Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 60 (2007) 5–29
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

On compressing frequent patterns ☆

Dong Xin, Jiawei Han *, Xifeng Yan, Hong Cheng

Department of Computer Science, University of Illinois at Urbana-Champaign, Rm 2132, Siebel Center for Computer Science, 201 N. Goodwin Avenue, Urbana, IL 61801, USA

Available online 3 March 2006

Abstract

A major challenge in frequent-pattern mining is the sheer size of its mining results. To compress the frequent patterns, we propose to cluster frequent patterns with a tightness measure δ (called δ-cluster), and select a *representative pattern* for each cluster. The problem of finding a minimum set of representative patterns is shown NP-Hard. We develop two greedy methods, RPglobal and RPlocal. The former has the guaranteed compression bound but higher computational complexity. The latter sacrifices the theoretical bounds but is far more efficient. Our performance study shows that the compression quality using RPlocal is very close to RPglobal, and both can reduce the number of closed frequent patterns by almost two orders of magnitude. Furthermore, RPlocal mines even faster than FPClose [G. Grahne, J. Zhu, Efficiently using prefix-trees in mining frequent itemsets, in: Proc. IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03)], a very fast closed frequent-pattern mining method. We also show that RPglobal and RPlocal can be combined together to balance the quality and efficiency.

Keywords: Data mining; Frequent pattern mining

1. Introduction

Frequent-pattern (or itemsets) mining has been a focused research theme in data mining due to its broad applications in mining association [2], correlation [5], causality [18], sequential patterns [3], episodes [14], multi-dimensional patterns [13], max-patterns [9], partial periodicity [11], and many other important data mining tasks.

The problem of frequent-itemsets mining can be defined as follows. Given a transaction database, let $\mathcal{O} = \{o_1, o_2, \ldots, o_d\}$ be the set of items that appear in the database, $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ be the transaction set, and $I(t_i) \subseteq \mathcal{O}$ be the set of items in transaction t_i. For any itemset P, let $T(P) = \{t \in \mathcal{T} | P \subseteq I(t)\}$ be the corresponding set of transactions. We say P is the *expression* of pattern P, and $|T(P)|$ is the *support* of

☆ The work was supported in part by the US National Science Foundation NSF IIS-03-08215/05-13678. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

* Corresponding author. Fax: +1 217 265 6494.
E-mail address: hanj@cs.uiuc.edu (J. Han).

0169-023X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2006.01.006

pattern P. An itemset P is *frequent* if $|T(P)| \geqslant min\_sup$, where *min_sup* is a user-specified threshold. The task of frequent-itemsets mining is to *find all the frequent itemsets*. Several extensions have been made to the original frequent itemsets problem. A frequent itemset P is *closed* if there is no itemset P' such that $P \subset P'$ and $T(P) = T(P')$, a frequent itemset P is *maximal* if there is no frequent itemset P' such that $P \subset P'$.

There have been many scalable methods developed for frequent-pattern mining [12]. However, the real challenge in frequent-pattern mining is the sheer size of its mining results. In many cases, a high *min_sup* threshold may discover only commonsense patterns but a low one may generate an explosive number of output patterns, which severely restricts its usage. To solve this problem, it is natural to explore how to "*compress*" the patterns, i.e., find a concise and succinct representation that describes the whole collection of patterns.

Two major approaches have been developed in this direction: lossless compression and lossy approximation. The former is represented by the *closed frequent itemsets* [16] and *non-derivable frequent itemsets* [6]. Their compression is lossless in the sense that the complete set of original frequent patterns can be recovered. However, the methods emphasize too much on the *supports* of patterns so that its compression power is quite limited. The latter is represented by the *maximal frequent itemsets* [9], as well as *boundary cover sets* proposed recently [1]. These methods only consider the *expressions* of patterns, while the *support* information in most of the itemsets is lost.

To achieve high-quality pattern compression, it is desirable to build up a pattern compression framework that concerns both the *expressions* and *supports* of the patterns. A motivation example is shown as follows.

Example 1. Table 1 shows a subset of frequent itemsets on *accidents* data set [8], where 39, 38, 16, 18, 12, 17 are the names of individual items. The closed itemsets cannot get any compression on this subset. The maximal itemsets will only report the itemset P_3. However, we observe that itemsets P_2, P_3 and P_4 are significantly different w.r.t. their supports, and the maximal itemset totally loses this information. On the other hand, the two pairs (P_1, P_2) and (P_4, P_5) are very similar w.r.t. both expressions and supports. We suggest a high-quality compression as P_2, P_3 and P_4.

A general proposal for high-quality compression is to cluster frequent patterns according to certain similarity measure, and then select and output only a *representative pattern* for each cluster. However, there are three crucial problems that need to be addressed: (1) how to measure the similarity of the patterns, (2) how to define quality guaranteed clusters where there is a representative pattern best describing the whole cluster, and (3) how to efficiently discover these clusters (and hence the representative patterns)? This paper addresses these problems.

First, we propose a distance measure between two frequent patterns, and show it is a valid distance metric. Second, we define a clustering criterion, with which, the distance between the representative pattern and every other pattern in the cluster is bounded by a threshold δ. The objective of the clustering is to minimize the number of clusters (hence the number of representative patterns). Finally, we show the problem is equivalent to set-covering problem, and it is NP-hard w.r.t. the number of the frequent patterns to be compressed. We propose two greedy algorithms: RPglobal and RPlocal. The former has bounded compression quality but higher computational complexity; whereas the latter sacrifices the theoretical bound but is far more efficient. Our performance study shows that the quality of the compression using RPlocal is very close to RPglobal, and both can reduce the number of patterns generated by about two orders of magnitude w.r.t. the original collection of

Table 1
A subset of frequent itemsets

Pattern ID	Itemsets	Support
P_1	{38, 16, 18, 12}	205227
P_2	{38, 16, 18, 12, 17}	205211
P_3	{39, 38, 16, 18, 12, 17}	101758
P_4	{39, 16, 18, 12, 17}	161563
P_5	{39, 16, 18, 12}	161576