

Introduction:

Association Rule discovery is very successful and important mining task which aims at finding frequent patterns among transactions composed of data attributes or items. Rules are produced from all frequent patterns. The larger the set of frequent item set more the number of rules presented to the user and so many of them are redundant. And that is also true that as we lower the frequency requirement (Support), the set of association rule can rapidly grow. The number of redundant rules is more exponential in the length of longest frequent item set.

Association rule discovery mainly used in departmental stores, to know customer's buying habits. By knowing the buying patterns of the customers, sales can be increased. Say customer who buys milk, buys bread. So combining milk and bread together, making some deal sales of milk and bread can be increased. Or say very few customers are buying orange juice and so many people are buying milk, so buy combining milk and oranges juice together, sale of orange juice can be increased.

Background:

There has been a lot of research in developing efficient algorithms for mining frequent itemsets. SO many algorithms were proposed to mine frequent item sets, like Apriori, Aclose, Maxminer, CHARM etc. Most of the algorithms enumerate all frequent item sets. Using this for rule generates so many redundant rules. Some of them enumerate maximal frequent item sets. We cannot use maximal frequent item set for rule generation because using them for rule generation we loose information.

So in this paper, a new framework is presented for association rule mining based on concept of **CLOSED frequent item sets**. The set of all closed frequent item set is lot smaller then the set of all frequent itemsets, in some cased by 3 or more orders of magnitude. Closed itemsets can be found in a fraction of time it takes to mine all frequent itemsets and the number of rules returned to the user can be smaller by a factor of 3000 or more (The gap widens for smaller frequency values).

The main purpose of this paper is to explain the concept of closed frequent item sets. Closed frequent item sets are very less in number compare to frequent item sets mined by traditional algorithms. So the number of rules produced from CLOSED frequent item set are very less in numbers compare to the rules generated by traditional methods.

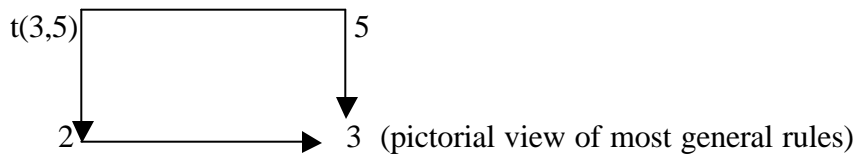
Problem explanation:

Lets say we have transaction set

1	2	3
1	3	5
2	3	5
2	3	
2	3	5
3	5	

Now in this case we have frequent item sets 13,23,25,35 and 235. We can generate following rules from 235 frequent item sets.

- 1) 2 ->3 (100% conf rule)
- 2) 5 ->3 (100 % conf rule)
- 3) 2, 5 ->3
- 4) 2 -> 3, 5
- 5) 5 -> 2, 3



We say that first two rules are more general and the rest of the rules are redundant because third rule we can find from augmentation for 1 and 2. The fourth and the fifth rule we can find by adding one item to consequent or antecedent. So the last three rules are redundant.

Now consider lets take another example to understand redundant rules.

2	3	5
2	3	5
3	5	
3	5	
2	3	5
3	5	

In this case we have frequent item sets 23, 25,35 and 235. So we have following rules from these item sets.

1)  $2 \rightarrow 3$  (100 % conf rule)

2)  $3 \rightarrow 5$  (100 % conf rule)

3)  $2 \rightarrow 5$

4)  $2 \rightarrow 3, 5$

5)  $3 \rightarrow 2, 5$

6)  $2,5 \rightarrow 3$

7)  $2, 3 \rightarrow 5$



Now in this case we can see that first two rules are more general than rest of the rules. We can find third rule by applying transitivity property between rule 1 and rule 2. And last 4 rules we can get by adding one more item to rules 1 or 2. So rules 1 and 2 are more specific than last 4 rules. Therefore we can say that rules 1 and 2 are more general and rests of the rules are redundant.

So to reduce the redundancy we find closed frequent item sets, instead of frequent item sets. CLOSED frequent item sets are very less in numbers as we will see later.

#### Definition of Redundant rule:

Let  $R_i$  denote the rule  $X1_i \rightarrow X2_i$ . We say that a rule  $R1$  is more general than a rule  $R2$ , denoted  $R1 \leq R2$  provided that  $R2$  can be generated by adding additional items to either the antecedent or consequent of  $R1$ .

#### CLOSED frequent item set:

Now lets look at following example to understand what is closed frequent item set.

<b>TID1</b>	<b>A</b>	<b>B</b>		
<b>TID2</b>	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>
<b>TID3</b>	<b>A</b>	<b>C</b>		
<b>TID4</b>	<b>A</b>	<b>C</b>	<b>D</b>	
<b>TID5</b>	<b>A</b>	<b>B</b>	<b>C</b>	

Now lets say  $T = \{1,2,3,\dots,m\}$  is a set of transaction and we call it tid set and  $I = \{A,B,C,\dots,Z\}$  is set of Items and we call it item set. Then input database is a binary relation  $\delta \subseteq I * T$ . If an item  $i$  occurs in transaction  $t$  then we say that  $(i, t) \in \delta$ , or alternatively  $i\delta t$ .

