Information Sciences 339 (2016) 254–273

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Revisiting bound estimation of pattern measures: A generic framework

Lei Zhang[a], Ping Luo[b,*], Enhong Chen[c,**], Min Wang[d]

[a] Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230039, China
[b] University of Chinese Academy of Sciences, Beijing 100080, China
[c] University of Science and Technology of China, Heifei 230026, China
[d] Google Research, Mountain View, CA 94043, USA

ARTICLE INFO

Article history:
Received 6 January 2015
Revised 26 November 2015
Accepted 26 December 2015
Available online 11 January 2016

Keywords:
Bound estimation
Utility
Occupancy
Constrained pattern mining

ABSTRACT

It is widely recognized that constrained pattern mining helps in the capture of a relatively large amount of semantics among different applications, and thus, increases the effectiveness of mining. One major challenge in this field is how the properties of pattern measures can be pushed deeply into the mining process to achieve improved efficiency. The usual solution to this challenge is to estimate the bound of a given pattern measure, $\mathcal{PM}$, for all the supersets of an itemset, X. However, in most previous studies, the authors estimated the bounds for their proposed pattern measures individually and a generic and unified framework that is applicable to any pattern measure has not been proposed. To this end, we revisit the problem of bound estimation and propose a general framework for it by summarizing the commonality among the estimation methods for different pattern measures. The basic idea is to maximize (or minimize) the measures by assigning any item labels to the items in the original supporting transactions. To achieve a balance between bound tightness and computational efficiency, we also propose techniques for addressing this tradeoff issue in order to improve the overall performance. As a case study, we applied this framework to two typical pattern measures: *utility* and *occupancy*. Additionally, we describe the application of our proposed techniques to other measures. The results of our extensive experimental evaluation on real and large synthetic datasets demonstrate the effectiveness of our proposed techniques.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

The mining of frequent patterns in large databases plays an essential role in many important data mining tasks, such as association rule-based classification and clustering. To improve the effectiveness and efficiency of mining, constrained pattern mining [4,11,14,16,21,22,28,30,36,38,41,45,49–51] is attracting increasing research interest, with the focus being on pushing the properties of the pattern measures deeply into the mining process to achieve better efficiency. To prune the search space, a usual approach is to estimate the bound on a given pattern measure $\mathcal{PM}$ for all supersets of an itemset X. If

* Corresponding author. Tel.: 010 62600537.
** Corresponding author.
E-mail addresses: zl@ahu.edu.cn (L. Zhang), luop@ict.ac.cn (P. Luo), cheneh@ustc.edu.cn (E. Chen), minwang@google.com (M. Wang).

http://dx.doi.org/10.1016/j.ins.2015.12.036
0020-0255/© 2016 Elsevier Inc. All rights reserved.

this bound violates the constraint, then all supersets of X should be pruned. Thus, the bound estimation of pattern measures becomes one of the key tasks for constrained pattern mining.

Previous studies on constrained pattern mining frequently followed the research paradigm, where each new pattern measure is proposed individually (motivated by real-world applications) with its corresponding bound estimation method. For example, the measures of pattern utility [27,46], pattern occupancy [39], block pattern measures [18], tiling [19], and weight confidence [47] are all introduced individually. When researchers address a new pattern measure, they need to invest effort in designing efficient solutions for bound estimation. Thus, we argue, it is urgent to analyze the commonalities among these estimation methods and determine whether it is feasible to find a general framework that is applicable to the bound estimation of any pattern measure. With the increasing interest in constrained pattern mining, this general framework is extremely necessary to guide the fast development of the new constrained pattern mining algorithms with the new pattern measures.

To this end, we propose a unified framework to estimate the bound of any pattern measure. Specifically, given an itemset X and a pattern measure $\mathcal{PM}$, the proposed framework can guide the efficient estimation of the bound of $\mathcal{PM}$ for all the frequent patterns containing X. The key point of this framework is that, to achieve efficient estimation, we consider only the weights of item occurrences and ignore the item labels in the supporting transactions. Then, the estimation of the pattern measure bound is equivalent to the process in which we assign the item labels to the item occurrences so that the pattern measure in the resultant transactions is maximized or minimized.

In this study, we focus on the pattern measures that have two inputs, namely, the supporting transactions of a pattern and the item weights for all item occurrences. To the best of our knowledge, this general form covers all pattern measures proposed so far. Here, we consider the constraint with the most commonly used form $\mathcal{PM}(X)\theta\beta$, where $\theta \in \{\geq, \leq\}$ and $\beta \in \mathcal{R}^+$. Note that the other forms of constraints are outside the scope of this study. We first consider the case of $\theta =$ "$\geq$" for the constraint $\mathcal{PM}(X)\theta\beta$. Thus, we need only to estimate the *upper bound* of the pattern measures for pruning the search space. Then, we show that a similar idea can also be applied to estimate the *lower bound* of the pattern measures for the constraints with $\theta =$ "$\leq$." The contributions of our work can be summarized as follows.

- We argue that the bound estimation methods used in previous constrained pattern mining methods are rather dispersive (each new pattern measure is proposed individually with its corresponding bound estimation method) and lack a general framework that is applicable to the bound estimation of any pattern measure.
- We formulate the bound estimation problem *without the item labels*, and then, we propose a generic framework to efficiently address this problem for any pattern measure. We theoretically prove that the proposed general framework can give the tightest bound when the item labels in the supporting transactions are ignored. Additionally, we also propose techniques that achieve the balance between bound tightness and computational efficiency that is needed to improve the overall performance.
- To show the manner in which our proposed framework can guide the development of efficient bound estimation, we apply the proposed framework to two typical pattern measures, namely, *utility* [46] and *occupancy* [39], as a case study. In addition, we extend the traditional SQL-like pattern measures [10,34], such as *min, max, avg*, and *var*, to the corresponding *relative pattern measures* (see Section 2.3), and we also explain the application of the proposed techniques to these pattern measures.
- We systematically evaluate the extent to which the proposed techniques of bound estimation can improve the overall efficiency for the task of *mining top-k constrained frequent closed patterns*. The experiments were conducted on real and large synthetic datasets under different data characteristics and different problem settings.

The rest of the paper is organized as follows. In Section 2, we first introduce the background of pattern mining, in which the properties of all the commonly used pattern measures are summarized and the problem of *top-k constrained frequent closed pattern mining* is proposed. Then, we give the formal problem formulation of bound estimation of pattern measures in Section 3 and present its general solution in Section 4. We give two typical examples of pattern measures, i.e., utility and occupancy, and their bound estimation methods in Sections 5 and 6, respectively. We report our empirical studies to show the efficiency improvements achieved by the proposed techniques in Section 7. We present related work in Section 8 and conclude the paper in Section 9.

2. Background of pattern mining

In this section, we first give some notions and notations of pattern mining. Then, we give a summary of the constraints on the commonly used pattern measures proposed thus far. In addition, we extend the traditional SQL-like pattern measures, such as *min, max, avg*, and *var* to the corresponding relative forms. Finally, we describe the task of *top-k constrained frequent closed pattern mining* to educe the key problem of the bound estimation of pattern measures.

2.1. Notions and notations for pattern mining

A transaction database is a set of transactions, where each transaction is a set of items. Let $\mathcal{I}$ be the complete set of distinct items and $\mathcal{T}$ be the complete set of transactions. Any non-empty set of items is called an *itemset*, while any set

Download English Version:

https://daneshyari.com/en/article/6857385

Download Persian Version:

https://daneshyari.com/article/6857385

Daneshyari.com