ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

Enhancement of Attributes of Apriori Algo in Association Rule Learning

Harveen Buttar[1], Rajneet kaur[2]

Research Scholar, Department of Computer Science Engineering, SGGSWU, Fatehgarh Sahib, Punjab, India[1]

Assistant Professor, Department of Computer Science Engineering, SGGSWU, Fatehgarh Sahib, Punjab, India[2]

ABSTRACT: The data mining of association rules is very important in order to determine the buying patterns of the customers. The basic goal of association rule mining is to detect the relationships or associations between specific values in large data sets. In this paper, we study the conventional method of mining association rules- Apriori algorithm and thus form new algorithm which is based upon the apriori Algorithm that will enhance the efficiency and reduce time attribute by making a model of prototype which will be beneficial in overcoming the shortcomings of apriori algorithm. We theoretically and experimentally analyze the apriori Algorithm which is the most established algorithm for frequent itemset mining. The work is focused on apriori Algorithm.

Keywords: Apriori algorithm, frequent items, association

I. INTRODUCTION

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.[1] These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance.

II. ASSOCIATION RULE MINING

Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [2]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc.

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. [3]The problem is usually decomposed into two subproblems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is Lk, Lk = {I1, I2, … , Ik}, association rules with this itemsets are generated in the following way: the first rule is {I1, I2, … , Ik-1}⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second subproblem is quite straight forward, most of the researches focus on the first subproblem.

The first sub-problem can be further divided into two sub-problems: candidate large itemsets generation process and frequent itemsets generation process. We call those itemsets whose support exceed the support threshold as large or frequent item-sets, those itemsets that are expected or have the hope to be large or frequent are called candidate itemsets.

In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. [4] Several strategies have been proposed to reduce the number of association rules, such as generating only "interesting" rules, generating only "non redundant" rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength.

Apriori is more efficient during the candidate generation process [5]. Apriori uses pruning techniques to avoid measuring certain itemsets, while guaranteeing completeness. These are the itemsets that the algorithm can prove will not turn out to be large.

III. APRIORI ALGORITHM

A basic algorithm Apriori, designed by Agrawal and others in 1993, generated all frequent item sets, Apriori uses the recursive method, the core algorithm is:

```
LI =find_frequent_I-itemsets( D);
for (k=2;LH :;t¢;k++){
Ck = apriori _gen(Lk-l,minsup).,
for each transaction tED {//scan L = UkL k for counts
C1 = subset( Ck, t ) ;//get the subsets of t that are
candidates
for each candidate C E C1
c.count+ +;
Lk = {CE Ck I c.count □ minsup}
}
returnL = ukLk; Iiall of Lk;
```

First, scan database once, resulting in frequent I item sets of LI ; and then loop, in the first k cycles, the first the frequent k -1 item sets through self-connection and pruning, to generate the candidate frequent k item sets Ck, and then use the Hash function to store Ck in a tree, scanning the database for each transaction T to use the same Hash functions to calculate the candidate frequent k item sets the transaction T contains, and make the support number of the candidate frequent k item sets plus 1, If the candidate frequent k item sets support number is greater than or equal to the number of minimum support, then the candidate frequent k item sets are frequent k item sets; the loop ends until the candidate frequent k item sets are not generated any longer. [6]

A. Limitations of Apriori Algorithm

The various limitations are as follows :

1) It only tells the presence and absence of an item in transactional database [7].
2) It is not efficient in case of large dataset.
3) The algorithm treat all items in database equally by considering only the presence and absence of an item within the transaction [8].it does not take into account the significance of item to user or business.

The algorithm has a lot of disadvantages .These can be removed by using attributes like weight and quantity, weight attribute will give user an estimate of how much quantity of item has been purchased by the customer, profit attribute will calculate the profit ratio and tell total amount of profit an item is giving to the customer.

IV. PROPOSED METHODOLOGY

Following are the steps involved in the proposed methodology. It will tell us how the proposed work has been done. Firstly a transactional database has been assumed on which the proposed methodology is to be applied.

