Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 2350–2359
Expert Systems with Applications
www.elsevier.com/locate/eswa

A novel approach for discovering retail knowledge with price information from transaction databases

Yen-Liang Chen *, Tony Cheng-Kui Huang, Sih-Kai Chang

Department of Information Management, National Central University, Chung-Li 320, Taiwan, ROC

Abstract

With the advances in information technology and the emergence of Internet commerce, analysis of transaction data has become a crucial technique for effective decision-making and strategy formation in business operations. It is especially critical for retail management, in both online and brick-and-mortar stores. Traditional research in mining retail knowledge, however, does not take into account the products' prices and how such settings can affect potential demand. This paper opens a new research dimension by treating products' prices as an important decision variable in mining retail knowledge. To the best of our knowledge, the problem addressed in this paper has never been dealt with in existing research papers. We propose a representation scheme to incorporate price information into historical transaction data. An efficient algorithm is developed to "dig" out implicit, yet meaningful, patterns with price information. In addition, an extensive and well-designed experiment is executed, showing that the algorithm is computationally efficient and that the proposed analysis is significant and useful.

Keywords: Association rules; Data mining; Price; Retailing management

1. Introduction

Data mining extracts implicit, previously unknown, and potentially useful information from databases. According to the classification scheme proposed in Chen, Han, and Yu (1996), major approaches to data mining include mining association patterns, clustering, classification, mining sequential patterns, data generalization and summarization, and traversal pattern analysis. Among them, mining association patterns is probably the most popular because of its widespread applications. This approach was first introduced in Agrawal, Imielinski, and Swami (1993), Agrawal and Srikant (1994), and can be stated as follows.

Given a database of sales transactions, an association pattern, denoted as X, is a set of items that frequently co-occur in databases. To find association patterns from databases, we first need to calculate the support of itemset X, where the support of X is the percentage of transactions in the database containing X. If its support is higher than the user-specified minimum support (*minsup*), we claim that itemset X is frequent. Otherwise, it is infrequent.

Since association patterns are useful and easy to understand, they have been used in many successful business applications, including finance, telecommunications, marketing, recommendation, retailing, and web analysis (Bose & Mahapatra, 2001; Changchien & Lu, 2001; Chen, Tang, Shen, & Hu, 2005; Lee, Kim, & Rhee, 2001; Lin, Chen, Chen, & Chen, 2003; Wang & Shao, 2004). The method has also attracted increased research interest, and many extensions have been proposed in recent years, including (1) algorithm improvements (Brin, Motwani, Ullman, & Tsur, 1997; Chen & Ho, 2005; Han, Pei, & Yin, 2000; Rastogi & Shim, 2002), (2) fuzzy patterns (Chen & Huang, 2005; Kuok, Fu, & Wong, 1998), (3) multi-level patterns (Clementini, Felice, & Koperski, 2000; Han & Fu, 1999), (4) quantitative association patterns (Park, Chen, & Yu,

* Corresponding author. Tel.: +3 4267266; fax: +3 4254604.
E-mail address: ylchen@mgt.ncu.edu.tw (Y.-L. Chen).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.03.006

1997; Srikant & Agrawal, 1996; Hsu, Chen, & Ling, 2004), (5) spatial association patterns (Clementini et al., 2000; Koperski & Han, 1995), (6) inter-transaction patterns (Lu, Feng, & Han, 2000), (7) interesting association patterns (Bayardo & Agrawal, 1999; Freitas, 1999), and (8) temporal association patterns (Ale & Rossi, 2000; Chen, Chiang, & Kao, 2003; Li, Ning, Wang, & Jajodia, 2001; Roddick & Spiliopoilou, 2002). (Chen et al., 1996) and (Han & Kamber, 2006) give brief literature reviews of association patterns.

