SOTPNARM : SIZE OF TRANSACTION-BASED ON POSITIVE AND NEGATIVE ASSOCIATION RULE

RAKOTOMANANA René[1], ANDRIAMANOHISOA Hery-Zo[2], ROBINSON Matio[3]

[1] *Student in Doctoral School of Science and Technic Engineering and Innovation ,Laboratory of Cognitive Sciences and Application, High School Of Polytechnical University in Antananarivo, Madagascar*

,

[2] *Professor, in Doctoral School of Science and Technic Engineering and Innovation ,Laboratory of Cognitive Sciences and Application, High School Of Polytechnical University in Antananarivo, Madagascar*

[3] *Doctor , in Doctoral School of Science and Technic Engineering and Innovation ,Laboratory of Cognitive Sciences and Application, High School Of Polytechnical University in Antananarivo, Madagascar*

ABSTRACT

Association rules mining has many algorithm and technic to solve problems. Apriori algorithm is used in market data base. Support and Confidence are two values to calculate and must be compared with threshold minsupp and threshold minconf proposed by user. Improved algorithm about association rules tries to optimize the numbers of itemsets , the quality of association rules. One of the improved algorithm is the Size Of Transaction-based Association Rule Mining(SOTARM) . This paper shows how to extract item sets frequents positives and negatives using the size of transaction, named SOTPNARM. The aim of our proposed solution is not only the using of negative item in the transaction data base when the original item is unfrequented but the negation of the rare item can be among of the association rules found. We begin with the introduction that tells what sets are item frequent, itemsets frequents. Then the proposed solution using the size of transaction with example illustration followed the discussion about the results and we terminate with the conclusion and the perspective.

Keyword: *Association rule, Item, Item-sets, Size of Transaction, Item-sets positive, Item-sets negative, SOTPNARM*

1. INTRODUCTION

Many authors present their research' result about Data Mining. One of them is the association rules based on items and item sets frequents.

Apriori algorithm [1] considered as a classical algorithm and the based on this topic.

To find association rules in data base the resolution has two separate steps :

- the first step is to find items frequents and item-sets frequents using the threshold value minsupp,
- the second step is to create association rules with item-sets found in the first step with the threshold value minconf.

The advantage of the Apriori algorithm is its easy implementation [1][8]However, there is a major disadvantage of Apriori algorithm that it requires too many scans over the entire data sets which comprises of the database to find

rules which leads to high usage of the system as more memory is required to complete the need of increasing Input and Output cost that's why many improved Apriori algorithm exist.
Using the Size of Transaction to look for Frequent item sets is presented in [2] and [3] but they began to enumerate positive frequent item sets and create the possible combination of items to have item sets why they have association rules positives after and [3] uses the relation to calculate support between positive and negative for one item such as : if X is an item positive then the item negative associates is $\bar{X}$.

With this paper, we proposed the use of the Size Of Transaction (SOT) [3]to find item and item sets frequents at the first step (creating the set of 1-item) we test if one item given is frequent positive else we test its negation and keep the item frequent (positive or negative) and transform our data base if necessary. why the association rules result has positive item sets or negative item sets in the premise or/and in the conclusion.

Working with the association rules mining needs some information about the data base:

The extraction context is a 3_upplet named K = (O, I, R) that O is a set of object and I is a set of item (or attribute) , R is a binary relation formed by $O \times I$. Let $o \in O$ and $i \in I$, $(o, i) \in R$ is synonym of item i contents the o object. One transaction is identified by identifier named TID [1][2]that means Tuple IDentifier (TID).

For example: Table 1 bellow shows an example. Let I={a, b, c, d, e, f} and O={T_1,T_2,T_3,T_4,T_5,T_6} where a,b,c,d,e,f are attributes and T_1,T_2,T_3,T_4,T_5,T_6 are transactions.

