Revisiting “An Apriori-based Approach for First-Order Temporal Pattern Mining”

Sandra de Amo¹, Daniel A. Furtado², Arnaud Giacometti², Dominique Laurent³

¹ Universidade Federal de Uberlândia, Computer Science Department - Uberlândia - Brazil
deamo@ufu.br, danielfurtados@yahoo.com.br

² LI-Université de Tours, UFR de Sciences - Blois - France
giaco@univ-tours.fr

³ ETIS-CNRS-ENSEA-Université de Cergy Pontoise - Cergy Pontoise - France
dominique.laurent@dept-info.u-cergy.fr

Categories and Subject Descriptors: Information Systems [Miscellaneous]: Databases

Keywords: Temporal Data Mining, Sequence Mining, Sequential Patterns, Temporal Data Mining, Frequent Patterns, Knowledge Discovery

1. INTRODUCTION

A lot of different approaches related to sequential pattern mining have been proposed in the literature, since 2004, when the original paper was published in the proceedings of SBBD 2004. Among these approaches, we distinguish five main directions of research: (1) development of more efficient methods for the classical sequential pattern mining problem, (2) sequential pattern mining with constraints, (3) multidimensional and multilevel sequential patterns, (4) temporal patterns specified by more general structures (tree and graph patterns), (5) temporal relational patterns with interval time attributes. The first approach addresses computational issues ([Pei et al. 2004; Ayres et al. 2002; Chiu et al. 2004; Yang et al. 2006]), rather than extensions of the type of sequential patterns that are mined. In particular, sequential patterns mined according to these approaches are standard sequential patterns that can be expressed in Propositional Temporal Logic. For that reason, in this essay we concentrate our attention on the four latter lines of research, focusing mainly on approaches developed by the authors of the original paper.

2. SEQUENTIAL PATTERN MINING WITH CONSTRAINTS

One of the main contributions of the original paper is that an Apriori-like technique based on candidate generation, pruning and validation can be used to mine first-order temporal patterns msp and that a pure first-order technique (SM-Miner) produces better results than a technique adapted from classical methods for propositional sequential pattern mining (PM-Miner).

Preliminary experiments on multi-sequential pattern mining showed an overwhelming volume of candidate patterns, most completely irrelevant to the users, resulting in a great and wasteful computational cost. This naturally led some of the authors of the original paper to consider specification formalisms to allow user focus during the msp mining process in order to prevent the generation of uninteresting and useless patterns. The proposed approach follows the idea introduced in [Garofalakis...
3. MULTIDIMENSIONAL AND MULTILEVEL SEQUENTIAL PATTERNS

Multidimensional Sequential Pattern Mining, where sequential patterns are mined in a multidimensional table has been originally introduced in [Pinto et al. 2001] and further investigated in [Rashad and Kantardzic 2007; Stefanowski 2007; Stefanowski and Ziembinski 2005; Zhang et al. 2007; Yu et al. 2005]. These approaches have all been generalized by some of the authors of the original paper in [Plantevit et al. 2005]. On the other hand, Multilevel rules have been studied in the framework of sequential patterns, based on the seminal work in [Agrawal and Srikant 1996], where hierarchical relations between the items are assumed. In this approach, sequential patterns are expressed using several levels of hierarchies defined over attributes in the dataset. Some of the authors of the present paper, generalized this approach in [Plantevit et al. 2006; Plantevit et al. 2010], based on their approach introduced in [Plantevit et al. 2005]. Here, the sequential patterns to be mined are not only multilevel, in the sense that these patterns are expressed using values occurring in hierarchies defined over attributes, but also multidimensional, in the sense that these patterns are expressed over several attributes.

In [Plantevit et al. 2010], sequential patterns are mined from a given relational table $D$ defined over an attribute (or dimension) set $U$, partitioned into four subsets as follows: $D_t$ contains a single dimension, called the temporal dimension, $D_A$ contains the analysis dimensions, $D_R$ contains the reference dimensions, and $D_T$ contains the ignored dimensions.

Moreover, hierarchies are assumed to be defined on attributes in $U$, so as to express patterns according to different levels of generalization, thus referring to the fact that multilevel patterns are mined.

