ISSN No. 0976-5697

Volume 2, No. 6, Nov-Dec 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Efficeint Algorithm to Mine Negative Regular Itemset using Vertical Database

NVS. Pavan Kumar*
School of Computing
K L University
Andhra Pradesh, India
nvspavankumar@gmail.com

M. Sreedevi
School of Computing
K L University
Andhra Pradesh, India
msreedevi_27@yahoo.co.in

G. Vijay Kumar
School of Computing
K L University
Andhra Pradesh, India
gvijay_73@yahoo.co.in

Abstract: Recently, mining negative patterns has received some attention from the researchers because there are huge numbers of such patterns that can be derived from transactional databases. Mining negative patterns are relatively tough compared to positive patterns. They are logically less in number and the constraints to find them are often complex. In this paper we propose an algorithm (RN2I) to find all possible itemsets of size two which are regular and negative from a transactional database using vertical data format. Our experiment results show the efficiency and effectiveness of the algorithm at a user given regularity threshold.

Keywords: Negative Regular patterns, Vertical databases, Transactional database, Regularity threshold.

I. INTRODUCTION

Like Frequent Patterns [1], infrequent patterns [2] also playing an important role in knowledge discovery in databases. Mining infrequent patterns is a challenging endeavor because there is an enormous number of such patterns that can be derived from a given data set. Infrequent itemsets are not extracted by standard frequent itemset generation algorithms such as Apriori [3], and FP-growth [1]. Previous technique like considering every item as a symmetric binary variable is found not suitable for large number of items. Infrequent patterns, negative patterns and negative correlated patterns [2] are closely related concepts. Infrequent patterns and negatively correlated patterns refers to itemsets those contain positive items where as negative patterns refer to itemsets those contain both positive and negative items.

Let $I = \{i_1, i_2, . . ., i_n\}$ be a set of items. A negative item, $\bar{i}_k$ denotes the absence of item i_k from a given transaction. For example, $\overline{cofee}$ is a negative item whose value is 1 if a transaction does not contain coffee. A negative itemset X is an itemset that has the following properties:(1) $X = A \cup \overline{B}$, where A is a subset of positive items, $\overline{B}$ is a set of negative items, $|\overline{B}| \geq 1$, and (2) s(X) ≥ *minsup*. A regular itemset is an itemset which occurs regularly in a database for a user given regularity threshold. A negative regular itemset is of the form $X = A \cup \overline{B}$, where A and $\overline{B}$ are regular itemset.

In this paper we focus on finding itemsets which are regular and also negative. In our observation there is no much effort made in this direction earlier. For finding regularity of an itemset we opted vertical format of the transaction database because it requires less memory and read/write operations and is proven in earlier research. Hence in our paper we are presenting an algorithm/pseudo-code to find regularity of an item and at the same time the negative regular items which are associated with this. Our algorithm successively eliminates the items which are not regular and finds all negative itemsets among regular itemsets. Our experiment results show the efficiency of the algorithm with different number of transactions and variable user given regularity threshold (λ).

II. RELATED WORK

Mining Negative Association Rules (NAR) and Positive Association Rules (PAR) in the form A(e.g., (a_1a_2)) ⇒ B, A ⇒ ¬B (e.g., ¬(b_1b_2)), ¬A⇒B and ¬A⇒¬B from frequent and infrequent itemsets are mainly discussed in [4] [5] [6]. The left side and right side of rules are all positive or all negative. These papers do not deal with Negative Frequent Itemsets (NFIS). Zhang et. al [7] introduced a concept called maximum support to decrease the number of frequent items and exclude meaningless association rules because search space become more time consuming while mining generalizing association rules with negative items. This paper used Apriori-Like method to get NFIS. In [8] they proposed the FP-growth by importing a bit string for each node of the frequent pattern tree to store the prefix. NFIS were mined through extending frequent patterns on the frequent pattern tree. In [9] they proposed a tree based algorithm called Free-PNP to find negative frequent patterns and complete negative association rules.

III. PROBLEM DEFINITION

Let I = {i_1, i_2, ...,i_n} be an item set of n items. DB is a database consisting of transactions T = {t_1, t_2, t_3… t_m}. The total number of transactions m=|DB|. A transaction is of the form t=(Tid, X) where X is a set of items called *pattern and* X ⊆ I. $T^X = \{t_j^X, \ldots, t_k^X\}$, j ≤ k and j, k ∈ [1, m]. Set of maximum regularity MR={mr_1, mr_2, mr_3, … mr_k} where mr_i is the maximum regularity of an item i. Maximum regularity mr_i of an item i is calculated by finding max{r/ $r_i = t_{j+1}^i – t_j^i$, ∀ t_i ∀ i, 1≤ j ≤ n }. When maximum regularity *mr* of an item set X is no greater than the user given threshold, X would be regular pattern. The user given threshold is denoted by λ and 1 ≤ λ ≤ |DB|.

