International Journal of Engineering &Technology, 6 (4) (2017) 141-144

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v6i4.8300
Research paper

A study of frequent itemset mining techniques

Sachin Sharma [1]*, Shaveta Bhatia [2]

[1]Research Scholar, Manav Rachna International University, Faridabad
[2]Assistant Professor, Manav Rachna International University, Faridabad
**Corresponding author E-mail:sachin.fca@mriu.edu.in*

Abstract

Frequent item set is the most crucial and expensive task for the industry today. It is the task of mining the information from different sources and a key approach in Data Mining. Frequent item sets satisfying the minimum threshold can be discovered. Association rules are extracted from frequent item sets. The Association rules are affected by the minimum support value entered by the user may be considered as Positive or negative. There may be some other Association rules, which involve the rare item sets. Various methods have been used by researchers for generating the Association Rules. In this paper, our aim is to study various techniques to generate the Association rules.

Keywords: *Association Rules; Frequent Item Sets; Rare Item Sets; Support Threshold.*

1. Introduction

The process of obtaining the hidden and useful knowledge from the large amount of data is called Data Mining.Increasing volume of data and computational power are the reasons to create the Data Mining techniques and algorithms to find the interesting knowledge from the data warehouse. There are different functions in Data Mining e.g Associations, prediction and clustering. The best example used in data mining is Market Basket Analysis, where a departmental manager wants to find the buying habits of the customer that is what the customer purchases with an item x or item y.This process helps the departmental manager to devise effective marketing strategies. A formal definition of data mining is given as “A process of non-trivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database” [5].

Association rules are statements to find the association between the different variables in large databases. One traditional approach to discover the relationships via support and confidence was proposed by Agrawal, Imielinski and Swami in 1993. Using this approach, item sets that can be observed frequently are identified. The user sets the minimum value called “minimum support threshold”to segregate the frequent item sets from infrequent item sets. Furthermore, another interestingness measure “confidence” is used to filter the interesting association rules.

a) Let X and Y be two different item sets in the transaction T, then the Support(s) of Association rule is defined as the ratio of records that contain X U Y to the total number of records in the database. The formula for support is as given below:

$$\text{Support}(\text{X}|\text{Y}) = \frac{\text{Support count of XY}}{\text{Total number of transactions}}$$

b) Confidence of an association rule is defined as the ratio of the records that contain X U Y to the total number of records that contain X, where if the percentage goes beyond the threshold of confidence; the strong association rule X=>Y can be generated.

$$\text{Confidence}(\text{X}|\text{Y}) = \frac{\text{Support count of XY}}{\text{Support count of X}}$$

c) Lift is introduced by S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. It measures how many times more often X and Y occur together than expected if they where statistically independent. A lift value of 1 indicates independence between X and Y. The Lift of a rule is defined as

$$\text{Lift}(\text{X} => Y) = \frac{\text{Support(X U Y)}}{\text{Support (X)} * \text{Support(Y)}}$$

d) Conviction is introduced by Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Turk.It was developed as an alternative to confidence. Conviction of a rule is defined as

$$\text{Conv}(\text{X} => Y) = \frac{1 - \text{Support (Y)}}{1 - \text{conf}(\text{X} => Y)}$$

2. Apriori algorithm

Apriori is a fast algorithm for mining association rules. This algorithm uses a “bottom-up” methodology, where frequent itemsets are extended one item at a time.

The basic algorithm[2] works as follows:

```
Algorithm: Apriori
Input: D, database of transactions
min_sup: Minimum Support Threshold
Output: L, Frequent itemsets in D
  1) Scan the database to get the support of each 1-itemset, compare with min_sup and get the frequent 1-itemset.
  2) Use $L_{K-1}$, Join $L_{K-1}$ to generate the candidate k-itemset.
  3) Scan the database to get the support of each candidate k-itemset, compare with min_sup and get the frequent k-itemset.
  4) Repeat the steps 2 to 3 until candidate itemset is null. If null, generate all subsets for each frequent itemset.
```

The diagrammatical representation for generating the Association Rules is as follow:

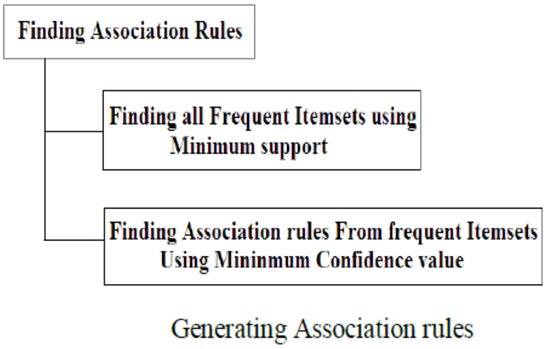
Finding Association Rules
Finding all Frequent Itemsets using Minimum support
Finding Association rules From frequent Itemsets Using Mininmum Confidence value
Generating Association rules

Fig. 1: Generating Association Rules [8].