Table 1 Example of Transaction data base

TID	Item list
T_1	ab
T_2	acde
T_3	cde
T_4	def
T_5	abcde
T_6	abc

Table 1 can be changed as shown in Table 2

Table 2: Interpretation

TID	a	b	c	d	e	f
T_1	x	x				
T_2	x		x	x	x	
T_3			x	x	x	
T_4				x	x	x
T_5	x	x	x	x	x	
T_6	X	X	X			

The x symbol in Table 2 marked that the attribute in the first line belongs to the TID in the same row.

Table 2 can be presented as binary data base where if item is present the value is 1 i.e. x becomes 1 else 0. We use Table 3 during our work. Table 3 bellow shows the binary data base from Table 1.

The size of the transaction is the number of the attributes that contains the transaction as for example T_1 has 2 attributes a and b the size of T_1 is 2 and T_1 is 2-items as a definition one transaction formed by k attributes is said k-items where $1 \leq k \leq |O|$ with $|O|$ is the size of the set O.

2. RELATED WORKS

To look for Association rules in data base the gait is divided on two problems whose the first is to find item sets frequents and the second is to build association rules;

Apriori algorithm was proposed by Agrawal et al.[1][2] and is used to find in the first step item sets frequents beginning to find 1-itemsets and create k-items sets candidates for k≥2 and k≤ |O| and in the second step the association rules .

Let X an item, the probability of transactions that contains X is the support: supp(X) :

$$\text{supp}(X) = \frac{|\{\text{Ti}/\ \text{X} \in\ \text{Ti}\}|}{|O|}$$

A minimal threshold value named minsupp is fixed by user, X is an item frequent if Supp(X) ≥ minsupp and X is a positive item. But If minsupp(X)<minsupp then X is unfrequent and the X complementary is a negative item using the notation $\bar{X}$

After obtaining item set frequents, authors continue to find Association Rules. An association rules is an implication between two items that X⟹ Y , X is the premise and Y is the conclusion. That means If X is present in the transaction then Y is present:

$$\text{supp}(\text{X} \Longrightarrow \text{Y}) = \text{supp}(X \cup Y)$$

The next step is to calculate the confidence of rule that

$$\text{conf}(\text{X} \Longrightarrow \text{Y}) = \frac{\text{supp}(\text{X} \Rightarrow \text{Y})}{\text{supp}(\text{X})}$$

As the same idea with support, a minimal threshold value named minconf is fixed by user.

One rule X⟹ Y is said valid or frequent if [1][2][3]

supp(X)≥minsupp and supp(Y)≥minsupp and supp(X⟹ Y) ≥minsupp and conf(X⟹ Y)≥ minconf

Association rules containing at least one item negative is said association rules negative as $X \Longrightarrow \bar{Y}$ or $\bar{X} \Longrightarrow Y$ or $\bar{X} \Longrightarrow \bar{Y}$. [6][8][7]

The use of the size of the transaction [3][4][5] is already presented by researchers but only determines the max size of item sets candidates not counting the recovery of negative items then they continue by looking for positive association rules and then treat the rules of negative associations by the use of the ratio supp (X) = 1-supp ($\bar{X}$) generally without the comparison with minsupp because a pattern X which is not frequent is not necessarily that $\bar{X}$ is frequent.

For example minsupp = 0.60, let X as an item supp (X) = 0.48 <0.60 and supp ($\bar{X}$) = 0.52 <0.60 in this case neither X nor $\bar{X}$are not frequent in order to decrease the number of rules of associations.

SOTARM proposed the use of the size of transaction with only the 1-item that has support greater than minsupp and at the end of the algorithm it test association rules infrequent using the negative item associate to find negative association rules that's why association rules based on rare item is missing in the result.

