International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Survey Report on Association Rules Mining for Frequent Item Set Generation

S. V. Subramanyam

Research Scholar, Department of Computer Science and Engineering

Abstract: ***Data mining is considered to deal with huge amounts of data which are kept in the database, to locate required is survey on association rule mining using information and facts. Innovation of association rules among the huge number of item sets is observed as a significant feature of data mining. The always growing demand of finding pattern from huge data improves the association rule mining. The main purpose of data mining provides superior result for using knowledge base system. Researchers presented a lot of approaches and algorithms for determining association rules. This paper discusses few approaches for mining association rules. Association rule mining approach is the most efficient data mining method to find out hidden or required pattern among the large volume of data. It is responsible to find correlation relationships among various data attributes in a huge set of items in a database. Studying Apriori algorithm, Apriori that is used to extract frequent itemsets from large itemsets. The Apriori algorithm has a limitation of wasting time for scanning the whole database searching on the frequent itemsets. Here we are using an improved apriori algorithm to reduce the time consumed in transaction scanning for candidate itemsets by reducing the number of transactions to be scanned by using smallest minimum support. Another approach discussed is regression technique for pairing the unpaired itemsets also reducing the time consuming of the itemsets.***

Keywords: Data mining, Apriori algorithm, minimum support threshold, multiple scan, frequent itemsets, regression

1. Introduction

Currently the world has a wealth of data, stored all over the planet (the Internet and Web are prime examples), but we need to understand that data. It has been stated that the amount of data doubles approximately every twenty months. This is especially true since the use of computers and electronic database packages. The amount or quantity of data easily exceeds what a human can comprehend on their own and thus if we wish to use and understand as much data as possible we need tools to help us. From this overwhelming state, the field of data mining has taken off and become hotly utilized.

The role of data mining is simple and has been described as "extracting knowledge from large amounts of data". Association rule mining is one of the dominating data mining technologies. Association rule mining is a process for finding associations or relations between data items or attributes in large datasets. It allows popular patterns and associations, correlations, or relationships among patterns to be found with minimal human effort, bringing important information to the surface for use. Association rule mining has been proven to be a successful technique for extracting useful information from large datasets. Various algorithms or models were developed many of which have been applied in various application domains that include telecommunication networks, market analysis, risk management, inventory control and many others. The success of applying the extracted rules to solving real world problems is very often restricted by the quality of the rules. However, the quality of the extracted rules has not drawn adequate attention. Measuring the quality of association rules is also difficult and current methods appear to be unsuitable, especially when multi - level (rules whose items / topics come from one taxonomy level, but the set of rules span more than one taxonomy level) and cross level (rules whose items / topics come from more than one taxonomy level) rules are involved.

Moreover, most of the successful applications are restricted to cases where the datasets involve only a single concept level and the success of the application is heavily dependent on the quality of the discovered rule set. Mining quality non redundant multi - level association rules from multi - level datasets is a challenge still needing to be worked on and it is a desired goal for helping to solve real world problems.

Association rule mining discovers the frequent patterns among the itemsets. It aims to extract interesting associations, frequent patterns, and correlations among sets of items in the data repositories. For Example, In a Laptop store in India, 80% of the customers who are buying Laptop computers also buy Data card for internet and pen drive for data portability. The formal statement of Association rule mining problem was initially specified by Agrawal [2]. Let I = $I_1, I_2, \ldots, I_m$ be a set of m different attributes, T be the transaction that comprises a set of items such that T $\subseteq$ I, D be a database with different transactions Ts. An association rule is an insinuation in the form of X $\Rightarrow$ Y, where X, Y $\subset$ I are sets of items termed itemsets, and X $\cap$ Y = ϕ. X is named antecedent. Y is called consequent. The rule means X implies Y.

The two significant basic measures of association rules are support(s) and confidence(c). Since the database is enormous in size, users concern about only the frequently bought items. The users can pre-define thresholds of support and confidence to drop the rules which are not so useful.

