ABSTRACT
Mining high utility itemsets is one of the most important research issues in data mining owing to its ability to consider nonbinary frequency values of items in transactions and different profit values for each item. Although a number of relevant approaches have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. In this paper, the authors propose an efficient algorithm, namely BAHUI (Bitmap-based Algorithm for High Utility Itemsets), for mining high utility itemsets with bitmap database representation. In BAHUI, bitmap is used vertically and horizontally. On the one hand, BAHUI exploits a divide-and-conquer approach to visit itemset lattice by using bitmap vertically. On the other hand, BAHUI horizontally uses bitmap to calculate the real utilities of candidates. Using bitmap compression scheme, BAHUI reduces the memory usage and makes use of the efficient bitwise operation. Furthermore, BAHUI only records candidate high utility itemsets with maximal length, and inherits the pruning and searching strategies from maximal itemset mining problem. Extensive experimental results show that the BAHUI algorithm is both efficient and scalable.

Keywords: Bitmap, Data Mining, High Utility Itemsets, Maximal Itemsets

INTRODUCTION
Frequent itemset mining (Agrawal & Srikant, 1994; Cho, Pei, & Wang, 2005; Han, Cheng, Xin, & Yan, 2007) is a fundamental and primary research topic in data mining. It started as a phase in the discovery of association rules (Ashrafi, Taniar, & Smith, 2007; Daly & Taniar, 2004; Jiang, Hu, Wei, Song, Han, & Liang, 2011; Taniar, Rahayu, Lee, & Daly, 2008; Tjioe & Taniar, 2005; Vo, Hong, & Le, 2013), but has since been generalized, independent of these, to many other patterns, for example, frequent sequences (Al-Mulla & Aghbari, 2011; Lu, Chen, Adjei, & Keech, 2008), episodes (Tatti & Cule, 2012), movement pattern (Taniar & Goh, 2007), and frequent subtree (Nguyen & Yamamoto, 2012; Xiao, Yao, & Yang, 2005).

One of the most important and popular applications for frequent itemset mining is
market basket analysis. However, the itemsets found without considering their quantity and profit/weight may mask the really important information. In fact, the quantity and profit of items are important factors that cannot be ignored absolutely. Therefore, frequent itemset mining has the inevitable limitations for users who desire to find itemsets with high profit. For example, in retail applications, frequent itemsets identified by the traditional frequent itemset mining algorithms may contribute only a small portion of the overall revenue or profit, as high margin or luxury goods typically do not appear frequently in transactions. A similar problem occurs when data mining is applied within an enterprise to identify the most valuable client segments, or product combinations that contribute most to the company’s bottom line. To address the limitation, utility mining (Chan, Yang, & Shen, 2003; Yu, Shao, Luo, & Zeng, 2009) emerges as an important topic in today’s data mining task.

In utility mining, each item in transactions has its own profit and can occur more than once. The utility of an itemset represents its importance or profitability to users. The utility values of items in a transaction database consist of two parts: one is the profit of distinct items, which is called external utility; while the other is the quantity of items in one transaction, which is called internal utility. The utility of an itemset is defined as the external utility multiplied by the internal utility. An itemset is called a high utility itemset (HUI) if its utility is no smaller than a user specified value of minimum utility.

Nevertheless, HUI mining is not an easy task. The difficulty is that it does not follow the downward closure property (Agrawal & Srikant, 1994) that plays an important role in frequent itemset mining, which means, a high utility itemset may consist of some low utility subsets. In other words, we cannot prune these itemsets with low utility directly, as it may cause a lack of actual high utility itemsets. To facilitate the task of HUI mining, Liu et al. (2005) defined high transaction-weighted utilization itemset, which satisfying downward closure property.

Earlier studies (Li, Yeh, & Chang, 2008; Yao, Hamilton, & Butz, 2004) suffered from the level-wise candidate generation-and-test problem, involving several database scans depending on the length of candidate itemsets. In view of this, some novel tree structures (Ahmed, Tanbeer, Jeong, & Lee, 2009; Tseng, Wu, Shie, & Yu, 2010) have been used for HUI mining. Based on the pattern growth approach (Han, Pei, Yin, & Mao, 2004), these algorithms are usually efficient than candidate generation-and-test methods.

Generally speaking, the discovery of HUIs is partitioned into two phases: mine high transaction-weighted utilization itemsets at first, then identify HUIs by scanning transaction database. The existing approaches suffer from the following two problems. On the one hand, there are always a huge set of potential HUIs generated by existing algorithms. The situation may become worse when the database contains many long transactions or low threshold is set. On the other hand, although show better performance than the level-wise approaches (Liu, Liao, & Choudhary, 2005; Li, Yeh, & Chang, 2008), the tree-based approaches (Lan, Hong, & Tseng, 2012; Tseng, Wu, Shie, & Yu, 2010) still generate a vast number of conditional trees with a corresponding high cost in terms of memory usage.

To address the above-mentioned two issues, in this paper, we propose a novel algorithm named BAHUI (Bitmap-based Algorithm for High Utility Itemsets). BAHUI uses a bitmap representation for searching and counting the itemset lattice, and quickens the phase of candidate HUIs identification based on maximal itemset (Bayardo, 1998; Song, Yang, & Xu, 2008) problem. The major contributions of this work are summarized as follows:

Firstly, BAHUI uses a bitmap representation for compressing the transactional database. Thus, the storage cost is reduced and the candidate enumeration efficiency is improved.

Secondly, BAHUI only stores potential high utility itemsets with maximal length. Thus, the number of candidate HUIs is reduced greatly.

Finally, both synthetic and real datasets are used in experimental evaluations. The experimental results show that BAHUI outperforms
Related Content

Applications of Data Mining in Software Development Life Cycle: A Literature Survey and Classification
Naveen Dahiya, Vishal Bhatnagar, Manjeet Singh and Neeti Sangwan (2014). *Data Mining and Analysis in the Engineering Field* (pp. 67-79).

Artificial Immune Systems: Using the Immune System as Inspiration for Data Mining
[www.igi-global.com/chapter/artificial-immune-systems/7591?camid=4v1a](http://www.igi-global.com/chapter/artificial-immune-systems/7591?camid=4v1a)

Influence of Domain and Model Properties on the Reliability Estimates’ Performance
[www.igi-global.com/article/influence-domain-model-properties-reliability/37405?camid=4v1a](http://www.igi-global.com/article/influence-domain-model-properties-reliability/37405?camid=4v1a)
Classifying Very High-Dimensional Data with Random Forests Built from Small Subspaces
Baoxun Xu, Joshua Zhexue Huang, Graham Williams, Qiang Wang and Yunming Ye (2012). *International Journal of Data Warehousing and Mining* (pp. 44-63). www.igi-global.com/article/classifying-very-high-dimensional-data/65573?camid=4v1a