Expert Systems with Applications 40 (2013) 1296–1304

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Mining generalized temporal patterns based on fuzzy counting

Francisco Guil [a,*], Antonio Bailón [b], José A. Álvarez [a], Roque Marín [c]

[a] High School of Engineering, University of Almeria, Almería, Spain
[b] High Technical School of Computer Science and Telecommunications, University of Granada, Granada, Spain
[c] Faculty of Computer Science, University of Murcia, Murcia, Spain

ARTICLE INFO

Keywords:
Temporal data mining
Fuzzy sets
Temporal patterns
Event-based sequences

ABSTRACT

Event-based sequences are a kind of pattern based on temporal associations with two essential characteristics: they are syntactically simple and have a great expressive power. For this reason, event-based sequence mining is an interesting solution to the problem of knowledge discovery in dynamic domains, mainly characterized by a time-varying nature. The inter-transactional model has led to the design of algorithms aimed to obtain this sort of patterns from time-stamped datasets. These algorithms extend the well-known *Apriori* algorithm, by explicitly adding the temporal context where associations among frequent events occurs. This leads to the possibility of extracting a larger number of patterns with a potential interest in decision making. However, its usefulness is diminished in those datasets where the characteristics of variability and uncertainty are present, which is a common issue in real domains. This is due to the rigidity of the counting method, which uses an exact measure of distance between temporal events. As a solution, we propose a generalization of the temporal mining process, which implies a relaxation of the counting method including the concept of approximate temporal distance between events. In particular, in this paper we present an algorithm, called *TSET^{fuzzy}-Miner*, which incorporates a fuzzy-based counting technique in order to extract general, flexible, and practical temporal patterns taking into account the particular characteristics of real domains.

1. Introduction

Data mining is an essential step in the process of knowledge discovery in databases that consists of applying data analysis and discovery algorithms that produce a particular enumeration of structures (models and patterns) over the data (Fayyad, Piatetky-Shapiro, & Smyth, 1996). Depending on the structure, data mining can be approached from two different perspectives: global and local methods (Mannila, 2002). In this work we are interested in local methods, commonly known as frequent pattern mining. The simplest case of pattern discovery is finding *association rules* (Agrawal, Imielinski, & Swami, 1993), a kind of pattern used as a means to help in the analysis of large transactional databases. Such association rules, when discovered, provide valuable knowledge for decision making. One approach is integrating the data mining process into the development of Knowledge Based Systems (Li, Xie, & Xu, 2011; Ordonez, Santana, & de Braal, 2000, 2011).

Since the problem of mining association rules was introduced by *Agrawal et al.* in Agrawal et al. (1993), many research work has been accomplished in a wide range of directions, including the improvement of the *Apriori* algorithm, mining generalized, multi-level, or quantitative association rules, mining weighted association rules, fuzzy association rules mining, constraint-based rule mining, efficient long patterns mining, maintenance of the discovered association rules, etc. In general, temporal data mining can be seen as an extension of this work.

Temporal data mining can be defined as the activity of searching for interesting correlations or patterns in large sets of temporal data accumulated for other purposes than those originally expected (Bettini, Wang, & Jajodia, 1996). It has the ability of mining activity, inferring associations of contextual and temporal proximity, some of which may also indicate a cause-effect association. This important kind of knowledge can be overlooked when the temporal component is ignored or treated as a simple numeric attribute (Roddick & Spiliopoulou, 2002). Data mining is an interdisciplinary field which has received contributions from a variety disciplines, mainly from databases, machine learning and statistic. However, in the case of temporal data mining techniques, the most influential field has been Artificial Intelligence, the reason why can be found in the extensive efforts in the temporal reasoning line that gave rise to the development of many of these techniques. In non-temporal data mining techniques, there are usually two different tasks, the description of the characteristics of the database (or analysis of the data) and the prediction of the evolution of the population. However, in temporal data mining this distinction is less appropriate, because the evolution of the population is

* Corresponding author. Tel.: +34 950015787.
E-mail addresses: fguil@ual.es (F. Guil), bailon@decsai.ugr.es (A. Bailón), jaberme@ual.es (J.A. Álvarez), roquemm@um.es (R. Marín).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.08.061

already incorporated in the temporal properties of the data being analyzed.

Literature regarding temporal data mining techniques is extensive and one can find myriad of references to serve as background. The most outstanding of these references are sequential pattern mining (Agrawal & Srikant, 1995; Chang, 2011; Wu, Lai, & Lo, 2012; Zaki, 2001), episodes in event sequences (Mannila, Toivonen, & Verkamo, 1997), temporal association rules mining (Lee, Lee, Kim, Hwang, & Ryu, 2002), discovering calendar-based temporal association rules (Li, Ning, Wang, & Jajodia, 2003), patterns with multiple granularities mining (Bettini et al., 1996), and cyclic association rules mining (Özden, Ramaswamy, & Silberschatz, 1998), among others. In line with this set of references provided, we want to highlight a relevant type of patterns that despite of their usefulness, they could not be discovered by means of any of these techniques. These are the inter-transactional associations presented in Lu, Feng, and Han (2000). The introduction of this type of association was motivated by the observation that many real-world associations happen under certain contexts, such as time, place, etc. In the case of temporal context, inter-transactional patterns represent associations among items along the dimension of time. As the number of potential frequent patterns may turn into an extremely large amount, the mining of these kinds of patterns becomes challenging as opposed to classical approaches. Several methods, as the ones proposed in Tung, Lu, Han, and Feng (2003), Feng, Yu, Lu, and Han (2002), Lee and Wang (2007) and Wang and Chu (2011) were developed to make this technique feasible in terms of computational resources.

