Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 66 (2008) 68–91
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

An efficient algorithm for mining closed inter-transaction itemsets

Anthony J.T. Lee *, Chun-Sheng Wang, Wan-Yu Weng, Yi-An Chen, Huei-Wen Wu

Department of Information Management, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan, ROC

Received 25 January 2007; received in revised form 5 February 2008; accepted 12 February 2008
Available online 20 February 2008

Abstract

In this paper, we propose an efficient algorithm, called ICMiner (Inter-transaction Closed patterns Miner), for mining closed inter-transaction itemsets. Our proposed algorithm consists of two phases. First, we scan the database once to find the frequent items. For each frequent item found, the ICMiner converts the original transaction database into a set of domain attributes, called a dataset. Then, it enumerates closed inter-transaction itemsets using an itemset–dataset tree, called an ID-tree. By using the ID-tree and datasets to mine closed inter-transaction itemsets, the ICMiner can embed effective pruning strategies to avoid costly candidate generation and repeated support counting. The experiment results show that the proposed algorithm outperforms the EH-Apriori, FITI, ClosedPROWL, and ITP-Miner algorithms in most cases.

Keywords: Data mining; Association rules; Inter-transaction itemsets; Closed itemset

1. Introduction

Mining association rules, which is a fundamental problem in the area of data mining, has been extensively studied in recent years [1,4,9,16,18,20,21,23,29,31–33,38,40,46,49]. Traditional association rule mining algorithms focus on association rules among itemsets within a transaction. Taking stock market databases as an example, association rule mining can be used to analyze the share price movements. Suppose a database records the price of every stock at the end of each trading day, an association rule might be: "if the stock prices of Microsoft and IBM go up, the price of Apple is likely to go up on the same day." This classical association rule expresses the associations among items within the same transaction, thus we call it *intra-transactional* association rule. However, the traditional approaches cannot capture a rule like: "if the stock prices of Microsoft and IBM go up, the price of Apple is likely to go up two days later." This rule represents some

* Corresponding author. Tel.: +886 2 33661188; fax: +886 2 33661199.
E-mail addresses: jtlee@ntu.edu.tw (A.J.T. Lee), d91725001@ntu.edu.tw (C.-S. Wang), r93725018@ntu.edu.tw (W.-Y. Weng), cookiemouse@gmail.com (Y.-A. Chen), wu.stars@gmail.com (H.-W. Wu).

0169-023X/$ - see front matter
doi:10.1016/j.datak.2008.02.001

association relationships among the itemsets from different transactions, thus we call it *inter-transaction* association rule.

A number of methods have been proposed for mining inter-transaction association rules [8,10,12,22,25,26,28,41]. Lu et al. [25] first used inter-transaction association rules to predict stock market movements. Subsequently, two algorithms, E-Apriori and EH-Apriori, for finding frequent inter-transaction itemsets were proposed in Ref. [26], where EH-Apriori adopts an additional pruning technique used in Ref. [34]. Feng et al. [10] used several optimization techniques, namely, joining, converging, and speeding, to enhance the EH-Apriori algorithm for mining inter-transactional association rules under rule templates. Then, Tung et al. [41] proposed the FITI algorithm, which is implemented in two phases. First, the frequent intra-transaction itemsets are discovered. Then, these itemsets are used to form the frequent inter-transaction itemsets. More recently, Lee et al. [22] proposed an algorithm, called ITP-Miner, which uses an ITP-tree to mine all frequent inter-transaction itemsets in a depth-first search manner. It has been shown that the ITP-Miner algorithm outperforms the previous inter-transaction mining algorithms.

Besides, Li et al. [24] extended the inter-transaction association rules to a more general form of association rules, called generalized multidimensional inter-transactional association rules, which expand rule contexts from point-wise (e.g. two days latter) to scope-wise (e.g. within three days). For example, such generalized inter-transaction association rules could be: "after McDonald and Burger King open branches, KFC will open a branch within two months and between one and three miles away."

In inter-transaction itemsets mining, there are a large number of frequent itemsets and the mining process could be extremely time-consuming. Thus, we incorporate the concept of closed itemsets into inter-transaction itemsets mining. That is, we only mine closed inter-transaction itemsets, instead of all frequent itemsets. A frequent itemset X is said to be closed if the database does not contain a superset of X with equal support [15,35,37,39,43,47], where the support of an itemset is defined as the number of transactions containing the itemset in the database. Generally speaking, mining closed itemsets is more efficient than mining a complete set of frequent itemsets. Therefore, we will mine closed ones in our proposed method.

In the last decade, many algorithms have been proposed for mining closed frequent itemsets [11,15,27,29,35–37,39,42,43,47]. These algorithms can be classified into four categories [44], namely "test-and-generate", "divide-and-conquer", "hybrid", and "hybrid without duplication".

The "test-and-generate" category includes CLOSE [36], and A-close [35] algorithms, which stress on the optimization of a level-wise process to discover the closed itemsets. The "divide-and-conquer" category includes CLOSET [37], TFP [15], CLOSET+ [43], and FP-CLOSE [11] algorithms, which use the highly compact data structure FP-tree (Frequent-Pattern tree) [14] within the closed itemset discovery process. Pei et al. [37] devised an algorithm, called CLOSET, which uses an FP-tree and a partitioning technique to mine frequent patterns. Subsequently, Han et al. [15] developed the TFP algorithm, for mining top-k closed frequent patterns without a minimum support constraint. Wang et al. [43] integrated more effective strategies and proposed an FP-tree-based method called CLOSET+ that builds conditional projected databases in two ways: bottom-up and top-down. FP-CLOSE [11] is a variant of CLOSET+ for mining closed itemsets.

The "hybrid" category includes CHARM [47], and CloseMiner [39] algorithms, which use properties of both "test-and-generate" and "divide-and-conquer" techniques. Zaki et al. [47] proposed an algorithm, called CHARM, which uses an itemset-tidset tree and four pruning properties to enumerate all closed itemsets. Based on the CHARM algorithm, Singh et al. [39] proposed an algorithm, called CloseMiner, which transforms the problem of discovering closed itemsets into a problem of clustering the itemsets with closed tidsets.

The "hybrid without duplication" category includes DCI-CLOSED [27], LCM [42], and PGMiner[30] algorithms, which can avoid the main drawback of the "hybrid" algorithms. These algorithms do not need to store, in main memory, the closed itemsets previously mined since they do not require performing the subsumption checking. The DCI-CLOSED [27] and LCM [42] algorithms traverse the search space in a depth-first search manner. The differences between DCI-CLOSED and LCM are the strategies for finding the closure and the data structures for storing the context extracted. The PGMiner [30] constructs a prefix graph structure and decomposes the database to bit vectors of variable lengths, which are assigned to the nodes of the prefix graph. Then, it uses both inter-node and intra-node pruning strategies to mine closed patterns.

To mine closed inter-transaction itemsets, Huang et al. [17] proposed the ClosedPROWL algorithm, which uses closed intra-transaction itemsets to form closed inter-transaction itemsets. The algorithm is implemented