

Association Rule Mining using Optimization Techniques

M. Sunitha¹, Sunita A Yadward²

M.Tech Student, CSE Department, Pragathi Engineering College, Peddapuram, Andhra Pradesh¹

Associate Professor, CSE Department, Pragathi Engineering College, Peddapuram, Andhra Pradesh²

Abstract: Data mining is a hot topic of research for past many years. Amongst the various algorithms popular and the researchers in progress in data mining Association rule mining (ARM) plays an important role due its tremendous publicity. It aims at extraction, discovery of hidden relation, exploration of interesting association between the existing items in a transactional database. It is used to generate frequent items and a set of rules to find frequent items. The entire purpose of this study is to highlight the fundamentals of association rule mining, compare the various modifications proposed to association rule mining approaches. The results generated by apriori algorithm can be further optimized using optimization algorithms. The study intends to determine the minimum support and minimum confidence values for mining association rules using the optimization algorithms. These algorithms are mainly defined for improving the performance of the Apriori algorithm. They minimize the quantization errors and fitness value to be improved. The algorithm improves the result produced by apriori algorithm. The major area of concentration is to optimize the rules generated by association rule mining (apriori method).

Keywords: Data mining ,Association RuleMining, Apriori Algorithm,Optimization Algorithms.

1. INTRODUCTION

Data Mining is the one of the most important research area in present era. In knowledge discovery process association rules mining play an important role to identify the rules and frequent items. Association rule mining was introduced by Agarwal in 1993 [1]. It aims to extract correlations, frequent patterns, associations among sets of items in the transaction databases.

Association rules have been extensively studied in the literature for the usefulness in many application domains such as market basket analysis, intrusion detection, diagnosis, decision support etc.

To select the best rule among all the available rules, association rule mining depends on two important constraints. They are minimum threshold on support and confidence. Since the database is large in size and customers are interested only on frequently purchased items. Users can predefine the threshold of support and confidence. To eliminate the rules which are not interesting to the user. These two thresholds are called as Minimal Support and Minimal Confidence.

Support(s):

Support of an association rule is defined as the percentage of the fraction of records that contain to the total number of records in the database. For example if support of an item is 0.1% then only 0.1 percent of the transaction contain purchasing of this item.

$$\text{Support}(XY) = \frac{\text{Support Count of}(XY)}{\text{Total Number of transactions in D}}$$

Confidence:

Confidence(C) of an association rule is defined as it is used to find the important relationships of the items. The confidence rule is given as $X \Rightarrow Y$, with respect to the total transactions in database D, It is the proportion of transactions that contain X along with Y. For example the confidence of the association rule $X \Rightarrow Y$ is 80% it means that 80% of the transactions that contain X along with Y.

$$\text{Confidence}(X/Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

2. LITERATURE SURVEY

2.1 Apriori Algorithm:

One of the common and most widely used algorithms of association rule mining is apriori algorithm. It was first introduced by R.Agrawal and S.Srikant in 1994. The principle of a priori algorithm is "if an item set is frequent then all the subsets of frequent item set must be frequent". It uses level wise search which is known as "breadth –first search". Here (k+1) item sets are generated from k item sets. First the database is scanned and support count for every individual item is calculated and stored as frequent 1-itemset and set as L1. Now from this L1 frequent 2-itemset is found, here all the items that are below the min sup will be removed, and is set as L2. This process continues until no frequent k item sets are found. A priori algorithm mainly consists of two steps they are Join: Ck is generated by joining Lk-1 with itself Prune: if any (k-1) itemset that is not frequent cannot be a subset of a frequent k-item set [2].

**Advantages:**

It is easy to implement and understand.

Disadvantage:

It requires more number of database scans which increases time and decreases efficiency and increases I/O cost and requires more memory.

To overcome all the above disadvantages many refinements are made on classical a priori algorithm.

2.2 Improvements in Apriori Algorithm

2.2.1 -Reduced Apriori Algorithm With Tag(Raat)-Refinement In Increasing Efficiency Of Support And Reducing Redundant Pruning: Wanjun Yu, Xiaochun Wang And Fangyi Wang, Erkang Wang And Bowen Chen (2008) Introduced "REDUCED APRIORI ALGORITHM WITH TAG(RAAT)". RAAT uses a priori operation to form candidate item sets which results in reducing the pruning operation. To increase speed of calculating the support, RAAT also uses the concept of Tag. So RAAT uses less time and has more efficiency^[3].

