A Survey on Frequent Patterns To Optimize Association Rules

[1] **B.Ramana Reddy**, [2]**R.Asritha,**

[1]Asst. Prof., Dept.of CSE, AITS, Tirupati, A.P., INDIA.

[2]PGScholar, Dept.of CSE, AITS, Tirupati, A.P., INDIA

ABSTRACT

Data mining will define hidden pattern present in data sets and association among the patterns. In case of data mining, the association rule mining is regarded as key technique to discover useful patterns taken from large amount of collection of data. The mining of frequent itemset is considered as a step of association rule mining. In order to gather itemsets after discovery of association rules the frequent item set mining is used. In this, the fundamentals regarding frequent itemset mining are explained. Current techniques are defined for frequent item set mining. Based on the more varieties of capable algorithms which have been established the most important ones are compared. The algorithms and investigation of their run time performance are organized.

1 INTRODUCTION

Data mining is regarded as a powerful new technology with a great potential in order to help the companies to focus on the important information present in the data that they have collected regarding the behavior of their customers and also potential customers . In Data mining, the extraction of hidden predictive information taken from large databases is considered as a new powerful technology with the great potential for helping companies to focus on the important information present in their data warehouses. The Data mining tools will predict future trends and also behaviors, by allowing businesses to make proactive and knowledge-driven decisions. The automated, prospective analysis which is offered by data mining wil move beyond the analysis of past events that are provided by retrospective tools typically of decision support systems.

Data mining is considered as the process of analyzing data from different perspectives and summarizes it into useful information.

The Data mining software is one of among number of analytical tools to analyze data. Sometimes, Data mining is also called as data or knowledge discovery. The Data mining is also regarded as knowledge discovery in databases which has been recognized as a new area for the database research. The area may be defined as efficiently discovering of interesting rules taken from large collections of data. Technically, the data mining is considered as the process of finding correlations or patterns among dozens of fields present in large relational databases. Typically, Correlation is association of rules.

In case of data mining research area the Frequent item set mining is regarded as one of the most significant and general topic of research for the purpose of association rule mining. Thus, it is required to mine frequent item set efficiently as the performance of association rule mining will depend upon the frequent itemsets mining. A frequent itemset is considered as an itemset which occurs more frequently. In case of frequent pattern mining in order to check whether an itemset will occur frequently or not a parameter is present which is called support of an itemset. An itemset is termed as frequent if its support count is higher compared to the minimum support count that is set up initially. The Association rule is faced by X→Y where X and Y are regarded as item sets and their intersection is null i.e. X∩Y= {}.The support of an association rule is defined as the support of the union of X and Y, i.e. XUY. X is known as the head or antecedent and Y is known as the tail or consequent of the rule .The confidence of an association rule is considered as the percentage of rows present in D containing itemset X that will also contain itemset Y, i.e, CONFIDENCE (X → Y) =P (X|Y) = SUPPORT (XY)/SUPPORT (X). A large number of algorithms were introduced by many of researchers to generate frequent itemsets, initially, Apriori, like algorithms were proposed but because of their large number of candidate generation, slow processing, more database scan and in some cases when the support threshold is low then generation of frequent patterns will become doubtful due to huge search space, high memory dependency and large I/O is required during these type of algorithms. New algorithms have been studied in this paper such as FP-Growth and their variations for reducing the memory requirements in order to decrease I/O dependencies and also for reducing the pruning strategies to efficient generation of frequent itemsets.

II LITERATURE REVIEW/SURVEY

Huan Wu et al. (2009) introduced an improved algorithm IAA which is based on the Apriori algorithm Goswami D.N. et al (2010) proposed three various frequent pattern mining approaches (Record filter, intersection and the Proposed Algorithm) which is based on classical Apriori algorithm. K. Vanitha and R. Santhi described regarding the implementation of Hash based Apriori algorithm. The principal data structure of their solution is analyzed, theoretically and experimentally. Sunil Kumar et al (2012) proposed a new algorithm that takes less number of scans for mining the frequent item sets obtained from the large database that will lead to mine the association rule among the database. Rehab H. Alwa and Anasuya V. Patil In 2013 proposed a novel approach in order to improve the Apriori algorithm with the creation of Matrix – File.

In 2013, Jugendra Dongre , Gend Lal Prajapati and S. V. Tokekar presented an approach for mining association rules by using apriori algorithm through calculation of various support and confidence values for each of the transaction in order to find frequent item sets. In 2013 Jaishree Singh et al introduced an Improved Apriori algorithm that reduces the scanning time by cutting down not only the required transaction records but also reduce the redundant generation of sub items in case of pruning the candidate items that can form directly the set of frequent item sets and also eliminate candidate having a subset which is infrequent. In 2014, Sallam Osman Fageeri, Rohiza Ahmad, Baharum B. Baharudi introduced a semi Apriori Algorithm for the purpose of mining association rules by using binary-based data structure which is used for discovering the frequent item sets as well as the association rules.