The two thresholds are named minimal support and minimal confidence. Support(s) is defined as the proportion of records that contain X $\cup$ Y to the overall records in the database. The amount for each item is augmented by one,

whenever the item is crossed over in different transaction in database during the course of the scanning.

$$Support\ (XY) = \frac{Support\ sum\ of\ XY}{Overall\ records\ in\ the\ database\ D}$$

Confidence(c) is defined as the proportion of the number of transactions that contain X ∪ Y to the overall records that contain X, where, if the ratio outperforms the threshold of confidence, an association rule X ⇒ Y can be generated.

$$Confidence\ (X \Rightarrow Y) = \frac{Support\ (X|Y)}{Support\ (X)}$$

Confidence, which assesses the degree of certainty of the detected association. if the confidence of the association rule X ⇒ Y is 80 percent, it infers that 80 per cent of the transactions that have X also comprise Y together, likewise to confirm the interestingness of the rules specified minimum confidence is also pre-defined by users.

Association rule mining is to discover association rules that fulfill the pre-defined minimum support and confidence. The problem is subdivided into two sub problems: The first one is to find the itemsets which existences surpass a predefined threshold, usually called frequent itemsets. The next one is to generate association rules from large itemsets with the limitations of minimal confidence. If one of the large itemsets is L_k, L_k= {I_1, I_2, …, I_{k-1}, I_k}, then association rules are generated with those itemsets. Checking the confidence with the rule {I_1, I_2, …, I_{k-1}} ⇒ {Ik}, it can be decided for interestingness. By deleting the last items, the other rules are created in the antecedent and placing it to the consequent, then the confidences of the new rules are checked to decide the interestingness. The processes iterated till the antecedent becomes empty. The main sub problem can be two folded into *candidate large itemsets* generation process and *frequent itemsets* generation process. Those itemsets whose support exceeds the support threshold called as *large* or *frequent itemsets*, those itemsets that are expected to be large or frequent are known *candidate itemsets*. An efficient model has classification rules with high confidence and large support.

2. Literature Survey

1) In 2008, He Jiang et al. [1] suggest the weighted association rules (WARs) mining are made because importance of the items is different. Negative association rules (NARs) play important roles in decision - making. But according to the authors the misleading rules occur and some rules.
2) In 2009, Yuanyuan Zhao et al. [2] suggest that the Negative association rules become a focus in the field of data mining. Negative association rules are useful in market - basket analysis to In 2012, Yihua Zhong et al. [3] suggest that association rule is an important model in data mining. However, traditional association rules are mostly based on the support and confidence metrics, and most algorithms and researches assumed that each attribute in the database is equal.
3) In 2012, Weimin Ouyang [6] suggest that traditional algorithms for mining association rules are built on the binary attributes databases, which has three limitations. Firstly, it cannot concern quantitative attributes; secondly, only the positive association rules are discovered; thirdly, it treats each item with the same frequency although different item may have different frequency. So he puts forward a discovery algorithm for mining positive and negative fuzzy association rules to resolve these three limitations.
4) In 2013, Luca Cagliero et al. [10] tackle the issue of discovering rare and weighted itemsets, i. e., the Infrequent Weighted Itemset (IWI) mining problem. They proposed two novel quality measures to drive the IWI mining process.
5) In 2014, Mohammed Al - Maloegi, Bassam Arkok proposed an improved apriori algorithm through reducing the time consumed itemsets by reducing the number of transactions to be scanned. Whenever the k of k - itemset increases, the gap between our improved Apriori and the original Apriori increases from view of time consumed, and whenever the value of minimum support increases, the gap between our improved Apriori and the original Apriori decreases from view of time consumed.
6) In 2013, Priyanka Asthana, Anju Singh and Diwakar Singh proposed a survey on association rule mining using apriori based algorithm and hash based method. Many algorithms share the same idea with Apriori in that they generate candidates. It include hash - based technique, partitioning, sampling and using vertical data format. Hash - based technique can minify the size of candidate itemsets. Further hash based methods can be combined with Apriori algorithm to reduce time and space complexity.

3. Problem Domain

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. In the real time environment we can subdivide the problem in two parts. First is to find the set exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second stage is the occurrence generation from association rules. So if we apply dynamic minimum support then level wise decomposition is easy. If we think of the most appropriate and efficient data mining algorithm then we always think about Apriori algorithm. However there are two bottlenecks of the Apriori algorithm. First is the process of candidate generation which can increase the time as well as the space. So the second thing is generated from the first that it needed multiple scan when it in the iteration process. Based on Apriori algorithm, many new algorithms were designed with some modifications or improvements.

