Mining Positive and Negative Association Rules: An Approach for Confined Rules

Maria-Luiza Antonie Osmar R. Zaïane

Department of Computing Science, University of Alberta
Edmonton, Alberta, Canada
{luiza, zaiane}@cs.ualberta.ca

Abstract. Typical association rules consider only items enumerated in transactions. Such rules are referred to as *positive association rules*. *Negative association rules* also consider the same items, but in addition consider negated items (i.e. absent from transactions). Negative association rules are useful in market-basket analysis to identify products that conflict with each other or products that complement each other. They are also very convenient for associative classifiers, classifiers that build their classification model based on association rules. Many other applications would benefit from negative association rules if it was not for the expensive process to discover them. Indeed, mining for such rules necessitates the examination of an exponentially large search space. Despite their usefulness, and while they were referred to in many publications, very few algorithms to mine them have been proposed to date. In this paper we propose an algorithm that extends the support-confidence framework with a sliding correlation coefficient threshold. In addition to finding confident positive rules that have a strong correlation, the algorithm discovers negative association rules with strong negative correlation between the antecedents and consequents.

1 Introduction

Association rule mining is a data mining task that discovers relationships among items in a transactional database. Association rules have been extensively studied in the literature for their usefulness in many application domains such as recommender systems, diagnosis decisions support, telecommunication, intrusion detection, etc. The efficient discovery of such rules has been a major focus in the data mining research community. From the original *apriori* algorithm [1] there have been a remarkable number of variants and improvements of association rule mining algorithms [2].

Association rule analysis is the task of discovering association rules that occur frequently in a given data set. A typical example of association rule mining application is the market basket analysis. In this process, the behaviour of the customers is studied when buying different products in a shopping store. The discovery of interesting patterns in this collection of data can lead to important marketing and management strategic decisions. For instance, if a customer buys

bread, what is the probability that he/she buys milk as well? Depending on the probability of such an association, marketing personnel can develop better planning of the shelf space in the store or can base their discount strategies on such associations/correlations found in the data.

All the traditional association rule mining algorithms were developed to find positive associations between items. By positive associations we refer to associations between items existing in transactions (i.e. items bought). What about associations of the type: "customers that buy Coke *do not* buy Pepsi" or "customers that buy juice *do not* buy bottled water"? In addition to the positive associations, the negative association can provide valuable information, in devising marketing strategies. Interestingly, very few have focused on negative association rules due to the difficulty in discovering these rules.

Although some researchers pointed out the importance of negative associations [3], only few groups of researchers [4], [5], [6] proposed an algorithm to mine these types of associations. This not only illustrates the novelty of negative association rules, but also the challenge in discovering them.

1.1 Contributions of This Paper

The main contributions of this work are as follows:

1. We devise a new algorithm to generate both positive and negative association rules. There are very few papers to discuss and discover negative association rules. Our algorithm differs from those in the sense that it uses a different interestingness measure and it generates the association rules from a different candidate set.
2. To avoid adding new parameters that would make tuning difficult and thus impractical, we introduce an automatic thresholding on the correlation coefficient. We automatically and progressively slide the threshold to find strong correlations.
3. We compare our algorithm with other existing algorithms that can generate negative association rules and discuss their performances.

The remainder of the paper is organized as follows: Section 2 gives an overview of the basic concepts involved in association rule mining. In Section 3 we introduce our approach for positive and negative rule generation based on correlation measure. Section 4 presents related work for comparison with our approach. Experimental results are described in Section 5 along with the performance of our system compared to known algorithms. We summarize our research and discuss some future work directions in Section 6.

2 Basic Concepts and Terminology

This section introduces association rules terminology and some related work on negative association rules.

