Expert Systems with Applications 36 (2009) 9747–9752

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An effective mining approach for up-to-date patterns ☆

Tzung-Pei Hong [a,b,*], Yi-Ying Wu [d], Shyue-Liang Wang [c]

[a] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan
[b] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan
[c] Department of Information Management, National University of Kaohsiung, Kaohsiung 811, Taiwan
[d] Department of Electrical Engineering, National University of Kaohsiung, Kaohsiung, 811, Taiwan

ARTICLE INFO

Keywords:
Data mining
Temporal patterns
Up-to-date patterns
Lifetime

ABSTRACT

Mining association rules is most commonly seen among the techniques for knowledge discovery from databases (KDD). It is used to discover relationships among items or itemsets. Furthermore, temporal data mining is concerned with the analysis of temporal data and the discovery of temporal patterns and regularities. In this paper, a new concept of up-to-date patterns is proposed, which is a hybrid of the association rules and temporal mining. An itemset may not be frequent (large) for an entire database but may be large up-to-date since the items seldom occurring early may often occur lately. An up-to-date pattern is thus composed of an itemset and its up-to-date lifetime, in which the user-defined minimum-support threshold must be satisfied. The proposed approach can mine more useful large itemsets than the conventional ones which discover large itemsets valid only for the entire database. Experimental results show that the proposed algorithm is more effective than the traditional ones in discovering such up-to-date temporal patterns especially when the minimum-support threshold is high.

1. Introduction

Knowledge discovery in databases (KDD) is to identify effective, coherent, and useful information in large databases (Frawley, Shapiro, & Matheus, 1991). Through years of research in KDD, a variety of data-mining techniques have been developed. According to the type of databases processed, the mining approaches may be classified as working on transaction databases, temporal databases, relational databases, multimedia databases, and stream database, among others. There are also many mining methods proposed in KDD, such as techniques for association rules, classification rules, clusters, sequential patterns and so on. In particular, association rules have been most used in KDD (Agrawal & Srikant, 1994; Agrawal, Srikant, & Vu, 1997; Han & Fu, 1995; Mannila, Toivonen, & Verkamo, 1994; Park, Chen, & Yu, 1997; Savasere, Omiecinski, & Navathe, 1995; Srikant & Agrawal, 1995). They are used to describe correlation relationships among items or itemsets in transactional databases and have successful applications in many areas.

Many algorithms based on the Apriori algorithm (Agrawal, Imielinksi, & Swami, 1993a), which generated candidate itemsets in a top–down level-wise process, were proposed to mine association rules. When the percentage of transactions containing a candidate itemset is greater than or equal to a user-specified minimum-support threshold, the itemset is called as a frequent (large) one and thought of as possessing correlation relationships among the items included.

Temporal data mining is another important topic attracting many researchers recently. It is concerned with the analysis of temporal data and the discovery of temporal patterns and the regularities in temporal datasets. It typically reveals ordered correlation of itemsets in transactions along with time. For example, consider a database from a retail store. The sales of ice cream in summer and the sales of mitten in winter should be higher than those in the other seasons. Such seasonal behavior of specific items can only be discovered when a proper window size is chosen for the mining process (Roddick & Spiliopoulou, 2002). But a fixed window size may also hide some information about the items.

In this paper, the concept for up-to-date patterns is described. Each itemset is attached an up-to-date lifetime, in which the user-defined minimum-support threshold must be satisfied. In some cases, an itemset may not be frequent (large) for an entire database but may be large up-to-date since the items seldom occurring early may often occur lately. The up-to-date patterns thus include the itemsets which are frequent for a flexible period of time from the current to the oldest past time. Up-to-date patterns are practical in the field of data mining because they can provide more useful information for the current usage than traditional ones. For example, the counts of be mined patterns by up-to-date approach are larger than traditional data mining, consequently,

☆ This is a modified and expanded version of the paper "Mining up-to-date knowledge from log data", presented at the IEEE International Conference on Granular Computing, 2008, China.

* Corresponding author. Address: Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan.
E-mail address: tphong@nuk.edu.tw (T.-P. Hong).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.02.029

we can find more effective association rules. Also, up-to-date information is usually important to making decisions. For example, a newly announced product such as i-phone may not be discovered as a frequent item from the whole market transaction database. It may, however, be mined in time by the proposed approach. Market managers can thus make effective decisions for marketing strategy.

The remainder of this paper is organized as follows. Related works are reviewed in Section 2. The up-to-date patterns from a log database and the proposed mining algorithm are described in Section 3. An example to illustrate the proposed algorithm is given in Section 4. Experiments for verifying the effectiveness of the above approach are stated in Section 5. Conclusion and future work are given in Section 6.

2. Review of related works

In this section, some related researches about mining association rules are briefly reviewed. They are data mining for association rules and temporal data mining.

