Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Effective Positive Negative Association Rule Mining Using Improved Frequent Pattern Tree

Ruchi Bhargava
I. T. Dept, R.K.D.F., Bhopal
India.

Shrikant Lade
I.T. Dept,R.K.D.F.,Bhopal
India.

Abstract— Association Rule is an important tool for today data mining technique. But this work only concern with positive rule generation till now. This paper gives study for generating negative and positive rule generation as demand of modern data mining techniques requirements. Here also gives detail of "A method for generating all positive and negative Association Rules" (PNAR). PNAR help to generates all unseen comparative association rules which are useful for interesting pattern finding. This work focus on determine positive and negative rules, generation of candidate set is key issue in these techniques. This paper also discussed existing techniques, such as frequent pattern growth (FP-growth) method it's a most efficient and scalable approach for rules generation. This method can generate rules without candidate ser generation. This main problem in FP tree growth is large number of conditional FP tree. This algorithm able to generates all positive and negative association rule mining. We also proposed new positive and negative association rule mining algorithm using improved frequent pattern tree for better and efficient association rules.

Keywords— FP tree, association rules, Classification.

I. INTRODUCTION

Data mining is the task to mining the useful meaningful information from data warehouse. It is the source of inexplicit, purely valid, and potentially useful and important knowledge from large volumes of natural data [7]. Data mining is the process of extracting implicit, previously unknown, and potentially useful information from large quantities of data. Through the accretion of current data with historical data, enterprises find themselves in possession of larger data sets in electronic form than at any time heretofore. Various techniques have been employed to convert the data into information, including clustering, classification, regression, association rule induction, sequencing discovery, and so forth. In general, an association rule represents a relationship between two sets of items in the same database. It can be written in the form $X \rightarrow Y$, where X and Y are item sets (i.e., values from stipulated domains) and $X \cap Y = \emptyset$. The left-hand side (LHS) of the rule is called the antecedent, while the right-hand side (RHS) is called the consequence. The selected knowledge must be not only precise but also readable, comprehensible and ease of understanding. There are a many of data mining task such as ARs, sequential patterns, Classification, clustering, time series, etc., and there have been lots of techniques and algorithms for these tasks and different types of data in data mining. When the data consist continuous values, it becomes hard to mine the data and some special techniques need to be prepared. Association rule basically use for finding out the useful patterns, relation between items found in the database of transactions. For example, consider the sales database of a Music CD store, where the records represent customers and the attributes represent Music CD. The mined patterns are the set of Music CDs most frequently bought together by the customer. An example could be that, 70% of the people who buy old song cds also buy guzzle cds. The store can use this information for future sales, self restore of records etc. There are many application areas for association rule mining techniques, which include catalog design, store layout, customer segmentation, and telecommunication alarm diagnosis and so on.

A. Applications of Association Rule Mining In E-Commerce

In this section we survey the articles that implemented association rule mining in e-commerce.

1) Web Personalization

Personalization is the use of customer information for delivering a customized solution to that customer thus satisfying personal needs [8]. In the e-commerce environment the available choices for the visiting customers are more. The search cost and time increase due to this overload. Personalization can aid the customer in decision making process. Personalization can also communicate appropriate messages to the right customers on the basis of customer profiles. Typical stages in personalization are discussed by Murthi & sarkar [8]. Common way to incorporate personalization in firm's interaction with customers is through the use of recommender system.

2) Recommender systems

Customers like to have a recommender system by which customer can see the feedback from other users who already purchased the products. E-commerce makes use of recommender system to not only show feedback from other users but

also suggest interesting and useful products to customers. Geyer et al [9] describes a recommender system that uses association rules derived from past purchases, for making recommendations to new anonymous customers. Diverse recommendation systems are proposed for different business which guides the customers to find products they would like to purchase. Most of them are based on either content filtering or collaborative filtering. Content based filtering (CBF) approach recommends products to target customers according to the preferences of their neighbors. The collaborative filtering (CF) approach recommends products to object customers based on their past preferences. The draw backs in these traditional approaches are rectified and an personalized system was proposed by Yiyang zhang e al[10]. Zhang Xizheng[11] propose a personalized recommendation system using association rule mining and classification. Set of association rules are mined from customer requirements database using Apriori algorithm and then he apply CBA-CB algorithm to produce best rules out of whole set of rules.