Association rule is a format in the form of X=>Y, where X, Y are called item sets. In general, it is a two-step process:

a) To find all frequent itemsets.
b) To generate strong association rules from the frequent item sets.

Apriori algorithm has various limitations. The main limitation is costly wasting of time to hold a vast number of candidate sets with much frequent itemsets, low minimum support or large itemsets. Also, it will scan database many times repeatedly for finding candidate itemsets. Apriori will be very low and inefficient when memory capacity is limited with large number of transactions.

A number of methods have been developed for searching these rules. In the next sections, we provide a brief introduction to this family of methods and the experimental evaluation of Apriori Algorithm.

3. Related work

In [9], the authors developed an improved Apriori version known as FP-growth method that will help to overcome the problems of traditional Apriori algorithm and provide more efficiency than exclusive one. The test results of the algorithm confirmed that the requirements have higher efficiency in relation to time, less storage space and CPU usage as compared to Apriori Algorithm.

In [4], the authors consider both the existence and non-appearance of items as a basis for generating rules. They measured the association rules using chi-square test and tested on census data.

In [12], the authors have generated the frequent item sets using Genetic Algorithm which is very simple and efficient. The major advantage of this algorithm is that they perform global search and its time complexity is less as compared to other algorithms.But this approach is not feasible for identifying the rare and negative Item sets.

In [1], the author gave the brief study of various Data Mining algorithms for mining frequent item sets using association rules. Among all algorithms, FP-Growth is the most capable method to find the frequent item sets. It constructed the conditional structure to find relevant item sets without candidate generation. FP growth consumes less memory so, the efficiency of the algorithm is effective but is unable to identify rare item sets. The major drawback of this method is that it generates the huge number of rules.

In [14], the authors derived a method known as FP-growth, based on FP-tree. The first probe of the database develops a list of frequent items in which items in the descending order are compacted into a frequent-pattern tree or FP-tree. The FP-tree is extracted to generate item sets. The limitation for this study is that it is based on the minimum support threshold.

In [6], the authors proposed a Hash-Based algorithm, which is specifically operational for the generation of candidate set for large item sets.In addition, the generation of smaller candidate sets enables the user to effectively trim the transaction database at much earlier stage of the iterations and thus reduce the computational cost for later stages.This algorithm is not effective for graphical data.

4. Techniques for generating frequent itemsets

There are various methods for generating the Association rules such as FP Growth, Hash-based Technique and so on. Following is the most efficient techniques which are most widely used in various application areas.

a) Genetic algorithm

Genetic Algorithms (GAs) are adaptive explorative algorithm based on the evolutionary ideas of natural selection and genetic. In 1970, John Holland [12] developed GA. It is stochastic search method demonstrated on the development of natural selection, which highlights biological evolution. It can be applied in many problems.

Genetic Algorithm uses an iterative approach by developing original populations of sequences from the past ones. Every sequence is the encoded version of a candidate solution. An estimation function links a fitness degree to every sequence specifying its fitness for the problem. It is most commonly used in both applied problems [13] as well as in Scientific Models. Some important terms used in Genetic Algorithm are as follows:

Chromosome: A chromosome is sometimes called a genome. It is a set of factors [14] which express a suggested method to the problem which the genetic algorithm is vexing to explain. The chromosome is frequently denoted as a simple string.

Gene: A Gene is a portion of chromosome. A gene contains a part of solution. For example, if 134569 is a chromosome then 1, 3, 4, 5, 6 and 9 are its genes.

Fitness: Fitness is a crucial idea in evolutionary theory. It can be defined either relating to a genotype or to a phenotype in a given surroundings. In either case, it defines the capability to both survive and imitate, and is equal to the average involvement [14] to the gene pool of the next generation that is prepared by an average individual of the specified genotype or phenotype.

There are the following numbers of generic operators to yield offspring having feature of the parents [12].

1) Selection.
2) Crossover.
3) Mutation.

