Chapter 11

Introduction of Item Constraints to Discover Characteristic Sequential Patterns

Shigeaki Sakurai
Toshiba Digital Solutions Corporation, Japan

ABSTRACT

This chapter introduces a method that discovers characteristic sequential patterns from sequential data based on background knowledge. The sequential data is composed of rows of items. This chapter focuses on the sequential data based on the tabular structured data. That is, each item is composed of an attribute and an attribute value. Also, this chapter focuses on item constraints in order to describe the background knowledge. The constraints describe the combination of items included in sequential patterns. They can represent the interests of analysts. Therefore, they can easily discover sequential patterns coinciding to the interests of the analysts as characteristic sequential patterns. In addition, this chapter focuses on the special case of the item constraints. It is constrained at the last item of the sequential patterns. The discovered patterns are used to the analysis of cause, and reason and can predict the last item in the case that the sub-sequence is given. This chapter introduces the property of the item constraints for the last item.

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 chapter focuses on a sequential pattern discovery method from discrete sequential data. The research expands the pattern discovery task (Agrawal & Srikant, 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. On the other hand, Sakurai et al. (2008b) proposes sequential interestingness as an evaluation criterion satisfying the Apriori property. It can discover sequential patterns including sub-patterns with relatively small frequency. The sequential patterns are regarded as rules which predict remaining sub-patterns in the case that the sub-sequential patterns are given. We can anticipate that the analysts are interested in the sequential patterns to some degree.

Also, the discovery methods tend to discover a large amount of sequential patterns when the thresholds of evaluation criteria are not appropriate. The thresholds depend on the sequential data. Therefore, the selection of appropriate thresholds are not easy for the analysts. Thus, 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. These methods can discover sequential patterns whose number is appropriate.

In addition, some methods use the background knowledge brought in order to discover sequential patterns corresponding to the interests of analysts (Garofalakis et al., 1999), (Pei et al., 2002), (Sakurai et al., 2008a). Garofalakis et al. (1999) describes the background knowledge based on regular expression constraints. Pei et al. (2002) proposes a framework related to 7 constraints such as the inclusion relation and the length of the patterns, and so on. Sakurai et al. (2008a) deals with seven constraints related to time interval between items in the sequential patterns.

This chapter focuses on item constraints representing the background knowledge (Sakurai et al., 2008c), (Sakurai et al., 2008d). Especially, it focuses on the sequential data described in tabular structured format. In the case of the data, we can deal with the relationships between attributes and attribute values. This chapter introduces the property of the constraints and the discovery method of sequential patterns based on the constraints.

BACKGROUND

This chapter explains basic terminology related to the discovery of sequential patterns. Sequential data is rows of item sets and a sequential pattern is a characteristic sub row 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 \( (l_{x1}, l_{x2}, \cdots, l_{xn}) \), where \( l_{xi} \) is an item set and \( nx \) is the number of the item sets included in the sequential pattern. The number \( nx \) is called length and the sequential pattern is called \( nx \)-th sequential pattern. Also, each \( l_{xi} \) is described as \( (v_{x1}, v_{x2}, \cdots, v_{xim}) \), where \( v_{xij} \) is an item that satisfies the following conditions: \( v_{xik} \neq v_{xik} \) and \( k_1 \neq k_2 \),