ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 6, June 2013

Association Rule Mining using Apriori Algorithm: A Survey

Charanjeet Kaur

Abstract— **Association rule mining is the most important technique in the field of data mining. Association rule mining finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. In this paper we present a survey of recent research work carried by different researchers. Of course, a single article cannot be a complete review of all the research work, yet we hope that it will provide a guideline for the researcher in interesting research directions that have yet to be explored.**

Keywords— **Association Rules, Confidence, frequent items, Item set, Minimum Support.**

.

I. INTRODUCTION

Association rules are one of the major techniques of data mining. The volume of data is increasing dramatically as the data generated by day-to-day activities. Therefore, mining association rules from massive amount of data in the database is interested for many industries which help in much business can decision making processes, such as cross marketing, Basket data analysis, and promotion assortment. It helps to find the association relationship among the large number of database items and its most typical application is to find the new useful rules in the sales transaction database, which reflects the customer purchasing behaviour patterns, such as the impact on the other goods after buying a certain kind of goods. These rules can be used in many fields, such as customer shopping analysis, additional sales, goods shelves design, storage planning and classifying the users according to the buying patterns, etc. The techniques for discovering association rules from

The data have traditionally focused on identifying relationships between items telling some aspect of

Human behaviour, usually buying behaviour for determining items that customers buy together. All

Rules of this type describe a particular local pattern. The group of association rules can be easily interpreted and communicated.

Manuscript received June, 2013.

***Charanjeet Kaur**, Department of Computer Engineering, Yadawindra College of Engineering, Talwandi Sabo, Punjabi University, Patiala, Bathinda, India.s*

II. ASSOCIATION RULES

In Association Rule mining find rules that will predict the occurrence of an item based on the occurrence of the other items in the transaction. Table shows Market-Basket Transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
2	Milk , Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules:
{Diaper} → {Beer},
{Bread, Milk} → {Egg, Coke},
{Bread, Beer} → {Milk},
Implication means co-occurrence, not causality. Association rule is an implication expression of the form $X \rightarrow Y$, where X and Y are itemsets.
Example: {Milk, Diaper} → {Beer}

Rule Evaluation:

- **Support (S):** Fraction of transactions that contain both X and Y.
- **Confidence (C):** Measures how often items in Y appear in transactions that contain X. Example:
  {Milk, Diaper} → {Beer}
  S = σ ({Milk, Diaper, Beer}) / [T]
  S = 2/5 S = 0.4
  C = σ ({Milk, Diaper, Beer} / σ ({Milk, Diaper}
  S = 2/3 S = 0.67
- **Itemset:** A collection of one or more items. Example {Milk, Diaper, Beer}. K-itemset that contains k-items.
- **Frequent Itemset:** An itemset whose support is greater than or equal to a min_sup threshold. In association rule mining task from a set of transactions T, the goal of association rule mining is to find all rules having Support >= min_sup threshold and Confidence>= min_conf threshold.

There are two phases in the problem of data mining association rules.

1. Find all frequent itemsets: i.e. all itemsets that have support s above a predetermined minimum threshold.
2. Generate strong association rules from the frequent itemsets: these association rules must have confidence c above a predetermined minimum threshold.

After the large item sets are identified, the corresponding association rules can be derived in a
relatively straightforward manner. Thus the overall
Performance of mining association rules is determined primarily by the first step. Efficient counting of large itemsets is thus the focus of most association rules mining algorithms.

III. APRIORI ALGORITHM

Apriori algorithm is, the most classical and important algorithm for mining frequent itemsets, proposed by R.Agrawal and R.Srikant in 1994. Apriori is used to find all frequent itemsets in a given database DB. The key idea of Apriori algorithm is to make multiple passes over the database. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-itemsets are used to explore (k+1)-itemsets. The working of Apriori algorithm is fairly depends upon the Apriori property which states that" All nonempty subsets of a frequent itemsets must be frequent". It also described the anti monotonic property which says if the system cannot pass the minimum support test, all its supersets will fail to pass the test. Therefore if the one set is infrequent then all its supersets are also frequent and vice versa. This property is used to prune the infrequent candidate elements. In the beginning, the set of frequent 1-itemsets is found. The set of that contains one item, which satisfy the support threshold, is denoted by L. In each subsequent pass, we begin with a seed set of itemsets found to be large in the previous pass. This seed set is used for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during
The pass over the data. At the end of the pass, we determine which of the candidate itemsets are actually large (frequent), and they become the seed for the next pass. Therefore, L is used to find L!, the set of frequent 2-itemsets, which is used to find L , and so on, until no more frequent k-itemsets can be found. The basic steps to mine the frequent elements are as follows: ·

