Chapter 2

Association Rule and Quantitative Association Rule Mining among Infrequent Items

Ling Zhou
University of Illinois at Chicago, USA

Stephen Yau
University of Illinois at Chicago, USA

ABSTRACT

Association rule mining among frequent items has been extensively studied in data mining research. However, in recent years, there is an increasing demand for mining infrequent items (such as rare but expensive items). Since exploring interesting relationships among infrequent items has not been discussed much in the literature, in this chapter, the authors propose two simple, practical and effective schemes to mine association rules among rare items. Their algorithms can also be applied to frequent items with bounded length. Experiments are performed on the well-known IBM synthetic database. The authors’ schemes compare favorably to Apriori and FP-growth under the situation being evaluated. In addition, they explore quantitative association rule mining in transactional databases among infrequent items by associating quantities of items: some interesting examples are drawn to illustrate the significance of such mining.

INTRODUCTION

The main goal of association rule mining is to discover relationships among set of items in a transactional database. Association rules have been extensively studied in the literature since Agrawal et al. (1993; 1994) first introduced them. A typical application of association rule mining is the market basket analysis. An association rule is an implication of the form $A \Rightarrow B$, where $A$ and $B$ are frequent itemsets in a transaction database and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ can be interpreted as “if itemset $A$ occurs in a transaction, then itemset $B$ will also likely occur in the same transaction”. By such information, market personnel can place itemset $A$ and $B$ within close proximity which may encourage the sale of them together and develop discount strategies based on such association/correlation found in the data. Therefore, association rule mining has received a lot of attention. For
example, Agrawal and Imielinski (1995; 1996) discussed mining sequential patterns, as well as mining quantitative association rules in large relational tables in [4], while Bayardo considered efficiently mining long patterns from databases and Dong and Li (1999) studied efficient mining of emerging patterns. On the other hand, Kamber et al. (1997) proposed to use data cubes to mine multi-dimensional association rules and Lent et al. (1997) used a clustering method. While most of the researchers focused on association analysis of rules (Agrawal, Imielinski & Swami, 1993; Chen, Han & Yu, 1996; Han, Pei & Yin, 2000; Mannila, Toivonen & Verkamo, 1994; Savasera, Omiecinski & Navathe, 1995; Srikant & Agrawal, 1995), Brin et al. (1997) analyzed the correlations of association rules. With the development of data mining technique, quite a few researchers worked on the alternative patterns, such as Padmanabhan et al. (2000) who discussed unexpected patterns, Liu et al. (1999) and Hwang et al. (1999) studied exception pattern in, and Savasere et al. (1998), Wu et al. (2004) and Yuan et al. (2002) discussed negative association respectively.

The traditional algorithms discover valid rules by exploiting support and confidence requirements, and use a minimum support threshold to prune their combinatorial search space. Two major problems may arise when applying such strategies. (1) If the minimum support is set too low, this may increase the workload significantly such as the generation of candidate sets, construction of tree nodes, comparisons and tests. It will also increase the number of rules considerably, which makes the traditional algorithms suffer from extremely poor performance problems. In addition, many patterns involving items with substantially different support levels are produced, which usually have a weak correlation and are not really interesting to users. (2) If the minimum support threshold is set too high, many interesting patterns involving items with low supports are missed. Such patterns are useful for identifying associations among rare but expensive items such as diamond necklace, ring and earring, as well as the identification of identical or similar web documents, etc.

Recently, there is some growing interests in developing techniques for mining association patterns without support constraints. The algorithms proposed by Cohen et al. (2000) are limited to deal with identifying pairs of similar columns. The approaches presented by Wang et al. (2001) and Xiong et al. (2003) employ a confidence-based pruning strategy instead of support-based pruning adopted in traditional association rule mining. The mining of support-free association discovers rules in the patterns with high support, cross-support where items have widely differing support levels, and low supports. In fact, patterns with a high minimum support level are often obvious and well known; patterns with cross-support level have extremely poor correlation, and patterns with low support often provide valuable new insights. J. Ding (2005) discussed association rule mining among rare items. He designed a new disk-based data structure, called Transactional Co-Occurrence Matrix (TCOM) to store the data. This structure combines the advantages of transactional oriented (horizontal) layout and item oriented (vertical) layout of the database. So any itemsets could be randomly accessed and counted without a full scan of the original database or the TCOM. He also constructed a compressed matrix structure, called Reduced Transactional Co-Occurrence Matrix (RTCOM) which resides in memory. This matrix only contains the items which are of interest in application. Then the infrequent patterns and the valid association rules among infrequent items can be mined out. Although this is a significant theoretical advancement in the subject, it is quite costly to implement this algorithm. Due to the specialties of patterns with low support and the lack of practical and efficient methods for mining rules among rare items, we propose two schemes (Matrix-based Scheme and Hash-based Scheme) to explore interesting associations among infrequent items with memory-resident data structure. In our both schemes, only two passes over data-
Related Content

Software Quality Modeling with Limited Apriori Defect Data
[www.irma-international.org/chapter/software-quality-modeling-limited-apriori/24898/](www.irma-international.org/chapter/software-quality-modeling-limited-apriori/24898/)

Examining University Retention Efforts of Non-Traditional Students
[www.irma-international.org/chapter/examining-university-retention-efforts-non/60850/](www.irma-international.org/chapter/examining-university-retention-efforts-non/60850/)

Heuristic Search-Based Stacking of Classifiers
Agapito Ledezma, Ricardo Aler and Daniel Borrajo (2002). *Heuristic and Optimization for Knowledge Discovery* (pp. 54-67).
[www.irma-international.org/chapter/heuristic-search-based-stacking-classifiers/22149/](www.irma-international.org/chapter/heuristic-search-based-stacking-classifiers/22149/)

Using "Blackbox" Algorithms Such AS TreeNET and Random Forests for Data-Mining and for Finding Meaningful Patterns, Relationships and Outliers in Complex Ecological Data: An Overview, an Example Using G
[www.irma-international.org/chapter/using-blackbox-algorithms-such-treenet/24212/](www.irma-international.org/chapter/using-blackbox-algorithms-such-treenet/24212/)

Strategic Tacit Knowledge-Based Competitiveness
[www.irma-international.org/chapter/strategic-tacit-knowledge-based-competitiveness/117847/](www.irma-international.org/chapter/strategic-tacit-knowledge-based-competitiveness/117847/)