1) Selection: It compacts with the probabilistic existence of the fittest, in that; additional genes are chosen to continue.
2) Crossover: This operation is accomplished by choosing an arbitrary genetic factor along with the size of the chromosomes and changing all chromosomes after that idea.
3) Mutation: This operation adjusts the fresh results to improve stochasticity in the search for improved results. Consequently, a bit within a gene will be reversed (0 turn into 1, 1 turn into 0).

The proposed algorithm [17]used for generating frequent itemsets using Genetic Algorithm is a hybrid of FP Tree creation algorithm i.e. FP Tree and Apriori algorithm. This proposed algorithm can be explained into two phases.

The first phase constructs the FP Tree and second phase involves mining the FP Tree created using the hybrid (FP+Apriori) algorithm.

Phase 1: Tree construction using FP Tree Algorithm.

Phase 2: Tree Mining using hybrid (FP+Apriori) Algorithm.

b) Coherent rules

Another method for finding the frequent itemsets is coherent rules. Traditional association mining techniques find the rules based on the input entered by the user. That is, minimum support threshold. To address this problem, Narra et al. [3] has proposed pattern discovery, which is based on logic Propositional equivalence. The rule is a statement that relates two components; antecedent A and Consequence C. Antecedent Statesthe condition at issue and Consequence C states the result realized from the condition.

An association rule is a type of condition that relates a set of items. Suppose i = {i_1, i_2 ... i_n} is a set of items. A task-relevant transaction record, t1, holds a subset of items such as that t1⊂ i. Let A and B are two sets of items where A ⊂I, B ⊂I. For each combination of items, A and B, there are following association rules:

a) A⇒B iff A⊂ t1and B⊂t1

b) A⇒ ¬B iff A⊂ t1and B⊄t1

c) ¬A⇒B iff A⊂ t1and B⊂t1

d) ¬A⇒¬B iff A⊄ t1 and B⊄ t1

Association rule A⇒B is drawn to an association and is motivating iff, both item sets are perceived from a single transaction. Similarly, all four mappings are given below:

a) A⇒B is mapped to association p→q, iff both p and q are observed.

b) A⇒¬B is mapped to association p→¬q, iff p is observed and q is not observed.

c) ¬A⇒Y is mapped to association ¬p→q, iff p is not observed, and q is observed.

d) ¬A⇒¬B is mapped to association ¬p→¬q, iff both p and q are not observed

The association rules [7] are mapped to equivalences subject to the conditions in the below manner:

In multiple transactions, association rules are mapped to implications as follows:

A→B is mapped to an implication p ≡ q if and only if

1) Sup (A, B) > Sup (¬A, B);
2) Sup(A, B) > Sup(A, ¬B);
3) Sup(A, B) > Sup (¬A, ¬B).

A→¬B is mapped to an implication p ≡ q if and only if

1) Sup(A, ¬B) > Sup (A, B);
2) Sup(X, ¬B) >Sup(¬A,B);
3) Sup(X, ¬B) > Sup (¬A, ¬B).

¬A→B is mapped to an implication p ≡ q if and only if

1) Sup(¬A, B) > Sup (A, B);
2) Sup(¬A, B) > Sup(A, ¬B);
3) Sup(¬A, B) > Sup (¬X, ¬B).

¬A→¬B is mapped to an implication p ≡ q if and only if

1) Sup(¬A, ¬B) > Sup (A, B);
2) Sup(¬X, ¬B) > Sup(A, ¬B);
3) Sup(¬A, ¬B) > Sup (¬A,B).

The proposed algorithm [18] for generating frequent itemsets using Coherent rules is based on Contingency Table. The above mentioned conditions are already in the Contingency. The perception of coherent rule is that if a rule X→Y exists, then the rules ~X →~Y should also exist.After the calculation of contingency table, the count value for that can be calculated according to the contingency table.

The total count will be calculated by using below formula

$$\text{Count} = \sum_{i=1}^{n} \text{I n}$$

According to the count value of each item, the infrequent item sets are mined easily compared to the association rule mining.

Table 1: Relationship between Equivalences and Association Rules [7]

Equivalences	p ≡ q	P ≡ ¬q	¬p ≡ q	¬p ≡ ¬q
Association Rules	A⇒B	A⇒¬B	¬A⇒B	¬A⇒¬B
True or False	Necessary Conditions (to map associations to Equivalences			
T	A⇒B	A⇒¬B	¬A⇒B	¬A⇒¬B
F	A⇒¬B	A⇒B	¬A⇒¬B	¬A⇒B
F	¬A⇒B	¬A⇒¬B	A⇒B	A⇒¬B
T	¬A⇒¬B	¬A⇒B	A⇒¬B	A⇒B

c) CH-square method

In this method, it is not needed to create frequent item set and association rule within each item set. In traditional algorithm, it is not possible to determine the negative item sets. Ch-Square method is also used to find the positive and negative item set.Furthermore, there is no need to enter the minimum value by the user who is it is not dependent on the minimum support threshold.

