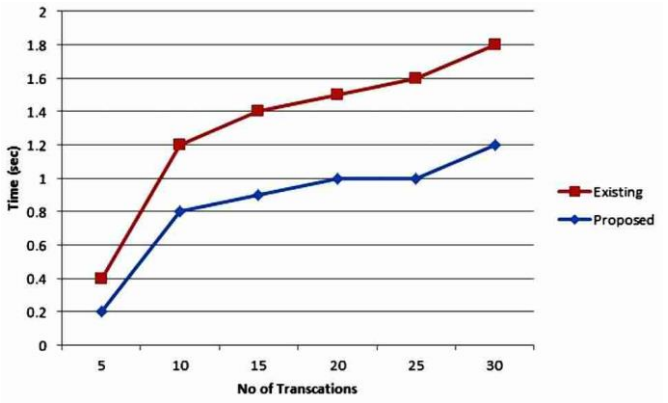

Fig.4. Graphical representation of Comparison between Existing System and Proposed System

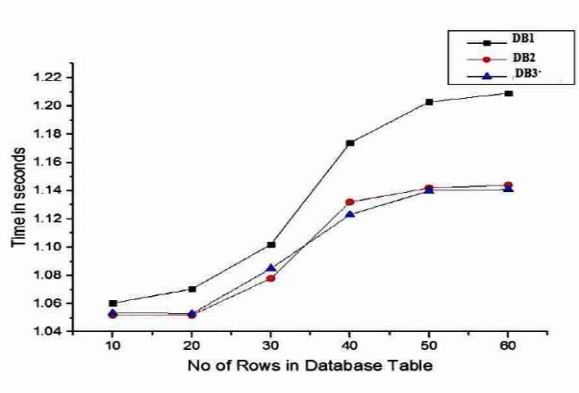

Fig.5. Graphical representation of Comparison between different dynamically added databases

Fig 4 represents the Graphical representation of Comparison between Existing System and Proposed System based on Table I. The time consuming in improved Apriori (DAA) in each group of transactions is less than it in the traditional Apriori. Fig 5 represents the Graphical representation of Comparison between different dynamically added databases based on number of rows in database table and time. DB1, DB2, DB3 represents we can take any type of databases (in traditional apriori algorithm processing is done by fixed data bases). In DAA we can create dynamic databases based on business databases (business database is not a fixed database), we can add rows and columns in the table according to our own interest. The algorithm adopts C# program and compiles in Visual Studio 2019 experiment environment with Microsoft SQL database. Correlation threshold discover its application in candidate item set generation. In improved Apriori algorithm, we include correlation threshold for discovering strong Association rules between the itemsets. The correlation threshold is a value between (0 and 1). If the value is 1, then the attributes are highly related to each Other. While a value close to zero displays the dataset as independent. This Correlation verifies the presence of all itemset appearing in traditional Apriori (classical Apriori) in proposed algorithm (DAA). The experiment compares the time consumed of original Apriori, and our improved algorithm by applying the six groups of transactions in the implementation. The time consuming in improved Apriori (DAA) in each group of transactions is less than it in the original Apriori, and the

difference increases more and more as the number of transactions increases.

V. CONCLUSION

The emergence of big data has come, in order to improve the speed and efficiency of big data mining. Design of using association rule mining algorithm is to discover the meaning and relationship of large datasets .The proposed method has run times and less memory space than Apriori algorithm method. The experiment analyzes the time consumed of original Apriori, our improved algorithm by applying the six groups of transactions in the implementation. The time consuming in improved Apriori (DAA) in each group of transactions is less than it in the original Apriori. The experiential results of the surveys show that the efficiency of the improved Apriori algorithm (DAA) is much higher than that of the traditional Apriori algorithm (Classical Apriori Algorithm).

REFERENCES

[1] Adie Wahyudi Oktavia Gama and Ni Made Widnyani, " Simple Modification For An Apriori Algorithm With Combination Reduction And Iteration Limitation ", 2020 https://doi.org/10.17977/um018v3i22020p89-98 ©2020 Knowledge Engineering and Data Science | W : http://journal2.um.ac.id/index.php/keds | E : keds.journal@um.ac.id

[2] Hongqin Wang and Lina Yuan, "Research on an Improved Algorithm of Apriori Based On Hadoop", 2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS)

[3] Prof. Dr. K. Rajeswari,"Improved Apriori Algorithm – A Comparative Study Using Different Objective Measures", IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3185-3191

[4] Mohit Ohri and Komal Thakur," An Enhanced Apriori And Improved Algorithm For Association Rules**"** International Research Journal of Engineering and Technology (IRJET**),** Volume: 03 Issue: 12 | Dec -2016

[5] **. Xiuli Yuan**, "An Improved Apriori Algorithm For Mining Association Rules " ,AIP Conference Proceedings **1820**, 080005 (2017); https://doi.org/10.1063/1.4977361 Published Online: 13 March 2017