Engineering Applications of Artificial Intelligence 69 (2018) 112–126

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Efficient mining of high utility itemsets with multiple minimum utility thresholds

Srikumar Krishnamoorthy

Indian Institute of Management, Ahmedabad, India

ARTICLE INFO

Keywords:
High utility mining
Multiple utility thresholds
Frequent itemset mining
Data mining

ABSTRACT

Mining high utility itemsets is considered to be one of the important and challenging problems in the data mining literature. The problem offers greater flexibility to a decision maker in using item utilities such as profits and margins to mine interesting and actionable patterns from databases. Most of the current works in the literature, however, apply a single minimum utility threshold value and fail to consider disparities in item characteristics. This paper proposes an efficient method (MHUI) to mine high utility itemsets with multiple minimum utility threshold values. The presented method generates high utility itemsets in a single phase without an expensive intermediate candidate generation process. It introduces the concept of suffix minimum utility and presents generalized pruning strategies for efficiently mining high utility itemsets. The performance of the algorithm is evaluated against the state-of-the-art methods (HUI-MMU-TE and HIMU-EUCP) on eight benchmark datasets. The experimental results show that the proposed method delivers two to three orders of magnitude execution time improvement over the HUI-MMU-TE method. In addition, MHUI delivers one to two orders of magnitude execution time improvement over the HIMU-EUCP method, especially on moderately long and dense benchmark datasets. The memory requirements of the proposed algorithm was also found to be significantly lower.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

High utility itemset (HUI) mining is one of the important problems in the data mining literature. It has numerous scientific and business applications and is one of the actively researched topics in the recent years (Fournier-Viger et al., 2014b; Li et al., 2008; Lin et al., 2016c; Liu et al., 2005; Liu and Qu, 2012; Liu et al., 2012; Yao and Hamilton, 2006; Zida et al., 2017). A high utility itemset (HUI) is an itemset whose utility value is above a user specified utility threshold (Liu et al., 2005). It can be considered as a generalized version of the frequent itemset (FI) (Agrawal and Srikant, 1994a). An itemset mining framework that is based on HUI offers greater flexibility to a decision maker in incorporating her/his notion of item utilities such as margins, profits and so on. The traditional frequent itemset mining framework, on the other hand, is limited to analysis of occurrence counts of items in the database. The use of generalized utility functions in HUI mining framework allows generation of more interesting and actionable patterns for decision making.

High utility itemset mining is a computationally challenging problem (Liu et al., 2005). This is due to the lack of anti-monotonic property that is commonly exploited in traditional frequent itemset mining. Several data structures and pruning strategies have been proposed in the literature (Fournier-Viger et al., 2014b; Krishnamoorthy, 2015; Liu et al., 2005; Liu and Qu, 2012; Liu et al., 2012; Zida et al., 2017) to improve the efficiency of mining HUIs. These methods can be broadly categorized as level-wise candidate generation and test (Li et al., 2008; Liu et al., 2005; Yao and Hamilton, 2006), tree-based or pattern growth based (Ahmed et al., 2011; Tseng et al., 2012), utility list based (Fournier-Viger et al., 2014b; Krishnamoorthy, 2015; Lin et al., 2016a; Liu and Qu, 2012), hyperlink based (Liu et al., 2012) and horizontal database projection based (Zida et al., 2017) methods.

The frequent and high utility itemsets finds application in wide variety of domains such as retail, warehouse or distribution center, electronic commerce, banking, insurance and health care. In the warehouse or distribution context, a manager is often interested in optimizing material movements and storage space. The frequent or high utility itemsets can help a manager effectively analyze the customer ordering or demand patterns and improve the operational efficiency of a warehouse (Chen and Wu, 2005; Chen et al., 2005). In the electronic commerce context, mining frequent or high utility itemsets can help discover interesting patterns of customer click or purchase patterns and generate suitable product recommendations.

We motivate the need for this research with the help of the most common retail market basket analysis scenario. A large retail firm holds

E-mail address: srikumark@iima.ac.in.

https://doi.org/10.1016/j.engappai.2017.12.012
Received 20 July 2017; Received in revised form 9 November 2017; Accepted 22 December 2017
0952-1976/© 2018 Elsevier Ltd. All rights reserved.

several hundred or thousands of items across diverse product categories. These items have distinct characteristics in terms of their purchase frequency (apparel vs bread), price (diamond rings vs sunglasses), margins (jewelry vs kitchen items) and so on. A retail manager is often interested to explore different combinations of items, discover actionable patterns and make key decisions related to pricing, promotions and product placements. While the use of HUIs can be useful to address such problems, it does not account for numerous variations in product characteristics. Moreover, the use of standard HUI mining frameworks require selection and tuning of a single minimum utility threshold value. Choosing too low a minimum utility threshold value (to handle disparate item characteristics) leads to combinatorial explosion of search space and the number of patterns generated can be very large. This imposes additional burden on the user to manually filter and identify actionable patterns. On the other hand, setting too high a minimum utility value can result in omission of several interesting patterns (e.g. rare item problem (Liu et al., 1999)). In essence, the use of basic HUI mining framework with a single minimum utility threshold value is often inadequate to effectively handle the decision making requirements of retail managers. Therefore, there is a need to design HUI mining frameworks that can support multiple (item level) minimum utility threshold values. Recent works in the literature have formalized and discussed the non-trivial nature of this new problem (Gan et al., 2016; Lin et al., 2016c) and highlighted the need for more efficient methods.