2.1 Association Rules

Formally, association rules are defined as follows: Let $\mathcal{I} = \{i_1, i_2, ...i_n\}$ be a set of items. Let $\mathcal{D}$ be a set of transactions, where each transaction T is a set of items such that $T \subseteq \mathcal{I}$. Each transaction is associated with a unique identifier TID. A transaction T is said to contain X, a set of items in $\mathcal{I}$, if $X \subseteq T$. An *association rule* is an implication of the form "$X \Rightarrow Y$", where $X \subseteq \mathcal{I}, Y \subseteq \mathcal{I}$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ has a *support* s in the transaction set $\mathcal{D}$ if $s\%$ of the transactions in $\mathcal{D}$ contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \Rightarrow Y$ holds in the transaction set $\mathcal{D}$ with *confidence* c if $c\%$ of transactions in $\mathcal{D}$ that contain X also contain Y. In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions $\mathcal{D}$ consists of generating the rules that have a *support* and *confidence* greater than given thresholds. These rules are called *strong rules*, and the framework is known as the *support-confidence framework* for association rule mining.

2.2 Negative Association Rules

Example 1. Suppose we have an example from the market basket data. In this example we want to study the purchase of organic versus non-organic vegetables in a grocery store. Table 1 gives us the data collected from 100 baskets in the store. In Table 1 "organic" means the basket contains organic vegetables and "$\neg$ organic" means the basket does not contain organic vegetables. The same applies for non-organic. On this data, let us find the positive association rules in the "support-confidence" framework. The association rule "non-organic $\rightarrow$ organic" has 20% support and 25% confidence (supp(non-organic $\wedge$ organic)/supp(non-organic)). The association rule "organic $\rightarrow$ non-organic" has 20% support and 50% confidence (supp(non-organic $\wedge$ organic)/supp(organic)). The support is considered fairly high for both rules. Although we may reject the first rule on the confidence basis, the second rule seems a valid rule and may be considered in the data analysis. Now, let us compute the statistical correlation between the *non-organic* and *organic* items. A more elaborated discussion on the correlation measure is given in Section 3.1. The correlation coefficient between these two items is -0.61. This means that the two items are negatively correlated. This measure sheds a new light on the data analysis on these specific items. The rule "organic $\rightarrow$ non-organic" is misleading. The correlation brings new information that can help in devising better marketing strategies.

The example above illustrates some weaknesses in the "support-confidence" framework and the need for the discovery of more interesting rules. The interestingness of an association rule can be defined in terms of the measure associated with it, as well as in the form an association can be found.

Brin *et. al* [3] mentioned for the first time in the literature the notion of negative relationships. Their model is chi-square based. They use the statistical

Table 1. Example 1 data

	organic	¬organic	$\sum_{row}$
non-organic	20	60	80
¬non-organic	20	0	20
$\sum_{col}$	40	60	100

Table 2. 2x2 contingency table

	Y	$\neg Y$	$\sum_{row}$
X	f_{11}	f_{10}	f_{1+}
$\neg X$	f_{01}	f_{00}	f_{0+}
$\sum_{col}$	f_{+1}	f_{+0}	N

test to verify the independence between two variables. To determine the nature (positive or negative) of the relationship, a correlation metric was used. In [6] the authors present a new idea to mine strong negative rules. They combine positive frequent itemsets with domain knowledge in the form of a taxonomy to mine negative associations. However, their algorithm is hard to generalize since it is domain dependant and requires a predefined taxonomy. A similar approach is described in [7]. Wu *et. al* [4] derived a new algorithm for generating both positive and negative association rules. They add on top of the support-confidence framework another measure called *mininterest* for a better pruning of the frequent itemsets generated. In [5] the authors use only negative associations of the type $X \rightarrow \neg Y$ to substitute items in market basket analysis.

We define as *generalized negative association rule*, a rule that contains a negation of an item (i.e a rule for which its antecedent or its consequent can be formed by a conjunction of presence or absence of terms). An example for such association would be as follows: $A \wedge \neg B \wedge \neg C \wedge D \rightarrow E \wedge \neg F$. To the best of our knowledge there is no algorithm that can determine such type of associations. Deriving such an algorithm is not an easy problem, since it is well known that the itemset generation in the association rule mining process is an expensive one. It would be necessary not only to consider all items in a transaction, but also all possible items absent from the transaction. There could be a considerable exponential growth in the candidate generation phase. This is especially true in datasets with highly correlated attributes. That is why it is not feasible to extend the attribute space by adding the negated attributes and use the existing association rule algorithms. Although we are currently investigating this problem, in this paper we generate a subset of the generalized negative association rules. We refer to them as *confined negative association rules*. A confined negative association rule is one of the follows: $\neg X \rightarrow Y$, $X \rightarrow \neg Y$ or $\neg X \rightarrow \neg Y$, where the entire antecedent or consequent must be a conjunction of negated attributes or a conjunction of non-negated attributes.

