Chapter 13
OLAP-Sequential Mining: Summarizing Trends from Historical Multidimensional Data using Closed Multidimensional Sequential Patterns

Marc Plantevit, Anne Laurent, and Maguelonne Teisseire

Abstract  Data warehouses are now well recognized as the way to store historical data that will then be available for future queries and analysis. In this context, some challenges are still open, among which the problem of mining such data. OLAP mining, introduced by J. Han in 1997, aims at coupling data mining techniques and data warehousing. These techniques have to take the specificities of such data into account. One of the specificities that is often not addressed by classical methods for data mining is the fact that data warehouses describe data through several dimensions. Moreover, the data are stored through time, and we thus argue that sequential patterns are one of the best ways to summarize the trends from such databases. Sequential pattern mining aims at discovering correlations among events through time. However, the number of patterns can become very important when taking several analysis dimensions into account, as it is the case in the framework of multidimensional databases. This is why we propose here to define a condensed representation without loss of information: closed multidimensional sequential patterns. This representation introduces properties that allow to deeply prune the search space. In this paper, we also define algorithms that do not require candidate set maintenance. Experiments on synthetic and real data are reported and emphasize the interest of our proposal.

13.1 Introduction

Data warehouses are now well spread over companies. They contain valuable information that can easily be queried and visualized with the OLAP tools, provided the fact that the user is able to design on-line his own queries. However, it is still challenging to provide the user with tools that can automatically extract relevant knowledge from such huge amounts of data. Data warehouses are indeed different
from usually mined databases as they contain aggregated data, described by means of several dimensions that can possibly be organized through hierarchies. In this paper, we thus try and extend existing methods that are now recognized for mining classical databases to this framework. As data are historized, we argue that sequential patterns are well-suited to this task. Sequential patterns have been studied for more than ten years [1], with a lot of research and industrial applications (e.g. user behavior, web log analysis, discovery of patterns from proteins’ sequences, security and music). Algorithms have been proposed, based on the APriori-based framework [23, 9, 2], or on other approaches [13, 8]. Sequential patterns have recently been extended to multidimensional sequential patterns by [14, 15, 22]. They aim at discovering patterns that take time into account and that involve several dimensions. For instance in [15], rules like A customer who bought a surfboard together with a bag in NY later bought a wetsuit in SF are discovered. Note that such sequences can also contain a wild-card item * instead of a dimension value. For instance, considering the previous example, if there is no frequent pattern in the database describing the fact that wetsuits were later bought in SF, or NY etc, but there are numerous wetsuits bought whatever the city, then the rule A customer who bought a surfboard together with a bag in NY later bought a wetsuit will be mined, represented as \((\text{surfboard, NY})(\text{wetsuit, *})\).

Sequential patterns are usually extracted from the simple schema: (e.g. products, customer_id and date) but the number of mined patterns can be very huge. This is why condensed representations were proposed for the itemset framework ([10, 12, 24, 5]) and for sequential patterns ([21, 18]). In both cases, the approaches allow a condensed representation and a pruning strategy in the search space.

However, these works are not suitable for multidimensional sequential patterns because they only consider a particular case for candidate generation. In our context, a super-sequence may indeed result from several cases (1) a longer sequence (more items) or (2) a more general sequence based on the relation between dimension values and the wild-card value * (more general items).

The main contributions of this paper are a theoretical framework for mining closed multidimensional sequential patterns and some algorithms called CMSP (Closed Multidimensional Sequential Pattern mining) to mine such patterns. When considering multidimensional data, the number of possible patterns is combinatorially explosive and the generate-and-prune methods are no more scalable for long sequences, as highlighted in [11, 16]. This is why we adopt the pattern growth paradigm ([13]) to propose a greedy approach for mining frequent sequences without candidate generation.

The paper is organized as follows. First we recall the related work in Section 13.2 and we detail why existing works are not suitable for mining data warehouse. Then we present the core of our proposition: the definitions are introduced in Section 13.3 while the CMSP algorithms are detailed in Section 13.4. Experiments on synthetic and real data, reported in Section 13.5, show that our method performs well both on runtime and number of extracted closed sequences. Finally, we provide some concluding remarks and suggestions for future work in Section 13.6.