The past studies in the literature that addresses the HUI mining problem with multiple minimum utility threshold values include HUI-MMU (Lin et al., 2016c, 2015) and HIMU (Gan et al., 2016). The underlying data structures and mining methods used in these methods are: two-phase level-wise candidate generation (Lin et al., 2016c, 2015) and utility list based tree enumeration (Gan et al., 2016) approach. All of these methods primarily sort the items based on their minimum utility threshold values to mine HUIs with varying threshold levels. The key intuition behind the use of sorting in these approaches is to support several key pruning properties proposed in the basic HUI mining literature. The pruning properties are adapted and specialized by imposing a specific ordering of items and using the concept of least minimum utility.

We argue that sorting items based on minimum utility threshold value is likely to be highly inefficient. This is due to the fact that the utility threshold values are determined by the decision maker and does not reflect the underlying dataset characteristics (e.g. itemset frequency, and utility distributions). It is quite intuitive to understand that small perturbations in preferences (specified through minimum utility threshold values) of a decision maker can dramatically affect the performance of HUI mining. Therefore, there is a need for better methods for mining HUIs in the new framework. The prior works in the literature (Gan et al., 2016; Lin et al., 2016c) have also shown the inapplicability of basic HUI mining properties and the consequent challenges for the new problem. Our primary objective in this paper is explore an alternate approach that is more efficient and scalable for mining HUIs with multiple minimum utility thresholds.

The core idea of the proposed approach is fundamentally different from existing works in the literature. We propose generalized pruning properties to improve the efficiency of mining HUIs with multiple minimum utility threshold values. We also introduce a new concept of Suffix Minimum Utility (SMU) and use it effectively as part of the proposed pruning strategies. We show that the proposed ideas are much more effective and efficient compared to the state-of-the-art methods through rigorous experimental evaluation.

The key contributions of this paper are as follows:

1. We present a new algorithm, named MHUI, for efficiently mining HUIs with multiple (item level) minimum utility thresholds. The presented algorithm utilizes a vertical database representation to efficiently store itemset information and mine HUIs.
2. A new concept of *Suffix Minimum Utility* (SMU) is introduced to efficiently mine HUIs. As the standard pruning properties are inapplicable to the new HUI mining problem, we adapt them using the concept of SMU. The pruning properties used in this paper are generic in nature and are independent of any specific ordering of items.
3. We conduct rigorous experimental evaluation of MHUI against the state-of-the-art methods on eight benchmark sparse and dense datasets to demonstrate its usefulness. It is to be noted that earlier works in the literature have primarily evaluated their methods on highly sparse benchmark datasets. Our experimental results show that the proposed method is two to three orders of magnitude faster than the most recent HUI-MMU-TE (Lin et al., 2016c) method. MHUI algorithm also shows one to two orders of magnitude execution time improvement over the state-of-the-art HIMU-EUCP (Gan et al., 2016) method, especially on moderately long and dense benchmark datasets.

The rest of the paper is organized as follows. The related works in the literature is reviewed in Section 2. The definitions, notations and problem statement are described in Section 3. The MHUI algorithm is presented in Section 4. The different pruning strategies employed for efficient HUI mining are also presented in this section. The experimental evaluation of MHUI against the state-of-the-art methods on eight benchmark datasets is made in Section 5. Finally, Section 6 presents concluding remarks and directions for further research.

2. Related literature

In this section, we review the literature on both frequent and high utility itemset mining with single and multiple threshold (support or utility) levels.

2.1. Frequent itemset mining: Single and multiple support thresholds

Association Rule Mining (ARM) (Agrawal and Srikant, 1994a) is one of the extensively studied problems in the data mining literature. Association rules are mined in two steps. In the first step, itemsets that frequently co-occur in transactions are mined. Subsequently, in the second step, rules are discovered from the generated frequent itemsets. The computationally expensive step in rule mining is the frequent itemset mining. Numerous algorithms have been proposed to efficiently mine frequent itemsets. Some of the popular algorithms include Apriori (Agrawal and Srikant, 1994a), FP-Growth (Han et al., 2000) and Eclat (Zaki, 2000).

The standard frequent itemset mining algorithms consider only a single support threshold and hence suffer from *rare item problem* (Liu et al., 1999). In order to address this problem, the use of multiple minimum support thresholds were explored in the literature. Some of the prominent works in the literature that consider multiple support thresholds include: MSApriori (Liu et al., 1999), CFP-Growth (Hu and Chen, 2006), CFP-Growth++ (Kiran and Reddy, 2011) and the more recent FP-ME (Gan et al., 2017b). These approaches primarily extend the basic frequent itemset mining methods such as Apriori & FP-Growth and allow specification of supports at an individual item level. But, these approaches cannot be directly adapted to the HUI mining problem that uses a generalized notion of utilities such as profits, margins, and purchase quantities (Lin et al., 2016c).

2.2. High utility itemset mining: Single utility threshold

HUI mining is one of the most actively researched topics in the last ten years. The problem was first introduced by Liu et al. (2005) in order to address the key limitations of the more common support based framework used in association rule mining. The two-phase algorithm (Liu et al., 2005) followed a level-wise candidate generation and

Download English Version:

https://daneshyari.com/en/article/6854254

Download Persian Version:

https://daneshyari.com/article/6854254

Daneshyari.com