- **Generate and test:** In this first find the 1-itemset frequent elements L by scanning the database and removing all those elements from C which cannot satisfy the minimum support criteria.
- **Join step:** To attain the next level elements C_k join the previous frequent elements by self join i.e. L_{k-1}* L_{k-1} known as Cartesian product of L_{k-1} . I.e. This step generates new candidate k-itemsets based on joining L_{k-1} with itself which is found in the previous iteration. Let C_k denote candidate k-itemset and L_k be the frequent k-itemset.
- **Prune step**: C_k is the superset of L_k so members of C_k may or may not be frequent but all K ' 1 frequent itemsets are included in C_k thus prunes the C_k to find K frequent itemsets with the help of Apriori property. I.e. This step eliminates some of the candidate k-itemsets using the Apriori property A scan of the database to determine the count of each candidate in C_k would result in the determination of L_k (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k). C_k, however, can be huge, and so this could involve grave computation. To shrink the size of C_k, the Apriori property is used as follows. Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k-1)-subset of candidate k-itemset is not in L_{k-1} then the candidate cannot be frequent either and so can be removed from C_k. Step 2 and 3 is repeated until no new candidate set is generated.

It is no doubt that Apriori algorithm successfully finds the frequent elements from the database. But as the dimensionality of the database increase with the number of items then:

- More search space is needed and I/O cost will increase.
- Number of database scan is increased thus candidate generation will increase results in increase in computational cost.

Therefore many variations have been takes place in the Apriori algorithm to minimize the above limitations arises due to increase in size of database. These subsequently proposed algorithms adopt similar database scan level by level as in Apriori algorithm, while the methods of candidate generation and pruning, support counting and candidate representation may differ. The algorithms improve the Apriori algorithms by:

- Reduce passes of transaction database scans
- Shrink number of candidates
- Facilitate support counting of candidates

IV. MATH

In **An improved Apriori Algorithm for Association Rules of Mining [1]** the basic concepts of association rule mining and the classical Apriori algorithm is discussed. The idea to improve the algorithm is also discussed. The new algorithm is made that works on the following technique, firstly, separate every acquired data according to discretization of data items and count the data while scan the database, secondly, prune the acquired item sets. After analysis, the improved algorithm reduces the system resources occupied and improves the efficiency and quality.

Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery [2] the paper presents the implementation of an association rules discovery data mining task using Grid

technologies. A result of implementation with a comparison of classic apriori and distributed apriori is also discussed. Distributed data mining systems provide an efficient use of multiple processors and databases to speed up the execution of data mining and enable data distribution. The main aim of grid computing is to give organizations and application developers the ability to create distributed computing environments that can utilize computing resources on demand. Therefore, it can help increase efficiencies and reduce the cost of computing networks by decreasing data processing time and optimizing resources and distributing workloads, thereby allowing users to achieve much faster results on large operations and at lower costs. In this paper distributed apriori association rule on grid based environment is mined and the knowledge obtained is interpreted.

Optimization of association rule mining and apriori algorithm Using Ant colony optimization [3].This paper is on Apriori algorithm and association rule mining to improved algorithm based on the Ant colony optimization algorithm. ACO was introduced by dorigo and has evolved significantly in the last few years. Many organizations have collected massive amount data. This data set is usually stored on storage database systems. Two major problems arise in the analysis of the information system. One is reducing unnecessary objects and attributes so as to get the minimum subset of attributes ensuring a good approximation of classes and an acceptable quality of classification. Another one is representing the information system as a decision table which shows dependencies between the minimum subset of attributes and particular class numbers without redundancy. In Apriori algorithm, is working process explained in steps. Two step processes is used to find the frequent item set to join and prune. ACO algorithm was inspired from natural behaviour of ant colonies. ACO is used to solve to numerous hard optimizations including the travelling salesman problem. ACO system contains two rules .One is local pheromone update rule, which is applied in constructing solution. Another one is global pheromone update rule which is applied in ant construction.ACO algorithm includes two more mechanisms, namely trail evaporation and optionally deamonactions.ACO algorithm is used for the specific problem of minimizing the number of association rules. Apriori algorithm uses transaction data set and uses a user interested support and confidence value then produces the association rule set. These association rule set is discrete and continues. Hence weak rule set are required to prune.

An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction [4] elaborates the basic ideas and the shortcomings of Apriori algorithm, studies the current major improvement strategies of it. The improved Apriori algorithm based on pruning optimization and transaction reduction is proposed. According to the performance comparison in the simulation experiment, the number of frequent item sets is much less and the running time is significantly reduced as well as the performance is enhanced then finally the algorithm is enhanced.

.

