Chapter 7

Graph–Based Modelling of Concurrent Sequential Patterns

Jing Lu
Southampton Solent University, UK

Weiru Chen
Shenyang Institute of Chemical Technology, China

Malcolm Keech
University of Bedfordshire, UK

ABSTRACT

Structural relation patterns have been introduced recently to extend the search for complex patterns often hidden behind large sequences of data. This has motivated a novel approach to sequential patterns post-processing and a corresponding data mining method was proposed for Concurrent Sequential Patterns (ConSP). This article refines the approach in the context of ConSP modelling, where a companion graph-based model is devised as an extension of previous work. Two new modelling methods are presented here together with a construction algorithm, to complete the transformation of concurrent sequential patterns to a ConSP-Graph representation. Customer orders data is used to demonstrate the effectiveness of ConSP mining while synthetic sample data highlights the strength of the modelling technique, illuminating the theories developed.

1. INTRODUCTION

The goal in patterns mining is to find useful patterns from very large databases. Frequent patterns mining is one of the most important knowledge discovery techniques, which includes frequent itemset mining (Agrawal et al., 1993), sequential patterns mining (Agrawal & Srikant, 1995; Pei et al., 2001; Zaki, 2001), graph mining (Cook & Holder, 2000; Huan et al., 2004) and tree mining (Asai et al., 2002; Zaki, 2005).

While frequent itemset mining aims to find frequent itemsets in a transaction database, sequential patterns mining aims to find sub-sequences that appear frequently (i.e. more than a given support threshold) in a sequence database. The problem
of discovering sequential patterns was first introduced by Agrawal and Srikant (1995) and their approach introduced some of the most important and basic definitions in sequential patterns mining. Since then, it has been studied extensively in the literature, resulting in algorithms such as GSP (Generalized Sequential Pattern; Srikant & Agrawal, 1996), FreeSpan (Frequent pattern-projected Sequential patterns mining; Han et al., 2000), PrefixSpan (Prefix-projected Sequential patterns mining; Pei et al., 2001) and SPADE (Sequential PAttern Discovery using Equivalence classes; Zaki, 2001).

In traditional sequential patterns mining, as the support threshold decreases the number of sequential patterns can increase rapidly, and it is difficult to explore so many patterns or get an overall view of them. As a result, there are some trends to mine a more condensed or constrained set of sequential patterns such as closed sequential patterns (Yan et al., 2003), compressed sequential patterns (Chang et al., 2006) and contiguous sequential patterns (Chen & Cook, 2007).

With the successful implementation of efficient and scalable algorithms for mining sequential patterns and their variations, it is natural to extend the scope of previous study to structured data mining – the process of finding and extracting useful information from semi-structured databases – such as graph mining and tree mining.

Graph mining here means either graph-transaction mining or single-graph mining (Ivancsy & Vajk, 2005). In graph-transaction mining the database to be mined is a set of graphs and the purpose of this mining task is to search for sub-graphs which occur at least in a given number of graphs (Inokuchi et al., 2003; Huan et al., 2004). On the other hand, in the single-graph format, the input data of the mining process is a single large graph and reoccurring sub-graphs are searched in the single graph (Cook & Holder, 2000; Kuramochi & Karypis, 2004).

Tree mining, being another instance of frequent patterns mining, extracts frequent sub-trees from a database of labelled trees (Zaki, 2005). Mining frequent trees is useful in applications like bio-informatics, computer vision, text retrieval, web analysis and so on. For example, Asai et al. (2002) modelled semi-structured data by labelled ordered trees and studied the problem of discovering all frequent tree-like patterns in a given collection of datasets.

In the context of frequent patterns mining, the term pattern refers to itemset, sequence, graph and tree patterns. There are some questions in this area, for example: is it possible to summarise and represent these patterns? Can any other patterns be discovered beyond these to extend the scope of frequent patterns mining?

For the first question, Zaki et al. (2005) introduced the Data Mining Template Library and provided a description of the graphical representation of these frequent patterns. In particular, with respect to sequential patterns summary and analysis, Lu, Wang et al. (2004) proposed a Sequential Patterns Graph (SPG) as the minimal representation of a collection of sequential patterns. These research areas are discussed further in the Related Work section below.

For the second question, there is some research on mining more general structured patterns such as partial orders, by summarising sequential data (Garriga, 2005), and structural relation patterns from Post Sequential Patterns Mining (Lu & Adjei et al., 2004).

Garriga (2005) addressed the task of summarising sequences by means of local ordering relationships on items. Their work goes beyond the idea of closed sequential patterns in that they generalised sequential patterns into closed partial orders and modelled the patterns using the concept of a lattice. They showed that post-processing of the closed sequences leads to the generalisation of closed partial orders from sequential patterns.