The computational cost of association rules mining can be reduced in the following ways:

- It can be reduced by reducing the passes.
- Need sampling
- Add extra constrains according to your requirements

In the traditional approach of discovering rules they need lot of search space, redundant data and dangling tuples. The

missing item set is called dangling tuples. To reduce the search space, and to improve the quality of the mined rules, it is fruitful to introduce additional measures apart from support and confidence which is achieved by negative associations. The traditional Apriori - based implementations are efficient but cannot generate all valid positive and negative ARs. So we try to solve that problem without paying too high a price in terms of computational costs and reducing space with less time.
The main challenges:

- How to effectively search for interesting item sets
- How to effectively identify negative association rules of interest.
- Identification of negative association rules exist in the non - frequent item set.

In the traditional algorithms, the process of data mining for association rules generally split in two parts: first, mining for frequent item sets; and second, generating strong association rules from the discovered frequent item sets. But to club both of the sequences and generate the classified rule on - the - fly while analyzing the correlations within each candidate or non - candidate item set is missing which also avoids evaluating item combinations redundantly.

- Criteria of the discovered rules for the user requirements may not be the same. Many uninteresting association rules for the user requirements can be generated when traditional mining methods are applied.
- If we think of the most appropriate and efficient data mining algorithm then we always think about apriori algorithm.
- There are two bottlenecks of apriori algorithm:
- First is the process of candidate generation which can increase the time as well as the space.
- Second thing is generated from the first that it needed multiple scan when it in the iteration process.

The algorithm [2] makes many searches in database to find frequent itemsets where k itemsets are used to generate k+1 - itemsets. Each k - itemset must be greater than or equal to minimum support threshold to be frequency. Otherwise, it is called candidate itemsets. In the first, the algorithm scan database to find frequency of 1 - itemsets that contains only one item by counting each item in database. The frequency of 1 - itemsets is used to find the itemsets in 2 - itemsets which in turn is used to find 3 - itemsets and so on until there are not any more k - itemsets.

Limitations of Apriori Algorithm:
Apriori algorithm suffers from some weakness in spite of being clear and simple. 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. For example, if there are 10^4 from frequent 1 - itemsets, it need to generate more than 10^7 candidates into 2 - length which in turn they will be tested and accumulate. Furthermore, to detect frequent pattern in size 100 (e. g.) v1, v2… v100, it have to generate 2^{100} candidate itemsets that yield on costly and wasting of time of candidate generation. So, it will check for many sets from candidate itemsets, also it will scan database many times repeatedly for finding candidate itemsets. Apriori will be very low and inefficiency when memory capacity is limited with large number of transactions.

4. Proposed Work

There is much scope in the domain of association rule mining, as in today era huge amount of data is generating day by day. Also synthesis of various techniques together to solve complex problems is proved very efficient. Hence province of association rule mining can also be made more efficient with the help of various optimization techniques for instance genetic algorithm, fuzzy logics, rough set, soft set etc.

The aforesaid techniques could improve the unidentified and constructive frequent patterns also minimize the negative, indifferent association rules.

The computational cost of association rules mining can be reduced in the following ways:

- It can be reduced by reducing the passes.
- Need sampling.
- Using regression technique
- Reduce time consuming with large data scan.
- Reduce again and again large data scan problem.

5. Methodology

The Improved Algorithm of Apriori
This section will address the improved Apriori ideas, the improved Apriori, an example of the improved Apriori, the analysis and evaluation of the improved Apriori and the experiments.

The improved Apriori ideas
In the process of Apriori, the following definitions are needed:
Definition 1: Suppose T={T1, T2, …, Tm}, (m_1) is a set of transactions, Ti= {I1, I2, …, In}, (n_1) is the set of items, and k - itemset = {i1, i2, …, ik}, (k_1) is also the set of k items, and k - itemset I.

Definition 2: Suppose _ (itemset), is the support count of itemset or the frequency of occurrence of an itemset in transactions.

Definition 3: Suppose Ck is the candidate itemset of size k, and Lk is the frequent itemset of size K.

