Information Sciences 181 (2011) 398–418

Contents lists available at ScienceDirect

Information Sciences

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

Discovering multi-label temporal patterns in sequence databases

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

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

ARTICLE INFO

Article history:
Received 4 July 2009
Received in revised form 3 September 2010
Accepted 12 September 2010

Keywords:
Sequential patterns
Point-based event sequence
Interval-based event sequence
Temporal patterns

ABSTRACT

Sequential pattern mining is one of the most important data mining techniques. Previous research on mining sequential patterns discovered patterns from point-based event data, interval-based event data, and hybrid event data. In many real life applications, however, an event may involve many statuses; it might not occur only at one certain point in time or over a period of time. In this work, we propose a generalized representation of temporal events. We treat events as multi-label events with many statuses, and introduce an algorithm called *MLTPM* to discover multi-label temporal patterns from temporal databases. The experimental results show that the efficiency and scalability of the *MLTPM* algorithm are satisfactory. We also discuss interesting multi-label temporal patterns discovered when *MLTPM* was applied to historical Nasdaq data.

1. Introduction

Data mining (DM) extracts implicit, previously unknown, and potentially useful information from databases. It has been successfully used in many applications, such as risk analysis, financial analysis, customer relationship management, churn prediction, fraud detection, and so on. Many data mining approaches have been proposed to discover knowledge from different types of data. Sequential pattern mining is one of the most important techniques in the data mining research field.

The sequential pattern mining problem was first introduced in the mid-1990s [4]. A typical example of a sequential pattern is a customer who, after buying a computer, returns to buy a scanner and a microphone. In this example, the sequence data is nothing but a sequence of events (items) that happen chronologically, and the pattern tells us which items were bought and in what order.

According to the classification scheme proposed by Wu and Chen [38], traditional sequential pattern mining methods can be classified into three categories: (1) *point-based methods*, which discover patterns from point-based event sequences, (2) *interval-based methods*, which discover patterns from interval-based event sequences, and (3) *hybrid-based methods*, which discover patterns from hybrid event sequences.

The first type assumes that events are point-based events, which is the common assumption adopted by most previous sequential pattern mining studies. An event is viewed as something that occurs at a certain point in time. For example, in analyzing web traversal behaviors, if we record when users visit pages but not how long they stay, then every visit can be treated as a point-based event. In the past, research on web usage mining has investigated how we can transform users' browsing data into point-based sequence data through data preprocessing techniques, and how interesting web traversal behaviors can be discovered by applying sequential pattern mining methods [10–13,32,33,40].

In many practical situations, however, events cannot be represented as points. In the web traversal application example, a visit becomes an interval event if we record when the page is visited and how long the user stays. Therefore, some previous

[*] Corresponding author. Tel.: +886 3 4267266; fax: +886 3 4254604.
E-mail address: ylchen@mgt.ncu.edu.tw (Y.-L. Chen).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.09.024

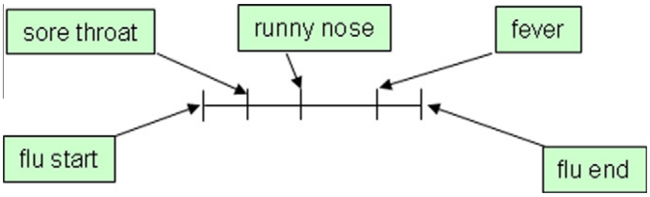

Fig. 1. The statuses that occur during the flu.

research extended the sequential pattern mining problem to include interval-based events [19,39]. When events are intervals, an event can be described with three major characteristics: event name, event starting time, and event ending time. The research on mining temporal patterns from interval-based events attempted to find the temporal relationships between these interval events. For example, in a hospital-related application, a temporal pattern may be that patients frequently start a *fever* when they start to *cough* and these symptoms all occur when they catch the flu.

In some applications, however, events are neither purely point-based nor purely interval-based; they are hybrid events. For example, phenomena in meteorology can be treated as hybrid events. Thunder and lightning are point-based events, while rain, snow, and sunshine are interval-based events. One common meteorological hybrid temporal pattern is "thunder (point-based event) occurred after a lightning strike, and they both occurred during rain (interval-based event)". This pattern, consisting of point- and interval-based events, is called a *hybrid temporal pattern*. The research on mining temporal patterns from hybrid events attempted to find the temporal relationships between these hybrid events.

As discussed above, there are three types of sequential pattern mining models, categorized by the type of events contained in the sequences. All three models can be summarized by a simple observation: point events are events with one label, interval events are events with two labels, and hybrid events have one or two labels. This work proposes the multi-label model of sequential pattern mining. The model is more capable of describing sequence data that could not be represented by previous models. In the following, we introduce the multi-label model.

In some applications, an event can have many different statuses; we call these events multi-label events. For example, as shown in Fig. 1, when a patient gets the flu (event start), he may have some symptoms (statuses), like a sore throat, runny nose, and a fever. He may be treated by a doctor (status), and at the same time, he may have another disease (another event), like pneumonia. During the entire "flu" period, the situation may change involving different statuses in the same event or different events. The evolution of these statuses, whether in the same or different events, may have some associations; multi-label temporal pattern mining attempts to discover the status relationships among these events.

There are other potential applications of multi-label temporal pattern mining. In traditional web management, the web log only records when a visitor visits a page. Nowadays, it is possible to record more information in the web log, such as the actions performed on a visited page, due to the Ajax technology [15]. Ajax is a group of inter-related web development techniques used for creating interactive web applications by exchanging small amounts of data with the server "behind the scenes". This means the entire web page does not have to be reloaded each time the user performs an action [1]. By recording users' actions in Ajax applications, we can collect information about what actions are performed by users on which pages. Fig. 2 shows a possible sequence with multi-actions. From this kind of sequence data, we may find a pattern like the figure on the right. It means a user browsed a "member register" page *a* (a_{start}) and clicked a button (a_1) to check if his user name was available. Then he clicked a link (a_2) to obtain some information from the popup window (b_{start}), closed it (b_{end}), and finally, he clicked the submit button (a_{end}) to finish the registration process.

The discussions above indicate that multi-label sequence data exist in practice. Unfortunately, since traditional models did not consider multi-label sequence data, they treated multi-label sequence data as either point-based or interval-based event data. This means a large portion of temporal relationships were unused when discovering knowledge, and as a result, the patterns discovered are incomplete and partial. This research gap suggests that traditional sequential pattern mining approaches should be extended to deal with multi-label sequence data.

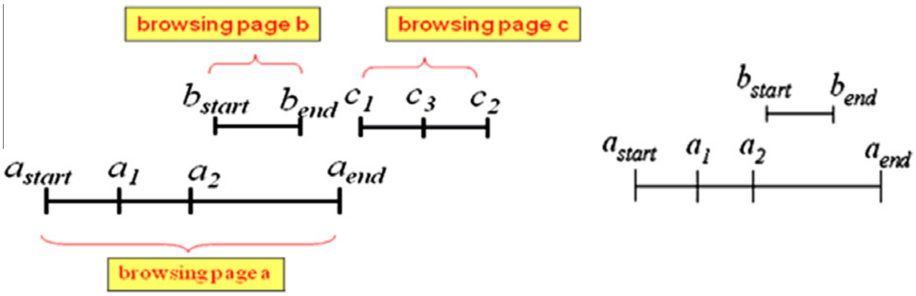

Fig. 2. A multi-label web sequence data and pattern.