Chapter 1

Uncovering Hidden Associations Through Negative Itemsets Correlations

Ioannis N. Kouris, University of Patras, Greece

Christos H. Makris, University of Patras, Greece

Athanasios K. Tsakalidis, University of Patras, Greece

Abstract

Most algorithms and approaches dealing with data mining in general and especially those focusing on the task of association rule mining have assumed all items to be only positively correlated, and looked only into the items that remained finally in a shopping basket. Very few works have proposed the existence of negative correlations between items, based though on the absence of items from transactions rather than on their actual removals. In this specific chapter we look into mining that takes into consideration valuable information from rejected items and propose various alternatives for taking the specific items into account efficiently. Finally we provide experimental evidence on the existence and significance of these items.
Introduction

In the last 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, and 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 as the processing power for evaluating and analyzing it and therefore much of the data accumulated remains unanalyzed still today. 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; Fayad, Piatetsky-Shapiro, & Smyth, 1996). Mining association rules especially from large databases of business data such as transactions records has been the “hottest” topic in the area of data mining, probably due to its large financial interest. 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). Practically, the goal of this 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, for example, with text documents (Holt & Chung, 2001), census data (Brin, Motwani, Ullman, & Tsur, 1997), telecommunication data, medical images, and more. In fact, any data set consisting of “baskets” containing multiple “items” can fit this model. Many solutions have been proposed in the last few years using a sequential or parallel paradigm, experimenting on factors such as memory requirements, I/O scans, dimensionality reduction, and so on.

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 a major reference point for all subsequent works. Most algorithms and approaches proposed thereafter focus on either decreasing the number of passes made over the data or at improving the efficiency of those passes (for example, by using additional methods for pruning the number of candidates that have to be counted). Toivonen (1996) proposed a sampling algorithm that required only a pass over the data but at the cost of generating a large number of candidate itemsets that have to be counted (false positives). Brin, Motwanti, Ullman, et al. (1997) proposed algorithm DIC that reduced the number of passes made over the database and at the same time...
A Survey of Approaches to Web Service Discovery in Service-Oriented Architectures


www.igi-global.com/article/survey-approaches-web-service-discovery/49725?camid=4v1a