III VARIOUS METHODS OF ASSOCIATION RULE MINING

Various methods for association rule mining to find the frequent itemsets are described below:

Apriori Algorithm: The Apriori algorithm is considered as one of the classical algorithms which is proposed by R. Srikant and R. Agrawal in 1994 to find frequent patterns for boolean association rules. The Apriori will employ an iteractive approach called level-wise search, in which k-itemsets were used to explore (k+1)-itemsets. Initially, the set of frequent 1-itemsets is obtained by scanning the database for accumulating the count of each item and then collecting those items which satisfy minimum support. The resulting set will be denoted by L1.Then, L1 is used for finding L2, the set of frequent 2-itemsets that is used for finding L3 and so on, until no frequent k-itemsets can be found. For finding each Lk it requires one full scan of the database. The algorithm will be executed in two steps: Prune and Join. In First step, it will retrieve all the frequent itemsets taken from the database in considering of those itemsets whose support is not less than the minimum support (called min_sup). In Second step, it will generate the association rules by satisfying the minimum confidence (called min_conf) taken from the frequent itemsets that are generated in first step. The first step will consist of join and pruning actions. During joining phase, the candidate set Ck is produced through joining Lk-1 with itself and then pruning of the candidate sets will be done by using the Apriori property i.e. all of the non-empty subsets for a frequent itemset should also be frequent.

FP-Growth Algorithm:

FP-growth algorithm is proposed by Jiawei Han that finds the association rules more efficiently compared to Apriori algorithm without the generation of candidate itemsets. The Apriori algorithm will require n+1 scans, where n is known as the length of the longest pattern. The FP-growth algorithm will require only two scans of the database for finding frequent patterns. FP-growth algorithm will adopts divide and conquer strategy. Initially, it will construct a FP-tree by using the data present in transactional database and then it mines all the frequent patterns taken from FP-tree. The association rules can be generated easily after mining of frequent patterns. Applications of Association Rule Mining.

IV APRIORI ALGORITHM

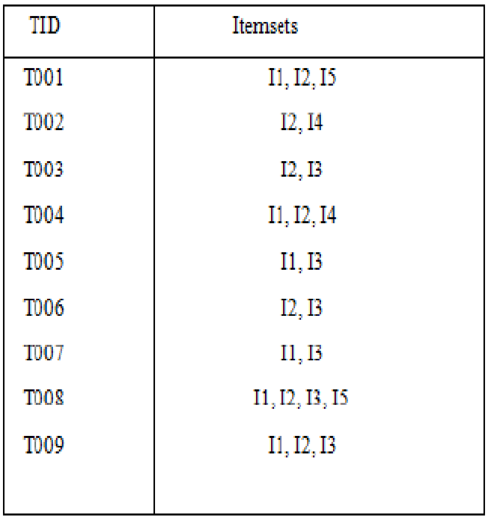
The Apriori algorithm is considered as one of the classical algorithms which is proposed by R. Srikant and R. Agrawal in 1994 to find frequent patterns for boolean association rules. The Apriori will employ an iteractive approach called level-wise search, in which k-itemsets were used to explore (k+1)- itemsets. Initially, the set of frequent 1-itemsets is obtained by scanning the database for accumulating the count of each item and then collecting those items which satisfy minimum support. The resulting set will be denoted by L1.Then, L1 is used for finding L2, the set of frequent 2-itemsets that is used for finding L3 and so on, until no frequent k-itemsets can be found. For finding each Lk it requires one full scan of the database. The algorithm will be executed in two steps: Prune and Join. In First step, it will retrieve all the frequent itemsets taken from the database in considering of those itemsets whose support is not less than the minimum support (called min_sup). In Second step, it will generate the association rules by satisfying the minimum confidence (called min_conf) taken from the frequent itemsets that are generated in first step. The first step will consist of join and pruning actions. During joining phase, the candidate set Ck is produced through joining Lk-1 with itself and then pruning of the candidate sets will be done by using the Apriori property i.e. all of the non-empty subsets for a frequent itemset should also be frequent.

1. Algorithm: The basic algorithm for mining association is given as follows

Let I = {I1, I2….In} be as a set of item and

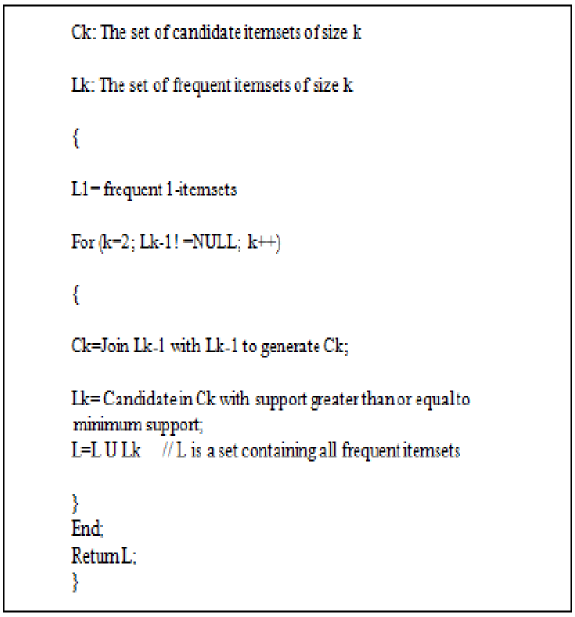
D= {T1, T2 …Tn} be as a set of transaction