2.2.2: Improved Apriori Algorithm-

Refinement In Reducing Memory Utilization And Transaction: Jaishree Singh, Hari Ram, Dr. J .S Sodhi (2003) introduced a new algorithm named "IMPROVED APRIORI ALGORITHM". This algorithm truncates scanning time of the database and minimizes all the transactions that are not related. A new attribute "size of transaction" is introduced. This carries number of items in a particular transaction. So I/O cost gets reduced and efficiency increases^[4].

2.2.3: Refinement Based on Customer Habits - Shuo Yang et al (2012) introduced several new theorems to improve traditional Apriori algorithm. This minimizes database scans based on customer habits. To find frequent item sets relative theorem is used. A shopping site has to be created to use improved Apriori algorithm on online. The reason is based on current purchasing item system will predict the next choice of the customer. So customers can easily generate association rules which save time and increases efficiency^[5].

2.2.4 - Refinement In Reducing Redundant Operation: Yanfei Zhou, Wanggen Wan, Junwei Liu and Long Cai(2010): Introduced three segments firstly in the generation of frequent item sets number of required to the database is reduced secondly pruning /reducing frequent item set and at last database gets optimized. When both classical and improved a priori algorithms are compared with each other based on different support degree, different number of trading services and on different items it is proved that improved i apriori algorithm is more efficient than classical algorithm. So redundant operations are reduced while generating strong association rules. The advantage is Performance and Efficiency is increased^[6].

2.2.5-Algorithm Basing on Function Interest: Wei-Min and Zhu-Ping Liu (2008)

The two conditions on which classical apriori algorithm depends on are minimum support and minimum confidence to generate association rules. Sometimes we may require strong association rules and sometimes we require less strong association rules. Two revised algorithms were introduced to beat this requirement they are AMS (Algorithm That Mines Strong Association Rules), AMLS (Algorithm That Mines Less Strong Association Rules). Both these algorithms depend on three constraints they are minimum support, minimum confidence, minimum interest and works in matrix form. The major advantage of these two algorithms is Decreases time while scanning the database^[7].

With the above discussion it is proved that all the limitations of classical apriori algorithm enhanced.

2.3 Frequent Pattern Growth Algorithm:

Fp growth algorithm Han et al 2000 this is another major and commonly used algorithm of association rule mining that uses tree structure. This mining technique generates frequent patterns without candidate generation. Fp growth technique is mainly implemented to overcome the limitations of apriori algorithm^[8]. To generate frequent patterns Fp growth technique requires two passes they are

Pass-1

- Data is scanned and support for every item is counted.
- Infrequent items are pruned.
- Based on support frequent items are stored based on decreasing order

Pass-2

- Reads one transaction at one time.
- As it uses a fixed order paths can be shared.
- Pointers can be used between the nodes that share common item.
- Infrequent items are taken out from list.

Advantages:

- Does not require candidate generation.
- Its execution is fast compared to a priori algorithm.

Disadvantages:

- Fp-tree requires more memory space.
- As it's a tree structure it is difficult to insert a new item into tree.

2.3.1-Improved FP-Growth Algorithm

Yi Zeng, Shiqun Yin, Jiangyue Liu, and Miao Zhang: To enhance the Fp-growth algorithm improved Fp-growth algorithm was introduced. Two kinds of improved Fp-growth algorithms were used they are N-Painting Growth Algorithm and Painting Growth Algorithm^[9].

2.3.2 N-Painting Growth Algorithm:

To find out the association sets of all frequent items this algorithm set up permutation sets of two items and according to these association sets, this algorithm digs up all the items that turn to be frequent.



Painting Growth Algorithm: to find association sets of all the frequent items this algorithm first construct association picture of the two item permutation and then digs all the items that are frequent according to association sets. These two algorithms scans database only once while Fp-growth requires two scans to the database. And it is very easy to construct and requires less memory space^[9].

Optimization gives a set of solutions to any problem, finding the best solution among all the solutions

3. OPTIMIZATION TECHNIQUES

3.1 Ant colony optimization

The association rules that are generated through apriori algorithm are optimized through Ant Colony Optimization algorithm. It uses a meta-heuristic technique to solve tough combinatorial optimization problems. To make better decisions good rules are helpful. To improve the rules that are generated through association rule mining (apriori algorithm) a new optimization technique was proposed based on ant colony that is probabilistic technique. It is introduced to optimize the rules that are generated through apriori algorithm. Here time is taken as the main factor. All the rules that are generated are compared with time, if rules are decreased it means that time of work decreases. So it's better than apriori algorithm as it takes more time in generating rules^[10].

