Category: Data Mining & Databases

Association Rules Mining for Retail Organizations

Ioannis N. Kouris
University of Patras, Greece

Christos Makris
University of Patras, Greece

Evangelos Theodoridis
University of Patras, Greece

Athanasios Tsakalidis
University of Patras, Greece

INTRODUCTION

In recent years, we have witnessed an explosive growth in the amount of data generated and stored from practically all possible fields (e.g., science, business, medicine, military just to name a few). However, the ability to store more and more data has not been followed by the same rate of growth in the processing power, and, therefore, much of the data accumulated remains today still unanalyzed. Data mining, which could be defined as the process concerned with applying computational techniques (i.e., algorithms implemented as computer programs) to actually find patterns in the data, tries to bridge this gap. Among others, data mining technologies include association rule discovery, classification, clustering, summarization, regression and sequential pattern discovery (Adrians & Zantige, 1996; Chen, Han, & Yu, 1996; Fayyad, Piatetsky-Shapiro, & Smyth, 1996). This problem has been motivated by applications known as market basket analysis which find items purchased by customers; that is, what kinds of products tend to be purchased together (Agrawal, Imielinski, & Swami, 1993).

BACKGROUND

The goal of the data mining task is to find all frequent itemsets above a user specified threshold (called *support*) and to generate all association rules above another threshold (called *confidence*) using these frequent itemsets as input. This type of information could be used for catalogue design, store layout, product placement, target marketing, and so forth. The prototypical application of this task has been the market basket analysis, but the specific model is not limited to it since it can be applied to many other domains (e.g., text documents [Holt & Chung, 2001], census data, [Brin et al., 1997], telecommunication data and even medical images, etc.). In fact, any data set consisting of "baskets" containing multiple "items" can fit this model. Many solutions have been proposed in the last years using a sequential or a parallel paradigm, experimenting on factors such as memory requirements, I/O scans, dimensionality reduction, and so forth.

The specific problem was first introduced by Agrawal et al. (1993) and an algorithm by the name AIS was proposed for effectively addressing it. Agrawal and Srikant (1994) have introduced a much more efficient solution and two new algorithms by the names Apriori and AprioriTid were proposed. Algorithm Apriori has been and still is the major reference point for all subsequent works. Most algorithms and approaches proposed thereafter (Toivonen, 1996; Brin et al., 1997; Park, Chen, & Yu, 1995; Han, Pei, & Yin, 2000) focus on either decreasing the number of passes made over the data or at improving the efficiency of those passes (i.e., by using additional methods for pruning the number of candidates that have to be counted). Among other things studied in association rule mining are: (1) incremental updating (Cheung, Han, Ng, & Wong, 1996; Lee, Lin &Chen, 2001), (2) mining of generalized and multi-level rules (Han & Fu, 1995; Srikant &Agrawal, 1995), (3) using multiple minimum supports (Liu, Hsu, & Ma, 1999), (4) mining of quantitative rules (Srikant & Agrawal, 1996), (5) parallel algorithms (Agrawal & Shafer, 1996; Park, Chen, & Yu, 1995).

Lately, it has been widely accepted that the model of association rules is either oversimplified or suffers from several omissions that cannot nowadays be considered insignificant especially in a retail environment. For example, treating the items as mere statistical probabilities and neglecting their real significance or handling them as Boolean variables and not taking into consideration their exact number of appearances in every transaction leads to fragmentary and dubious results (Liu, Hsu, & Ma, 1999). In this article we present a collection of works trying to solve all these problems as

well as to address new ones. The main technical contribution of these works is the technically challenging assignment of weight values to different items in a given sell-period independently from other items; we call this process association rules mining since it tries with this careful selection of weights to mine suitable association rules. The focus of all these works is retail data and organizations.

IDENTIFYING THE "HOT" ITEMS WITHOUT GETTING BURNED

The idea behind association rule mining is to search a considerable amount of data collected over a long period and to apply various techniques to discover some more or less unseen knowledge hidden inside the data. However, one of the biggest omissions of the approaches used up to now was that they discovered long existing relations and rather ignored the emerging ones. All approaches up to now followed rather than kept up with the sales or, more generally, the appearances of the itemsets. What we need is an approach that finds emerging trends in the bud along with the long established ones. This situation can be explained better in the example next.

Let's suppose that there exists a product that is sold in a retail store for many years, with moderate sales as compared to all other products and another product that just entered the market (e.g., is on sale about a year) but has tremendous sales. Applying the classical statistical model of association rule mining, where we simply measure the number of appearances of every itemset and if it is above a user specified threshold it is considered as frequent, would unavoidably doom the new product. A product that is on sale for so little can not practically come even near the sales of a product that is on sale for so long. So one must either wait for so long as for the new products to sum enough sales, which could well take months or even years, or find a way to effectively take them into consideration as soon as possible. The same situation can take place with products that were on the market but with very low sales and suddenly begun to present abnormally high sales because they are currently under heavy promotion, they suddenly became in fashion, some external circumstances or factors (e.g., weather conditions) promoted their sales, and so forth. The specific situation is very usual, especially at retail stores where the items on sale present such behaviors. After all, this kind of bursty sales behavior in a retail organization is probably far more interesting than that of high selling but stable products.

Therefore, the notion of "hot" items has been introduced by Kouris, Makris, and Tsakalidis (2004b), where as "hot" is considered any item that presents very high sales in a certain period of time. More formally, for every 1-itemset the, so called, *interest ratio* is calculated, which is defined as the number of sales of an item in the last period of sales (i.e., from the last time the algorithm has run again) divided by the mean number of sales of all items in the same period.

$$ir_i = \frac{sales_i}{\overline{sales}}$$

Every item that has its interest ratio above a user defined threshold called minimum interest threshold is considered as "hot" for the specific period. Of course, if an item has a number of sales above the support threshold it is treated as a frequent item. In essence, the proposed algorithm searches for items that have sales below the support threshold, but their interest ratio is above the minimum interest threshold. The user has, of course, the option of giving to that threshold any value depending on the desired output.

A logical question could be what happens with an item that was "hot" in the previous period but is no longer "hot" or frequent in the next one. One option would be to treat these items as infrequent items in the new period since they are obviously no longer interesting for the users. Another one could be to give these items a grace period, and treat them as "hot", to see if they will come back to high sales. Either one is possible and acceptable and depends solely on the needs of the data miner. One, though, could claim that such items were wrongly considered as interesting since their subsequent trend showed that they are no longer "hot". Nevertheless, the algorithm managed to immediately identify the period that they became interesting, took them into consideration, and promoted their sales when they were actually very interesting and this was exactly the goal. If, on the other hand, a "hot" item becomes frequent in the next period then the algorithm managed to successfully predict its future performance early enough and to take it into consideration in advance rather than having to wait for it to actually become frequent.

ASSESSING THE IMPORTANCE OF ITEMS IN A RETAIL ORGANIZATION

In contrast to the assumption upon which all association rules approaches work, that is that the correct support value is already known, in practice the correct minimum support value is not known and can only be estimated based on domain knowledge or on previous experience. Also when talking about the "correct" value, this is judged after the completion of the mining process based on the number of discovered frequent itemsets (i.e., too few or too many itemsets as compared to what has been anticipated), or in other words through completely subjective and rather trivial criteria. Consequently, if the support threshold changes, then the mining has to be repeated from the beginning requiring