Table 3 : Binary Data Base

TID	a	b	c	d	e	f
T_1	1	1	0	0	0	0
T_2	1	0	1	1	1	0
T_3	0	0	1	1	1	0
T_4	0	0	0	1	1	1
T_5	1	1	1	1	1	0
T_6	1	1	1	0	0	0

3 SIZE OF TRANSACTION POSITIVE AND NEGATIVE ASSOCIATION RULE MINING

3.1 Proposed Solution

Notations:

SOTPNARM Size Of Transaction-based on Positive and Negative Association Rule Mining

NNEG: Number NEGative
ST: Size of Transaction
NbREP: Number of REPetition
SITEM: Set of ITEM
SRULE: Set of RULE
{Ti} =Set of items that belong to the transaction Ti
{Tii}=Set of items extract from {Ti} and having max support

The proposed solution has several steps:
At the first time we need to have the value of minsupp and the value of minconf
SITEM= ∅
SRULE= ∅
Step 1: Reading the Transaction Database (Table 1 and Table 2 in our Example)

Step 2: Transform into binary (Table 3 of our example) but add an array variable named NNEG of a row and each element of the array is initialized to 0.

After we calculate the support of each attributes i.e. the sum of the column divided by number of data base row, and realize a test with minsupp . if supp(X) ≥ minsupp then X is frequent and add X to the SITEM : SITEM = SITEM ∪ {X} else we take the negation of the column where is X and modify the 1 value to 0 value and vice versa i.e. we use the value of $\bar{X}$ and, with the constraint supp($\bar{X}$)≥ minsupp we keep $\bar{X}$ in the data base and NNEG for X becomes 1 finally SITEM =SITEM ∪ { $\bar{X}$ }. At the end of this step SITEM has the list of 1-item frequent positive or negative

Step 3: Count the size of each Transaction Ti .

From Table 3 we will count the size of each transaction Ti i.e. we will add a new column named ST which is the sum of the value present in the corresponding row. The newly obtained Table 4 is to be sorted compared to the last column. The first line of the sorted array then gives us the maximum size of the possible item sets (candidate items according to the term used in the Data Mining domain).

NB : If one row has ST<1 then we delete this row

Step 4: Transaction grouping of the same size: counting of transaction numbers same presentation, that is to say, patterns of the same composition of attributes. We then obtain a new Table 5 whose the last column is named NbREP (Number of REPetition).

Step 5: For each line of Table 5 we calculate the support of each line by dividing ST with the Sum of the transaction and we keep the line whose calculated support is greater than or equal to minsupp : SITEM =SITEM ∪ {Ti} and go to Step 6 else we create subset from the current line transaction having size ST - 1 to 2 and calculate each support and take the combination that has the max value SITEM= SITEM ∪ {Tii}.

Step 6: Creating Association Rule

After finding an item sets A frequents in step 5 , our calculate is used with three values where the first is A and the second is minconf and the third is minsupp to look for a sub-pattern Y without being equal to A such that supp (Y) = max {supp (T) / T ⊂A} (use of the Antimonotonie of the support) in this case the rule is Y ⟹(A-Y) subject to trust Conf(Y⟹ (A-Y))≥ minconf in this case the rule is Y⟹ (A-Y) and the rule is conjunctive type , to known if is it a negative rule , just search in NNEG if exist one at least among the constituent of A is 1 else it's a positive rule.

SRULE= SRULE ∪ {Y,A-Y}

Step 7 Listing Of Association rules

3.2 Illustration based example

We use Table 1 as the base of our illustration example
minsupp = 0.60
minconf = 0.40
Step 1: We already have Table 1 and Table ***2***

Step2: Table 3 and NNEG

Table 2: Binary Database with NNEG

NNEG	0	0	0	0	0	0
TID	a	b	c	d	e	f
T_1	1	1	0	0	0	0
T_2	1	0	1	1	1	0
T_3	0	0	1	1	1	0
T_4	0	0	0	1	1	1
T_5	1	1	1	1	1	0
T_6	1	1	1	0	0	0

Step 3 Compute support and transaction size

Table 3 Support for each item and ST

NNEG	0	0	0	0	0	0	
TID	**a**	**b**	**c**	**d**	**e**	**f**	**ST**
T_1	1	1	0	0	0	0	**2**
T_2	1	0	1	1	1	0	**4**
T_3	0	0	1	1	1	0	**3**
T_4	0	0	0	1	1	1	**3**
T_5	1	1	1	1	1	0	**5**
T_6	1	1	1	0	0	0	**3**
Support	**4/6**	**3/6**	**4/6**	**4/6**	**4/6**	**1/6**	

