Survey Paper for Enhancement of Apriori Algorithm

Prof. Savita Rathod¹, Ashutosh Sharma²
¹,²Department of Computer Science & Engineering, Truba College of Engineering & Technology, Indore, INDIA

Abstract
As with the advancement of the IT technologies, the amount of accumulated data is also increasing. It has resulted in large amount of data stored in databases, warehouses and other repositories. Thus the Data mining comes into picture to explore and analyze the databases to extract the interesting and previously unknown patterns and rules known as association rule mining. In data mining, Association rule mining becomes one of the important tasks of descriptive technique which can be defined as discovering meaningful patterns from large collection of data. Mining frequent item set is very fundamental part of association rule mining. As in retailer industry many transactional databases contain same set of transactions many times, to apply this thought, in this thesis present an improved Apriori algorithm that guarantee the better performance than classical Apriori algorithm.

Keywords--- Apriori, Machine learning, Data Mining

I. INTRODUCTION

With the enhance in Information Technology, the size of the databases created by the organizations due to the accessibility of low-cost storage and the development in the data capturing technologies is also increasing. These association sectors include retail, fuel, telecommunications, utilities, manufacturing, transport, credit cards, insurance, banking and many others, extracting the valuable data, it required to explore the databases completely and efficiently. Knowledge discovery in databases (KDD) helps to identifying valuable information in such huge databases.

1.1 DATA MINING

Data mining is the main part of KDD. Data mining normally involves four classes of task; classification, clustering, regression, and association rule learning.

Data mining as a field of study involves the integration of ideas from many domains rather than a pure discipline the four main disciplines [1], which are contributing to data mining include:

- Statistics: it can make available tools for measuring importance of the given data, estimating probabilities and many other tasks (e.g. linear regression).
- Machine learning: it provides algorithms for inducing knowledge from given data (e.g. SVM).
- Data management and databases: in view of the fact that data mining deals with huge size of data, an efficient way of accessing and maintaining data is needed.
- Artificial intelligence: it contributes to tasks involving knowledge encoding or search techniques (e.g. neural networks).

1.1.1 The Primary Methods of Data Mining

Data mining addresses two basic tasks: verification and discovery. The verification task seeks to confirm user’s hypotheses. While the finding task searches for unidentified knowledge hidden in the data. In general, discovery task can be further divided into two categories, which are descriptive data mining and predicative data mining.

- Descriptive data mining describes the data set in a summary manner and presents interesting general properties of the data.
- Predictive data mining constructs one or more models to be later used for predicting the behavior of future data sets.
- There are a number of algorithmic techniques existing for each data mining tasks, with features that must be weighed against data characteristics and additional business requirements. Among all the techniques, in this research, we are focusing on the association rules mining technique, which descriptive mining technique, with transactional database system.

1.1.2 Fundamental Components of Data Mining Technology

It is fundamentally important to declare that the prime key to understand and realize the data mining
technology is the ability to make different between data mining, operations, applications and techniques [2], as shown in Fig 1.

1.2 APRIORI ALGORITHM

1.2.1 Introduction of Apriori algorithm

Apriori is a classic algorithm for learning association rules in data mining. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules [11]. The Apriori algorithm is a classical data mining method for association rule discovery typically applied to market basket data, such as the study of what products tend to be purchased together in an online market place (e.g. Amazon etc).

There are two properties: “all nonempty subset of a frequent itemset must also be frequent; all superset of non frequent itemset must also be non-frequent” the properties is used in Apriori algorithm to scanning the database, resulting in Boolean association rules frequent itemsets.

Specifically, Apriori uses an iterative search method layer by layer, where k-dimensional itemsets are used to explore (k+1)-dimensional itemsets. First, the set of frequent 1-dimensional itemsets is found and denoted L1. Next, L1 is used to find L2, the set of L2 frequent 2-itemsets, which is used to find L3, and so on until no more frequent k-dimensional itemsets can be found[13]. Finally, getting the rules from large set of data items. How Li-1 is used to find Li is consisting of two step process, join and prune actions as followed [14]:

1. The join step: Join Lk-1 with itself, than combine the same extension item appeared to generate a possible candidate k-dimensional itemsets, this set of candidates is denoted Ck, Ck ⊇ Lk.
2. The prune step: Scan the database to determine the count of each candidate in Ck. When the count is less than the minimum support count, it should be delete from the candidate itemsets.