3.2 Genetic Algorithm:

To optimize the rules that are generated through association rule mining **Manish Saggur et.al** used genetic algorithm. Normally in apriori algorithm, when rules are generated they do not contain any negative attributes, so when genetic algorithm is applied to these rules the system has the ability to predict negative attributes. This improvement mainly helps the rules under classification. To generate rules that contain negative attributes the authors first generated rules using apriori algorithm and then applied genetic algorithm to the generated rules. Database they used is produced synthetically. When the proposed technique is applied on synthetic database ,all the rules that contain negative attributes and rules generated through association rule mining are included. When this algorithm is used on distributed computing some changes are to be done to reduce its complexity^[11].

3.3 Particle swarm optimization

Particle swarm optimization is an optimization technique proposed by the social behavior of fish schooling and bird flocking. It was first initiated by Dr.Eberhart and Dr.Kennedy in the year 1995 and is a population based Meta -heuristic algorithm. It depends on swarm intelligence. It's easy to implement PSO as it was a bio-inspired optimization technique and has high convergence rate to find the best solution. With the help of PSO we can derive best solution for extremely tough problems. This algorithm consists of swarm of particles that is group of random particles in the search space where every single solution is a bird. Optimized solution for every particle

can be calculated by fitness function. Every Particle has a velocity that directs the sight of flight. Every particle will have a random velocity that is adjusted according to its flying experience and experience of its neighbor. Every particle will have two best values they are pbest(personal best) and gbest(global best).pbest is nothing but the best solution of the particle that is calculated so far with fitness function, gbest is nothing but the best solution that is obtained so far by any particle. Now particles will update with new values^[12].

4. CONCLUSION

The paper is a thorough survey of papers on association rule mining, rule generation and optimization. The algorithmic aspects of association rule mining are reviewed in this paper. We discuss different methods of apriori and frequent set item generation and optimization algorithm. We also survey related research in the direction of rule optimization and provide the analysis that Rule optimization can be applied using ACO, Genetic, and PSO algorithm and compare with the genetic algorithm. Through the Application of rule optimization, we can obtain reduced rules which are helpful .This kind of surveyed approach may be lead to various possibilities of architectural alternatives in future and these methods are play a very useful role in Data Mining to minimize the harmful impacts and maximizing the possible benefits.

REFERENCES

- [1] Rakesh Agrawal , et al., "Mining Association Rules between Sets of Items in Large Databases" Proceedings of the ACM SIGMOD Conference,1993.
- [2] Herbert A. Edelstein "Introduction to Data Mining and Knowledge Discovery" Third Edition by Two Crows Corporation,2000.
- [3] Wanjun Yu et al., "The Research of Improved Apriori Algorithm for Mining Association Rules" 2008 11th IEEE International Conference on Communication Technology Proceedings.
- [4] Jaishree Singh, et al., "Improving efficiency of Apriori algorithm using Transaction Reduction" International Journal of Scientific and Research Publications, Volume 3, Issue 1,January 2013.
- [5] Shuo Yang, "Research and Application of Improved Apriori Algorithm to Electronic Commerce" 2012 11th International Symposium on Distributed Computing and Applications to Business, 2012.
- [6] Yanfei Zhou, et al., "Mining Association Rules Based on an Improved Apriori Algorithm", IEEE Conference in 2010.
- [7] Wei-Min Ma, Et Al.,"Two Revised Algorithms Based On Apriori For Mining Association Rules", Proceedings of the Seventh IEEE International Conference on Machine Learning and Cybernetics, 2008.
- [8] Pang-Ning Tan, et al.,Introduction to Data Mining, Addison-Wesley. Chapter 6: Association Analysis: Basic Concepts and Algorithms, 2003.
- [9] Yi Zeng, et al., "Research of Improved FP-Growth Algorithm in Association Rules Mining " Hindawi Publishing Corporation Scientific Programming Volume 2015 (2015), 2015.
- [10] Shivika, et al., " Optimization of Association Rule Using Heuristic Approach", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 4 Issue: 8, 2016.
- [11] Manish Saggur, et al., "Optimization of Association Rule Mining using Improved Genetic Algorithms", IEEE International Conference on Systems, Man and Cybernetics, pp. 3725- 3729, 2004.
- [12] Kennedy, J, et al., "Particle Swarm Optimization ". Proceedings of IEEE International Conference on Neural Networks, 1995.