3) Cross selling analysis

The association rule mining is a powerful tool to realize cross selling. Cross selling is a marketing strategy to sell a new product or service to the customer who already used the products of the same enterprise. To introduce a new product or service to a new customer and an old customer, the old customer is more likely to accept it and the success rate is higher. Cross selling is a marketing method which can improve customer value, for the more the relations between the enterprise and the customer, the more dependent the customer will be on the enterprise, and the higher the loyalty will be. Xiao-li Yin [12] discusses how the banks analyze the association relations between the deal activities and other properties like customer age, gender, education, occupation, etc, and can get the result which bank services or financial products the customer will be interested in.

4) Purchasing and travelling behaviour of customers.

In e-commerce environment finding association rules between purchasing items is very important. There are many algorithms devised in this field. Path traversal pattern mining is the technique that finds most of the navigation behaviors of customers in the e-commerce environment. This information can be used to improve the website design and performance. Navigational suggestions can also be suggested to customers using this information. Many works are carried out in this field [13]. Yeu-shi-lee et al [14] propose an algorithm IPA that considers both purchasing behavior and travelling patterns of customer at the same time.

II. RELATED WORK

Paper [1] presents an efficient algorithm (PNAR) for mining both positive and negative association rules in databases. The algorithm extends traditional association rules to include negative association rules. When mining negative association rules, author use same minimum support threshold to mine frequent negative itemsets. With a Yule's Correlation Coefficient measure and pruning strategies, the algorithm can find all valid association rules quickly and overcome some limitations of the previous mining methods. Traditional algorithms for mining association rules are built on the binary attributes databases, which has three limitations [2]. Firstly, it cannot concern quantitative attributes; secondly, only the positive association rules are discovered; thirdly, it treat each item with the same frequency although different item may have different frequency. Mining the negative patterns has also attracted the attention of researchers in this area. The aim of paper [3] is to develop a new model for mining interesting negative and positive association rules out of a transactional data set. The proposed model in [3] is integration between two algorithms, the Positive Negative Association Rule (PNAR) algorithm and the Interesting Multiple Level Minimum Supports (IMLMS) algorithm, to propose a new approach (PNAR_IMLMS) for mining both negative and positive association rules from the interesting frequent and infrequent itemsets mined by the IMLMS model. Against the above shortcomings, paper [4] proposes an FP-tree-based algorithm MMFI optimized with array and matrix for mining the maximal frequent itemsets. It not only reduces the quantity of the FP-trees constructed, but also saves the time in traversing the FP-trees. Paper [5] introduced the concept of weighted dual confidence, a new algorithm which can mine effective weighted rules is proposed in this paper, which is on the basis of the dual confidence association rules used in algorithm. The case studies show that the algorithm can reduce the large number of meaningless association rules and mine interesting negative association rules in real life. In practical applications, the occurrence frequency of each itemset is different. Author set different minimum support for itemsets. In association rule mining, if the given minimum support is too high, then the items with low frequency of appearance couldn't be mined. Otherwise, if the given minimum support is too low, then combination explosion may occur. Authors support the technique that allows the users to specify multiple minimum supports to reflect the natures of the itemsets and their varied frequencies in the database. It is very effective for large databases touse algorithm of association rules based on multiple supports. The existing algorithms are mostly mining positive and negative association rules from frequent itemsets. But the negative association rules from infrequent itemsets are ignored. Furthermore, Authors set different weighted values for items according to the importance of each item. Based on the above three factors, an algorithm for mining weighted negative association rules from infrequent itemsets based on multiple supports(WNAIIMS) is proposed in paper [6].

III. PROBLEM STATEMENT

The search for exception rules will be based on the knowledge about strong association rules in the database. An example: we discover a strong association rule in the database, for instance shares of companies X and Y most times go up together X=>Y. Then those cases when shares of the companies X and Y do not go up together, X=>~Y or ~X=>Y, we

call exceptions when satisfying the exceptionality measure explained in the next section. An algorithm for mining exception rules based on the knowledge about association rules will be proposed in the following sections.