3 Discovering Positive and Negative Association Rules

The most common framework in the association rules generation is the "support-confidence" one. Although these two parameters allow the pruning of many associations that are discovered in data, there are cases when many uninteresting rules may be produced. In this paper we consider another framework that adds to the support-confidence some measures based on correlation analysis.

Next section introduces the correlation coefficient, which we add to the support-confidence framework in this work.

3.1 Correlation Coefficient

Correlation coefficient measures the strength of the linear relationship between a pair of two variables. It is discussed in the context of association patterns in [8]. For two variables X and Y, the correlation coefficient is given by the following formula:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad . \tag{1}$$

In Equation 1, $Cov(X,Y)$ represents the covariance of the two variables and σ_X stands for the standard deviation. The range of values for ρ is between -1 and +1. If the two variables are independent then ρ equals 0. When $\rho = +1$ the variables considered are perfectly positive correlated. Similarly, When $\rho = -1$ the variables considered are perfectly negative correlated. A positive correlation is evidence of a general tendency that when the value of X increases/decreases so does the value of Y. A negative correlation occurs when for the increase/decrease of X value we discover a decrease/increase in the value of Y.

Let X and Y be two binary variables. Table 2 summarizes the information about X and Y variables in a dataset in a 2x2 contingency table. The cells of this table represent the possible combinations of X and Y and give the frequency associated with each combination. N is the size of the dataset considered.

Given the values in the contingency table for binary variables, Pearson introduced the ϕ *correlation coefficient* which is given in the equation 2:

$$\phi = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{+0}f_{+1}f_{1+}f_{0+}}} \quad . \tag{2}$$

We can transform this equation by replacing f_{00}, f_{01}, f_{10}, f_{0+} and f_{+0} as follows:

$$\phi = \frac{f_{11}(N - f_{10} - f_{01} - f_{11}) - f_{10}f_{01}}{\sqrt{f_{+0}f_{+1}f_{1+}f_{0+}}} \tag{3}$$

$$\phi = \frac{f_{11}N - f_{11}f_{10} - f_{11}f_{01} - f_{11}^2 - f_{10}f_{01}}{\sqrt{f_{+0}f_{+1}f_{1+}f_{0+}}} \tag{4}$$

$$\phi = \frac{f_{11}N - (f_{11} + f_{10})(f_{11} + f_{01})}{\sqrt{f_{+0}f_{+1}f_{1+}f_{0+}}} \tag{5}$$

$$\phi = \frac{Nf11 - f_{1+} * f_{f+1}}{\sqrt{f_{1+}(N - f_{1+})f_{+1}(N - f_{+1})}} \quad . \tag{6}$$

The measure given in Equation 6 is the measure that we use in the association rule generation.

Cohen [9] discusses about the correlation coefficient and its strength. In his book, he considers that a correlation of 0.5 is large, 0.3 is moderate, and 0.1 is small. The interpretation of this statement is that anything greater than 0.5 is large, 0.5-0.3 is moderate, 0.3-0.1 is small, and anything smaller than 0.1 is insubstantial, trivial, or otherwise not worth worrying about as described in [10].

We use these arguments to introduce an automatic progressive thresholding process. We start by setting our correlation threshold to 0.5. If no strong correlated rules are found the threshold slides progressively to 0.4, 0.3 and so on until some rules are found with moderate correlations. This progressive process eliminates the need for manually adjusted thresholds. It is well known that the more parameters a user is given, the more difficult it becomes to tune the system. Association rule mining is certainly not immune to this phenomenon.