IV. MINING NEGATIVE REGULAR ITEMSET

As this algorithm is developed based on vertical database, data from transaction database (Table 1) is to be converted into vertical format (Table 2). The format in transactional database is of type { Tid: Itemset} is been converted in to the vertical format of type {item : Tid-set}. Using this algorithm we can successfully find all negative regular itemsets of the form $A \cup \overline{B}$ and $\overline{A} \cup B$. i.e. Transactions which contains A does not contain B and vice versa using and operation on Tid-set of A and B.

In this algorithm we first find whether an item is regular using FindMaxReg function. In this function we find Tid-set of an item. Using a loop we calculate the regularity set *R* for all transactions, by finding the difference of the two consecutive transactions in which the item is present and then find the maximum regularity *MR* of the item (Table 2). If *MR* exceeds the user given threshold (e.g., λ =3) delete this item from database as it is not a regular item. Thus we find the regular items for which maximum regularity is no greater than the user given threshold, and we continue with the remaining items using an inner loop.

Table 1. Transactional Database

TID	Itemsets
1	a, c , f, g, h
2	b, d, e, f, g
3	a, e, f, g, h
4	b, e, d, g
5	c, f, g , h
6	a, f
7	d, c, f
8	b, d, g
9	a, c, h

In this process only regular items remains in the database. Here without scanning database again, we proceed to find the negative regular itemsets by doing intersection operation (25) of the Tid-Sets of two items. If the resultant set contains no transactions, then this pair is of the required form and is added to ResultSet (26). In this process only regular items remains in the database. Here without scanning database again, we proceed to find the negative regular itemsets by doing intersection operation (25) of the Tid-Sets of two items. If the resultant set contains no transactions, then this pair is of the required form and is added to ResultSet (26).

Pseudo code:

Algorithm NR2I :

```
a.   {
b.   ResultSet ← ϕ;
c.   m=|I|;
d.   V=|VDB|;
e.   for ( i=1 ; i<=m ; i++)
f.   { // finding regularity of items
g.      mr=0;
h.      mr=  FindMaxReg( item(i) );
i.       if (mr > λ )
j.       {
k.           delete item (i) from VDB;
l.         }
m.       else
n.   {// finding negative pairs
o.   for ( x=i+1; x<=m ; i++)
p.   {   mr=0;
q.      if mr of x is not calculated earlier then
r.       {
s.         mr=FindMaxReg(x);
t.         if (mr > λ )
u.          {
v.             delete item (x) from VDB;
w.          }
x.         else
y.          if { $t^i \cap t^X = \phi$ ; ∀ $v^i$  , ∀ $v^x$} then
z.               ResultSet ← (i , x);
aa.    }  }  } } }
     Algorithm:
bb.  FindMaxReg(item i )
cc.  {
dd.      mr=0;
ee.      $v^i$= number of transactions containing item i;
ff.      $T^i$= Transaction set containing i;
gg.         for( j=1; j<=v1; j++)
hh.         {
ii.              if (mr> $t_{j+1}^i – t_j^i$ )
jj.             {
kk.                  mr ← $t_{j+1}^i – t_j^i$;
ll.             }
mm.       return ( mr );
nn.  }
```

Table 2. Vertical Format of Itemset with R and MR

Itemset- I	Tid-Set	R	MR
A	1, 3, 6, 9	1, 2, 3, 3	3
B	2, 4, 8	2, 2, 4, 1	4
C	1, 4, 5, 7, 9	1, 3, 1, 2, 2	3
D	2, 4, 7, 9	2, 2, 3, 2	3
E	2, 3, 6	2, 1, 3, 3	3
F	1, 2, 3, 5, 6, 7	1, 1, 1, 2, 1, 1, 2	2
G	1, 2, 3, 4, 5, 8	1, 1, 1, 1, 1, 3,1	3
H	1, 3, 5, 9	1, 2, 2, 4	4

In this process we need not do repeated scans of database. Also we calculate maximum regularity only once for each item and would be skipped in the next iterations.

Table 3. Regular Items with R and MR

Itemset	TID-Set	R	MR
a	1, 3, 6, 9	1, 2, 3, 3	3
c	1, 4, 5, 7, 9	1, 3, 1, 2, 2	3
d	2, 4, 7, 9	2, 2, 3, 2	3
e	2, 3, 6	2, 1, 3, 3	3
f	1, 2, 3, 5, 6, 7	1, 1, 1, 2, 1, 1, 2	2
g	1, 2, 3, 4, 5, 8	1, 1, 1, 1, 1, 3, 1	3

While finding regularity *R*, we should have special approach to handle two special cases, first and last transactions of an item. For the first transaction of an item i , $r_1^i = t_{first}^i - 0$. For the last transaction of an item i , $r_{last}^i = m - t_{last}^i$. This algorithm successfully eliminates the items which are not regular and not negative. Finally all the remaining item sets in Resultset are negative regular itemset.

