UDC 681.3.06

**A. U. Krakovetskiy, post graduate student**

# SEARCH METHOD OF ASSOCIATIVE RULES USING STRONG ITEMSETS AND FP-TREE

*The paper presents strong itemsets conception which solves the problem of a huge number of candidates generation in solving the task of knowledge search in the kind of associative rules. It had been suggested the search method of strong itemsets which do not intersect as well as the search method of associative rules on the basis on strong itemsets.*

***Keywords:*** *associative rules, FP-tree, Data Mining.*

## Preliminaries

Nowadays candidates generation is one of the most actual task in associative rules mining and knowledge discovery researches. There total number of rules candidates can be very large [1, 2]. There are some methods and approaches with are used for reducing their total number, such as an antimonotone rule [2, 3], experts knowledge, some assumptions etc. The approach based on the frequent-pattern tree (FP-tree) or transactions tree construction is also one of the methods for the candidates number reduction [4]. This method allows to find frequent itemsets but it has some shortcomings such as itemset confidence ignoring, lack of intuitive logic relationship with items in itemsets. Thus, solving a candidates generation problem is an actual task.

The given paper presents the conception of *strong* itemsets which solve the described problem, suggests the searching method of strong itemset which do not intersect, as well as the associative rules mining method on the basis on strong data sets.

## Problem Statement

The definition of the associative rules mining problem requires to find the number of itemsets with support and confidence greater than the minimum support $Supp_{\min}$, minimum confidence $Conf_{\min}$ and with improvement greater than 1 [1, 2]:

$$L = \{F \mid Supp(F) > Supp_{\min}, Conf(F) > Conf_{\min}, impr(F) > 1\}.$$

## Concept of Strong Itemsets

Let *the itemset* $X$ is a subset of elements of $A$ where $A$ is an element set (itemset):

$$X \neq \varnothing, X \subset A, \tag{1}$$

$LSI$ ($left-side\ itemset$) is a not-empty itemset which forms the left part of associative rule and $RSI$ ($right-side\ itemset$) is a not-empty itemset which forms the right part correspondingly:

$$LSI \Rightarrow RSI, LSI \neq \varnothing, RSI \neq \varnothing, LSI \cup RSI = \varnothing. \tag{2}$$

Let's assume that $LSI = X$ and $RSI = Y$, $0 \le \sigma, \tau \le 100$, where $\sigma, \tau$ are the minimum support and the minimum confidence [1, 2]. Associative rule is *possible* for specific $\sigma$ and $\tau$ if the following conditions are satisfied:

$$\begin{aligned} &Support(X) \ge \sigma, \\ &Support(X \cup Y) \ge \sigma, \\ &\frac{Support(X \cup Y)}{Support(X)} \ge \tau. \end{aligned} \tag{3}$$

If any itemset $X$ and $Y$ of the set $A$ ($X \cup Y = A, X \cap Y = \varnothing, |A| > 2$) form only possible

associative rules $X \Rightarrow Y$ for specific $\sigma$ and $\tau$ then $A$ is the *strong itemset*.

Let's consider an example. Let $A = \{a, b, c\}$ is a set which consists of three elements. The set of all possible variants of division $A$ to subsets which are the templates of associative rules consists of 12 elements (see table 1).

Table 1

**Data sets and appropriate rules for a set $A = \{a, b, c\}$**

| Itemset | Rule | Itemset | Rule |
|---|---|---|---|
| $\{a\}, \{b\}$ | $a \Rightarrow b$ | $\{a\}, \{b, c\}$ | $a \Rightarrow b, c$ |
| $\{b\}, \{a\}$ | $b \Rightarrow a$ | $\{b, c\}, \{a\}$ | $b, c \Rightarrow a$ |
| $\{a\}, \{c\}$ | $a \Rightarrow c$ | $\{a, b\}, \{c\}$ | $a, b \Rightarrow c$ |
| $\{c\}, \{a\}$ | $c \Rightarrow a$ | $\{c\}, \{a, b\}$ | $c \Rightarrow a, b$ |
| $\{b\}, \{c\}$ | $b \Rightarrow c$ | $\{a, c\}, \{b\}$ | $a, c \Rightarrow b$ |
| $\{c\}, \{b\}$ | $c \Rightarrow b$ | $\{b\}, \{a, c\}$ | $b \Rightarrow a, c$ |

If *all* rules which can be formed with these sets are possible then the set $A$ is strong.