Therefore, the approach of [Plantevit et al. 2010] can be seen as a generalization of the original paper because of the following: (1) In the two approaches, supports are counted with respect to some partitioning of the dataset, defined by one attribute in the original work (namely $IdG$ in our example), and defined by several attributes (namely, attributes in $D_R$) in [Plantevit et al. 2010]. (2) In [Plantevit et al. 2010], sequential patterns are expressed using sequences of sets of tuples (defined over the attributes in $D_A$), whereas in the original approach, sequential patterns are expressed using sequences of atomic values (defined over the attribute Item in our example). (3) Hierarchies over attributes are used in [Plantevit et al. 2010] in order to mine patterns at different levels of generalization, which is not the case in the original approach.

The mining algorithms proposed in [Plantevit et al. 2010] follows a level-wise strategy without any decomposition, based on the Spade Algorithm ([Zaki 2001]), instead of GSP as in the original article. Thus the way patterns are mined in [Plantevit et al. 2010] is close to Algorithm SM, apart from the first step, during which hierarchies are taken into account.

4. TREE PATTERN MINING

Some recent applications dealing with complex data require sophisticated data structures (trees or graphs) for their specification. As pointed out in the pioneer work [Zaki 2002], sequential pattern mining can be applied in the Web Mining context, but tree pattern mining methods yield more informative patterns, highlighting the usefulness of mining complex patterns represented by trees and graphs. In [de Amo et al. 2010; 2007], some of the authors of the original paper focused on extending the constraint-based problem they treated in [de Amo and Furtado 2005; 2007] to the tree mining
context. They proposed to use tree automata as a mechanism to specify user constraints over tree patterns and developed the algorithm CobMiner allowing user constraints specified by a tree automata to be incorporated in the mining process. The algorithm has been tested throughout an extensive set of experiments executed over synthetic and real data (XML documents and Web usage logs). These tests allowed to conclude that incorporating constraints during the tree mining process is far better effective than filtering the frequent and interesting patterns after the mining process. In [de Amo and Felicio 2007], the authors introduced a visual language allowing users to specify tree automata constraints in the context of tree pattern mining with constraints.

5. TEMPORAL RELATIONAL INTERVAL PATTERN MINING

In [Höppner 2001], Allen’s Propositional Interval Logic [Allen and Ferguson 1994] has been used for the first time, to treat the problem of discovering association rules over time series. In [de Amo et al. 2007], some of the authors of the original paper extended the sequential pattern mining in order to deal with temporal patterns where the time domain is a set of intervals instead of a set of points. They proposed a new temporal pattern defined as a set of atomic first order formulae where time is explicitly represented by an interval variable, together with a set of interval relationships (before, during, starts, finishes, overlaps, meets) described in terms of Allen’s First Order Interval Logic [Allen and Ferguson 1994]. An example of such temporal pattern is \{\text{Med}(x, \text{penicillin}, i_1), \text{Symp}(x, \text{dizziness}, i_2), \text{during}(i_2, i_1)\}, meaning that “patients who take penicillin during the time interval \(i_1\) will feel dizzy during this period, but eventually the dizziness will stop”. The algorithm MILPRIT for mining temporal interval patterns has been introduced in [de Amo et al. 2007], which uses variants of the classical level-wise search algorithms. MILPRIT allows a broad spectrum of constraints over temporal patterns to be incorporated in the mining process. The algorithm has been generalized in [de Amo et al. 2008], in order to deal with both point and interval time attributes.

REFERENCES


S. Amo, D. A. Furtado, A. Giacometti and D. Laurent


PLANTEVIT, M., LAURENT, A., LAURENT, D., TEISSEIRE, M., AND CHOONG, Y. Mining multidimensional and multilevel sequential patterns. ACM Transactions on Knowledge Discovery from Data vol. 4, pp. 4:1–4:37, 2010.


YU, C., CHEN, Y., AND Y.-L. Mining sequential patterns from multidimensional sequence data. IEEE Transactions on Knowledge and Data Engineering vol. 17, pp. 136–140, 2005.