3.2 Our Algorithm

Traditionally, the process of mining for association rules has two phases: first, mining for frequent itemsets; and second, generating strong association rules from the discovered frequent itemsets. In our algorithm, we combine the two phases and generate the relevant rules on-the-fly while analyzing the correlations within each candidate itemset. This avoids evaluating item combinations redundantly. Indeed, for each generated candidate itemset, we compute all possible combinations of items to analyze their correlations. At the end, we keep only those rules generated from item combinations with strong correlation. The strength of the correlation is indicated by a correlation threshold, either given as input or by default set to 0.5 (see above for rational). If the correlation between item combinations X and Y of an itemset XY, where X and Y are itemsets, is negative, negative association rules are generated when their confidence is high enough. The produced rules have either the antecedent or the consequent negated: ($\neg X \rightarrow Y$ and $X \rightarrow \neg Y$), even if the support is not higher than the support threshold. However, if the correlation is positive, a positive association rule with the classical support-confidence idea is generated. If the support is not adequate, a negative association rule that negates both the antecedent and the consequent is generated when its confidence and support are high.

The algorithm generates all positive and negative association rules that have a strong correlation. If no rule is found, either positive or negative, the correlation threshold is automatically lowered to ease the constraint on the strength of the correlation and the process is redone. Figure 1 gives the detailed pseudo-code for our algorithm.

Initially both sets of negative and positive association rules are set to empty (line 1). After generating all the frequent 1-itemsets (line 2) we iterate to generate all frequent k-itemsets, stored in F_k (line 8). F_k is verified from a set of candidate C_k computed in line 4. The iteration from line 2 stops when no longer frequent itemsets are possible. Unlike the join made in the traditional *Apriori* algorithm, to generate candidates at level k, instead of joining frequent $(k-1)$-itemsets, we join the frequent itemsets at level $k-1$ with the frequent 1-itemsets (line 4). This is because we want to extend the set of candidate itemsets and have the possibility to analyze the correlation of more item combinations. The rational will be explained later. Every candidate itemset generated this way is on one hand tested for support (line 7), and on the other hand used to analyze possible correlations even if its support is below the minimum support (loop from line 9 to 22). Correlations for all possible pair combinations for each candidate itemset are

Algorithm Positive and Negative Association Rules Generation
Input TD, $minsupp$, $minconf$, and ρ_{min}, respectively Transactional Database, minimum support, minimum confidence, and correlation threshold.
Output AR: Positive and Negative Association Rules.
Method:

```
(0)  if ρ_min is undefined then ρ_min = 0.5
(1)   positiveAR ← ∅; negativeAR ← ∅ /*positive and negative AR sets*/
(2)   scan the database and find the set of frequent 1-itemsets (F_1)
(3)   for (k = 2, F_{k-1} ≠ ∅, k++){
(4)      C_k= F_{k-1} ⋈ F_1
(5)      foreach i ∈ C_k {
(6)         s=support(TD,i) /*support of item i is computed*/
(7)         if s≥ minsupp then
(8)            F_k ← F_k ∪ {i} /*item i is added to F_k*/
(9)         foreach X,Y (i = X ∪ Y) {
(10)           ρ=correlation(X,Y) /*correlation btw X and Y is computed*/
(11)           if ρ ≥ ρ_min then
(12)             if s≥ minsupp then
(13)               if confidence(X → Y) ≥ minconf then
(14)                 positiveAR ← positiveAR ∪ {X → Y}
(15)             else if confidence(¬X → ¬Y) ≥ minconf and
                          supp(¬X¬Y) ≥ minsupp then
(16)               negativeAR ← negativeAR ∪ {¬X → ¬Y}
(17)           if ρ ≤ −ρ_min then /*ρ < 0 and |ρ| ≥ ρ_min */
(18)             if confidence(X → ¬Y) ≥ minconf then
(19)               negativeAR ← negativeAR ∪ {X → ¬Y}
(20)             if confidence(¬X → Y) ≥ minconf then
(21)               negativeAR ← negativeAR ∪ {¬X → Y}
(22)         }
(23)      }
(24)   }
(25)   AR ← positiveAR ∪ negativeAR
(26)   if AR = ∅ then {
(27)     ρ_min = ρ_min − 0.1
(28)     if ρ_min ≥ 0 then go to step (3)
(29)   }
(30)   return AR
```

