Discovering Association Rules in Temporal Databases

Juan M. Ale  
*Universidad de Buenos Aires, Argentina*

Gustavo H. Rossi  
*Universidad Nacional de La Plata, Argentina*

**INTRODUCTION**

The problem of the discovery of association rules comes from the need to discover interesting patterns in transaction data in a supermarket. Since transaction data are temporal we expect to find patterns that depend on time. For example, when gathering data about products purchased in a supermarket, the time of the purchase is stamped in the transaction.

In large data volumes, as used for data mining purposes, we may find information related to products that did not necessarily exist throughout the complete data-gathering period. So we can find a new product, such as a DVD player, that was introduced after the beginning of the gathering, as well as a product, like a 5 1/4-inch flexible disk unit, that had been discontinued before the ending of the same gathering. It would be possible that that new product could participate in the associations, but it may not be included in any rule because of support restrictions. Suppose we have gathered transactions during 10 years. If the total number of transactions is 10,000,000 and we fix as minimum support 0.5 %, then a particular product must appear in, at least, 50,000 transactions to be considered frequent. Now, take a product that has been sold during these 10 years and has just the minimum support: It appears on average in 5,000 transactions per year. Consider now another product that was incorporated two years ago and that appears in 20,000 transactions per year. The total number of transactions in which it occurs is 40,000; for that reason, it is not frequent, even though it is four times as popular as the first one. However, if we consider just the transactions generated since the product appeared in the market, its support might be above the stipulated minimum. In our example, the support for the new product would be 2 %, relative to its lifetime, since in two years the total of transactions would be about 2,000,000 and this product appears in 40,000 of them. Therefore, these new products should appear in interesting and potentially useful association rules. Moreover, we should consider the case of some products that may be frequent just in some subintervals strictly contained in their period of life but not in the entire interval corresponding to their lifespan.

We solve this problem by incorporating time in the model of discovery of association rules. We call these new rules *general temporal association rules*.

One by-product of this idea is the possibility of eliminating outdated rules, according to the user’s criteria. Moreover, it is possible to delete obsolete sets of items as a function of their lifetime, reducing the amount of work to be done in the determination of the frequent items and, hence, in the determination of the rules.

The temporal association rules introduced in Ale and Rossi (2000) are an extension of the nontemporal model. The basic idea is to limit the search for frequent sets of items, or *itemsets*, to the lifetime of the itemset’s members. On the other hand, to avoid considering frequent an itemset with a very short period of life (for example, an item that is sold once), the concept of temporal support is introduced. Thus, each rule has an associated time frame, corresponding to the lifetime of the items participating in the rule. If the extent of a rule’s lifetime exceeds a minimum stipulated by the user, we analyze whether the rule is frequent in that period. This concept allows us to find rules that with the traditional frequency viewpoint, it would not be possible to discover.

The lifespan of an itemset may include a set of sub-intervals. The subintervals are those such that the given itemset: (a) has maximal temporal support and (b) is frequent. This new model addresses the solution of two problems: (1) Itemsets not frequent in the entire lifespan but just in certain subintervals, and (2) the discovery of every itemset frequent in, at least, subintervals resulting from the intersection of the lifespans of their components, assuring in this way the anti-monotone property (Agrawal & Srikant, 1994). Because of this we call “general” the rules formed from these kinds of frequent itemsets.

**BACKGROUND**

Previous work about data mining that includes temporal aspects is usually related to the sequence of events’ analysis (Agrawal & Srikant, 1995; Bettini, Wang, Jajodia, & Lin, 1998; Mannila, Toivonen, & Verkamo, 1995). The
Discovering Association Rules in Temporal Databases

usual objective is to discover regularities in the occurrence of certain events and temporal relationships between the different events. In particular, in Mannila et al. the authors discuss the problem of recognizing frequent episodes in an event sequence; an episode is defined there as a collection of events that occur during time intervals of a specific size. Meanwhile Agrawal and Srikant (1998) review the problem of discovering sequential patterns in transactional databases. The solution consists in creating a sequence for every customer and to look for frequent patterns into each sequence. In Bettini et al. the authors consider more complex patterns. In these cases temporal distances with multiple granularities are treated. Chakrabarti, Sarawagi, and Dom (1998), in a totally different approach, use the minimum description length principle together with an encoding scheme to analyze the variation of inter-item correlation along time. That analysis, whose goal is extracting temporally surprising patterns, is an attempt to substitute the role of knowledge domain in searching interesting patterns.