We explain a few terms that will be used along the paper. Itemset is a set of database items. Association rule is an implication of the form X=>Y, where X and Y are database itemsets. The rule X => Y has support s, if s% of all transactions contain both X and Y. The rule X => Y has confidence c, if c% of transactions that contain X, also contain Y.

In association rules mining user-specified minimum confidence (minconf), minimum support(minsup) are given. Association rules with support>=minsup and confidence>=minconf are referred to as strong rules. Itemsets that have support at least equal to minsup are called frequent itemsets. Negative itemsets are itemsets that contain both items and their negations (for example XY~Z). ~Z means negation of the item Z (absence of the item Z in the database record).

Negative association rule is an implication of the form X=>~Y, ~X=>Y, ~X=>~Y, where X and Y are database items, ~X, ~Y are negations of database items. Examples of negative association rules could be Meat=>~Fish, which implies that when customers purchase meat at the supermarket they do not buy fish at the same time, or ~Sunny=>Windy, which means no sunshine implies wind, or ~OilStock=>~PetrolStock, which says if the price the oil shares is falling, petrol shares price will be falling too.

IV. FP-TREE (FREQUENT PATTERN TREE)

FP-tree structure for the first time proposed for store information on frequent items, thus transforming the issue of mining frequent itemsets into that of mining FP-trees. In an FP-tree, every node is composed of three domains: an item name, designated as item_name, a node count, designated as count, and a link, designated as node_link; besides, in order to facilitate the traversing of the FP-trees, an item header table is created to make every time point to its presence in the tree through a node link, and the header table is made up of two domains: an item name, designated as item_name, and a node link head, designated as head, and the head points to the first node in the FP-tree with the same name with it.

A tree structure in which all items are arranged in descending order of their frequency or support count. After constructing the tree, the frequent items can be mined using FP-growth.

A. Creation of FP-Tree

1) First Iteration

Consider a transactional database which consists of set of transactions with their transaction id and list of items in the transaction. Then scan the entire database. Collect the count of the items present in the database. Then sort the items in decreasing order based on their frequencies (no. of occurrences).

2) Second Iteration

Now, once again scan the transactional database. The FP-tree is constructed as follows. Start with an empty root node. Add the transactions one after another as prefix subtrees of the root node. Repeat this process until all the transactions have been included in the FP-tree. Then construct a header table which consists of the items, counts and their head-of-node links. Consider the transactional database shown in Table 1 with 5 transactions.

TABLE I. EXAMPLE OF TRANSACTIONAL DATABASE

Tran. ID	Items
T1	A,B,D,E
T2	A,C,D
T3	E,F,H,I
T4	A,B
T5	C,E,F

The frequent itemlist for the above database is given in Table 2.

TABLE II. FREQUENT ITEMLIST FOR THE TRANSACTIONAL DATABASE IN TABLE1

Items	Count
A	3
B	2
C	2
D	2
E	3
F	2
H	1
I	1

The items that does not meet the minimum threshold has been eliminated. The frequent itemlist that support the minimum support threshold is given in Table 3.

TABLE III. FREQUENT ITEMLIST FOR THE TRANSACTIONAL DATABASE THAT SUPPORT MINIMUM THRESHOLD

Items	Count
A	3
E	3
B	2
C	2
D	2
F	2

The transactional database according to the frequent item list is given in Table 4.