STEP 1:Firstly a Database will be assumed which will consist of number of ITEMS to be purchased by the customer and total Profit achieved by the items .Profit ratio for each item will be calculated by applying Q-Factor .
STEP 2:Now consider a Transactional Database will be assumed which will consist of number of ITEMS to be purchased by the customer and total number of transactions in which customer purchase the items.
STEP 3:Now apply Apriori algorithm of Association rule mining in order to determine the frequency of each Itemset.
STEP 4:Calculate the CONFIDENCE measure of each Itemset.
STEP 5 :Sort itemsets based on user specified minimum Confidence.
STEP 6 : Now Apply Profit-Weighing Factor on the sorted itemset.
STEP 7 :Output will be the frequent itemsets which are giving maximum profit to the business.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

A. Example of Efficient Frequent itemset generation using attributes

In the following Example , it illustrate how these steps are performed on a particular transactional Dataset. The first step of the proposed approach is the computation of the profit ratio for all items based on Profit which can be calculated by using the formula in Fig.1, q-factor as :

$$Q - Factor = \frac{P}{\sum P_i}$$

Fig.1 Fomula of Q-Factor

Where i = 1 to n and n = number of items and P = profit of an item.

a) Calculation of profit ratio: Following table 1 shows the five Items A,B,C,D,E. Each Item has a unique Transactional ID and Profit Gained. When the particular Item is purchased by the customer. Profit Ratio is calculated by using the Q-Factor for each item and the profit associated with each item. By using this Profit Ratio The P-W Factor is calculated in the Final Step of Proposed Methodology.

Table 1: Calculation of Q-Factor

TID	ITEMS	PROFIT	Q-FACTOR
1	A	60	0.28571428571428
2	B	10	0.04761904761904
3	C	30	0.1428571428571
4	D	90	0.4285714285714
5	E	20	0.0952380952380

b) Applying Apriori algorithm on transactional database: One of the contemporary approaches for frequent item set mining is the Apriori Algorithm. Table 2 Shows the Transactional database with a minimum support = "3". Minimum support is termed as the user specified support assumed in order to determine the frequency of each itemset occurrence within transactional database. As shown in the Table 2 , we have 10 number of transactions which represent the Presence and absence of an item in the transactional database. It tells in how many transactions a particular item is purchased or not purchased by the customer . From this transactional database we will calculate the frequency of each item's occurrence within the transactional database using Support measure of apriori algorithm, as shown in Table 3.

Table 2: Given Transactional Database and min supp = 3

TID	A	B	C	D	E
1	1	1	1	1	0
2	0	1	0	1	0
3	1	0	0	1	0
4	0	0	1	0	0
5	1	1	0	1	1
6	1	1	1	1	1
7	0	1	1	0	1
8	0	0	0	1	1
9	1	0	1	1	0
10	0	1	1	1	1

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

Table 3 shows final Frequent Itemsets satisfying Minimum support that is assumed. In the Diagram red marked Itemsets are not frequent Itemsets as they are not satisfying minimum support. So in the Final Output they are excluded.

Table 3: Determining Frequency of Patterns using Apriori algorithm

Patterns	Frequency
AD	5
D	8
ACD	3
BD	5
ABD	3
CD	4
ED	4
CBD	3
EBD	3
A	5
AC	3
AB	3
C	6
ECB	3
CB	4
EC	3
EB	4
E	5
B	6

c) Frequent Pattern Selection using Confidence measure : Table 4 shows the selected frequent patterns using confidence measure of apriori algorithm.confidence is assumed to be 60% .Sorting of frequent patterns is done and those patterns are selected having confidence ≥ 60%