Fig. 1. Discovering positive and negative confined association rules

computed. For an itemset i and a pair combination (X, Y) such that $i = X \cup Y$, the correlation coefficient is calculated (line 10). If the correlation is positive and strong enough, a positive association rule of the type $X \rightarrow Y$ is generated, if the $supp(X \cup Y)$ is above the minimum support threshold and the confidence of the rule is strong. Otherwise, if we still have a positive and strong correlation but the support is below the minimum support, a negative association rule of the type $\neg X \rightarrow \neg Y$ is generated if its confidence is above the minimum confidence threshold (lines 15-16). On the other hand, if the correlation test gives a strong negative correlation, association rules of the types $X \rightarrow \neg Y$ and $\neg X \rightarrow Y$ are generated and appended to the set of association rules if their confidence is adequate. The result is compiled by combining all discovered positive and negative association rules. Lines 26 onward, illustrate the automatic progressive thresholding for the correlation coefficient. If no rules are generated at a given correlation level, the threshold is lowered by 0.1 (line 27) and the process re-iterated.

4 Related Work in Negative Association Rule Mining

In this section, we discuss two known algorithms that generate negative association rules. We compare our approach with them later in the experiments section.

4.1 Negative Association Rule Algorithms

We give a short description of the existing algorithms that can generate positive and negative association rules. For more details, please refer to [4] and [5].

First, we discuss the algorithm proposed by Wu *et. al* [4]. They add on top of the support-confidence framework another measure called *mininterest* (the argument is that a rule $A \rightarrow B$ is of interest only if $supp(A \cup B) - supp(A)supp(B) \geq mininterest$). The authors consider as itemsets of interest those itemsets (positive or negative) that exceed minimum support and minimum interest thresholds. Although, [4] introduces the "mininterest" parameter, the authors do not discuss how to set it and what would be the impact on the results when changing this parameter. The approach differs from our algorithm in that in our algorithm we use the correlation coefficient as measure of interestingness, which was thoroughly studied in the statistics community. In addition, the value of our parameter is well defined and it is not as sensitive to the dataset as the *mininterest* parameter. In our algorithm (line 9) we compute the correlation coefficient for every pair X,Y of an item i where $i = X \cup Y$. As described earlier, when such a pair is found correlated an association rule is generated from it. In [4], they compute the interest for every pair X,Y of the item i where $i = X \cup Y$. However, they extract rules from itemset i only if any expression $i = X \cup Y$ exceeds the minimum interest threshold. We claim that by adding this condition they are loosing some potential interesting association rules. In addition, in our algorithm the candidate set C_k is generated as a join between F_{k-1} *and* F_1. In [4] the candidate set C_k is generated as a union of two frequent itemsets in F_i *for* $1 \leq i \leq k-1$. This turns out to be expensive. Since we all make the assumption that a k-itemset must have all its subsets in F_{k-1} we prove in the next theorem that our join generates the same itemsets as in [4].

Theorem All candidate items $c \in C_k$ generated by $F_i \bowtie F_j, 1 \leq i, j \leq k-1$ for which $\exists t \in c$ such that $t \in F_{k-1}$, can be discovered by $F_{k-1} \bowtie F_1$.