TABLE IV. SORTED AND ELIMINATED TRANSACTIONS OF THE DATABASE IN TABLE 1

Tran. ID	Items
T1	A,E,B,D
T2	A,C,D
T3	E,F
T4	A,B
T5	E,C,F

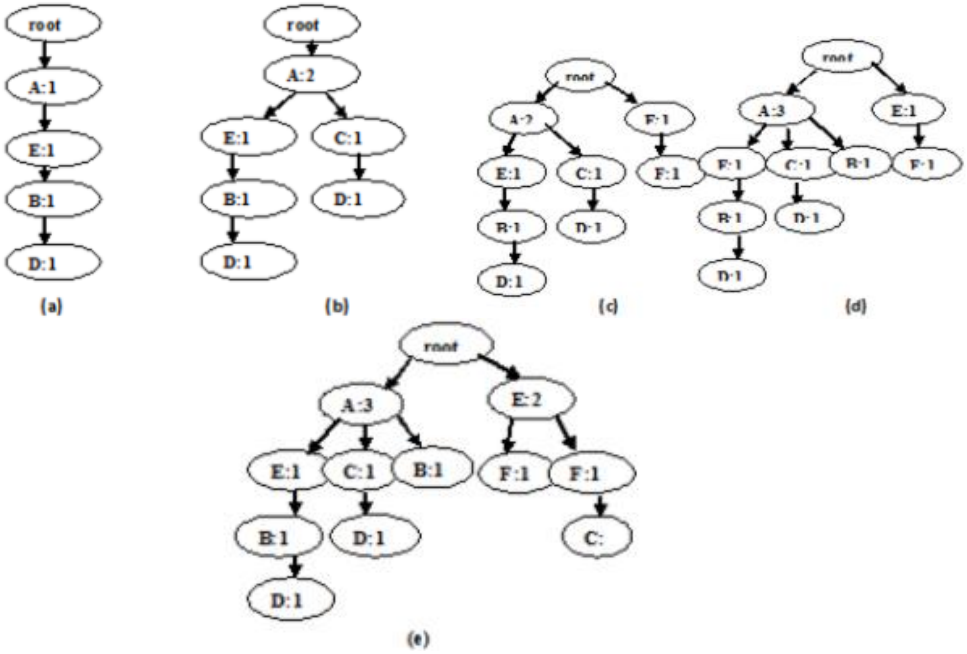

Fig. 1. **Steps in Creating the FP-Tree.**

B. Finding Frequent Patterns from FP-Tree

After the construction of FP-tree, the frequent patterns can be mined using a iterative approach FP-growth. This approach looks up the header table shown above and selects the items that supports the minimum support. It removes the infrequent items from the prefix-path of a existing node and the remaining items are considered as the frequent itemsets of the specified item.

Consider the item D. Its prefix paths are {((A, E, B): 1), ((A, C):1)}. After removing the infrequent items, (A: 2). so the frequent item set for D is A.

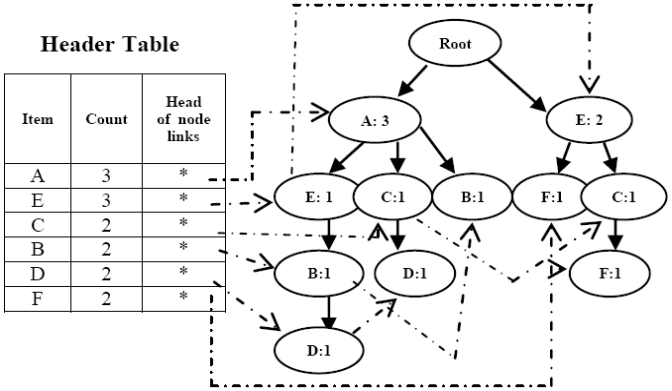

Fig. 2. **FP-Tree with Header Table.**

C. Advantages and Disadvantages

This method is advantageous because, it doesn't generate any candidate items. It is disadvantageous because, it suffers from the issues of special and temporal locality issues.

V. PROPOSED WORK

A. Positive and Negative Association rules

Corresponding to a positive association rule such as A=> B, there are three possible negative association rules, A=>~B, ~A=>B and ~A=>~B. For a negative association rule A=>~B and a certain transaction T, if $A \subseteq T$ and $\sim B \not\subset T$, we say that the transaction T supports A=>~B. Assume there is a negative association rule such as ({i1}, ~{ i2 , i3}), which means that if i1 is in a transaction T, i2 and i3 would not appeared in the transaction T at same time, but there is a possibility that one of the i2 and i3 is in transaction T. To discover negative association rule, we need to consider all of the possible item sets in transaction databases. If A=>~B is a negative association rule, it will hold that sup(A∪~B)>= minsup. A higher value for minsup possibly means sup(A∪ B)<minsup, i.e., (A, B) is infrequent sequence. However, there are too many infrequent sequences in database. If A is a frequent itemset while B is a infrequent item set with support 1, we will have: sup(A)>=minsup, sup(B)=0, sup(A∪~B) = sup(A)>=minsup. Therefore, it seems that A=>~B is a negative association rule. In fact, this kind of sequences is rather prevalent in real database, for example, a set of the goods rarely bought by customers in supermarket is an infrequent item set. In practice, since the task of data mining is to find all kinds of valuable correlations, we usually more focus on the correlations between the well-sold goods, which are based on the frequent sequence. In other word, if A=>~B, ~A=>B and ~A=>~B are negative association rules, A and B would be frequent sequence. In generally speaking, we only focus on the frequent sequence whether the association rules are positive or negative.