2.1. Data mining for association rules

Data-mining technology has become increasingly important in the field of large databases and data warehouses. This technology helps discover non-trivial, implicit, previously unknown and potentially useful knowledge, thus being able to help managers make good decisions. Among the various types of databases and mined knowledge, mining association rules from transaction databases is the most interesting and popular (Agrawal et al., 1993a, 1997; Agrawal, Imielinksi, & Swami, 1993b; Han & Fu, 1995; Li & Deogun, 2005; Mannila et al., 1994; Savasere et al., 1995; Srikant & Agrawal, 1995). In general, the process of mining association rules can roughly be decomposed into two tasks:

(1) finding frequent (large) itemsets satisfying a user-specified minimum-support threshold from a given database, and
(2) generating interesting association rules satisfying a user-specified minimum confidence threshold from the frequent itemsets found.

A variety of mining approaches based on the Apriori (Agrawal & Srikant, 1994) algorithm were proposed, such as DIC (Brin, Motwani, Ullman, & Tsur, 1997), DHP (Park et al., 1997), Sampling, and FP-Growth (Han, Pei, & Yin, 2000). Each of them was designed for a specific problem domain, a specific data type, or for improving its efficiency.

2.2. Temporal data mining

Association-rule mining discovers unordered correlations between items from a given database. However, temporal data mining reveals ordered correlations from databases with time. More formally, temporal data mining is concerned with the analysis of temporal data to find out temporal patterns and regularity from a set of data with time. Temporal patterns can be discovered in a variety of forms, like sequential association rules (Agrawal & Srikant, 1995), periodical association rules (Li & Deogun, 2005), cyclic association rules (Ozden, Ramaswamy, & Silberschatz, 1998), and calendar association rules (Li, Ning, Wang, & Jajodia, 2003).

The calendar time expression (Verma, Vyas, & Vyas, 2005) is widely used to specify the features of temporal patterns. A calendar time expression is composed of calendar units in a specific calendar and may represent different time features, such as an absolute time interval over the time domain, a periodic time over the time domain, or a periodic time within a specific time period. However, it is very difficult to choose a right period of time such that the associations of particular interest can be found from the transaction data. Besides, since different items may have different exhibition periods in a log database, considering a fixed window size of each item might not lead to a fair measurement. Therefore in this paper, we propose a flexible algorithm with different effective lifetimes for items, focusing on the most recent itemsets.

3. The proposed approach for mining up-to-date patterns from log databases

In this section, a new concept for up-to-date patterns is proposed. Up-to-date information is usually important to making decisions. In traditional data mining, an itemset may not be frequent (large) for an entire database but be large up-to-date since the items occurring early may not occur lately. Finding up-to-date knowledge is thus very interesting and practical in the field of data mining. The concept of up-to-date knowledge is shown in Fig. 1.

For example, supposed user set a minimum support is 5%, and the database are consist of 100 transactions, thus, an item can be extracted if the occurrence of the item have to be larger than 5, in other words, if an item a occur only 4 times in the database, it would not be extracted. On the contrary, we may extract the pattern like $(\{a\}, \langle 2008/4/21\ 10{:}00,\ 2008/4/22\ 13{:}30\rangle)$ in this algorithm.

In Fig. 1, an up-to-date pattern is a frequent (or called large) pattern with a valid lifetime, in which the end point is the current time. Its start time will make the lifetime as long as possible. It is a little like the concept of slide windows that it only care the most recent itemsets in a fixed length. In this algorithm, we not only care about the whole database, but also care the most recent itemsets in a non-fixed length. Formally, an up-to-date pattern is defined as follows.

Definition. An up-to-date pattern is a pair $(\{Itemset\}, \langle Lifetime\rangle)$, where the first term *Itemset* is a set of items and is large from a database duration the lifetime which is the second term in the pair. The end value of the lifetime is the current time and no other lifetime for the itemset may last longer than it.

According to the above definition, an algorithm is proposed in this section to find all the up-to-date patterns from a given log database. It first translates the log database into an item-oriented bitmap representation to speed-up the execution in the later mining process and then extracts large itemsets valid with the longest lifetime from the past to the current time. The proposed approach can mine more useful large itemsets than the conventional ones which discover large itemsets valid only for the entire database. Before the algorithm is described, the notation used in this paper is first defined below.

3.1. Notation

D	the log database
n	the number of transactions in D
I	an item or an itemset

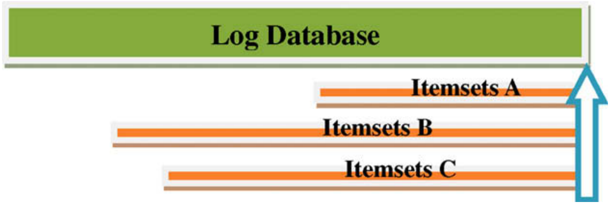

Fig. 1. The concept of up-to-date knowledge.