Table 4. ResultSet

Itemset
(a, d)
(c, e)
(d, h)

In above Table 4 we got negative regular itemset at a user givern regularity threshold (λ=3).

V. EXPERIMENT RESULTS

We performed experiments on various databases with different regularity threshold values and found interesting results. Here we are giving the results of various parameters while working with *T1014D100K*. Ideal conditions like more number of items and variable number of items present in each transaction suits well for our algorithm. The time taken in seconds for execution of three different number of transactions are presented in Figure 1. The way in which the number of negative regular itemset varies with maximum regularity threshold is shown in Figure 2.

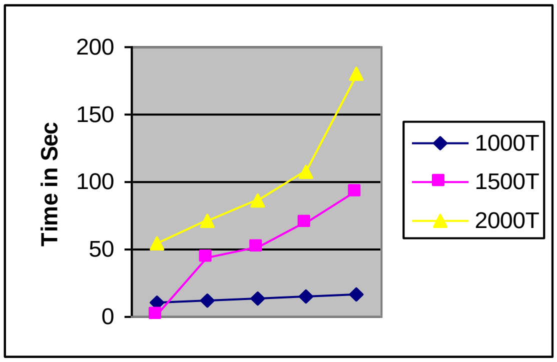

Figure 1. Execution time over T1014D100K

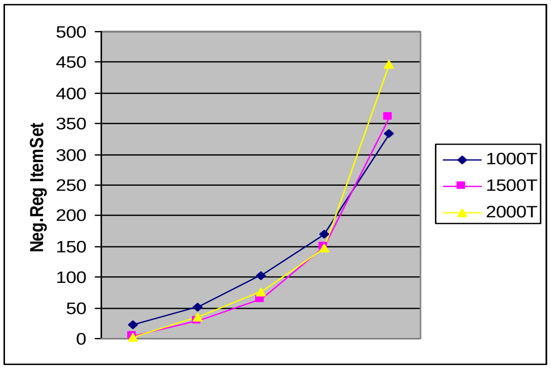

Figure 2. Number of negative regular itemsets found with different maximum regularity threshold values

VI. CONCLUSION

In this paper we presented an algorithm RN2I, works efficiently to find negative regular itemset of size two. Our algorithm is designed in a simple way, needs no complex structures. This algorithm first finds the regularity of an item and proceeds with the second to the last item, finding both regularity and negative itemset. Vertical data format is used because of the flexibility in finding intersection of Tid-set.

There is very much scope for expanding this algorithm to find negative regular itemset of size n. As our focus is to find negative regularity of itemset, using this algorithm we can find all item set of kind (A,¬B) and (¬A, B). We ignored (¬A,¬B) as it is not going to effect our search.

VII. REFERENCES

[1] Han, J., Yin, Y. Yin, “Mining Frequent Patterns without candidate generation”, In Proc. ACM SIGMOD international Conference on management of Data, PP. 1-12 (2000).

[2] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Introduction To Data Mining“ , Pearson Education, pp. 457-469.

[3] Jiawei Han, Micheline Kamber, “Data Mining : Concepts and Techniques”, 2nd ed. An Imprint of Elsevier, Morgan Kaufmann publishers, pp. 232-248, 2006.

[4] X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of both Positive and Negative Association Rules," ACM Transactions on Information Systems, 22, 2004, pp. 381-405.

[5] X Dong, Z Niu, X Shi, X Zhang, D. Zhu, "Mining Both Positive and Negative Association Rules from Frequent and lnfrequent Itemsets," Proceedings of the Third International Conference on Advanced Data Mining and Applications (ADMA 2007), Harbin, 519 China, August 6-8, 2007. Lecture Notes in Computer Science 4632, Springer 2007, pp. 122-133.

[6] M.L. Antonie, O. Zaiane, "Mining Positive and Negative Association Rules An Approach for Confined Rules," Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD04), LNCS 3202, Springer-Verlag Berlin Heidelberg, Pisa, Italy, 2004, pp.27-38.

[7] Y Zhang, C. Wang and Z. Xiong, "Improved algorithm of mining association rules with negative items," computer engineering and appl ications, 44(20), 2008, pp.169-171.

[8] Y Zhang, Z. Xiong, C. Wang and C. Liu, "Study on association rules with negative items based on bit string," Control and Decision, vo1.25 no.l, .Ian, 2010. pp.37-42.

[9] B. Yuan, L. Chen and Y . .lin, "Mining of negative frequent patterns in databases," Computer Engineering and Applications, 46(8), 2010, pp.I17-119.