In order to find valuable association rules, Piatetsky- Shapiro had put forward an interestingness measurement of association rules [15]. If sup(X∪Y) = sup(X) × sup(Y), X=>Y is considered as uninteresting rules. In other words, we can say that as the association rule X=>Y is interesting only if sup(X∪Y) = sup(X) × sup(Y) is no less than a specified minimum interesting value, mininterest. Similarly, we adopt the same method to measure the interesting of negative association rules.

Definition 1: an interesting negative association rule

A=>~B as:

(1) $A \cap B = \phi$;

(2) sup(A)>=minsup, sup(B) >= minsup, sup(A∪~B) >= minsup;

(3) sup(A∪~B) = sup(A) × sup(~B)>= mininterest.

Noted it need to satisfy condition sup(B)>=minsup since we are only interested in frequent itemset in association rules. By the same way we can define conditions of negative association rules forms as ~A=>B and ~A=>~B If A=>~B is a negative association rule, A□ B will be an interesting infrequent itemset. If i is an interesting infrequent itemset, there exists at least one expression, i=A□ B, which makes one of A=>~B, ~A=>B and ~A=>~B be interesting negative association rule hold.

B. PNAR Algorithm

As mentioned before, the process of mining both positive and negative association rules can be decomposed into the following three sub problems, in a similar way to mining positive rules only.

1) Generate the set PL of frequent itemsets and the set NL of infrequent itemsets.

2) Extract positive rules of the form A=>B in PL.

3) Extract negative rules of the forms A=>¬B and ¬A=>B in NL.

Noted it need to satisfy condition sup(B)>=minsup since we are only interested in frequent itemset in association rules. By the same way we can define conditions of negative association rules forms as ~A=>B and ~A=>~B If A=>~B is a negative association rule, A∪B will be an interesting infrequent itemset. If i is an interesting infrequent itemset, there exists at least one expression, i=A∪B, which makes one of A=>~B, ~A=>B and ~A=>~B be interesting negative association rule hold.

1) Algorithm for PNAR

As mentioned before, the process of mining both positive and negative association rules can be decomposed into the following three sub problems, in a similar way to mining positive rules only.

(1)Generate the set PL of frequent itemsets and the set NL of infrequent itemsets.

(2)Extract positive rules of the form A=>B in PL.

(3)Extract negative rules of the forms A=>¬B and ¬A=>B in NL.

Let DB be a database, and ms, mc, dms and dmc given by the user. Our algorithm for extracting both positive and negative association rules with a correlation coefficient measure and pruning strategies is designed as follows:

2) Algorithm: Positive and Negative Association Rules

Input: TDB-Transactional Database

MS-Minimum Support

MC-Minimum Confidence

Output: Positive and Negative Association Rules

Method:

1. P←ø, N←ø

2. Find F1← Set of frequent 1- itemsets


```
3. for ( k=2;Fk-1!= ø; k++)
4. {
5.    Ck= Fk-1 join Fk-1
6.    // Prune using improve FP-Tree
7.    for each i ε Ck, any subset of i is not in
Fk-1 then Ck = CK - { i }
8.    for each i ε Ck find Support(i)
9.    for each A,B (A U B= i)
10.   {
11.      QA,B= Association(A,B);
12.      if Q>0
13.      if(supp(A→B)>=MS &&conf(A→B)>=MC) then
14.            P←P U { A→B}
15.      if Q<0
16.     {
17.            if(supp(A→¬B)>=MS&&conf(A→¬B)>=MC)
      then
18.      N←N U { A→¬B}
19.      if(supp(¬A→B)>=MS&&conf(¬A→B)>=MC)
         then
20.      N←N U { ¬A→B}
21.      }
22.   }
23. }
24. AR← P U N
```

PNAR generates not only all positive association rules in PL, but also negative association rules in NL. When mining negative association rules, we use same threshold to improve the usability of the frequent negative itemsets. With a Yule's correlation coefficient measure and pruning strategies, the algorithm can find all valid association rules quickly. An example of mining positive and negative itemsets is given below for illustrative purposes.