Proof Let us suppose $\exists c \in C_k$ such that $c \in F_i \bowtie F_j, 1 \leq i, j \leq k-1$ and $c \notin F_{k-1} \bowtie F_1$. Given the condition stated in theorem $\exists t \in c$ such that $t \in F_{k-1}$. Since $c \notin F_{k-1} \bowtie F_1$ and $t \in F_{k-1}$ it follows that $c - t \notin F_1$. This is false as $c - t$ is of length one and $c \in C_k$ was generated from frequent itemsets. Thus $\forall c \in C_k, c \in F_{k-1} \bowtie F_1$. Q.E.D

Second, we present the algorithm proposed in [5]. The algorithm is named by the authors SRM (substitution rule mining). We refer to it in the same way throughout the paper. The authors develop an algorithm to discover negative associations of the type $X \rightarrow \neg Y$. These association rules can be used to discover to which items are substitutes for others in market basket analysis. Their

Table 3. TD (a)

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

Table 4. TD (b)

TID	Items	Equivalent bit vector
1	$A, \neg B, C, D, \neg E, \neg F$	(101100)
2	$\neg A, B, C, \neg D, \neg E, \neg F$	(011000)
3	$\neg A, \neg B, C, \neg D, \neg E, \neg F$	(001000)
4	$A, B, \neg C, \neg D, \neg E, F$	(110001)
5	$A, \neg B, C, D.\neg E, \neg F$	(101100)
6	$\neg A, \neg B, \neg C, \neg D, E, \neg F$	(000010)
7	$\neg A, B, \neg C, \neg D, \neg E, F$	(010001)
8	$\neg A, B, C, \neg D, \neg E, F$	(011001)
9	$A, B, \neg C, \neg D, E, \neg F$	(110010)
10	$A, \neg B, \neg C, D, \neg E, \neg F$	(100100)

algorithm discovers first what they call *concrete items*, which are those itemsets that have a high chi-square value and exceed the expected support. Once these itemsets are discovered, they compute the correlation coefficient for each pair of them. From those pairs that are negatively correlated, they extract the desired rules (of the type $X \rightarrow \neg Y$). This paper, although interesting for the substitution items application, it is limited in the kind of rules that can discover.

Using the next example, which is an extension of the example presented in [5], we present some of the differences among the three algorithms.

Example 2. Let us consider a small transactional table with 10 transactions and 6 items. In Table 3 a small transactional database is given. To illustrate the challenges in mining negative association rules we create another transactional database where for each transaction, the complement of each missing item is appended to it. The new created dataset is shown in Table 4. This new database can be mined with the existing association rule mining algorithms. However, there are a few drawbacks of this naive approach. In practice, the data collections are very large, thus adding all the complemented items to the original database requires a large storage space. Not only the storage space has to increase considerably, but the execution times as well, in particular when the number of unique items in the database is very large. In addition, many association rules would be generated, many of them being of no interest to the applications at hand.

Using a minimum support of 0.2, the following itemsets are discovered using the three discussed algorithms. For this example the *correlation coefficient* was set to 0.5, and the *minimum interest* to 0.07.

In Table 5 and Table 6, the first column presents the results when our approach was used. The second column uses the algorithm from[4], while in the third one the results are obtained using the approach in [5]. In both tables the positive itemsets are separated by the negative ones by a double horizontal line. The positive itemsets are in the upper part of the tables. As it can be seen, for the 2-itemsets all three algorithms find the same positive ones. The differences occur for the negative itemsets. The itemset DF has a *minimum interest* of 0.09, but it has a *correlation* of only 0.42. That is why it is not found by our approach

Table 5. 2-itemsets

Correlation	Interest	Concrete
AD	AD	AD
BF	BF	BF
BD	BD	BD
CE	CE	
	DF	

Table 6. 3-itemsets

Correlation	Interest	Concrete
ACD		ACD
ABC	ABC	
ABD		ABD
BCD		

or by the SRM algorithm [5]. The itemset CE is not found by SRM because their condition is that the itemset should have higher correlation than the minimum value. In our approach the condition is to be greater or equal. Since the itemset CE has a *correlation* of 0.5 it is discovered by our algorithm, but not by SRM.

In Table 6 there are differences for both, the positive and the negative ones. The algorithm that uses the *minimum interest* parameter discovers only the ABC itemset because it is the only one that has all the pairs X,Y of the item ABC where $ABC = X \cup Y$ above the parameter. Although. all the other itemsets discovered by the other algorithms have at least two strong pairs they are not considered of interest. Our approach and SRM generate the same positive 3-itemset. The itemsets BCD and ABC are not discovered by SRM because none of its subsets of two items are generated as concrete during the process.

