Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 176 (2006) 1003–1031
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Looking into the seeds of time: Discovering temporal patterns in large transaction sets

Yingjiu Li [a,*], Sencun Zhu [b], X. Sean Wang [c], Sushil Jajodia [d]

[a] *School of Information Systems, Singapore Management University, Singapore 259756, Singapore*
[b] *Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA*
[c] *Department of Computer Science, University of Vermont, Burlington, VT 05405, USA*
[d] *Center for Secure Information Systems, George Mason University, Fairfax, VA 22030, USA*

Received 19 February 2004; received in revised form 4 November 2004; accepted 26 January 2005

Abstract

This paper studies the problem of mining frequent itemsets along with their temporal patterns from large transaction sets. A model is proposed in which users define a large set of temporal patterns that are interesting or meaningful to them. A temporal pattern defines the set of time points where the user expects a discovered itemset to be frequent. The model is general in that (i) no constraints are placed on the interesting patterns given by the users, and (ii) two measures—*inclusiveness* and *exclusiveness*—are used to capture how well the temporal patterns match the time points given by the discovered itemsets. Intuitively, these measures indicate to what extent a discovered itemset is frequent at time points included in a temporal pattern p, but not at time points not in p. Using these two measures, one is able to model many temporal data mining problems

* Corresponding author. Tel.: +65 6822 0913; fax: +65 0822 0919.
E-mail addresses: yjli@smu.edu.sg (Y. Li), szhu@cse.psu.edu (S. Zhu), xywang@cs.uvm.edu (X.S. Wang), jajodia@gmu.edu (S. Jajodia).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2005.01.019

appeared in the literature, as well as those that have not been studied. By exploiting the relationship within and between itemset space and pattern space simultaneously, a series of pruning techniques are developed to speed up the mining process. Experiments show that these pruning techniques allow one to obtain performance benefits up to 100 times over a direct extension of non-temporal data mining algorithms.

Keywords: Knowledge discovery; Temporal data mining; Temporal pattern

1. Introduction

A large amount of data is collected every day in the form of event time sequences. Common examples are customer transactions in supermarkets, logs of network connections, bank transactions, or events related to manufacturing in industry. These sequences are valuable sources to analyze not only the frequency of certain events, but also the temporal patterns with which the events happen. For example, from data consisting of web clicks, one may discover that a large number of web browsers that visit washingtonpost.com in *morning hours* also visit cnn.com. As another example, in market basket analysis, a supermarket manager may discover that turkey and pumpkin pie are frequently sold together *in November*. Discovering those patterns may reveal interesting information that can be used for understanding of the behavior of customers, market, or the monitored process in different time periods.

However, these types of patterns cannot be discovered by traditional non-temporal data mining approaches that treat all the data as one large segment, with no attention paid to utilizing the time information of the transactions. Returning to the supermarket example, suppose the data set consists of sales data over several years. If look into the entire data set rather than the transactions that occur in November, it is likely that one will not be able to discover the pattern of turkey and pumpkin pie since the overall support for them (i.e., the percentage of all transactions that contain the items turkey and pumpkin pie) will be evidently small.

Mining temporal patterns together with frequent itemsets (or events) is challenging since the sizes of itemset space and temporal pattern space are extremely large in practice. In the supermarket example, the itemset space consists of all combinations of items (i.e., itemsets) and the pattern space all subsets of time points which are considered in data mining. Given 1000 items and 365 time points (e.g., days), the size of itemset space and pattern space are $2^{1000} - 1$ and $2^{365} - 1$, respectively.

Although mining frequent itemsets alone has been extensively studied in the past ten years (see, e.g., [1,18,5,20,7,9,14,11,3]), mining temporal patterns together with frequent itemsets has not attracted much effort until recently (see