We eliminate the item f because its support supp(f) = 0.16 <0.6 supp($\bar{f}$) = 1- 0.16 = 0.84> 0.60 then we replace it with its negation $\bar{f}$. The main table of our work will be modified in turn because there will be changes and the value of NNEG = 1 to find that the column is a negative reason

Similarly for the item b, supp(b) = 0.50 <0.60, we eliminate b from the table and supp($\bar{b}$) = 1-supp (b) = 1-0.50 = 0.50 <0.60 from where there is no replacement.

After updating our main table, we notice:

- there is a change in the column ST,
- the most common reason is $\bar{f}$
- The absence of the pattern b
- SITEM={ {a},{c},{d},{e},{$\bar{f}$}}

Table 4 Data base sorting the ST column

NNEG	0	0	0	0	1	
TID	**a**	**c**	**d**	**e**	$\bar{f}$	**ST**
T_2	1	1	1	1	1	**5**
T_5	1	1	1	1	1	**5**
T_3	0	1	1	1	1	**4**
T_6	1	1	0	0	1	**3**
T_1	1	0	0	0	1	**2**
T_4	0	0	1	1	0	**2**
Support	**4/6**	**4/6**	**4/6**	**4/6**	**5/6**	

Step 4: Calculate the number of repetitions of each transaction of the same size and the same constituent items

Table 5: Presentation of transaction with repetition numbers

TID	a	c	d	e	$\bar{f}$	ST	NbREP
T_2	1	1	1	1	1	**5**	**2**
T_3	0	1	1	1	1	**4**	**1**
T_6	1	1	0	0	1	**3**	**1**
T_1	1	0	0	0	1	**2**	**1**
T_4	0	0	1	1	0	**2**	**1**
Support	**4/6**	**4/6**	**4/6**	**4/6**	**5/6**		

Step 5 Calculation support for each line

The maximum size of our pattern is of value 5 i.e. 5-items having a negative pattern.

Each line is now an item sets of our treatment the support of a line is none other than the nbRep/6. The {acde$\bar{f}$} pattern has 2/6 support and the rest is 1/6, hence none of these patterns are common. In this case we must look for the sub reasons that are frequent from this information to find the rules of association.

Because {acde$\bar{f}$} isn't an frequent item sets , our algorithm extracts the sub sets and take the support ≥minsupp.

Table 6 : Result of subset's support from T2 ={acde$\bar{f}$}

Subset with 4-items	Subset with 3-items	Subset with 2-items
{acde }supp=2/6	{acd} supp =2/6	{ac} supp= 3/6
{acd$\bar{f}$}supp=2/6	{ace} supp =2/6	{ad} supp=2/6
{ace$\bar{f}$} supp=2/6	{ac$\bar{f}$} supp =3/6	{ae} supp=2/6
{acde}supp =2/6	{ade} supp =2/6	**{a$\bar{f}$} supp=4/6**
{cde$\bar{f}$}supp =3/6	{ad$\bar{f}$} supp =2/6	{cd} supp=3/6
	{cde} supp =3/6	{ce} supp=3/6
	{cd$\bar{f}$} supp =3/6	***{c$\bar{f}$} supp=4/6***
	{ce$\bar{f}$} supp =2/6	**{de} supp=4/6**
	{de$\bar{f}$} supp =3/6	{d$\bar{f}$} supp=3/6
	{ae$\bar{f}$} supp =2/6	{e$\bar{f}$} supp=3/6

With minsupp=0.60 item sets frequents are SITEM={{a},{c},{d},{e},{$\bar{f}$}, {a$\bar{f}$},{c$\bar{f}$},{de}}

Step 6 With minconf = 0.40 Association Rules are (from the Table 6 bold italic) :

conf(a⟹ $\bar{f}$)= 1 >0.40

conf(c ⟹ $\bar{f}$)=4/5=0.80 >0.40

conf(d ⟹e)=1 >0.40

SRULE= { a⟹ $\bar{f}$, c⟹ $\bar{f}$, d ⟹e}

3.3 Discussion

We have just presented the result of our future system and this section is for comparison between Apriori, SOTARM and SOTPNARM

From the beginning, with Apriori, negatives items are not presented or in our example item b and item f are not frequent but we traits their negations items ant test them if they are frequent or no.