From the itemsets that were shown in Table 5 and Table 6 a set of association rules can be generated. Here we show, some of the rules that were generated from the itemsets that were discovered by one algorithm, but not by others. From itemset CE, the association rule $negE \rightarrow C$ can be found with support 0.5 and confidence of 62%. This rule seems to be strong, but it is missed by the SRM algorithm. From itemset DF, which is discovered only by the minimum interest algorithm, the association rules $negD \rightarrow F$ and $D \rightarrow \neg F$ can be discovered. However, both rules have support 0.3 and confidence of 42%. These rules could have been eliminated when the confidence threshold is set to 50%, thus our approach and SRM do not miss much by not generating them. In addition, our approach generates the 3-itemset BCD. From this itemset the rule $B \rightarrow \neg C \neg D$ is discovered and it has support of 0.2 and confidence of 60%.

5 Experimental Results

We conducted our experiments on a real dataset to study the behaviour of the algorithms compared. We used the Reuters-21578 text collection [11]. Reuters dataset had 6488 transaction, when only the ten largest categories were kept.

We compare the three algorithms discussed in the sections above. For each algorithm a set of values for their main interestingness measure was used in the experiments. Our algorithm and SRM [5] had the correlation coefficient set to 0.5, 0.4 and 0.3. In [4] the authors used the value 0.07 in their examples. We used this value and two others in its vicinity (0.05, 0.07 and 0.09). Each algorithm was run to generate a set of association rules. For lack of space the results are

Table 7. Results for Reuters text collection

(a) Results for rules of type $X \rightarrow Y$								
		#rules	supp	conf	PS	Q	IS	J
corr	0.4	235	**0.23±0.03**	**0.79±0.16**	**0.14±0.02**	0.84±0.28	**0.78±0.08**	**0.63±0.12**
int	0.07	219	**0.23±0.03**	0.79±0.18	0.13±0.02	**0.85±0.27**	0.76±0.09	0.61±0.14
SRM	0.4	297	0.22±0.03	0.76±0.20	0.12±0.03	0.82±0.27	0.73±0.10	0.57±0.15
(b) Results for rules of type $X \rightarrow \neg Y$								
		#rules	supp	conf	PS	Q	IS	J
corr	0.4	6	**0.33±0.10**	**0.99±0.0**	**0.11±0.01**	**0.99±0.0**	**0.72±0.05**	**0.52±0.08**
int	0.07	4	0.25±0.01	0.98±0.02	0.08±0.01	0.70±0.47	0.62±0.03	0.39±0.03
SRM	0.4	6	**0.33±0.10**	**0.99±0.0**	**0.11±0.01**	**0.99±0.0**	**0.72±0.05**	**0.52±0.08**
(c) Results for rules of type $\neg X \rightarrow Y$								
		#rules	supp	conf	PS	Q	IS	J
corr	0.4	6	0.33±0.10	**0.49±0.08**	**0.11±0.01**	**0.99±0.0**	**0.72±0.05**	**0.52±0.08**
int	0.07	4	**0.34±0.06**	0.46±0.09	0.08±0.01	0.70±0.47	0.67±0.06	0.45±0.08
(d) Results for rules of type $\neg X \rightarrow \neg Y$								
		#rules	supp	conf	PS	Q	IS	J
corr	0.4	1474	0.31±0.09	0.41±0.13	0.15±0.02	**0.84±0.20**	0.80±0.06	0.66±0.10
int	0.07	148	**0.49±0.09**	**0.67±0.13**	**0.16±0.05**	0.81±0.39	**0.87±0.08**	**0.77±0.11**

reported only for correlation coefficient 0.4 and minimum interest 0.07. For all the results, please see [12].

For these association rules a number of measures were computed: support (supp), confidence (conf), Piatetsky-Shapiro measure (PS), Yule's Q (Q), cosine measure (IS) and the Jaccard measure (J). These measures evaluate the interestingness of the discovered pattern. For more details on these measures for frequent patterns see [13]. In [13] a set of measures are compared and discussed. The measures are clustered with respect to their similarity. We chose to compute a few measures from different clusters to ensure the diversity of evaluation.

