Discovery of Sequential Patterns Based on Sequential Interestingness and Constraint Conditions

Shigeaki Sakurai
Toshiba Corporation, Japan

INTRODUCTION

We can easily collect a large amount of data due to the progress of the computer and network environments. It is anticipated that the data includes knowledge leading to a smart society. The discovery task of the knowledge from the data is put on a significant position in the information science. Many researchers try to discover the knowledge. Recently, the amount of data expands more and more, leading to the creation of a buzzword such as BigData in the information communication technology field. Even if BigData is not always well-defined, it is characterized by three 'V': Volume, Variety, and Velocity. It means that large and complex data is speedily brought up from our world. Many companies and institutions aggressively try to activate BigData.

This article focuses on sequential data because the sequential data is explosively expanding according to the progress of Twitter and YouTube, the high interest for smart grid and smart community, and so on. The sequential data is important parts of BigData and is a set of sequences. Each sequence is a row of item sets. In the case of retail field, an item is goods, an item set is a receipt, and a sequence is receipts which are gathered and sequentially arranged per a customer. Also, this article focuses on the discovery task of characteristic sequential patterns. The patterns are characteristic subsequences extracted from given sequential data. The task evaluates whether a subsequence is characteristic or not based on given evaluation criteria of sequential patterns. The discovered sequential patterns are activated for floor design, order planning, goods recommendation, and so on. Recently, techniques developed for the task are applied into various application fields. For example, they are applied to the analysis of word sequences and healthcare data. In the case of the word sequences, a word corresponds to an item. In the case of the healthcare data, a discretized test value or its change does so. Initial researches for the task have regarded frequent sequential patterns as characteristic ones. That is, the characteristic sequential patterns are sequential patterns with high frequency in given sequential data. However, the analysts are not always interested in the patterns because they still know the patterns. Therefore, many researchers have developed techniques discovering characteristic sequential patterns that are not simply frequent.

This article introduces an evaluation criterion, sequential interestingness (Sakurai, Kitahara, & Orihara, 2008). The evaluation criterion can simultaneously evaluate the frequency and conditional probability of a sequential pattern. Also, this article introduces the activation method of background knowledge. The knowledge can represent constraints related to the time between items (Sakurai, Ueno, & Orihara, 2008). The criterion and the knowledge can acquire other types of characteristic sequential patterns. On the other hand, this article introduces an application task based on these techniques in order to clarify the necessity of the discovery task in real world. The task is an analysis task of periodical medical examination (Sakurai, Kitahara, Orihara, Iwata, Honda, & Hayashi, 2008). This task discovers change of health situation as characteristic sequential patterns. The patterns are activated for the healthcare guidance by an industrial doctor and the improvement of healthcare situation. Lastly, this article introduces future research directions in this field.

DOI: 10.4018/978-1-4666-5888-2.ch169

BACKGROUND

The discovery task of sequential pattern has been formularized by Agrawal and Srikant (1995). It expands a discovery task of frequent item sets (Agrawal & Srikant, 1994) from a data set by introducing sequential concept. The set is composed of item sets such as receipts in a retail field. It is activated for decision making. Initial researches of the task discover frequent sequential patterns. Srikant and Agrawal (1996) introduced the Apriori property to efficiently discover all frequent sequential patterns. The property states that the frequency of a sequential pattern is smaller than or equal to the frequency of its sequential subpatterns. The proposed discovery method generates candidate sequential patterns whose numbers of item sets are larger due to combinations of frequent sequential patterns. It evaluates whether the candidate sequential patterns are frequent. The generation and the evaluation are repeated until the method discovers all frequent sequential patterns. The use of the Apriori property can avoid generating redundant candidate sequential patterns. Also, Pei, Han, Mortazavi-Asl, Pinto, Chen, Dayal, and Hsu (2001) introduced the concept of prefixes to sequential patterns. The prefixes are sequential subpatterns that appear in the former part of sequential patterns. The proposed discovery method generates projection databases corresponding to the prefixes. Each database is composed of sequences including the common prefix. The method discovers frequent sequential patterns related to the prefixes from their projection databases. These initial researches regard the frequent sequential patterns as patterns coinciding with analysts' interests. However, the patterns do not always coincide with the interests. This is because the patterns are common and are not necessarily a source of new knowledge for the analysts. The patterns coinciding with the interests may be buried in a large number of discovered sequential patterns.

For this problem, two approaches are used. One is the use of other evaluation criteria and the other is the use of background knowledge. Silberschatz and Tuzhilin (1996) proposed an evaluation criterion measuring the interestingness of a pattern. They assume that analysts are interested in unexpected patterns or patterns accompanying with actions. The criterion is calculated by a Bayesian model representing the patterns. Blanchard, Guillet, Briand, and Gras (2005) proposed an evaluation criterion based on a probabilistic model. The evaluation criterion measures the deviation from the maximum uncertainty of the consequent in the case that the antecedent is true. Brin, Motwani, and Silverstein (1997) proposed an evaluation criterion measuring significance of associations. The evaluation criterion uses χ^2 test for correlation from classical statistics. Shimazu, Momma, and Furukawa (2003) and Suzuki and Zytkow (2005) proposed evaluation criteria that discover exceptional patterns.

On the other hand, Garofalakis, Rastogi, and Shim (1999) proposed a method based on the regular expression constraint. The constraint uses user-specified regular expressions as the background knowledge. The expressions are described by referring to the well-defined format. The proposed method applies the expressions to sequential patterns and extracts only sequential patterns that satisfy the expressions. Srikant and Agrawal (1996) proposed a method that introduces time constraints, a time window, and taxonomy. Then, the time constraints specify the minimum and the maximum time interval between adjacent item sets. The time window specifies items included in the same item set. The taxonomy is used in order to specify or generalize items included in sequential patterns by using hierarchy relation among items. The proposed method has good scale-up properties with respect to data size. Pei, Han, and Wang (2002) introduced seven kinds of constraints. The constraints include an item constraint, a superpattern constraint, and a regular expression constraint. The item constraint can extract sequential patterns that include or do not include specific items. The superpattern constraint can extract sequential patterns including specific sequential subpatterns. This research also investigated properties of the constraints and proposed a new framework describing the constraints.

This article focuses on the sequential interestingness which is one of evaluation criteria and the time constraints in order to overcome the discovery problem of characteristic sequential patterns.

SEQUENTIAL DATA AND SEQUENTIAL PATTERN

A sequential pattern is a subsequence satisfying selected evaluation criteria and selected constraints, and is extracted from given sequential data. It is composed of a row of item sets. Here, each item set is composed