Chapter 10

Mining Unexpected Sequential Patterns and Implication Rules

Dong (Haoyuan) Li
LGI2P, École des Mines d’Alès, France

Anne Laurent
LIRMM, Université Montpellier II, France

Pascal Poncelet
LIRMM, Université Montpellier II, France

ABSTRACT

As common criteria in data mining methods, the frequency-based interestingness measures provide a statistical view of the correlation in the data, such as sequential patterns. However, when the authors consider domain knowledge within the mining process, the unexpected information that contradicts existing knowledge on the data has never less importance than the regularly frequent information. For this purpose, the authors present the approach USER for mining unexpected sequential rules in sequence databases. They propose a belief-driven formalization of the unexpectedness contained in sequential data, with which we propose 3 forms of unexpected sequences. They further propose the notion of unexpected sequential patterns and implication rules for determining the structures and implications of the unexpectedness. The experimental results on various types of data sets show the usefulness and effectiveness of our approach.

INTRODUCTION

Most real world applications process the data stored in sequence format, where the elements in data are sequentially ordered with temporal or spatial relation. For instances, in a customer retail database, a sequence can be all purchases of a customer ordered by the time of transaction; in a Web access log file, a sequence can be all of those resources accessed during a user session ordered by the time of request; in a telecommunication network monitoring database, a sequence can be all events during a period ordered by the time of occurrence; in a DNA segment, a sequence is a succession of nucleotide subunits with spatial order, etc. In order to discover the knowledge hidden in such sequential data, sequence data mining techniques (Dong & Pei, 2007; Han & Kamber, 2006) have been highly

DOI: 10.4018/978-1-60566-754-6.ch010
Mining Unexpected Sequential Patterns and Implication Rules

developed and widely applied in many application domains.

As one of the most important models of sequence data mining, the sequential pattern proposed by Agrawal and Srikant (1995) provides a statistical frequency based view of the correlations between the elements in sequential data. The problem of mining sequential patterns can be formally described as follows.

Example 1. Let D be a customer retail database, with the minimum support \( \sigma_{\text{min}} = 0.5 \), we may find the sequential pattern \( s = (\text{login})(\text{msglist})(\text{msgread})(\text{logout}) \) where \( \sigma (s, D) = 0.8 \), which can then be interpreted as “80% of users visit the login page, then visit the message list page, then read messages, and at last logout”. +

Up to now, a great deal of research work focuses on effectively mining sequential patterns (Ayres et al, 2002; Li et al, 2007; Masseglia et al, 1998; Pei et al, 2004; Srikant & Agrawal, 1996; Zaki, 2001) and the variances (Garofalakis et al, 1999; Lo et al, 2007; Mannila et al, 1997; Wang & Han, 2004; Yan et al, 2003). With sequential pattern mining, we can extract the sequences that reflect the most general behaviors within the context of sequential data, which can be further interpreted as domain knowledge for different purposes. However, although sequential patterns are essential for behavior recognition, when we consider domain knowledge within the mining process, the unexpected sequences that contradict existing knowledge on the data have never less importance than the frequent sequences. On the other hand, such unexpected sequences do not mean that they cannot be frequent, so that there exist following problems in discovering the unexpectedness in data with the frequency-based interestingness measures.

First, the redundancy problem of frequency-based data mining methods undermines many real world applications where the exponential pattern or sequence sets generated by mining processes make the post analysis extremely hard. Hence, the identification of unexpected information might be impossible when the support of such unexpected sequences, within the context of sequence data mining, is very low such that the unexpectedness may be hidden in millions of sequential patterns.

Example 2. Let \( D \) be a Web access log database, with the minimum support \( \sigma_{\text{min}} = 0.5 \), we may find the sequential pattern \( s = (\text{login})(\text{msglist})(\text{msgread})(\text{logout}) \) where \( \sigma (s, D) = 0.8 \), which can then be interpreted as “80% of users visit the login page, then visit the message list page, then read messages, and at last logout”. +

Example 3. Let us consider the instance illustrated in Example 1. Assume that in the
Related Content

The LBF R-Tree: Scalable Indexing and Storage for Data Warehousing Systems
Todd Eavis (2010). Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications (pp. 1-27).
www.igi-global.com/chapter/lbf-tree-scalable-indexing-storage/39585?camid=4v1a

Beyond Classification: Challenges of Data Mining for Credit Scoring
Anna Olecka (2007). Knowledge Discovery and Data Mining: Challenges and Realities (pp. 139-161).
www.igi-global.com/chapter/beyond-classification-challenges-data-mining/24905?camid=4v1a

A New Approach to Classification of Imbalanced Classes via Atanassov's Intuitionistic Fuzzy Sets
www.igi-global.com/chapter/new-approach-classification-imbalanced-classes/24213?camid=4v1a

Design and Development of Knowledge Bases for Forward and Reverse Mappings of TIG Welding Process
www.igi-global.com/chapter/design-development-knowledge-bases-forward/24219?camid=4v1a