Now we will analyze how the present work is related to others, specifically in mining temporal association rules. All of them have the same goals as ours: the discovery of association rules and their periods or interval time of validity. Our proposal was formulated independently of the others but shares with them some similarities. In Ozden, Ramaswamy, and Silberschatz (1998), the authors study the problem of association rules that exist in certain time intervals and display regular cyclic variations over time. They present algorithms for efficiently discovering what they called “cyclic association rules.” It is assumed that time intervals are specified for the user.

In Ramaswamy, Mahajan, and Silberschatz (1998), the authors study how the association rules vary over time, generalizing the work in Ozden et al. (1998). They introduce the notion of calendar algebra to describe temporal phenomena of interest to the users and present algorithms for discovering “calendric association rules;” that is, association rules that follow the temporal patterns set forth in the user-supplied calendar expressions.

The third study (Chen, Petrounias, & Heathfield, 1998) also suggests calendar time expressions to represent temporal rules. They present only the basic ideas of the algorithms for discovering the temporal rules. The fourth study (Li, Ning, Wang, & Jajodia, 2001) is the most elaborated expression within the calendar approach. The authors define calendar schemas and temporal patterns in these schemas. They also define two types of association rules: precise-match and fuzzy-match. They try to find association rules within the time intervals defined by the schemas.

Finally, in Lee, Lin, and Chen (2001) and Lee, Chen, and Lin (2003), the authors introduce some kind of temporal rules and propose an algorithm to discover temporal association rules in a publication database. The basic idea is to partition the publication database in light of exhibition periods of items. Possibly, this work is the most similar to ours. The notion of exhibition period is similar to our lifespan (Ale & Rossi, 2000) for an itemset, and the same happens with the concept of the maximal common exhibition period that we have called again lifespan when applying the concept to an itemset. The differences with the present work (Ale & Rossi, 2002) are more significant because we are defining frequent subintervals within an itemset’s lifespan and, in this way, we get the a priori property holds.

Our approach is based on taking into account the items’ period of life, or lifespan, this being the period between the first and the last time the item appears in transactions in the database. We compute the support of an itemset in the interval defined by its lifespan or subintervals contained in it, and define temporal support as the minimum interval width. We consider the history of an itemset as a time series, having the possibility of performing on it different kinds of analysis, based on its wavelet transform. However, this last part is not included in this article because of space restrictions. Our approach differs from the others in that it is not necessary to impose intervals or calendars since the lifespan is intrinsic to the data. Even more, we find association rules and the time when they hold, so the calendar approach becomes a special case. Moreover, our model satisfies the downward closure property, which a-priori-based algorithms are based on.

THE GENERAL TEMPORAL MODEL

Let \( T = \{ t_0, t_1, t_2, \ldots \} \) be a set of times over which a linear order \( \prec \) is defined, where \( t \prec t' \) means \( t \) occurs before or is earlier than \( t' \) (Tansel et al., 1993). We will assume that \( T \) is isomorphic to \( \mathbb{N} \) (natural numbers) and restrict our attention to closed intervals \( [t, t] \).

Let \( R = \{ A_1, \ldots, A_p \} \), where the \( A_i \)'s are called items, a transaction database \( d \) is a collection of subsets of \( R \).

Each transaction \( s \) in \( d \) is a set of items such that \( s \subseteq R \). Associated to \( s \) we have a time stamp \( t_s \), which represents the valid time of transaction \( s \).

Example 1.1.a: \( R = \{ A, B, C, D, E, F, G, H, I \} \). \( d \) is the collection of 10 transactions with tids 100, ..., 1000 and time stamps 1, ..., 10, respectively; see Figure 1(a).

We consider \( d \) is temporally ordered. Every item has a period of life, or lifespan, in the database, which explicitly represents the temporal duration of the item information, i.e., the time in which the item is relevant to the user. The lifespan of an item \( A_i \) is given by an interval \( [t_1, t_2] \), with
4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage: www.igi-global.com/chapter/discovering-association-rules-temporal-databases/11145?camid=4v1


Related Content

E-Government Databases

A Literature Overview of Fuzzy Database Modeling

Information Technology in the Practice of Law Enforcement
Susan Rebstock Williams and Cheryl Aasheim (2006). Cases on Database Technologies and Applications (pp. 81-103). www.igi-global.com/chapter/information-technology-practice-law-enforcement/6206?camid=4v1a

Data Mining: New Arsenal For Strategic Decision-making