Expression $C_N^m$ is a total number of choice variants of $m$ elements from $N$ ones which will make left part of associative rule. Expression $C_{N-m}^n$ is a total number of possible choice variants of $n$ elements from residual elements $N - m$. Since the left and the right part of associative rule must be available in the rules, the total number of combinations equals $C_N^m C_{N-m}^m$. In our case $m$ changes from 1 to $N-1$ and $n$ - from 1 to $N - m$. Let $m = |LSI|$ is the number of elements in the left rule side and $n = |RSI|$ - element number in the right rule side, $N = m + n$ is a total number of elements in a associative rule. Then the total number of rules which can be formed from the strong itemset $A = LSI \cup RSI$, which consist of $N$ elements can be determined from the following expression:

$$\eta = \sum_{m=1}^{N-1} \sum_{n=1}^{N-m} C_N^m C_{N-m}^m . \quad (4)$$

*Necessary and sufficient condition that itemset is strong.* Let's assume that we have a set $A = \{a_1, a_2, ..., a_n\}$, $n \geq 2$, where $s_1, s_2, ..., s_n$ are supports of elements $a_1, a_2, ..., a_n$ appropriately, $s_0$ is support of set $A$, $\sigma$ and $\tau$ - minimum support value and minimum confidence value appropriately. Then $A$ is a strong itemset if the following conditions are satisfied:

$$mn \geq \sigma \text{ and } \frac{mn}{mx} \geq \tau \quad (5)$$

where $mx = \max\{s_0, s_1, ..., s_n\}$.

*Proving.* (Necessary condition). Let's assume that $mn \geq \sigma$ and $\frac{mn}{mx} \geq \tau$. Set $A$ is strong if $X \Rightarrow Y$ - possible associative rule for any not-empty item sets $X$ and $Y$, $X \cup Y = A$. It is necessary to prove that *a)* $Support(X) \geq \sigma$, *b)* $Support(X \cup Y) \geq \sigma$, *c)* $\frac{Support(X \cup Y)}{Support(X)} \geq \tau$. It is clear that $Support(X) \geq mn$. As $mn \geq \sigma$ it follows $Support(X) \geq \sigma$. Similarly we prove that $Support(X \cup Y) \geq \sigma$. $\frac{Support(X \cup Y)}{Support(X)} \geq \frac{mn}{Support(X)} \geq \frac{mn}{mx}$ because of $\frac{mn}{mx} \geq \tau$ so all conditions are satisfied. So $X \Rightarrow Y$ is a possible associative rule. We proved that $A$ is a strong itemset.

(Sufficient condition). Let's assume that set $A$ is strong. We have to prove that *a)* $mn \geq \sigma$, *b)*

$\frac{mn}{mx} \geq \tau$. As $mn = \min\{s_0, s_1, ..., s_n\}$, $mx = \max\{s_0, s_1, ..., s_n\}$, let $X = \{a_k\}$ and $Support(X) = Support(\{a_k\}) = mx$. Let $Y = H \setminus X$ is set of all elements of $H$ except for $a_k$. As $H$ is a strong itemset so all conditions *1)* $Support(X) \geq \sigma$, *2)* $Support(X \cup Y) \geq \sigma$, *3)* $\frac{Support(X \cup Y)}{Support(X)} \geq \tau$ are satisfied for the rule $X \Rightarrow Y$.

In the worst case $Support(X \cup Y) = s_0 = mn$ so it follows that $mn \geq \sigma$. Then $\frac{Support(X \cup Y)}{Support(X)} \geq \frac{mn}{mx}$ so $\frac{mn}{mx} \geq \tau$. It follows from *3)*.

## Disjoint Strong Itemsets Searching Method

*Disjoint Strong Itemsets Searching Method* presents the strong searching method of itemsets which do not disjoint with each other in the database of transactions.