Following this same argument, we have presented an algorithm, named *TSET-Miner* (Guil & Marín, 2012), based on the inter-transactional framework for mining frequent sequences (also called event-based sequences or frequent temporal patterns) from time-stamped datasets. The enhancement of our proposed relied on the use of a unique structure to store all frequent sequences. The data structure used is the well-known set-enumeration tree, commonly used in the data mining field, in which the temporal semantic is incorporated. On one hand, although the algorithm can extract very interesting temporal patterns, the strictness of the process leads to a method that is not very useful in a plenty of application domains where variability and uncertainty are presented. For example, the 2-sequence: "*It is likely* (*s* = 90%) *that event b occurs 1 temporal unit after the occurrence of event a*", shows that in 90% of transactions, *b* always happens after *a exactly* 1 temporal unit later. The original inter-transactional scheme requires that (after a preprocessing task) the temporal dimension consists of a number of consecutive positive integers. On the other hand, it is difficult to find in real life cases where associations between events occur with exactly the same temporal distance in the strict sense. However, in real cases, the uniformity criterion is not always possible due to the possible presence of data gaps in the temporal dimension.

Summarizing, the problem can be described by stating that, in general, associations between events occur within a certain time interval. Li, Feng, and Wong (2005), proposed a line tightly related to this. It was a generalization of the inter-transactional mining based on *context expansion*, which relies on the generation of rules based on intervals (instead of time points), obtaining a more general form of association rules. The rules they proposed were composed in this way: "*It is likely* (*s* = 90%) *that event a occurs between 1 and 3 temporal units after the occurrence of event b*", obtained through the linguistic interpretation of the sequence $S = (a_{[0,0]}, b_{[1,3]})$, with support equal to 90%. Although with this proposal it is possible to extract more comprehensive and interesting patterns, but under our criteria, both the model and the algorithmic solution are very complex. Dealing with a solution based on intervals leads to a significant increment of the number of potential candidates, resulting in very high execution times when it is compared with the point-based solution.

The aim of this paper is to extend the algorithm introducing a fuzzy set-based technique into the mining process in order to make it more expressive and flexible. This improvement leads to a general framework capable of extracting general, practical and flexible temporal patterns. The goal is to obtain sequences similar to: "*It is likely* (*s* = 90%) *that event b occurs, approximately, 1 temporal unit after the occurrence of event a*", obtained following the linguistic interpretation of the event-based 2-sequence $S = \{a_0, b_1\}$, with *support* = 90%. The linguistic term "approximately" implies that, for example, the event *b* could have occurred 0, 1 or even 2 temporal units after *a*. The proposed solution consists of the inclusion of a reference fuzzy set in the counting method. This user-defined fuzzy set represents the meaning of the linguistic term "approximately equal to", defined by a 0-centered fuzzy number characterized by the membership function μ_z. Moreover, and with the aim of increasing versatility, if the user wants to obtain sequences defined over a temporal unit that is different from the temporal unit specified in the input dataset, the algorithm allows setting a user-defined parameter, denoted as *g*, which defines the granularity of the temporal dimension. From the experimental point of view, we have carried out a series of experiments to show how *TSET*^{fuzzy}*-Miner* algorithm behaves with both, synthetic and real-life datasets.

The remainder of the paper is organized as follows. Section 2 gives a formal description of the problem. Section 3 introduces briefly the basic principles of the algorithm and presents an example to illustrate the fuzzy approach proposed. Experimental evaluation is discussed and exposed in Section 4. Conclusions are finally drawn in Section 5.

2. Problem definition

This section introduces the notation and basic definitions needed to provide a detailed insight of our proposed algorithm for mining generalized frequent sequences from time-stamped transactional datasets, named *TSET*^{fuzzy}*-Miner*.

Definition 1 (*Transactional dataset*). Let $\Sigma = \{te_1, te_2, \ldots, te_n\}$ be a set of items. Let $\mathcal{T}$ be an attribute and $dom(\mathcal{T})$ the domain of $\mathcal{T}$. A transactional dataset D is a set that contains r transactions, $D = \{D[0], D[1], \ldots, D[r-1]\}$, where $D[i] = (t, E)$, with $t \in dom(\mathcal{T})$, and $E \subseteq \Sigma$.

The $\mathcal{T}$ attribute is called dimensional attribute, which in our case is the temporal attribute associated with the transaction. This attribute describes the temporal context in which transactions occur.

Example 1. Let D be a (toy) transactional dataset extracted from a supermarket database. Assuming that the type of product sold is $\sum = \{a, b, c, d, e, f\}$, an example of dataset is shown in Table 1. Each transaction reflects the list of products purchased on the day indicated by the *Date* attribute.

Table 1
Transactional dataset.

Date	Item list	Date	Item list	Date	Item list
0	a d	10	b f	20	b f
1	b f	11	a e	21	b e
2	c e	12	c e	22	b f
3	a e	13	b e	23	c e
4	b d	14	a d	24	b e
5	b e	15	b e	25	a d
6	b d	16	a e	26	b e
7	b e	17	b f	27	c e
8	b f	18	a e	28	b d
9	b e	19	c e	29	a e