Preference-Based Frequent Pattern Mining

Moonjung Cho, University of Buffalo, USA
Jian Pei, Simon Fraser University, Canada
Haixun Wang, IBM, T.J. Watson Research Center, USA
Wei Wang, Fudan University, China

ABSTRACT

Frequent pattern mining is an important data-mining problem with broad applications. Although there are many in-depth studies on efficient frequent pattern mining algorithms and constraint pushing techniques, the effectiveness of frequent pattern mining remains a serious concern: It is non-trivial and often tricky to specify appropriate support thresholds and proper constraints. In this paper, we propose a novel theme of preference-based frequent pattern mining. A user simply can specify a preference instead of setting detailed parameters in constraints. We identify the problem of preference-based frequent pattern mining and formulate the preferences for mining. We develop an efficient framework to mine frequent patterns with preferences. Interestingly, many preferences can be pushed deep into the mining by properly employing the existing efficient frequent pattern mining techniques. We conduct an extensive performance study to examine our method. The results indicate that preference-based frequent pattern mining is effective and efficient. Furthermore, we extend our discussion from pattern-based frequent pattern mining to preference-based data mining in principle and draw a general framework.

Keywords: data mining; frequent pattern mining

INTRODUCTION

Frequent pattern mining (Agrawal et al., 1993) is an important data-mining task with broad applications. In the last decade, there have been many in-depth studies on frequent pattern mining, which can be grouped into two categories. The first category of studies (Agrawal & Srikant, 1994; Han et al., 2000; Zaki et al., 1997a, 2001) focus on developing efficient algorithms for frequent pattern mining. With the exciting progress, mining frequent patterns from large databases becomes highly feasible,
although the development of faster algorithms is still demanding to catch up with the speed of data accumulation. On the other hand, it has been well recognized that the effectiveness of frequent pattern mining is a critical concern (Zheng et al., 2001). In many cases, frequent pattern mining may return such a huge number of frequent patterns that a user cannot handle it. To tackle this problem, the constraint-based frequent pattern-mining framework is proposed (Ng et al., 1998). Various constraints can be raised, and only the frequent patterns satisfying the constraints should be mined. Efficient algorithms have been developed to push different kinds of constraints deep into the mining (Kifer et al., 2003; Lakshmanan et al., 1999; Ng et al., 1998; Pei & Han, 2000; Pei et al., 2001).

With the current frequent pattern-mining algorithms and constraint pushing techniques, is frequent pattern mining effective and efficacious enough?

**Example 1 (Motivating Example 1).** Suppose a manager in a large supermarket wants to find frequent patterns containing expensive items from customer transactions. In the constraint-based mining framework, to make the mining more selective, the manager may specify some constraints, such as each pattern should contain an item with price more than $100, and the total amount of each pattern should be at least $500.

The quality of this mining largely depends on whether the user can specify some proper constraints. However, the constraint specification is often challenging. For example, without a real test, it is hard to quantify expensive item(s) in a constraint. In practice, a user often has to adopt a make-do-and-mend approach.

Finding the appropriate values for parameters in constraints by running the mining algorithms again and again is usually time consuming.

**Example 2 (Motivating Example 2).** Consider mining frequent patterns for classification. Several recent studies (Dong & Li, 1999; Li et al., 2001; Liu et al., 1998; Wang et al., 2000) have shown that classification based on frequent patterns, such as associative classification and classification by emerging patterns, can achieve high accuracy and good understandability. Intuitively, given a training data set where the records are grouped into two classes, positive samples $C_+$ and negative samples $C_-$, we want to find patterns frequent in one class and infrequent in the other. Moreover, the longer a frequent pattern (i.e., the more features a pattern covers), the better the predictability of a pattern. Thus, a user may specify a constraint on the length of the patterns. Again, quantifying the constraints is tricky. Without many real tests and fine tuning, there is no guarantee on the quality of the mining.

As shown in the previous examples, even with the progress on efficient frequent pattern mining algorithms and constraint pushing techniques, the effectiveness of frequent pattern mining still remains a serious concern. The major bottleneck is that a user has to specify the appropriate constraints, which is often beyond the user’s knowledge about the data.

In this study, we propose preference-based frequent pattern mining, a novel theme of frequent pattern mining. Instead of specifying solid constraints, a user can simply specify preferences. In Example 1, the manager in the supermarket can write a preference:
A Survey of Multidimensional Modeling Methodologies
[www.igi-global.com/article/survey-multidimensional-modeling-methodologies/1824?camid=4v1a](www.igi-global.com/article/survey-multidimensional-modeling-methodologies/1824?camid=4v1a)

FINGERPRINT: Summarizing Cluster Evolution in Dynamic Environments
Eirini Ntoutsi, Myra Spiliopoulou and Yannis Theodoridis (2012). *International Journal of Data Warehousing and Mining* (pp. 27-44).
[www.igi-global.com/article/fingerprint-summarizing-cluster-evolution-dynamic/67572?camid=4v1a](www.igi-global.com/article/fingerprint-summarizing-cluster-evolution-dynamic/67572?camid=4v1a)