Efficient Algorithm for Mining High Utility Pattern Considering Length Constraints

Kuldeep Singh, IIT (BHU), Varanasi, India
https://orcid.org/0000-0001-5289-362X

ABSTRACT

High utility itemset (HUI) mining is one of the popular and important data mining tasks. Several studies have been carried out on this topic, which often discovers a very large number of itemsets and rules, which reduces not only the efficiency but also the effectiveness of HUI mining. In order to increase the efficiency and discover more interesting HUIs, constraint-based mining plays an important role. To address this issue, the authors propose an algorithm to discover HUIs with length constraints named EHIL (Efficient High utility Itemsets with Length constraints) to decrease the number of HUIs by removing tiny itemsets. EHIL adopts two new upper bound named sub-tree and local utility for pruning and modify them by incorporating length constraints. To reduce the dataset scans, the proposed algorithm uses transaction merging and dataset projection techniques. The execution time improvements ranged from a modest five percent to two orders of magnitude across benchmark datasets. The memory usage is up to twenty-eight times less than state-of-the-art algorithm FHM+.

KEYWORDS

Constraint-Based Mining, Frequent Pattern Mining, High Utility Itemset Mining, Length Constraint, Pattern Mining, Utility Mining

INTRODUCTION

Frequent itemset mining (FIM) (Agrawal & Srikant, 1994; Han, Pei, & Yin, 2000; Zaki, 2000; Uno, Kiyomi, & Arimura, 2004; Singh, Shaky, & Biswas, 2015; Singh, Shaky, & Biswas, 2016b) is one of the fundamental research topics in data mining. Traditional FIM, assumes that item cannot appear more than once in each transaction, and each item has same importance like weight, unit profit, etc. Hiding importance and quantity of an item may also hide some important or relevant information. Hence, FIM not only loses valuable and important information of the itemsets but also generates many irrelevant and unimportant frequent itemsets. To address the issue of quantity, multi-frequency based frequent itemset mining algorithm are proposed Singh, Shaky, and Biswas, 2016a, but only quantity-based mining loose importance of the items. Hence, high utility itemsets (HUIs) (Liu & Qu, 2012; Fournier-Viger, Wu, Zida, & Tseng, 2014; Liu, Wang, & Fung, 2016; Singh, Singh, Kumar, Shaky, & Biswas, 2018) mining emerges as an important topic in data mining. The main goal of HUIs mining is to discover itemsets having a high utility. In HUIs mining, each item can appear more than once in each transaction (e.g., quantity). It also assumed that each item has a utility value (e.g., price or profit).

HUIs mining is essential in several business and scientific applications such as commerce (Shie, Hsiao, Tseng, & Yu, 2011), web-click stream analysis (Chu, Tseng, & Liang, 2008), mobile e-commerce (Li, Huang, Chen, Liu, & Lee, 2008), cross marketing (Yen, & Lee, 2007), user behavior analysis (Shie, Yu, & Tseng, 2013). HUIs are also used to earn the better profit by promoting the

DOI: 10.4018/IJDWM.2019070101

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
sales. However, HUIs mining is more difficult and complex problem compare to FIM mining because HUIs mining does not support downward closure property. Hence, prune the search space in HUIs mining is the difficult task. To tackle this problem, Ying Liu et al. proposed the concept of transaction weighted utility (TWU) to facilitate the performance of the mining task (Liu, Liao, & Choudhary, 2005). However, the TWU is used for overestimation of true or actual utility of an itemset. However, TWU model leads to waste lots of time to compute the utility for itemsets because their level-wise candidate generation. Therefore, algorithms for HUIs mining are generally slower than FIM algorithms.

To overcome the drawbacks of traditional HUIs algorithms, concise HUIs mining algorithms have been introduced. Concise HUIs are more interesting and actionable because they are lossless representation of HUIs. Concise HUIs can be found using closed or length constraints. In length-based concise HUIs mining, the additional length-based thresholds are introduced. This idea of length constraints with minimum utility ($\text{min\_util}$) is applicable in many scenarios. Minimum length ($\text{min\_length}$) and maximum length ($\text{max\_length}$) constraints with $\text{min\_util}$ HUIs mining serves as a promising solution for the limitation of traditional HUIs mining algorithms. Therefore, length-based HUIs mining is essential for many applications. However, developing efficient algorithms for mining such itemsets is not an easy task. It poses four major challenges as discussed below.

First, the itemsets in HUIs mining neither support to monotonic property nor to anti-monotonic property. FIM techniques with length constraints rely on anti-monotonicity property. Hence, FIM techniques cannot be directly applied to HUIs mining with length constraints.

Second, HUIs mining algorithms consume more time and memory compare to FIM algorithms. Hence, a compact dataset storage structure and an efficient pruning strategy are required.

Third, how to incorporate the concept of length constraints with the HUIs mining. And how to overestimate the upper bounds with length constraints.

Lastly, how to incorporate the length constraints ($\text{min\_length}$ and $\text{max\_length}$) with $\text{min\_util}$ threshold.

To address these challenges, an efficient utility mining algorithm is needed. The key contributions of this paper are as follows:

- The proposed algorithm relies on two dataset compaction techniques to speed up the execution. The dataset projection and transaction merging techniques reduce the dataset size;
- Two prune strategies are utilized to prune a large number of unpromising candidates. These pruning strategies are calculated using depth first search;
- To speed up the utility counting process, an efficient array-based utility counting technique is employed to compute the upper-bounds. The array-based utility counting technique calculates the utility of itemsets without scanning the original dataset;
- We demonstrate the usefulness of the proposed algorithm through rigorous experiments. The results show that the EHIL outperform the state-of-the-art FHM+ algorithm in term of execution time and memory usages.

The rest of the paper is organized as follows. Section 2 describes the related work on HUIs mining. Section 3 defines the problem and describes the key definitions and notations used in this paper. Section 4 presents the EHIL algorithm and pruning strategies. Section 5 provides detailed experimental analysis and results. Finally, Section 6 gives concluding remarks.

**RELATED WORK**

In recent years, several algorithms have been introduced to mine HUIs. First time Chan et al. introduced the concept of utility mining in 2003 (Chan, Yang, & Shen, 2003). Later Liu et al. proposed TWU model based two-phase algorithm to mine HUIs (Liu, Liao, & Choudhary, 2005). The two-phase
Coupling Materialized View Selection to Multi Query Optimization: Hyper Graph Approach
www.igi-global.com/article/coupling-materialized-view-selection-to-multi-query-optimization/125651?camid=4v1a

OLEMAR: An Online Environment for Mining Association Rules in Multidimensional Data
www.igi-global.com/chapter/olemar-online-environment-mining-association/7511?camid=4v1a