Improved Version of Apriori Algorithm Using Top Down Approach

Mr.kailash Patidar[1] Mr.Gajendra singh[2] Jatin Khalse[3]
[1]Associate Professor [2]Head of the Department [3]M.Tech Scholar
[1,2,3]Department of Computer Science & Engineering
[1,2,3]Sri Satya Sai Institute of Science & Technology, Sehore, 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.

Key words: Data Mining, Apriori algorithm, Novel Approach

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.

A. Data Mining:

Data miniing 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) 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 summery 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 is descriptive mining technique, with transactional database system.

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, application and techniques [2], as shown in Fig 1.

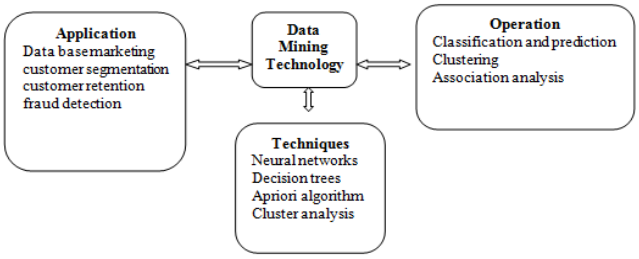

Fig1. Components of Data mining

B. Association Rule Mining:

The association rule of data mining is a elementary topic in mining of data [3]. Association rule mining discovery frequent patterns, associations, correlations, or fundamental structures along with sets of items or objects in transaction databases, relational databases, and other information repositories [4].

A lot of studies have been done in the region of association rules mining. First introduced the association rules mining in [5,6]. Many studies have been conducted to address various conceptual, implementation, and application issues relating to the association rules mining task.

The overall performance of mining association rules is determined primarily by the first step. The second step is easy. After the large itemsets are identified, the corresponding association rules can be derivative in straightforward manner. Our main consideration of the this

is is First step i.e. to find the extraction of frequent itemsets [9].

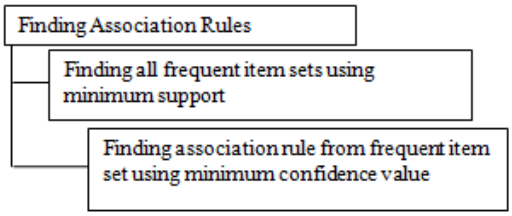

Fig 2. Generating Association rules

C. Apriori Algorithm:

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 on-line 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 frequentitemset 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

2) Classical Apriori Algorithm:

Apriori employs an iterative approach known as a level-wise search [15], where k-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=UkLk;
```

All nonempty subsets of a frequent itemsets 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.

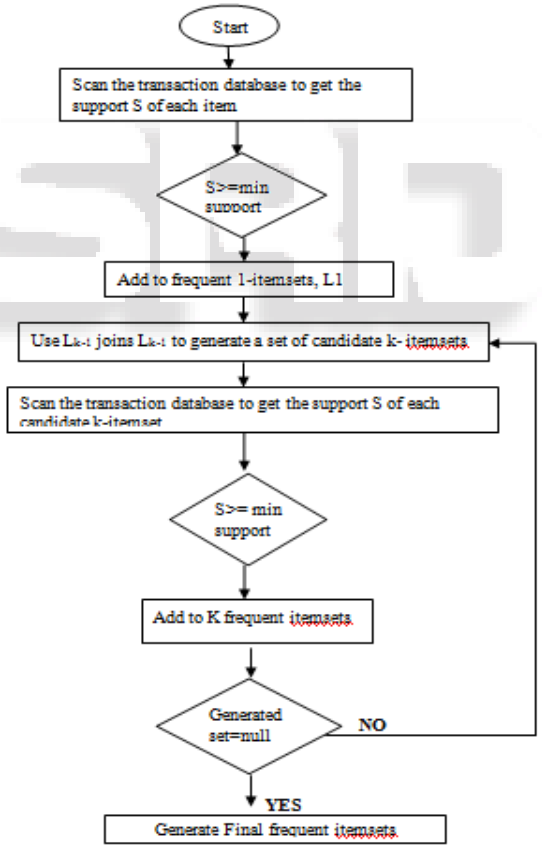

Fig 3. Flow chart of classical Apriori algorithm

3) 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].

4) 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, Apriorialgorithm 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. Implementation of Novel 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. Figure 4.depicts the method to mine frequent itemsets from the transactional data set using the new method.

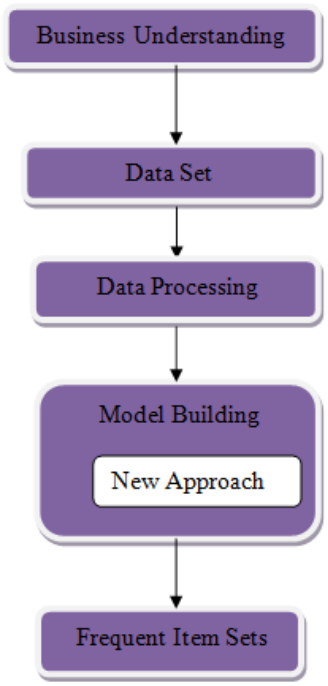

Figure 4.: Method used to mine frequent itemsets

A. Implementation of Improved Apriori Algorithm:

The improved Apriori algorithm is usually used for association mining technique by using top down approach. The top down Apriori algorithms requirements to large frequent item sets and generates frequent candidate item sets. The improved Apriori algorithm which reduce unnecessary data base scan. This algorithm is useful for large amount of item set. Therefore, improved top down algorithm uses less space, less number of iteration.

Pseudo Code:

```
Input: database (D), minimum support (min_sup).
Output: frequent item sets in D.
L1= frequent item set (D)
j=k; /* k is the maximum number of element in a
transaction from the database*/
for k= maxlength to 1
{
fori=k to 2
{
for each transaction Ti of order i
{
if (Ti has repeated)
{
Ti.count++;
}
m=0;
while (i<j-m)
{
if (Ti is a subset of each transaction Tj-m of order
j-m)
Ti.count++; m++; }
If (Ti.count>=min_sup)
Rule Ti generated

}
```


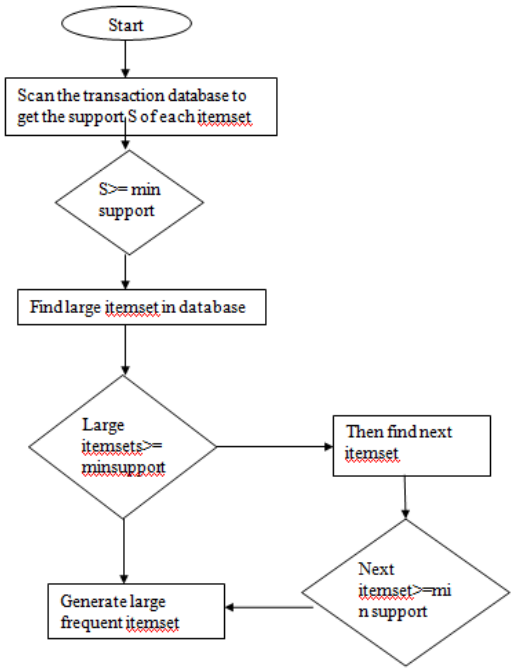

Figure 5 Flow chart of improved Apriori algorithm

IV. 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 time saving and reduce the no of itertionThus 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

[1] Tan P.N., Steinbach M., and Kumar V: Introduction to data mining, Addison Wesley Publishers, 2006.

[2] Han J. &Kamber M.: Data Mining Concepts and Techniques, First edition, Morgan Kaufmann publisher, USA 2001.

[3] Ceglar, A., Roddick, J. F: Association mining ACM Computing Surveys, volume 38(2) 2006.

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

[5] A. Savasere, E. Omiecinski and S. Navathe. : An efficient algorithm for mining Association rules in large databases, In Proc. Int‟l Conf. Very Large Data Bases (VLDB), Sept. 1995, p.p 432–443.

[6] Agrawal.R andSrikant R.: Fast algorithms for mining association rules,In Proc. Int‟l Conf. Very Large Data Bases (VLDB), Sept. 1994, p.p 487–499.

[7] Lei Guoping, Dai Minlu, Tan Zefu and Wang Yan: The Research of CMMB Wireless Network Analysis Based on Data Mining Association Rules, IEEE conference on Wireless Communications, Networking and Mobile Computing (WiCOM),ISSN :2161- 9646 Sept. 2011,p.p 1-4.

[8] Divya Bansal, LekhaBhambhu : Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, ISSN: 2277 128X September 2013 .

[9] Shweta, Dr. Kanwal Garg: Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms International Journal of Advanced Research in Computer Science and Software Engineering 3(6), June – 2013, pp. 306-312.

[10] SurajP . Patil1, U. M. Patil2 and SonaliBorse: The novel approach for improving Apriori algorithm for mining association Rule, World Journal of Science and Technolog 2(3), ISSN: 2231 – 2587, 2012, p.p75- 78.

[11] Toivonen .H. :Sampling large databases for association rules,In Proc. Int‟lConf Very Large Data Bases (VLDB), Bombay, India, Sept. 1996, p.p 134–145.

[12] Yanfei Zhou, Wanggen Wan, Junwei Liu, Long Cai: Mining Association Rules Based on an Improved Apriori Algorithm 978-1-4244-585 8- 5/10/ IEEE 2010 .

[13] Luo Fang: The Study on the Application of Data Mining Based on Association Rules, International Conference on Communication Systems and Network Technologies (IEEE) ,may 2012,p.p 477 - 480 .