5. Discussion

The study of the techniques mentioned above is shown in the following table. The techniques are systemized and their performance is analyzed based on the theoretical and runtime considerations. Various data sets are used for experimental analysis. In first example, the experiment is done on Abalone dataset obtained from UCI machine learning repository. The data set has 4177 samples. It is composed of one nominal attribute and 7 continuous attributes and one integer attributes.

The setting of parameter: the size of population N=500, maximum generation=100, crossover rate=0.006, mutation rate=0.001. The experiment was executed using software MATLAB 8.1.0.604(R2013a), Microsoft Windows XP Professional 2002 operating system.

Table 2: Analysis Using Genetic Algorithm [17]

Att	Min Support	Apriori with GA (No of rules generated	Apriori +FP with GA (No of rules generated)
A	2	76	42
B	3	77	43
C	5	79	44
D	9	83	48
E	10	84	49

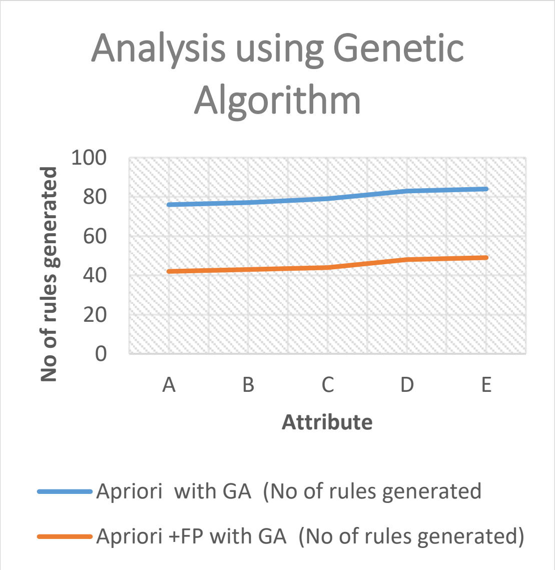

Fig. 2: Analysis Using Genetic Algorithm.

The number of rules generated in hybrid (Apriori+ FP) with GA algorithm is approximately 60% less than the number of rules generated in Apriori algorithm with GA. Thus, the proposed hybrid (Apriori+ FP) with GA algorithm gives the better results.

In another case, two data sets are considered for infrequent itemset mining. There are Real-life data sets and Synthetic data sets. Real-life data sets to validate the usefulness of the proposed algorithms technique we analyzed 10 collections, each composed of 31 real-lives weighted data sets. In Synthetic data sets we also exploited a synthetic data set generator to estimate algorithm performance and scalability. The data generator is based on the IBM data generator. The description is given in Table 4

Table 3: Analysis Using Coherent Rules [18]

No of transactions	Association Rule Method (Accuracy in %)	Coherent Rule Method (Accuracy in %)
1000	10	14
2000	20	23
3000	35	39
4000	58	61
5000	70	74
6000	85	90

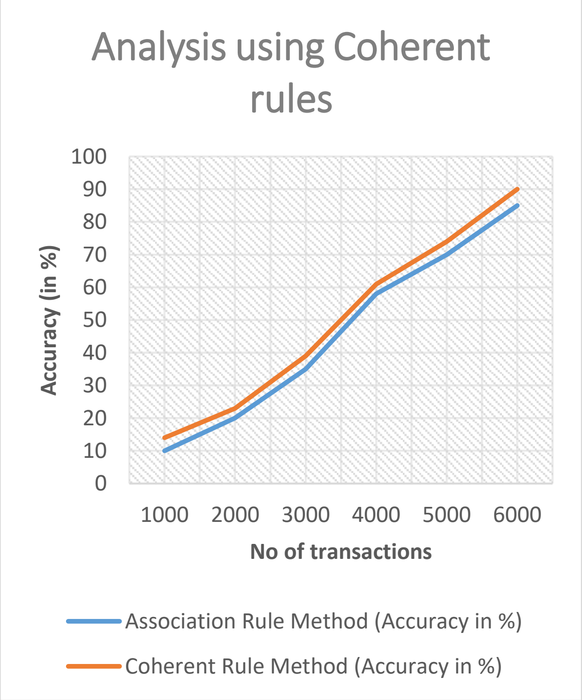

Fig. 3: Analysis Using Coherent Rules.