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

Table 4: Frequent Pattern selection based on Confidence

Frequent pattern	Confidence
AD	100%
D	100%
ACD	100%
BD	83.3%
ABD	100%
CD	83.3%
ED	80%
CBD	75%
EBD	75%
A	100%
AC	60%
AB	60%
C	100%
ECB	100%
CB	66.6%
EC	60%
EB	80%
E	100%
B	100%

d) Calculation of profit and weighing factor(PW-factor): The next step in Frequent Pattern mining is the calculation of PW-Factor for each frequent patterns selected based on confidence. It is calculated by using equation in Fig.2 .Table 5 shows the values calculated by applying the PW-factor on each Frequent itemset. Here frequency is the support calculated for each itemset in Table 3.Q-factor is the profit ratio calculated in Table 1.

$$PW = \sum_{i=1}^{n} frequency * Q - factor_i$$

Fig.2 Formula for Profit and Weighing Factor

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

Table 5: Calculating PW-Factor

Frequent pattern	P-W Factor
AD	3.5714
D	3.4285
ACD	2.5714
BD	2.3809
ABD	2.2857
CD	2.2857
ED	2.0952
CBD	1.8571
EBD	1.7142
A	1.2857
AC	1.2855
AB	1.0
C	0.8571
ECB	0.8571
CB	0.7619
EC	0.7142
EB	0.5714
E	0.4761
B	0.2857

e) Efficient frequent pattern selection: Table 6 shows the Sorting of Frequent patterns whose PQ-Factor ≥ 2.0. It is an efficient frequent pattern selection step consisting of frequent patterns giving maximum profit to the business. A discussion of the results obtained from our approach is presented here. In general, standard association rule mining algorithms result in enormous patterns , and users are expected to shortlist or select the patterns that are interesting to their own businesses. However, from the results, it is seen that this generates only number of interesting association patterns that are both statistically and semantically important for business development. The high utility patterns discovered in historical buying patterns certainly signify the importance of the items in the growth of the enterprise.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

Table 6: Efficient Frequent Pattern Selection

Frequent pattern	P-W Factor
AD	3.5714
D	3.4285
ACD	2.5714
BD	2.3809
ABD	2.2857
CD	2.2857
ED	2.0952

V. CONCLUSION

The conclusion to this work is that Apriori algorithm is applied on the transactional database. By using measures of apriori algorithm, frequent itemsets can be generated from the database. Also apriori algorithm is associated with certain limitations. Association rule mining efficiency can be improved by using attributes like profit ,quantity which will give the valuable
information to the customer as well as the business. In contradiction to traditional association rule mining algorithms, this generates only number of interesting association patterns that are both statistically and semantically important for business development. This in turn affects business utility because, in most cases , frequency plays a vital part in business development including sales backup and more.

REFERENCES

1. Pieter Adriaans and Dolf Zantinge , "*Introduction to Data Mining and Knowledge Discovery"*, Two Crows Corporation, *Third Edition* (Potomac, MD: Two Crows Corporation, 1999), *Data Mining* (New York: Addison Wesley, 1996).
2. Agrawal, R., Imielinski, T., and Swami, A. N," Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216, 1993
3. Hegland, M.," Algorithms for Association Rules, Lecture Notes in Computer Science", Volume 2600, Pages 226 - 234 , Jan 2003.
4. Techapichetvanich, K., Datta, A.," Visual Mining of Market Basket Association Rules, Lecture Notes in Computer Science", Volume 3046, Pages 479 - 488, Jan 2004.
5. Agrawal, R. and Srikant, R, " Fast algorithms for mining association rules ", In Proc. 20th Int. Conf. Very Large Data Bases, 487-499, 1994.
6. Kong Fang , Qian Xue-zhong, " Research of improved apriori algorithm in mining association rules ",Computer Engineering and Design, v29, n16, p4220-4223, 2008.
7. Tang, P., Turkia, M., " Parallelizing frequent itemset mining with FP-trees ", Technical Report , titus.compsci.ualr. edu/~ptang/papers/par-fi.pdf, Department of Computer Science, University of Arkansas at Little Rock, 2005.
8. Han, J., Pei, J. "Mining frequent patterns by pattern growth: methodology and implications". ACM SIGKDD Explorations Newsletter 2, 2, 14-20, 2000