The Research of Improved Association Rules Mining Apriori Algorithm [6] points out the bottleneck of classical Apriori's algorithm, presents an improved association rule mining algorithm. The new algorithm is based on reducing the times of scanning candidate sets and using hash tree to store candidate item sets. According to the running result of the algorithm, the processing time of mining is decreased and the efficiency of algorithm has improved.

An Improved Apriori Algorithm [5] called APRIORI-IMPROVE is proposed based on the limitations of Apriori. APRIORI-IMPROVE algorithm presents optimizations on 2-items generation, transactions compression and uses hash structure to generate L2, uses an efficient.

An Improved Apriori-based Algorithm for Association Rules Mining [7] elaborates that because of the rapid growth in worldwide information, efficiency of association rules mining (ARM) has been concerned for several years. In this paper, based on the original Apriori algorithm, an improved algorithm IAA is proposed. IAA adopts a new count-based method to prune candidate itemsets and uses generation record to reduce total data scan amount. Experiments demonstrate that our algorithm outperforms the original Apriori and some other existing ARM methods...
In this paper, an improved Apriori-based algorithm IAA is proposed. Through pruning candidate itemsets by a new count-based method and decreasing the mount of scan data by candidate generation record, this algorithm can reduce the redundant operation while generating frequent itemsets and association rules in the database. Validated by the experiments, the improvement is notable. This work is part of our Distributed Network Behavior Analysis System, though we have considered C-R problem in our algorithm, for specific dataset, more work is still needed. We also need further research to implement this algorithm in our distributed system.

Optimization of Association Rule Mining through Genetic Algorithm [8] explains the Strong rule generation is an important area of data mining. In this paper authors design a novel method for generation of strong rule. In which a general Apriori algorithm is used to generate the rules after that authors use the optimization techniques. Genetic algorithm is one of the best ways to optimize the rules. In this direction for the optimization of the rule set they design a new fitness function that uses the concept of supervised learning then the GA will be able to generate the stronger rule set.

V. CONCLUSION

Association rule mining is an interesting topic of research in the field of data mining. We have presented a survey of most recent research work. However association rule mining is still in a stage of exploration and development. There are still some essential issues that need to be studied for identifying useful association rules. We hope that data mining researchers can solve these problems as soon as possible.
Some problems for association rule mining are suggested below:

1. To make frequent pattern mining an essential task in data mining, much research is needed.
2. Most approaches are based on some strict assumptions. They should be generalized so that they can be more widely used.
3. More efficient and scalable methods for Association Rule mining should be developed.
4. Single scan and online mining methods should be developed.
5. Database-independent measurements should be established.
6. Deep-level association rules should be identified.
7. Techniques for mining association rules in multi-databases should be explored.
8. Effective techniques for Web Usage Mining should be developed.
9. New applications of association rule mining should be explored.

REFERENCES

[1] WEI Yong-qing, YANG Ren-hua, LIU Pei-yu, *"An Improved Apriori Algorithm for Association Rules of Mining"* IEEE(2009)

[2] Mrs. R. Sumithra, Dr (Mrs). Sujni Paul, "*Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery",* 2010 Second International conference on Computing, Communication and Networking Technologies, IEEE.

[3] Badri patel ,Vijay K Chaudahri,Rajneesh K Karan,YK Rana, "*Optimization of association rule mining apriori algorithm using Ant Colony optimization*" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-1, March 2011.

[4] Zhuang Chen, Shibang CAI, Qiulin Song and Chonglai Zhu, "*An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction", IEEE 2011.*

[5] Rui Chang, Zhiyi Liu, *"An Improved Apriori Algorithm",* 2011 International Conference on Electronics and Optoelectronics (ICEOE 2011)

[6] Huiying Wang, Xiangwei Liu, "*The Research of Improved Association Rules Mining Apriori Algorithm"* 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).

[7] Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, Wenbao Jiang," *An Improved Apriori-based Algorithm for Association Rules Mining*", Sixth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE Society community, 2009.

[8] Rupali Haldulakar, Prof. Jitendra Agrawal," *Optimization of Association Rule Mining through Genetic Algorithm*", International Journal on Computer Science and Engineering (IJCSE), Vol. 3, Issue. 3, Mar 2011

BIBLIOGRAPHIES:-

Charanjeet Kaur received her B.Tech degree in Computer Engineering from the Yadawindra College of Engineering, Guru Kashi Campus, Talwandi Sabo(Bathinda) affiliated to Punjabi University , Patiala(Punjab) in 2011, and pursuing M.Tech degree in Computer Engineering from Yadawindra College of Engineering, Talwandi Sabo (Bathinda). Currently, she is doing her thesis work on Optimization of Apriori Algorithm using Ant Colony Optimization. Her topic of interest is Data Mining.