Chapter LXVII
Sequential Pattern Mining from Sequential Data

Shigeaki Sakurai
Corporate Research & Development Center, Toshiba Corporation, Japan

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 methods proposed by 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.

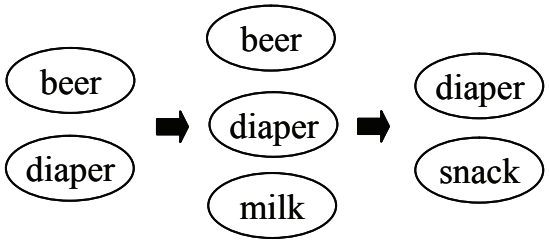
The problem has been pointed out in connection with the discovery of associative rules. Blanchard et al. (2005), Brin et al. (1997), Silberschatz et al. (1996), and Suzuki et al. (2005) propose other criteria in order to discover other kinds of characteristic patterns. The patterns discovered by the criteria are not always frequent but are characteristic of 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, methods that use the background knowledge of analysts have been proposed in order to discover sequential patterns corresponding to the interests of analysts (Garofalakis et al., 1999; Pei et al., 2002; Sakurai et al., 2008b; Yen, 2005).

This chapter focuses on sequential interestingness, which is an evaluation criterion of sequential patterns (Sakurai et al., 2008c). Also, this chapter focuses on 7 types of time constraints that are the background knowledge corresponding to the interests of analysts (Sakurai et al., 2008a). Lastly, this chapter introduces a discovery method based on the sequential interestingness and the time 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 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. Formally, a sequential pattern s_x is described as $(l_{x1}, l_{x2}, \cdots, l_{xn_x})$, where l_{xi} 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-th sequential pattern. Also, each l_{xi} is described as $(v_{xi1}, v_{xi2}, \cdots, v_{xim_i})$, where v_{xij} is an item that satisfies the following conditions: $v_{xik_1} \neq v_{xik_2}$ and $k_1 \neq k_2$, and m_i is the number of the items included in the item set l_{xi}. For example, ({"beer", "diaper"}, {"beer", "milk", "snack"}, {"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", "milk", "snack"}, and {"diaper", "snack"}. The pattern shows that a person buys "beer" and "diaper" on the first day, buys "beer", "milk", and "diaper" 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, 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.

Figure 1.



It is necessary to define the frequency of sequential patterns in order to discover the sequential patterns. In advance of this definition, this chapter explains the inclusion of sequential patterns. Let two sequential patterns $s_1 (= (l_{11}, l_{12}, \cdots, l_{1n_1}))$ and $s_2 (= (l_{21}, l_{22}, \cdots, l_{2n_2}))$ be given. s_2 is included in s_1, if s_1 and s_2 satisfy the following conditions: $\exists y \{y_1, y_2, \cdots, y_{n_2}\}$ satisfying the conditions $y_1 < y_2 < \cdots < y_{n_2}$, and $l_{21} \subseteq l_{1y_1}$, $l_{22} \subseteq l_{1y_2}, \cdots$, and $l_{21} \subseteq l_{1y_{n_2}}$. 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"}) is included in $s_{example}$, because {"diaper"} is included in the first item set of $s_{example}$ and {"milk"} is included in the second item set of $s_{example}$. On the other hand, ({"diaper", "milk"}, {"beer"}) is not included in $s_{example}$, because {"diaper", "milk"} is included in the second item set of $s_{example}$ but {"beer"} is not included in the item set after the second item set.

Each sequential pattern is evaluated to determine whether it is included in each row of sequential data. The number of rows including the sequential pattern is regarded as the frequency of the sequential pattern. For example, sequential data is given as shown in Table 1. The frequency of ({"beer", "diaper"}, {"diaper", "snack"}) is 3, because the sequential pattern is included in D1, D3, and D4. Also, the frequency of ({"diaper", "milk"}, {"beer"}) is 2, because the pattern is included in D1 and D2.

Next, this chapter explains the Apriori property, which is the most important property related to the discovery of sequential patterns. The property requires that if a sequential pattern and its sequential subpattern are given, a value of an evaluation criterion of the sequential pattern is smaller than or equal to a value of an