In the case of SOTARM first of all the two items b and f are not figured in the list of the items concerned because they are not frequent then even if we continue with the negative reasons are not presented only at the creation of the association rules: b and f are not part of the item to create the frequent item sets so the associated negatives will not be presented.

In addition the same rule of positive association is presented in the three methods.

Table 7 Use Apriori the remaining patterns for minsupp = 0.60

TID	a	c	d	E
T_1	1	0	0	0
T_2	1	1	1	1
T_3	0	1	1	1
T_4	0	0	1	1
T_5	1	1	1	1
T_6	1	1	0	0
support	4/6	4/6	4/6	4/6

Table 8 SOTARM Usage for minsupp = 0.60

TID	a	c	d	e	ST	Nbrep
T_2	1	1	1	1	4	2
T_3	0	1	1	1	3	1
T_4	0	0	1	1	2	1
T_6	1	1	0	0	2	1
~~T_1~~	~~1~~	~~0~~	~~0~~	~~0~~	~~1~~	~~1~~
support	4/6	4/6	4/6	4/6		

For the last two tables we have: supp (acde) = 2/6, supp (cde) = 3/6, supp (de) = 4/6, supp (ac) = 3/6.

4. CONCLUSION

By way of conclusion, the approach presented here shows us the means of treating the obtaining of positive or negative item sets and the corresponding association rules. The approach was based on the implementation of the management of the size of each transaction by adding the negative item in case the value of their corresponding media is greater than minsupp.

As a perspective, the implementation of a framework is necessary to visualize the results then to extend the result obtained in the search of the generalized association rules having in premise and / or in conclusion conjunction or disjunction of the positive or negative item sets.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in massive databases. In Proc. 1993 SIGMOD, pp. 207-216.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 1994 VLDB*, pp. 487-499.

[3] Asha Pandian et al. SOTARM: Size of transaction-based association rule mining algorithm, Department of Computer Science and Engineering, Sathyabama University, Chennai, Tamil Nadu, India, in Turkish Journal of Electrical Engineering & Computer Sciences, 2017 page 278-291

[4] Shalini Dutt, Naveen Choudhary & Dharm Singh. An Improved Apriori Algorithm based on Matrix Data Structure, Maharana Pratap University of Agriculture and Technology, India, Global Journal of Computer Science and Technology: C Software & Data Engineering Volume 14 Issue 5 Version 1.0 2014

[5] Tao Wan — Karine Zeitouni, "Mining Association rules with Multiple Min-supports Application to Symbolic Data " ,PRISM Laboratory, University of Versailles, 45, avenue des Etats-Unis 78035 Versailles Cedex, France, 2014

[6] Mohamed Anis Bach Tobji, Boutheina Ben Yaghlane, Extraction des itemsets fréquents à partir de données évidentielles : application à une base de données Éducationnelles, Laboratoire LARODEC, Université de Tunis, Institut Supérieur de Gestion 41 avenue de la Liberté, Cité Bouchoucha, Le Bardo 2000, Tunisie

[7] Xiangjun Dong, Liang Ma, and Xiqing Han, e-NFIS: Efficient Negative Frequent Itemsets Mining only based on Positive Ones, School of Information Science and Technology, Shandong Polytechnic University, Jinan 250353, China, IEEE 2011

[8] Komal Parmar1, Asst. Prof. Ravi Shukla, Enhanced Data Using Positive Negative Association Mining,in *IJARIIE-ISSN(O)-2395-4396, Vol-2 Issue-3 2016*