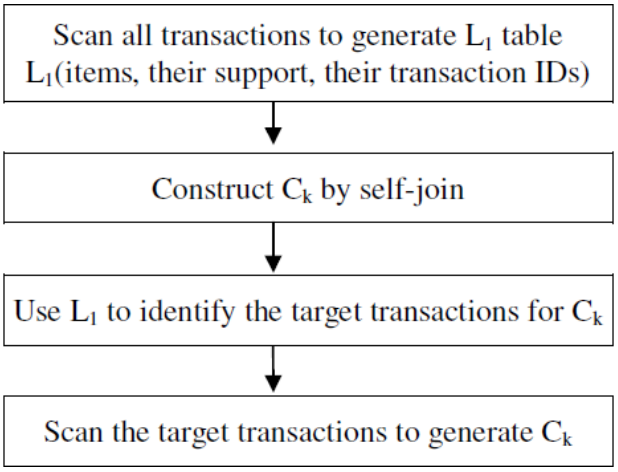

Figure 1: Steps for C_k generation

In our proposed approach, we enhance the Apriori algorithm to reduce the time consuming for candidates itemset generation. We firstly scan all transactions to generate L1 which contains the items, their support count and Transaction ID where the items are found. And then we use L1 later as a helper to generate L2, L3. . . Lk. When we want to generate C2, we make a self - join L1 * L1 to construct 2 - itemset C (x, y), where x and y are the items of C2. Before scanning all transaction records to count the support count of each candidate, use L1 to get the transaction IDs of the minimum support count between x and y, and thus scan for C2 only in these specific transactions. The same thing for C3, construct 3 - itemset C (x, y, z), where x, y and z are the items of C3 and use L1 to get the transaction IDs of the minimum support count between x, y and z, then scan for C3 only in these specific transactions and repeat these steps until no new frequent itemsets are identified. The whole process is shown in the Figure 1.

The improved Apriori

The improvement of algorithm can be described as follows:
//Generate items, items support, their transaction ID
(1) L1 = find_frequent_1_itemsets (T);
(2) For (k = 2; Lk - 1 __; k++) {
//Generate the Ck from the LK - 1
(3) Ck = candidates generated from Lk - 1;
//get the item Iw with minimum support in Ck using L1, (1_w_k).
(4) x = Get _item_min_sup (Ck, L1);
// get the target transaction IDs that contain item x.
(5) Tgt = get_Transaction_ID (x);
(6) For each transaction t in Tgt Do
(7) Increment the count of all items in Ck that are found in Tgt;
(8) Lk= items in Ck _ min_support;
(9) End;
(10) }

Implementation

For this problem, we can use weka tool as well as mat lab for pattern matching.
Through tables; we find out paired data set.

General steps:

1) In the first pass, the support of each individual item is counted, and the large ones are determined
2) In each subsequent pass, the large itemsets determined in the previous pass is used to generate new itemsets called candidate itemsets.
3) The support of each candidate itemset is counted, and the large ones are determined.
4) This process continues until no new large itemsets are found.

6. Conclusion

Association mining rules are very useful in applications going beyond the standard market basket alanysis. We have shown here various Apriori algorithms used to find frequent items in a given transaction of database. Since Apriori algorithm was first introduced and as experience was pile up, there have been many attempts to devise more efficient algorithms of frequent itemset mining. Many algorithms share the same idea with Apriori in that they generate candidates. Mining association rule is one of the most used functions in data mining. whenever the k of k - itemsets increases; the time span of apriori algorithm also increases. The time consumed to generate itemset. An improved apriori is used to reduce the time consumed in transactions scanning for candidate itemsets by reducing the transactions to be scanned. By using regression time consumption is more or less than the time consuming in the improved apriori. Association rule set used by improved apriori algorithm with regression technique is future aspect. Unpaired data set arranged with utilizing regression technique.

References

[1] He Jiang; Yuanyuan Zhao; Xiangjun Dong, "Mining Positive and Negative Weighted Association Rules from Frequent Itemsets Based on Interest, " Computational Intelligence and Design, ISCID '08. International Symposium on, vol.2, no., pp.242, 245, 17 - 18 Oct.2008.
[2] Yuanyuan Zhao, He Jiang; Runian Geng; Xiangjun Dong, "Mining Weighted Negative Association Rules Based on Correlation from Infrequent Items, " Advanced Computer Control, ICACC '09. International Conference on, vol., no., pp.270, 273, 22 - 24 Jan.2009.
[3] Luca Cagliero and Paolo Garza, “Infrequent Weighted Itemset Mining using Frequent Pattern Growth”, IEEE Transactions on Knowledge and Data Engineering, 2013.
[4] Johannes K. Chiang and Sheng - Yin Huang, “Multidimensional Data Mining for Healthcare Service Portfolio Management”, IEEE 2013.
[5] Mohammad Al. Maolegi, Bassam Arkok, An improved apriori algorithm for association rules.
[6] Charu C. Aggarwal, Philip S. Yu. Anew framework for itemset generation in IBM T J Watson Research Centre, 1998.