In our example  $I = \{A,B,C,D\}$  and  $T = \{1,2,3,4,5\}$

The fourth transaction can be represented as  $\{A\delta 4, c\delta 4, D\delta 4\}$ .

A set  $X \subseteq I$  is called an Item set and a set  $Y \subseteq T$  is called a tid set.

Now let the binary relation  $\delta \subseteq I * T$  be the input database for association mining. Let  $X \subseteq I$  and let  $Y \subseteq T$ .

The mappings

$$t: I \rightarrow T, t(X) = \{y \in T / \forall x \in X, x\delta y\}$$

$$i: T \rightarrow I, I(y) = \{x \in I / \forall y \in Y, x\delta y\}$$

Defines Galois connection.

For example  $t(AB) = t(A) \cap t(B) = 12345 \cap 125 = 125$

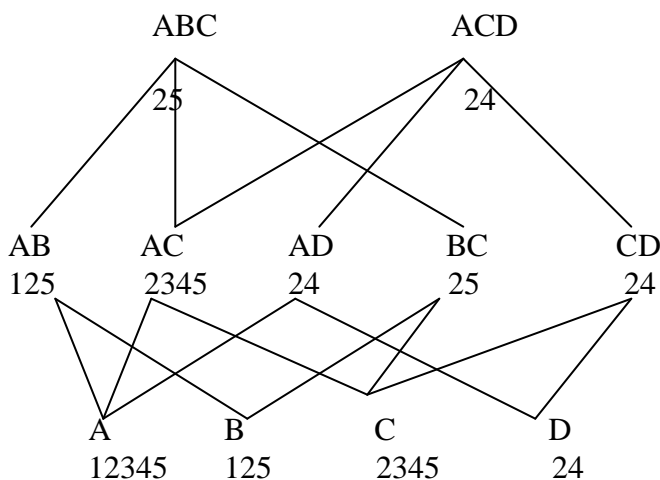
And  $I(235) = t(2) \cap t(3) \cap t(5) = ABCD \cap AC \cap ABC = AC$

Now we define a closed item set as an  $X$ , that is that same as its closure. i.e.  $X = \text{Cit}(X)$ .

And closed tidset is a tidset  $Y = \text{Cti}(Y)$ .

For example **ABC is closed item set** because  $\text{Cit}(ABC) = I * t(ABC) = I(25) = ABC$

While BC is not closed item set because  $\text{Cit}(BC) = I * t(BC) = I(25) = ABC$



So in this example **CLOSED frequent item sets are ABC, ACD, AB, AC, A** total of 5.

While **frequent item sets are ABC, ACD, AB, AC, AD, BC, CD, A,B,C,D** total of 11.

So we can see that we reduced the number of frequent item set by the concept of CLOSED frequent item set. So we can reduce the number of rules generated from closed frequent item sets and at the same time we don't lose any information.

Complexity of rule generation

Traditional Framework:  $O(f \cdot 2^l)$

Where  $f$  = total number of frequent itemsets

$l$  = length of longest frequent itemsets

In Example,

$f = 11, l = 3$

Number of rules generated =  $O(f \cdot 2^l) = 88$

With new Framework

$f = 5, l = 3$

Number of rules generated =  $O(f \cdot 2^l) = 40$

Results:

Experiments say that number of closed frequent item sets can be much smaller than the set of all frequent item sets. And time taken for closed frequent item set is also very less compared to Apriori algorithm which mines all frequent item sets. At the same time the third and the main advantage of CHARM Algorithm, which generates CLOSED frequent item sets, is reduction in number of rules.

Say for the chess database with 80% of minimum support we find 8227 frequent item sets Using Apriori Algorithm and 5083 Closed frequent item sets using CHARM, which is almost half. And we can see time taken and rules generated are also less compared to Apriori.

Database	Sup	Length	No of frequent item sets	No. of closed frequent item sets	ratio	Time taken Apriori	Time taken CHARM	Ratio
chess	80%	10	8227	5083	1.6	18.54	1.92	9.7

(Number of item sets and running time)

Database	Sup	Length	All possible rule Traditional	Closed	Ratio
chess	80%	10	552564	27711	20

(Number of rules)

Conclusion:

This paper has demonstrated in a formal way , supported with experiments on several databases like chess, mushroom etc., the well known fact that the traditional association rule framework produces too many rules, most of which are redundant. This paper proposed a new frame work based on closed itemsets that can drastically reduce the rule set and that can be presented to the user in succinct manner.

And this paper opens a lot of interesting direction for future work. For example we plan to use the concept of lattice for interactive visualization and exploration of a large set of mined associations.