Previous research on mining association patterns in transaction databases usually assumed that a transaction is formed from a set of items bought in that transaction (Agrawal et al., 1993; Agrawal & Srikant, 1994). In other words, the research ignored items' quantities and prices. Although this assumption is widely used, two difficulties may arise. First, in a practical situation, a transaction not only records purchased items, but also their quantities and prices. Therefore, if we view a transaction as only a set of items, a large portion of stored data is unused. Second, the association patterns found in conventional transaction databases only indicate if items are related or not; they do not tell us their quantity and/or price relationships. Without the quantity and price information, it is difficult to design a competitive package for sales promotions because we do not know how the prices and quantities of items influence one another. For example, we may have an association pattern, such as {milk, cheese}. This pattern would be more informative if it was more specific, such as {milk with high price, cheese with medium price}. The former only indicates that these two items are frequently bought together, but the latter tells us that this association happens when milk is at a high price and cheese at a medium price.

Some readers may wonder why we did not simply view price and quantity as numerical attributes and use methods for mining quantitative association patterns to deal with them (Hsu et al., 2004; Park et al., 1997; Srikant & Agrawal, 1996). Using this method, we can partition the prices or quantities of milk and cheese into multiple intervals. For example, we can partition price into five levels, where these five levels represent the list price, 0–10% off the list price, 10–20% off the list price, 20–30% off the list price, and more than 30% off the list price. Consequently, we can have a pattern like {milk with price level 2, cheese with price level 3}.

Although this idea seems reasonable, it may result in the following problems. Suppose the time span of the database is 12 months, and the prices of milk and cheese are at levels 2 and 3 only in June and July, respectively. Further assume that there are 200,000 total transactions in the database, and 50,000 of those transactions occurred in June and July. If there were 2000 transactions in June and July containing milk and cheese, what is the support of the pattern {milk with price level 2, cheese with price level 3}? Obviously, the answer should be $2000/50{,}000 = 4\%$, rather than $2000/200{,}000 = 1\%$. This is because [June, July] is the only period when this pattern can possibly occur, and the base of the support computation should be based on these two months rather than the entire year.

The above discussion reveals that we need a new method of defining the supports of patterns with price information. Accordingly, a new type of support, called local support, is proposed to measure the frequency of itemsets with price labels. There are still other problems, however, that we may encounter. Suppose we have three combinations of milk and cheese prices as follows: {milk: *p*-level 2, cheese: *p*-level 3}, {milk: *p*-level 2, cheese: *p*-level 1}, and {milk: *p*-level 3, cheese: *p*-level 3}. Suppose their local supports are 4%, 8%, and 2%, respectively. It is a reasonable conjecture that {milk: *p*-level 2, cheese: *p*-level 1} may be good for sales and {milk: *p*-level 3, cheese: *p*-level 3} may not, because the former seems to increase sales while the latter decreases sales. For good or bad, they both represent important information that deserves further analysis. This simple example illustrates that we need a new way to define important patterns. In the past, a pattern with high frequency was deemed important. What we are interested in now, however, are those patterns with frequencies deviating substantially from the average, either larger or smaller.

In addition to price information, we have to include other important information in the patterns, such as quantity information. Therefore, we define the new patterns with price and quantity information, such as {milk: *p*-level 2, average-qty 2.3; cheese: *p*-level 1, average-qty 3.5}. This means that when the price levels of milk and cheese are 2 and 1, respectively, the average quantities of milk and cheese in transactions are 2.3 and 3.5, respectively. By comparing all the patterns with the same product combination, we can understand how items' prices influence one another and how items' prices influence quantities. For example, assume that we have the following patterns:

{milk: *p*-level 2, average-qty 2.4; cheese: *p*-level 3, average-qty 3.3}, sup = 4%
{milk: *p*-level 2, average-qty 2.3; cheese: *p*-level 1, average-qty 3.5}, sup = 8%
{milk: *p*-level 1, average-qty 1.4; cheese: *p*-level 1, average-qty 2.0}, sup = 2%

Using the first pattern as the basis for comparison, we find that the price levels set in the second pattern may increase the frequency of purchase (the support increases from 4% to 8%), but does not change the average purchasing quantity. On the other hand, the third pattern not only decreases the purchasing frequency but also decreases the average purchasing quantity.

This paper defines two new metrics in order to measure a pattern's level of interest. The frequency strength (FS) measures whether a pattern's price levels affect its frequency significantly when compared to the average frequency of the same product combination. When the value of FS is greater than 1, the larger the value is, the higher the possibility of an increase in frequency in the price levels of the pattern will be. When the value of FS is smaller than