A Novel Approach on Negative Association Rules

Ioannis N. Kouris
University of Patras, Greece

INTRODUCTION

Research in association rules mining has initially concentrated in solving the obvious problem of finding positive association rules; that is rules among items that exist in the stored transactions. It was only several years after that the possibility of finding also negative association rules became especially appealing and was investigated. Nevertheless researchers based their assumptions regarding negative association rules on the absence of items from transactions. This assumption though besides being dubious, since it equated the absence of an item with a conflict or negative effect on the rest items, it also brought out a series of computational problems with the amount of possible patterns that had to be examined and analyzed. In this work we give an overview of the works having engaged with the subject until now and present a novel view for the definition of negative influence among items.

BACKGROUND

Association rule mining is still probably the prominent method for knowledge discovery in databases (KDD), among all other methods such as classification, clustering, sequential pattern discovery etc. The discovery of association relationships among items in a huge database has been known to be useful in various sectors such as telecommunications, banking, transport and particularly in retail. Also it has been applied to various data sets such as census data, text documents, transactional data, medical images and lately biological data. In fact any data type that is collected and constitutes a large database of "baskets", each containing multiple "items" can fit this model. The prototypical application of association rules mining and probably until today the most extensively studied and popular one is the analysis of supermarket sales or basket data; hence the problem of finding association rules is often referred to as the "market-basket" problem. In this problem, we are given a set of items and a large collection of transactions which are subsets (baskets) of these items. The objective is to find relationships correlating various items within those baskets. More formally the specific task can be stated as follows:

Let $I=\{i_1, i_2,\ldots,i_n\}$ be a set of items and D be a database organized in multiple transactions T where each transaction $\mathbf{T} \in \mathbf{D}$ is a set of items such that $\mathbf{T} \subseteq \mathbf{I}$. An association rule is an implication of the form $X \rightarrow Y$, where $\mathbf{X, Y} \subseteq \mathbf{I}$ and $\mathbf{X} \cap \mathbf{Y} = \varnothing$, and expresses the possibility that whenever we find a transaction that contains all items in X, then this transaction is likely to also contain all items in Y. Consequently X is called the body of the rule and Y the head. The validity and reliability of association rules is expressed usually by means of support and confidence, succored by other measures such as for example implication, lift etc. An example of such a rule is {cell_phone, hands_free➔case (sup=70%, conf=90%)}, which means that 90% of the customers that buy a cell phone and a hands free device buy a cell phone case as well, whereas 70% of all our customers buy all these three things together.

The task of finding association rules was first formulated by Agrawal, Imielinski and Swami (1993). Since then there have been a series of works trying to improve various aspects such as the efficiency of Apriori based algorithms, implementation of new measures for finding strong – alternative rules, application of association rules framework into other domains etc. (for more details interested reader is referred to the work of Tan, Steinbach & Kumar, 2005).

On the other hand negative association rules have been initially either neglected or considered unimportant. It was not only but after several years since the first introduction of association rules that the problem of negative association rules was brought up. Also despite the fact that the existence and significance of

negative association rules has been recognized for quite some time, comparably a very small percentage of researchers have engaged with it.

MAIN FOCUS

Positive association rules take into consideration only the items that remained and finally appear in the stored transactions. Negative association rules on the other hand, at least as they have been traditionally defined in the literature up to now, take into consideration also the items that do not appear in a transaction. For example a positive association rule would consider all transactions containing pasta and cheese and would generate from them all corresponding association rules (i.e. **pasta → cheese**). An approach dealing with negative association rules on the other hand would also consider rules such as: those that buy pasta buy also cheese but not refreshments (**pasta → cheese** $\wedge \neg$ **refreshment**), or those that buy pasta but not bread buy also cheese (**pasta** $\wedge \neg$ **Bread → cheese**). The measures used for determining the strength of a correlation and for pruning the insignificant ones are again the support and confidence metrics.

A so called generalized negative association rule can include various negated and positive items (i.e. items existing and items absent from transactions) either in its antecedent or in its consequent. An example of such a rule would be: $\mathbf{A} \wedge \neg \mathbf{C} \wedge \neg \mathbf{F} \wedge \mathbf{W} \rightarrow \mathbf{B} \wedge \neg \mathbf{D}$ (i.e. people that buy items A and W but not C and F are likely to buy B but not D). However the obvious insurmountable obstacle created by such a contemplation is the number of possible itemsets that would have to be generated and counted as well as the number of rules created, since apart from the existing items we would have to consider all possible absent items and their combinations. Most approaches thus far have made various assumptions in order to come to approximate solutions, working on subsets of the problem of generalized association rules. For example considering rules where the entire consequent or antecedent could be a conjunction of similar items (i.e. negated or non-negated), considering negative rules only in the infrequent items, considering rules only between two items etc.. Nevertheless the absence of an item from a transaction does not necessarily imply direct negative correlation between items, but could be attributed to various other factors.

Previous Works on Negative Association Rules

To our knowledge Brin, Motwani & Silverstein (1997) were the first that have introduced and addressed the existence of negative relationships between items, by using a chi-squared test for correlations derived from statistics. In the specific work though there were mined negative associations only between two items. Subsequently Savasere, Omiecinski & Navathe (1998) and Yuan, Buckles, Yuan & Zhang (2002) have proposed two similar approaches, which tried to discover strong negative associations depending heavily on domain knowledge and predefined taxonomies. In the same context was the work of Daly & Taniar (2004), who organized items hierarchically following an ancestor – descendant scheme in order to avoid considering all negative items when no additional knowledge would be extracted from lower levels. As an example if customers do not buy refreshments after having bought pasta then we need not examine each and every kind of refreshment. Another work was that of Wu, Zhang & Zhang (2002), which used in addition to the support and confidence measure, a measure for more efficient pruning of the frequent itemsets generated called mininterest. Thiruvady & Webb (2004) proposed an extension of GRD - Generalized Rule Discovery algorithm (Webb, 2000) for mining negative rules. The proposed algorithm did not require the use of any minimum support threshold, but rather the specification of some interestingness measure along with a constraint upon the number of rules that would be finally generated. Finally Antonie & Zaiane (2004) proposed an algorithm for finding generalized association rules, where the entire antecedent or consequent consists of similar items (i.e. only negative or only positive items), thus generating only a subset of negative rules that they refer to as confined negative association rules. In their work they used in addition to the minimum support and minimum confidence measures, the correlation threshold as a third parameter. In the same context can also be considered to some extent works dealing with unexpected patterns (also known as surprising patterns), where the negative items can be thought as such patterns (e.g. Liu, Lu, Feng & Hussain, 1999; Hwang, Ho & Tang, 1999; Hussain, Liu, Suzuki & Lu, 2000; Suzuki & Shimura,