1.3 Classical Apriori Algorithm

Apriori employs an iterative approach known as a level-wise search [15], where k-dimensional itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found. This set is denoted L1. L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of the database. In order to find all the frequent itemsets, the algorithm adopted the recursive method. The main idea is as follows [16]:

L1 = {large 1-itemsets};
for (k=2; Lk-1 ≠Φ; k++) do
{
Ck=Apriori-gen (Lk-1); // the new candidates
for each transactions t∈ D do // scan D for counts
{
Ct=subset (Ck, t);
// get the subsets of t that are candidates
for each candidates c∈ Ct do
c.count++;
}
Lk={c∈ Ck |c.count≥minsup}
}
Return=∪k Lk;

All noneempty subsets of a frequent item sets must also be frequent. To reduce the size of Ck, pruning is used as follows. If any (k-1)-subset of a candidate k-itemsets is not in Lk-1, then the candidate cannot be frequent either and so can be removed from Ck. The prune step reduces the cost of calculating all the support of candidate sets by reducing the size of candidate sets, which significantly improves the performance of finding frequent itemsets.

---

Fig. 1. Components of Data mining

---

Fig. 2. Flow chart of classical Apriori algorithm
1.3.1 Advantages of Apriori
- Easy implementation.
- Initial Information- transaction database D and user-defined minimum support threshold Min_supp.
- Algorithm uses information from previous steps to produce the frequent itemsets [18].

1.3.2 Limitations of Apriori
- In case of large dataset, this algorithm is not efficient [19].
- Apriori algorithm requires large no of scans of dataset [19].
- In case of large dataset, Apriori algorithm produce large number of candidate itemsets. Algorithm scan database repeatedly for searching frequent itemsets, so more time and resource are required in large number of scans so it is inefficient in large datasets [20].

II. RELATED WORK

One of the most well known and popular data mining techniques is the Association rules or frequent item sets mining algorithm. The algorithm was originally proposed by Agrawal et al. [21] [22] for market basket analysis. Because of its important applicability, many revised algorithms have been introduced since then, and Association rule mining is still a widely researched area.

Agrawal et. al. [22] developed various versions of Apriori algorithm such as Apriori, AprioriTid, and AprioriHybrid. Apriori and AprioriTid generate item sets using the large item sets found in the preceding pass, without consider the transactions

Park. J. S et.al [23] find out that different versions of Apriori were available, the problem with Apriori was that it generates too many 2-item sets that were not frequent.

Scalability is a different important area of data mining because of its huge size. Hence, algorithms should be able to “scale up” to handle large amount of data. Eui-Hong et. al [24] tried to create data distribution and candidate distribution scalable by Intelligent Data Distribution (IDD) algorithm and Hybrid Distribution (HD) algorithm respectively.

An further scalability study of data mining was reported by introducing a light-weight data structure called Segment Support Map (SSM) with the purpose of reduces the number of candidate item sets required for counting [25].

The problem of mining with Association rules is a natural fit. in addition Association rule mining Evolutionary algorithms were also reported that can generate association rules [26]. It allows overlapping intervals in different item sets.

III. PROBLEM DEFINITION

The traditional Apriori algorithm is generally used for association mining technique by using bottom up approach. This bottom-up Apriori- algorithms needs to combine small frequent item sets and generates various candidate item sets. Every candidate item set needs to compare with the all item sets of the database. The basic Apriori algorithm which suffers from increased number of data base scan. This algorithm is not useful for large amount of item set. Therefore, it is low efficiency and only suited to the database contain smaller item sets.

IV. PROPOSED APPROACH

The major objective of the research is to develop and propose a new idea for mining the association rules out of transactional data set. The proposed method is based on Improved Apriori approach. The proposed method is more efficient than classical Apriori algorithm. To achieve the research objective successfully, a series of sequence progresses and analysis steps have been adopted.

V. CONCLUSION

In this thesis, we measured the following factors for creating our new idea, which are the time and the no of iteration, these factors, are affected by the approach for finding the frequent itemsets. Work has been done to develop an algorithm which is an improvement over Apriori with using an approach of improved Apriori algorithm for a transactional database. According to our clarification, the performances of the algorithms are strongly depends on the support levels and the features of the data sets (the nature and the size of the data sets). Therefore we employed it in our scheme to guarantee the timsaving and reduce the no of itertion Thus this algorithm produces frequent itemsets completely. Thus it saves much time and considered as an efficient method as proved from the results. . We can summarize the main contribution of this research as follows:

- To study and examine various existing approaches to extract frequent itemsets.
- To devised a new better scheme than classical Apriori algorithm approach for mining frequent itemsets

REFERENCES

[4] Jiawei Han, Micheline Kamber, Morgan Kaufmann : Data mining Concepts and Techniques , 2006.