[14] Chengyu and Xiong Ying: Research and improvement of Apriori algorithm for Rules, In Intelligent Systems and Applications (ISA), 2nd International Workshop on, may 2010, p.p 1 -4.

[15] Jaishree Singh, Hari Ram, Dr. J. S. Sodhi: Improving Efficiency of Apriori Algorithm Using Transaction Reduction, In proceeding of International Journal of Scientific And Research Publication (IJSRP), ISSN 2250-3153, Volume 3, Issue 1, January 2013,p.p1-4.

[16] Partibha Parikh ,Dinesh Waghela: Comparative Study of Association Rule Mining Algorithms. In: Proceeding of UNIASCIT, ISSN 2250-0987, Vol. 2, Issue 1, 2012, p.p170-172.

[17] NiklasOlofsson :Implementation of the Apriori algorithm for effective item set mining, In Vigi Base TM august 2010,p.p 1-29.

[18] K. Geetha, Sk. Mohiddin: An Efficient Data Mining Technique for Generating Frequent Item Sets, In: Proceeding of IJARCSSE, ISSN 2277-128X, Vol. 3, Issue 4, April 2013,p.p 571-575.

[19] MamtaDhanda, SonaliGuglani, Gaurav Gupta: Mining Efficient Association rules ThroughApriori Algorithm Using Attributes, In: Proceeding of IJCST, ISSN 0876-8491, Vol. 2, Issue 3, September 2011,p.p342-344.

[20] SuhaniNagpal :Improved Apriori Algorithm Using Logrithmic Decoding and Pruning, In: Proceding of International Journal of Engineering Research and Applications, ISSN 2248-9622, Vol. 2, Issue 3, May-June 2012, pp. 2569-2572.

[21] Agrawal, R., Imielinski, T., and Swami, A. N.: Mining Association Rules Between Sets of Items in Large Databases, Proceedings of the ACM SIGMOD, International Conference on Management of Data, 1993, pp.207- 216.

[22] Agrawal. R. andSrikant. R.: Fast Algorithms for Mining Association Rules, Proceedings of 20th International Conference of Very Large Data Bases ,1994, pp.487-499 .

[23] Park. J. S, M.S. Chen, P.S. Yu.:An effective hash-based algorithm for mining association rules ,In Proc. ACM- SIGMOD Int‟lConf. Management of Data (SIGMOD), San Jose, CA, , May 1997, p.p 175–186.

[24] Eui-Hong Han, George Karypis and Kumar, V. : Scalable Parallel Data Mining for Association Rules, IEEE Transaction on Knowledge and Data Engineering, 12(3), 2000, pp.728- 737.

[25] Lakshmanan, V., S., Carson Kai-Sang, L., and T. Raymond: The Segment Support Map : Scalable Mining of Frequent Itemsets, Journal of ACM SIGKDD Explorations Newsletter, 2(2), 2000, pp.21-27.

[26] Mata, J., Alvarez, J. L., and Riquelme, J. C. Evolutionary: An Evolutionary Algorithm to Discover Numeric Association Rules, Proceedings of ACM Symposium on applied Computing, 2002, pp. 590-594.

