Order Preserving Data Mining

Ioannis N. Kouris
University of Patras, Greece

Christos H. Makris
University of Patras, Greece

Kostas E. Papoutsakis
University of Patras, Greece

INTRODUCTION

Data mining has emerged over the last decade as probably the most important application in databases. To reproduce one of the most popular but accurate definitions for data mining; "it is the process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as rules, constraints and regularities) from massive databases" (Piatetsky-Shapiro & Frawley 1991). In practice data mining can be thought of as the "crystal ball" of businessmen, scientists, politicians and generally all kinds of people and professions wishing to get more insight on their field of interest and their data. Of course this "crystal ball" is based on a sound and broad scientific basis, using techniques borrowed from fields such as statistics, artificial intelligence, machine learning, mathematics and database research in general among others. Applications of data mining range from analyzing simple point of sales transactions and text documents to astronomical data and homeland security (*Data Mining and Homeland Security: An Overview*). Usually different applications may require different data mining techniques. The main kinds of techniques that are used in order to discover knowledge from a database are categorized into association rules mining, classification and clustering, with association rules being the most extensively and actively studied area. The problem of finding association rules can be formulated as follows: Given a large data base of item transactions, find all frequent itemsets, where a frequent itemset is one that occurs in at least a user-specified percentage of the data base. In other words find rules of the form X→Y, where X and Y are sets of items. A rule 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. An example of such a rule is {smoking, no_workout→heart_disease (sup=50%, conf=90%)}, which means that 90% of the people that smoke and do not work out present heart problems, whereas 50% of all our people present all these together.

Nevertheless the prominent model for contemplating data in almost all circumstances has been a rather simplistic and crude one, making several concessions. More specifically objects inside the data, like for example items within transactions, have been attributed a Boolean hypostasis (i.e. they appear or not) with their ordering being considered of no interest because they are considered altogether as sets. Of course similar concessions are made in many other fields in order to come to a feasible solution (e.g. in mining data streams). Certainly there is a trade off between the actual depth and precision of knowledge that we wish to uncover from a database and the amount and complexity of data that we are capable of processing to reach that target.

In this work we concentrate on the possibility of taking into consideration and utilizing in some way the order of items within data. There are many areas in real world applications and systems that require data with temporal, spatial, spatiotemporal or ordered properties in general where their inherent sequential nature imposes the need for proper storage and processing. Such data include those collected from telecommunication systems, computer networks, wireless sensor networks, retail and logistics. There is a variety of interpretations that can be used to preserve data ordering in a sufficient way according to the intended system functionality.

BACKGROUND

Taking into consideration and using the order of items within transactions has been considered in another form and for an alternate goal; that is with the task of sequential pattern mining. Essentially in association rule mining we try to find frequent items that appear together in the same transactions. In sequential pattern mining we are interested in sets of items that appear frequently in different transactions. Thus association rules mining can be thought as an intra-transactional search process whereas sequential pattern mining as an inter-transactional.

A formulation of the task of sequential pattern mining is as follows (Dunham, 2003; Tan, Steinbach & Kumar, 2006): Let D be a database of customer transactions, and $I=\{i_1, i_2, ..., i_m\}$ be a set of distinct attributes called items. A transaction or else called an event is a non-empty collection of items (in most cases ordered), denoted as $(i_1, i_2, ..., i_k)$. A sequence α on the other hand is a collection of events such that $(\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_r)$, where α_i is an event. The support or frequency of a sequence, denoted as $\sigma(\alpha, D)$, is the total number of sequences in the database D that contain α. Consequently the process of sequential pattern mining concerns with extracting those sequential patterns whose support exceed a user predefined minimum support value; though the items within every transaction were reordered, thus destroying the initial purchase order.

Sequential pattern mining can be used in a wide range of areas from biology and medicine, to business processes and web sessions. The problem of mining sequential patterns was first introduced in (Agrawal and Srikant, 1995), where three algorithms were presented with algorithm AprioriAll having the best performance. An enhanced algorithm by the name GSP was proposed in (Srikant & Agrawal, 1996) that outperformed AprioriAll by up to 20 times. In the same period Mannila, Toivonen and Verkamo (1997) proposed a work for mining frequent episodes, which was further extended in (Mannila & Toivonen, 1996) in order to discover generalized episodes. Algorithms MEDD (Multi-Event Dependency Detection) and MSDD (Multi-Stream Dependency Detection) in (Oates, Schmill, Jensen, & Cohen, 1997) discover patterns in multiple event sequences, by exploring instead of the sequence space directly the rule space.

Other works include Zaki's SPADE (Zaki, 2001), a method that employs lattice based search techniques and simple joins in order to decompose the search space into sub-lattices small enough to be processed in main memory, thus reducing the number of database scans. PrefixSpan (Pei et al., 2001) employs an internal representation of the data made of database projections over sequence prefixes. Finally a family of works where time plays the dominant role are (Vautier, Cordier, & Quiniou, 2005) and the most recent work in (Giannotti, Nanni & Pedreschi, 2006). Interested reader can refer to (Zhao & Bhowmick, 2003) for a broader and more detailed view of the specific area.

ORDERED DATA: A NEW PERSPECTIVE

Traditional algorithms and approaches for association rules mining work on the rather naive assumption that every item inside a transaction has an influence on all the rest items inside the same transaction. Nevertheless items may appear together in several transactions but this does not necessarily mean they all influence one another. For example suppose we have a transaction that contains the following items in the exact order as they were purchased:

[V, G, A, S, C]

Then according to current approaches item V is considered to have influenced the purchase of all other items (i.e. items G, A, S and C). Items G, A, S, C are also considered to have influenced the purchase of all other items inside the transaction. However the reasonable question that arises from this assumption is how an event that occurred in the future could have influenced an event in the past. More specifically how can we claim for example that the choice of item A was influenced by the purchase of item S or C since when A was purchased items S and C were not even in the basket. It is fairly logical to presume that there exists the possibility that an item purchased might have an influence on the decision to purchase or not an item afterwards, but how can we claim the opposite? This scenario is especially evident in on-line stores where the purchases are made in a chain-like fashion (i.e. one item at a time and usually without following some predefined shopping list), resembling a lot the web browsing behaviour of the world wide web users. Hence in our opinion every item inside a transaction influences only the subsequent items inside the same transaction. This leads us to the following observation: the choice