Where ti is a set of transaction ti∈ I, An association rule is transaction of the form X→Y Where X, Y⊂ I and X∩Y=Ø .The rule X→Y holds in the set D with Support and Confidence. Example for All Electronics Transactional Database D is presented below in Table 1.1 in order to specify the process of Apriori algorithm. Let min_sup=2 and min_conf as 70%. By Apriori algorithm, the process of generating frequent itemsets is shown below in Table 1.1



TID	Itemsets
T001	I1, I2, I5
T002	I2, I4
T003	I2, I3
T004	I1, I2, I4
T005	I1, I3
T006	I2, I3
T007	I1, I3
T008	I1, I2, I3, I5
T009	I1, I2, I3

Table 1.1: All Electronics Transactional Database (D)



```
Ck: The set of candidate itemsets of size k

Lk: The set of frequent itemsets of size k

{

L1= frequent 1-itemsets

For (k=2; Lk-1! =NULL; k++)

{

Ck=Join Lk-1 with Lk-1 to generate Ck;

Lk= Candidate in Ck with support greater than or equal to
minimum support;
L=L U Lk    // L is a set containing all frequent itemsets

}
End;
Return L;
}
```

Figure 1.2: Pseudo code of Apriori Algorithm

V PROPOSED ALGORITHM

Applied Algorithm Algorithmic Structure: The proposed method for the generation of association rule through GA is as follows:

Step 1: Start

Step 2: A sample of records are loaded from the database which fits in the memory.

Step 3: Apriori algorithm is applied for finding the frequent item sets with the help of minimum support. Suppose A is the set of frequent item set that is generated by Apriori algorithm.

Step 4: Set Z= 0 in which Z is the output set that contains the association rule.

Step 5: Input the termination condition of the GA.

Step 6: Each frequent item set of A is represented as a binary string by using the combination of representation.

Step 7: The two members from the frequent item set are selected using Roulette Wheel sampling method.

Step 8: The crossover and mutation are applied on the selected members for generating the association rules.

Step 9: The fitness function is found for each rule X◊Y and the following condition is checked.

Step 10: If (fitness function > min confidence)

Step 11: Z = Z U {X ◊Y} is set.

Step 12: Go to Step 3 if the desired number of generations is not completed Step 13: Stop.

VI CONCLUSION

Data mining and knowledge discovery are considered as new emerging disciplines with the important applications present in Science, health care, engineering education and business. In data mining, the Association rule mining is one of the key fields. Several researchers are trying in order to develop efficient methods for finding the frequent patterns and to optimize the association rules. The Association rule mining is regarded as an important topic in case of data mining and it is receiving an increasing attention. In this research work, an efficient algorithm for the purpose of optimization of association rule mining has been proposed. Different association rule mining algorithms known as Apriori will suffer from limitation of large number of association rules generation. To find the minimized number of association rules an efficient method is developed in this research work.

REFERENCES

[1] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan- Kaufmann Publishers, 2000.

[2] R. Agrawal, R. Srikant, “Fast Algorithm for Mining Association Rules”, Proc. of the Int. Conf on Very Large Database, pp. 487- 499, 1994.

[3] J. Han, J. Pei, Y. Yin, “Mining Frequent Patterns without Candidate Generation”. Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp.1-12, 2000.

[4] J. Han, J. Pei, Y. Yin and R. Mao, “Mining Frequent Patterns without Candidate Generation: A Frequent Pattern Tree Approach”, In Data mining and Knowledge Discovery, Vol. 8, pp.53-87, 2004.

[5] S. Rangaswamy, Shobha G., “Optimized Association Rule Mining Using Genetic Algorithm,” Journal of Computer Science Engineering and information Technology Research (JCSEITR), Vol.2, Issue 1, pp 1-9, 2012. [6] S. Jain, S. Kabra. "Mining & Optimization of Association Rules Using Effective Algorithm," International journal of Emerging Technology and Advanced Engineering (IJETAE), Vol.2, Issue 4, 2012.

[7] Jun Gao, “A New Association Rule Mining algorithm and Its Applications”, IEEE 3rdInt. Conf. on Advanced Computer Theory and Engineering (ICACTE), vol 5, pp. 122-125,2010.

[8] Li Juan and Ming De-ting, “Research of an association rule mining algorithm based on FP tree”, IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), Vol. 1, pp. 559-563,2010.

[9] Zhi Liu, Mingyu Lu, Weiguo Yi, and Hao Xu, “An Efficient Association Rules Mining Algorithm Based on Coding and Constraints”, Proceedings of the 2nd International Conference on Biomedical Engineering and Informatics, pp. 1-5, 2009.

[10] Wanjun Yu , XiaochunWang, and Fangyi Wang, “The Research of Improved Apriori Algorithm for Mining Association Rules”, 11th IEEE International Conference on Communication Technology Proceedings, pp. 513-516, 2008.