The experimental results show that the proposed system achieves high accuracy compared to the existing system.

The study concludes that Coherent rules and Ch-Square methods are approaches for fulfilling our objectives that is generation of frequent itemsets without minimum support threshold.

6. Conclusion

The problem of mining association rules in large databases is not straightforward. There are various methods to generate the frequent item sets. In this paper, few methods have been proposed for the same. Genetic Algorithm Algorithm is not comfortable for identifying the positive and negative item sets and is fully dependent on minimum support value.

Our objective is to generate the frequent item sets without minimum support value and be able to identify the rare item sets. To fulfill these objectives, Coherent rules and Ch-square method is the greatest approaches for identifying the same as it is based on mathematical propositional logic and correlation coefficient threshold. In the future work, we propose the new algorithm for eliminating the disadvantages as stated above and may also apply the described approaches in our algorithm.

References

[1] Arpan Shah et al. (2014),'A Collaborative Approach of Frequent Item Set Mining: A Survey', International Journal of Computer Applications, Vol 107, No 8 https://doi.org/10.5120/18775-0088.

[2] Agrawal, R. &Srikant R. (1994), 'Fast Algorithms for Mining Association rules', Proc. 20th VLDB conference, Santiago, Chile

[3] Narra S. et al. (2014), 'An efficient algorithm for mining coherent association rules', International Journal of Computer Applications, Vol 6, No 2 https://doi.org/10.5120/16769-6336.

[4] Brin S. et al. 'Beyond Market Baskets: Generalizing Association Rules to Correlations', in Proceedings of the ACM SIGMOD Conference, pp. 265-276 https://doi.org/10.1145/253260.253327.

[5] Fayyad U, (1997),'Data Mining and Knowledge Discovery in Databases: Implications from scientific databases', In Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, pp. 2-11 https://doi.org/10.1109/SSDM.1997.621141.

[6] Park J. et al. (1995) ,'Effective Hash-Based Algorithm for Mining Association', Proceedings of ACM SIGMOD International Conference on Management of Data, San Jose, CA, pp. 175 – 186 https://doi.org/10.1145/568271.223813.

[7] Kumbhar S. L. et al. (2015), 'Pattern discovery using Apriori and Ch- Search Algorithm', International Journal of Computational Engineering Research, Vol 5, No 3

[8] K.Raja & Shiva Prasad T., 'Association Rule Mining using Apriori algorithm for food dataset', available at https://www.academia.edu/8387094/ Association_Rule_ Mining_ using_ Apriori_algorithm_For_food_dataset

[9] Manisha Kundal & Dr Parminder Kaur (2015), 'Various Frequent itemset based on Data Mining Technique', International Research Journal of Engineering and Technology, Vol 02, No 3

[10] Chen C. et al. (2014), 'A Projection-based Approach for mining highly coherent Association rules', Springer Publishing Switzerland https://doi.org/10.1007/978-3-319-07776-5_8.

[11] Pei M. et al. , 'Feature Extraction using genetic algorithm', Case-Center for Computer-aided Engineering and Manufacturing W., Department of Computer Science

[12] Ghosh S. et al. (2010), 'Mining frequent item sets using Genetic Algorithm', International Journal of Artificial Intelligence and Applications, Vol 1, No 4

[13] Sharma S. et al. (2011), 'Efficiency of Spiral Model by applying Genetic Algorithm', International Journal of Computer Science and Technology, Vol 2, No 2

[14] Sharma A. et al. (2012), 'A Survey of Association Rule Mining Using Genetic Algorithm', International Journal of Computer Applications and Information Technology, Vol 1, No 2

[15] Dandu S. et al. (2013),' Improved Algorithm for Frequent Item sets Mining Based on Apriori and FP-Tree', Global Journal of Computer Science and Technology Software & Data Engineering, Vol 13, No 2

[16] Kavya T & Sasi kumar R (2015), 'Finding of repeated itemsets using Genetic Algorithm', International Journal of Engineering Sciences & Research, Vol 4, No 7

[17] Patel U.K. (2016) 'Optimization of Association Rule Mining using Genetic Algorithm', Conference Proceeding of International Conference on Recent Innovation in Science, Technology and Management.

[18] Akilandeswari S. et al (2015),'A novel approach to mine infrequent weighted itemset using coherent rule mining algorithm, Indian Journal of Innovations and Developments, Vol 4, No 3