Tables 7 presents the results obtained for the Reuters dataset. We conducted the experiments with support 20% and confidence 0%. In each table a subset of the obtained rules are compared. Table 7 (a) compares rules of the type $X \rightarrow Y$, Table 7 (b) rules of the type $X \rightarrow \neg Y$, Table 7 (c) rules of the type $\neg X \rightarrow Y$ and Table 7 (d) rules of the type $\neg X \rightarrow \neg Y$. In each table the average of the measurement and the standard deviation are reported. The value in bold represents the best value for each measure.

Table 7 (a) shows that for positive association rules our approach tends to generate a more interesting set of rules compared to the other methods.

For rules of type $X \rightarrow \neg Y$ (Table 7 (b)) our approach and SRM perform best. They produce the same set of rules for correlation values of 0.4.

Table 7 (c) and Table 7 (d) compare our approach with the one in [4] only, since SRM algorithm does not generate this kind of rules.

In Table 7 (c) the symmetric rules of the ones in Table 7 (b) are generated, since the confidence is set to 0% and the correlation and minimum interest are computed for XY itemset.

However, for the rules of type $\neg X \rightarrow \neg Y$ (Table 7 (d)) the method in [4] generates a smaller set of rules, but with higher values for the measures.

6 Conclusions and Future Research Directions

In this paper we introduced a new algorithm to generate both positive and negative association rules. Our method adds to the support-confidence framework the correlation coefficient to generate stronger positive and negative rules. We compared our algorithm with other existing algorithms on a real dataset. We discussed their performances on a small example for a better illustration of the algorithms and we presented and analyze experimental results for a text collection. The results prove that our algorithm can discover strong patterns. In addition, our method generates all types of confined rules, thus allowing to be used in different applications where all these types of rules could be needed or just a subset of them.

Acknowledgements: This work was partially supported by Alberta Ingenuity Fund, iCORE and NSERC Canada.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. of SIGMOD. (1993) 207–216
2. Goethals, B., Zaki, M., eds.: FIMI'03: Workshop on Frequent Itemset Mining Implementations. Volume 90 of CEUR Workshop Proceedings series. (2003) http://CEUR-WS.org/Vol-90/.
3. Brin, S., Motwani, R., Silverstein, C.: Beyond market basket: Generalizing association rules to correlations. In: Proc. of SIGMOD. (1997) 265–276
4. Wu, X., Zhang, C., Zhang, S.: Mining both positive and negative association rules. In: Proc. of ICML. (2002) 658–665
5. Teng, W., Hsieh, M., Chen, M.: On the mining of substitution rules for statistically dependent items. In: Proc. of ICDM. (2002) 442–449
6. Savasere, A., Omiecinski, E., Navathe, S.: Mining for strong negative associations in a large database of customer transactions. In: Proc. of ICDE. (1998) 494–502
7. Yuan, X., Buckles, B., Yuan, Z., Zhang, J.: Mining negative association rules. In: Proc. of ISCC. (2002) 623–629
8. Tan, P., Kumar, V.: Interestingness measures for association patterns: A perspective. In: Proc. of Workshop on Postprocessing in Machine Learning and Data Mining. (2000)
9. Cohen, J.: Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum, New Jersey (1988)
10. Hopkins, W.: A new view of statistics. http://www.sportsci.org/resource/stats/ (2002)
11. Reuters-21578: (The Reuters-21578 text categorization test collection) http://www.research.att.com/~lewis/reuters21578.html.
12. Antonie, M.L., Zaïane, O.: Mining positive and negative association rules: An approach for confined rules. Technical Report TR04-07, Dept. of Computing Science, University of Alberta (2004) ftp://ftp.cs.ualberta.ca/pub/TechReports/2004/TR04-07/TR04-07.ps.
13. Tan, P., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proc. of SIGKDD. (2002) 32–41