VI. CONCLUSION

This paper gives overall review for association rules generation and positive negative rules finding from large data. A new algorithm is presented to generate all positive and negative class association rules using improved FP-tree and to build an accurate classification. The scheme has several features:

(1) Its classification is performed based on positive and negative class association rules, which leads to better overall classification accuracy;

(2) It prunes contradictory positive and negative class association rules effectively based on correlation between item sets.

(3) An improved FP tree applies for mining association rules. This algorithm mines all possible frequent item set without generating the conditional FP tree. It also provides the frequency of frequent items, which is used to estimate the desired association rules.

This proposed method expected that the algorithm would be highly effective at classification and has better average classification accuracy and efficiency. In future this algorithm is performed for various large dataset and measurements will be taken for proof effectiveness of proposed algorithm.

REFERENCES

[1] CH.Sandeep Kumar K.Srinivas Peddi Kishor T.Bhaskar, "An Alternative Approach to Mine Association Rules", Page 420-424, 2011 IEEE.

[2] Weimin Ouyang, "Mining Positive and Negative Fuzzy Association Rules with Multiple Minimum Supports", International Conference on Systems and Informatics (ICSAI 2012), 2012, IEEE.

[3] Idheba Mohamad Ali O. Swesi, Azuraliza Abu Bakar, Anis Suhailis Abdul Kadir, "Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets", 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012), 2012, IEEE.

[4] PENG Hui-ling, SHU Yun-xing, "A New FP-tree-based Algorithm MMFI for Mining the Maximal Frequent Itemsets", Page 61-64, 2012 IEEE.

[5] Yihua Zhong, Yuxin Liao, "Research of Mining Effective and Weighted Association Rules Based on Dual Confidence", Fourth International Conference on Computational and Information Sciences, 2012, IEEE.

[6] He Jiang, Xiumei Luan, Xiangjun Dong, "Mining Weighted Negative Association Rules from Infrequent Itemsets Based on Multiple Supports", International Conference on Industrial Control and Electronics Engineering, 2012, IEEE.

[7] Olafsson Sigurdur, Li Xiaonan, and Wu Shuning.Operations research and data mining,in: European Journal of Operational Research 187 (2008) pp:1429–1448.

[8] Murthi B P S, Sarkar S 2003 ."The role of management sciences in research on personalilzation. Manage. Sci 49:1344-1362

[9] Liu Guo-rong，Zhang Xi-zheng, "Collaborative Filtering Based Recommendation system for Product bundling", Proceeding of ICMSE'06, Lille, France, pp.251-254, 2006.

[10] Zhan Xizheng "Building personalized recommendation system in E-commerce using Association rule-based mining and classification". Proceedings of 8th ACIS international conference on software engineering, Artificial intelligence, Networking and parallel/Distributed computing, 2007.

[11] Xiao-li Yin and Xu-sheng Fang ," Data Mining Technology Application in Bank CRM", Economic Research Guide, 2009, pp. 112- 113.

[12] M. S. Chen, J. S. Park and P. S. Yu. "Efficient Data Mining for Path Traversal Patterns in a Web Environment." IEEE Transaction on Knowledge and Data Engineering, Vol. 10, No. 2, pp. 209-221, 1998.

[13] S. J. Yen. "An Efficient Approach for Analyzing User Behaviors in a Web-Based Training Environment. "International Journal of Distance Education Technologies, Vol. 1, No. 4, pp.55-71, 2003.

[14] Hong Yu, Xiaolei Huang, Xiaorong Hu, Changxuan Wan, " Knowledge management in E-commerce: A data mining perspective", proceedings of international conference on Management of e-commerce and e-government, 2009.

[15] Liu, B., Hsu, W., and Ma, Y. Mining Association Rules with Multiple Minimum Supports. SIGKDD Explorations, 1999,pp.337-341.

