Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 59 (2006) 603–626
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Mining itemset utilities from transaction databases

Hong Yao, Howard J. Hamilton *

Department of Computer Science, University of Regina, 3737 Wascana Parkway, Regina, SK, Canada S4S 0A2

Received 13 October 2005; accepted 13 October 2005
Available online 18 November 2005

Abstract

The rationale behind mining frequent itemsets is that only itemsets with high frequency are of interest to users. However, the practical usefulness of frequent itemsets is limited by the significance of the discovered itemsets. A frequent itemset only reflects the statistical correlation between items, and it does not reflect the semantic significance of the items. In this paper, we propose a utility based itemset mining approach to overcome this limitation. The proposed approach permits users to quantify their preferences concerning the usefulness of itemsets using utility values. The usefulness of an itemset is characterized as a utility constraint. That is, an itemset is interesting to the user only if it satisfies a given utility constraint. We show that the pruning strategies used in previous itemset mining approaches cannot be applied to utility constraints. In response, we identify several mathematical properties of utility constraints. Then, two novel pruning strategies are designed. Two algorithms for utility based itemset mining are developed by incorporating these pruning strategies. The algorithms are evaluated by applying them to synthetic and real world databases. Experimental results show that the proposed algorithms are effective on the databases tested.

Keywords: Utility mining; Data mining; Semantic significance; User preference; Itemset

1. Introduction

Frequent itemset mining plays an essential role in the theory and practice of many important data mining tasks, such as mining association rules [1,2,17], long patterns [4], emerging patterns [10] and dependency rules [22]. It has been applied in fields such as telecommunications [3], census analysis [5], and text analysis [22]. An *itemset* is a set (i.e., a group) of items. The goal of frequent itemset mining is to identify all *frequent itemsets*, i.e., itemsets that have at least a specified minimum *support*, the percentage of transactions containing the itemset. The rationale behind using support is that only itemsets with high frequency are of interest to users.

The practical usefulness of the frequent itemset mining is limited by the significance of the discovered itemsets. There are two principal limitations. First, a huge number of frequent itemsets that are not interesting to the user are often generated when the minimum support is low. For example, there may be thousands of

* Corresponding author. Tel.: +1 3065854079; fax: +1 3065854745.
E-mail addresses: yao2hong@cs.uregina.ca (H. Yao), hamilton@cs.uregina.ca (H.J. Hamilton).

0169-023X/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2005.10.004

combinations of products that occur in 1% of the transactions. If too many uninteresting frequent itemsets are found, the user is forced to do additional work to select the itemsets that are indeed interesting. Secondly, support, as defined based on the frequency of itemsets, is not an adequate measure of a typical user's interest. Suppose that the goal of a sales manager is to find the itemsets that can generate a profit higher than a threshold. The following example shows that support based itemset mining may lead to some most profitable itemsets not being discovered due to their low support.

Example 1. Consider the small transaction database shown in Table 1 and the unit profit for the items shown in Table 2. Each value in the transaction database indicates the quantity sold of an item. Using Tables 1 and 2, the support and profit for all itemsets can be calculated (see Table 3). For example, since for the 10 transactions in Table 1, only two transactions, t_8 and t_9, include both items B and D, the support of the itemset BD is $2/10 = 20\%$. Since t_8 includes one B and one D, and t_9 includes one B and 10 Ds, a total of two Bs and 11

Table 1
A transaction database

Transaction ID	Item A	Item B	Item C	Item D
t_1	4	0	1	0
t_2	2	0	0	6
t_3	0	0	1	30
t_4	3	0	0	5
t_5	1	0	0	6
t_6	4	0	2	10
t_7	2	0	0	8
t_8	1	1	1	1
t_9	0	1	0	10
t_{10}	5	0	0	9

Table 2
The unit profit for the items

Item name	Profit ($)
Item A	5
Item B	100
Item C	38
Item D	1

Table 3
The support, and profit for all itemsets

Itemsets	Support (%)	Profit ($)
A	**80**	110
B	20	**200**
C	**40**	190
D	**90**	85
AB	10	105
AC	30	**197**
AD	**70**	135
BC	10	138
BD	20	**211**
CD	30	**193**
ABC	10	143
ABD	10	106
ACD	20	150
BCD	10	139
$ABCD$	10	144