Discovery of Characteristic Sequential Patterns Based on Two Types of Constraints

Shigeaki Sakurai, Toshiba Digital Solutions, Kawasaki, Japan

ABSTRACT

This article proposes a method for discovering characteristic sequential patterns from sequential data by using background knowledge. In the case of the tabular structured data, each item is composed of an attribute and an attribute value. This article focuses on two types of constraints describing background knowledge. The first one is time constraints. It can flexibly describe relationships related to the time between items. The second one is item constraints, it can select items included in sequential patterns. These constraints can represent the background knowledge representing the interests of analysts. Therefore, they can easily discover sequential patterns coinciding the interests as characteristic sequential patterns. Lastly, this article verifies the effect of the pattern discovery method based on both the evaluation criteria of sequential patterns and the background knowledge. The method can be applied to the analysis of the healthcare data.

KEYWORDS

Healthcare Data, Item Constraint, Sequential Patterns, Time Constraints

INTRODUCTION

Owing to the progress of computer and network environments, it is easy to collect data with time information such as daily business reports, weblog data, and physiological information. This is the context in which methods of analyzing data with time information have been studied. This paper focuses on a sequential pattern discovery method from discrete sequential data. The research expands the pattern discovery task (Agrawal & Srikanth, 1994). The methods proposed by (Garofalakis et al., 2010), (Pei et al., 2001), (Srikant & Agrawal, 1996), and (Zaki, 2001) efficiently discover the frequent patterns as characteristic patterns. However, the discovered patterns do not always correspond to the interests of analysts, because the patterns are common and are not a source of new knowledge for the analysts.

The problem has been pointed out in connection with the discovery of associative rules. Blanchard et al. (2005), Brin et al. (1997), Silberschatz et al. (1996), and Suzuki et al. (2005) propose other evaluation criteria in order to discover other kinds of characteristic patterns. The patterns discovered by the criteria are not always frequent but are characteristic with some viewpoints. The criteria may be applicable to discovery methods of sequential patterns. However, these criteria do not satisfy the Apriori property. It is difficult for the methods based on the criteria to efficiently discover the patterns. Also, methods that use the background knowledge brought by analysts have been proposed in order to discover sequential patterns corresponding to their interests (Garofalakis et al., 1999).
(Pei et al., 2002), (Sakurai et al., 2008b), (Yen, 2005). In addition, methods that limit the number of sequential patterns (Fournier-Viger et al., 2013), (Hathi & Ambasana, 2015), (Maciag, 2017), (Sakurai & Nishihizawa, 2015), (Tzvetkov et al., 2003) have been proposed in order to avoid discovering large amounts of patterns.

This paper focuses on both sequential interestingness (Sakurai et al., 2008b) and two kinds of constraints representing the background knowledge (Sakurai et al., 2008a), (Sakurai et al., 2008c). This paper introduces a discovery method based on them in order to discover characteristic patterns corresponding to the interests of analysts, where the number of patterns are reasonable. Also, the effect of the method is verified through numerical experiments based on the data collected from real world environment.

**BACKGROUND**

This paper explains basic terminology related to the discovery of sequential patterns. Sequential data is rows of item sets and a sequential pattern is a characteristic subrow extracted from the sequential data. Here, an item is an object, an action, or its evaluation in the analysis target. For example, “beer”, “diaper”, “milk”, and “snack” are items in retail business. Each item set has some items that occur at the same time, but each item set does not have multiple identical items. That is, the data focuses on only whether the items are bought or not by customers. It does not consider the price, the number of buying items, and so on. Formally, a sequential pattern \( s_x \) is described as \( (l_{s1}, l_{s2}, \ldots, l_{sn_x}) \), where \( l_{si} \) is an item set and \( n_x \) is the number of the item sets included in the sequential pattern. The number \( n_x \) is called length and the sequential pattern is called \( n_x \)-th sequential pattern. Also, each \( l_{si} \) is described as \( (v_{z1}, v_{z2}, \ldots, v_{zn_{si}}) \), where \( v_{zj} \) is an item that satisfies the following conditions: \( v_{zik} \neq v_{zik} \) and \( k_1 \neq k_2 \), and \( n_{si} \) is the number of the items included in the item set \( l_{si} \). For example, \{“beer”, “diaper”\}, \{“beer”, “diaper”, “milk”\}, \{“diaper”, “snack”\} is an example of the sequential pattern \( s_{example} \) in the retail business. The pattern is a third sequential pattern and is composed of three item sets: \{“beer”, “diaper”\}, \{“beer”, “diaper”, “milk”\}, and \{“diaper”, “snack”\}. The pattern shows that a person buys “beer” and “diaper” on the first day, buys “beer”, “diaper”, and “milk” on the second day, and buys “diaper” and “snack” on the third day. The sequential pattern is depicted in Figure 1. In this figure, each circle shows an item, each circle with the same textile shows the same kind of items, items separated by arrow lines show item sets, and this figure shows that an item set at the left side occurs before an item set at the right side.

It is necessary to define the frequency of sequential patterns in order to discover the sequential patterns. In advance of this definition, this paper explains the concept of the inclusion for sequential patterns. Let two sequential patterns \( s_1(= (l_{11}, l_{12}, \ldots, l_{1n_1})) \) and \( s_2(= (l_{21}, l_{22}, \ldots, l_{2n_2})) \) be given. \( s_2 \) is included in \( s_1 \), if \( s_1 \) and \( s_2 \) satisfy the following conditions: \( \exists y \{ y_1, y_2, \ldots, y_{n_y} \} \) satisfying the conditions \( y_1 < y_2 < \cdots < y_{n_y} \), and \( l_{21} \subseteq l_{y_1}, l_{22} \subseteq l_{y_2}, \ldots; \) and \( l_{2n_2} \subseteq l_{y_{n_y}} \). The inclusion is described as \( s_2 \subseteq s_1 \). For example, a sequential pattern \{“beer”, “diaper”\}, \{“diaper”, “snack”\} is included in \( s_{example} \) because \{“beer”, “diaper”\} corresponds to the first item set of \( s_{example} \) and \{“diaper”, “snack”\} corresponds to the third item set of \( s_{example} \). Also, another pattern \{“diaper”\}, \{“milk”\})
Related Content

Safe Implementation of Research into Healthcare Practice through a Care Process Pathways Based Adaptation of Electronic Patient Records
[www.igi-global.com/chapter/safe-implementation-research-into-healthcare/46523?camid=4v1a](www.igi-global.com/chapter/safe-implementation-research-into-healthcare/46523?camid=4v1a)

A Multi-Agent Based Modeling and Simulation Data Management and Analysis System for the Hospital Emergency Department

Smart Medication Management, Current Technologies, and Future Directions
[www.igi-global.com/chapter/smart-medication-management-current-technologies-and-future-directions/192746?camid=4v1a](www.igi-global.com/chapter/smart-medication-management-current-technologies-and-future-directions/192746?camid=4v1a)
Knowing How Intranets Enable Knowledge Work: An Exploratory Study in Public Health
www.igi-global.com/chapter/knowing-intranets-enable-knowledge-work/7238?camid=4v1a