Data & Knowledge Engineering 68 (2009) 1309–1330

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Discovering hybrid temporal patterns from sequences consisting of point- and interval-based events

Shin-Yi Wu [a,*], Yen-Liang Chen [b]

[a] Industrial Technology Research Institute, Hsinchu 310, Taiwan, ROC
[b] Dept. Information Management, National Central University, Chung-Li 320, Taiwan, ROC

ARTICLE INFO

Article history:
Received 2 April 2008
Received in revised form 25 June 2009
Accepted 26 June 2009
Available online 5 July 2009

Keywords:
Data mining
Hybrid temporal pattern
Temporal pattern
Sequential pattern
Hybrid event sequences

ABSTRACT

Previous sequential pattern mining studies have dealt with either point-based event sequences or interval-based event sequences. In some applications, however, event sequences may contain both point-based and interval-based events. These sequences are called hybrid event sequences. Since the relationships among both kinds of events are more diversiform, the information obtained by discovering patterns from these events is more informative. In this study we introduce a hybrid temporal pattern mining problem and develop an algorithm to discover hybrid temporal patterns from hybrid event sequences. We carry out an experiment using both synthetic and real stock price data to compare our algorithm with the traditional algorithms designed exclusively for mining point-based patterns or interval-based patterns. The experimental results indicate that the efficiency of our algorithm is satisfactory. In addition, the experiment also shows that the predicting power of hybrid temporal patterns is higher than that of point-based or interval-based patterns.

1. Introduction

Sequential pattern mining is an important data mining technique that can be used to help make decisions in a variety of applications [17,16,14]. This technique can be utilized to discover the sequential patterns that occur frequently in a huge sequence database [2,27]. For example, a typical sequential pattern that shows up in video rentals [2] is that customers will rent a series of movies in a certain order, for example "Star Wars", followed by "The Empire Strikes Back", and finally "Return of the Jedi." As defined in sequential pattern mining, this pattern is supported by a customer sequence when the customer rents the items in the above-mentioned order (although not necessarily consecutively). When a pattern is supported by at least a *min_sup* (minimum support, a user-specified threshold) percentage of customer sequences, we say that this pattern is *frequent* (or *large*). In other words, frequent sequential patterns are patterns which have occurred frequently in past experience. Sequential patterns can be helpful in making crucial decisions and used to predict future events.

Since sequential pattern mining is so valuable, it has been studied by many researchers. Recent studies include: (1) improved algorithms [31,34,18,42,23]; (2) constraint-based sequential pattern mining [15,28,37,6,33,10]; (3) incremental sequential pattern mining [36,45,26,8,41]; (4) mining variants of sequential patterns, including maximum sequential patterns [2,40], similar sequential patterns [30,4] closed sequential patterns [39,35,9,41,5], iterative patterns [25], and fuzzy sequential patterns [21,19,24,7]; (5) mining sequential pattern from different sources [46]; (6) storage and querying

* Corresponding author. Tel.: +886 3 5914010; fax: +886 3 5820085.
E-mail address: sywu@itri.org.tw (S.-Y. Wu).

0169-023X/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2009.06.010

methods for sequential patterns [29]; (7) mining patterns from interval-based event sequences [22,38,20]; (8) mining patterns from sequences with point- and interval-based events [12,11] and this work; and many others.

The seven types of researches mentioned above can be further divided into three categories: (1) mining patterns from point-based event sequences; (2) mining patterns from interval-based event sequences; and (3) mining patterns from hybrid event sequences (sequences consisting of point- and interval-based events). The point-based category includes types (1) through (6). In this category, researchers deal with point-based event data, wherein a sequence is a series of events or items that occur at specific time points.

The second category, the interval-based category, includes type (7). In this category, the patterns are discovered from interval-based event sequences. Events in this domain last for periods of time instead of happening at specific points in time, and the starting and ending times of these interval-based events are known and stored in databases. *Temporal patterns*, which are frequent subsequences consisting of interval events, can be discovered from these interval-based event sequences [22,38]. For example, in a hospital database, each disease a certain patient suffers from can be viewed as an interval-based event, and a temporal pattern may be that patients frequently start a "fever" when they start to "cough" and all these symptoms occur when they catch the flu.

Hoppner [20] for once worked on interval-based event sequences, aimed at discovering temporal pattern rules rather than temporal patterns, as was done in Refs. [22,38]. These rules answer the question, how often will pattern B occur in a given sequence if pattern A has already occurred. This method has been shown to be useful in time series problems.

Type (8) is included in the third category. As mentioned previously, sequential patterns are discovered using either point-based or interval-based approaches. In some applications, however, events are neither purely point-based nor purely interval-based; there may include both kinds of events in data sequences, for example, meteorological phenomena. Thunder and lightning are point-based events, while rain, snow, and sunshine are interval-based events. One common example for meteorological hybrid temporal pattern is that of "lightning (point-based event) followed by thunder (point-based event), both of which happen during a rain storm (interval-based event)". This pattern, consisting of point- and interval-based events, can be called a *hybrid temporal pattern*. If it is supported by sufficient hybrid event sequences, it is a *frequent* hybrid temporal pattern. In this study we develop an algorithm that can be used to discover all frequent hybrid temporal patterns from a set of hybrid event sequences given a threshold *min_sup*. These patterns are more informative than the patterns discovered by either point-based methods or interval-based methods alone.

There has been some previous research on hybrid event sequences. For example Amo et al. proposed a constraint-based method to discover hybrid temporal patterns [12,11]. Constraint-based methods, such as *SPIRIT* [15], in the point-based category and *MILPRIT* [*] [12,11] in the hybrid category, allow the user to reduce the search space by setting regular expression constraints on the patterns. In our study, we focus on discovering the hybrid temporal pattern by employing the embedding store technique, which has been used in other mining problems [1,42,43]. Since doing so can reduce the number of database scans, it is expected that our proposed method could have a better performance than *MILPRIT** when no pattern constraint is imposed. Further comparisons between *MILPRIT** and our method are discussed in Section 2.3.1.

1.1. Applications of hybrid temporal patterns

Hybrid temporal patterns have many applications. In meteorology, we can use the discovered hybrid temporal patterns to predict typhoons, earthquakes, or even tsunamis. In the financial domain the relationships among similar, contiguous subsequences were established in a recent research [13]. In this paper, fluctuations of stock indexes are treated as interval-based events, while announcements of cash dividends and stock splits can be treated as point-based events. The financial hybrid temporal patterns are discovered to help people determine when to buy or sell stocks. Similarly, in the world of medicine, most diseases are interval-based, while treatments are often point-based. Medical hybrid temporal patterns describe the relations between diseases and treatments. Without hybrid temporal pattern mining, we can only find the relations among point-based events or among interval-based events, which may lead to incorrect decisions due to incomplete knowledge. Clearly, hybrid temporal pattern mining is useful and necessary in diverse applications.

1.2. Paper organization

Although mining hybrid temporal patterns is a significant problem, to the best of our knowledge, very few researches have considered this problem. In Section 2, related work is introduced. The difference between these and the hybrid temporal pattern mining problem are explained. In Section 3, we formally define the hybrid temporal pattern mining problem. Based on the definitions, we propose the algorithm, named *HTPM*, for mining hybrid temporal patterns in Section 4. Performance evaluations using both synthetic and real data sets are given in Section 5 while conclusions are drawn in Section 6.

2. Related works

The sequential pattern mining (point-based), temporal pattern mining (interval-based), and hybrid temporal pattern mining (hybrid) problems are quite different. In this section, we discuss these mining problems and the main methods used to resolve them and point out the differences among them.