The basic idea of this method is to find all frequent patterns itemsets which satisfy condition (5). If the strong data set is found than it is removed from the original database and the same searching process is repeated for a new base.

Let $D$ is database which consists of $N$ transactions $T$, $\sigma$ is minimum support, $\tau$ is minimum confidence.

Lets assume that $F = \{f_1, f_2, ..., f_n\}$ is a frequent elements set i.e. the following conditions are satisfied for them:

$$Support(f_i) \geq \sigma, Confidence(f_i) \geq \tau, f_i \in F, i = 0..|F|. \quad (6)$$

Clearly that the strong itemsets consist only of elements of the set $F$:

$$f_i \in s_j, f_i \notin s_k, j \neq k, f_i \in F. \quad (7)$$

As a result we will have a set of disjoint strong itemsets $S$:

$$\begin{gathered} \forall(s_i \cap s_j) = \varnothing, \ i \neq j, s_i, s_j \in S, \\ Support(s_i) \geq \sigma, \ Confidence(s_i) \geq \tau, \end{gathered} \quad (8)$$

and a set of "superfluous" elements $F'$ which do not exist in the strong itemsets:

$$F' = \{f_i \notin (\forall s_j \in S)\}. \quad (9)$$

This method can be used as an alternative frequent itemsets searching method. However this method can be used as an initial stage for the associative rules mining method on the basis of strong itemsets which is considered below.

## Associative Rules Mining Method Using Strong Itemsets

This method is a modification of Apriori algorithm [5]. It allows to find the associative rules without generating a huge number of candidates. The Input data for this method are the set of disjoint strong itemsets $S = \{s_1, s_2, ..., s_n\}$ and set of frequent elements which do not get into the strong itemsets $F = \{f_1, f_2, ..., f_n\}$.

As a result we receive a set of strong itemsets (which can contain the same elements) and associative rules formed on the basis of these sets.

In this method the candidates with the length $k$ generate on the basis of intersection of strong itemsets and frequent elements:

$$C_k = \{\{u, v\} \mid (1 \le i, j \le k, u \in s_i \wedge v \in s_j \wedge i \ne j) \vee$$
$$\vee (1 \le i \le m, 1 \le j \le k, u \in f_i \wedge v \in s_j) \vee \quad (10)$$
$$\vee (1 \le i, j \le m, u \in f_i \wedge v \in f_j \wedge i \ne j)\},$$

where $m = |F|$ is a total number of frequent elements.

Among all the candidates using (5) allows to find the set of strong itemsets $L_k$ concatenated with the set $S$:

$$S = S \cup L_k,$$

and the elements which do not exists in the $L_k$ are removed from the set $F$:

$$F = F \setminus \{L_k\}.$$

This process is repeated for all the candidates of longer length until $L_k \ne \varnothing$. The improvement calculation is a final stage for the associative rules mining process.

Described method allows considerably decrease a total number of candidates which are generated and the resulting strong itemsets allow to show relationships between individual itemsets and elements that exist in them.

## Conclusions

The given paper considers the conception of strong itemsets which solve the problem of generation of huge number of ineffective rules. A disjoint string itemsets searching method and the associative rules mining method on the basis of strong itemsets are suggested. The developed methods can be used for the knowledge discovery and associative rules mining problem in the economy, biology, engineering and scientific researches. In addition, the strong itemsets searching method can be used independently for logical relationships representation finding both between the individual sets and individual elements in them.

## REFERENCES

1. Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.

2. Дюк В.А., Самойленко А.П. Data Mining: учебный курс. – СПб.: Питер, 2001.

3. Чубукова И.А. Data Mining БИНОМ. Лаборатория знаний, Интернет-университет информационных технологий - ИНТУИТ.ру, 2006.

4. Han, J., J. Pei, Y. Yin, “Mining Frequent Patterns without Candidate Generation”, Proc. ACM SIGMOD 2000.

5. Agrawal R., T. Imielinski, A. Swami, “Mining Associations between Sets of Items in Massive Databases”, Proc. ACM SIGMOD 1993. - p. 207 – 216.

***Krakovetskiy Alexandr Uriyovitch*** **–** graduate student of the department computer controlled systems.

Vinnitsa National Technical University.