Rare Correlated High Utility Itemsets Mining: An Experimental Approach

P. Lalitha Kumari, S. G. Sanjeevi and T. V. Madhusudhana Rao

Abstract High utility itemsets are having utility more than user-specified minimum utility. These itemsets provide high profit but do not exhibit correlation between them. High utility itemsets mining generate huge number of itemsets considering only single interesting criteria. Existing algorithm mines correlated high utility itemsets mining extracts itemsets that provide high utility with correlation between them. The limitation of this algorithm is that it does not consider the rarity of itemsets. To overcome this limitation, this proposed algorithm mines rare correlated high utility itemsets. Firstly, it mines correlated high utility itemsets. Secondly, it determines whether the itemsets support is no greater than minsup specified by the user. It can be shown by experimental results that the proposed algorithm reduces considerably runtime and number of candidate itemsets.

Keywords Correlated itemsets · Rare itemsets · Correlated high utility itemsets
Frequent itemsets

1 Introduction

Frequent itemsets mining methods generate a large number of itemsets that may not be necessarily applicable for decision-making applications. Rare itemsets mining generate important itemsets for some special applications such as inventory

P. Lalitha Kumari (✉) · S. G. Sanjeevi
Department of CSE, National Institute of Technology, Warangal 506004, Telangana, India
e-mail: lalitharam.p@gmail.com

S. G. Sanjeevi
e-mail: sgs@nitw.ac.in

T. V. Madhusudhana Rao
Department of CSE, Sri Sivani College of Engineering, Srikakulam 532402,
Andhra Pradesh, India
e-mail: madhu11211@gmail.com

P. K. Pattnaik et al. (eds.), *Progress in Computing, Analytics and Networking*,
Advances in Intelligent Systems and Computing 710,
https://doi.org/10.1007/978-981-10-7871-2_73

systems, biomedical systems, marketing analysis. Rare itemsets mining result in itemsets with less minimum support threshold. High utility itemset mining approaches were solely considering only the utility, but not considering the correlation between them. Some research works have proposed effective algorithms for mining correlated high utility itemsets that generate the high utility itemsets with correlation. Those algorithms still generate more number of itemsets which do not consider the frequency or rarity of itemsets. High utility itemsets with rarity are the itemsets that are having utilities no less than user-given minimum utility and considers itemsets rarity.

Many algorithms [1, 7–10, 12, 13, 22, 23] have been proposed for mining the frequent itemsets in transactional databases. Apriori [1] is a well-known algorithm which works on the basis of candidate generation and test approach for mining frequent itemsets [1, 9, 12, 22, 23]. Apriori algorithm level-wise compares itemsets that support with user-specified minimum threshold. It returns all itemsets which satisfy minimum support threshold. Apriori-based approach algorithms need many scans for mining frequent itemsets. FP-growth algorithm [7, 8] as proposed to mine frequent itemsets efficiently than Apriori that considerably reduces number of scans. This approach needs two scans to find frequent itemsets using FP-tree.

Rare itemsets mining [11, 19, 21] has been experimented in recent years to generate all rare items that have low frequencies. Correlated itemsets mining is proposed in [4–6, 18]. For instance, let us consider medical database which consists the various symptoms as items: high temperature, headache, muscle pain, fatigue. If we consider the symptoms for fever and compare with the symptoms of dengue fever, there are several itemsets that appear to be the same as the symptoms for dengue fever. It is very hard to differentiate normal fever from dengue fever. In this case, identification of rare symptoms is necessary which causes the dengue fever. To determine the rare itemsets, that are more valuable, rare itemsets must consider with different values. This rare itemset must exhibit inherently the correlation with other. Consider two itemsets such as:

A: (high temperature, headache, muscle pain), B: (high temperature, headache, fatigue) are rare itemsets those are extracted from medical database. From these, itemsets (high temperature, headache) are common in more number of transactions and are correlated with each other. The itemsets (high temperature, headache, muscle pain) should be considered as more important than (high temperature, headache, fatigue). These itemsets are having low frequencies. These itemsets are treated as more serious symptoms for determining dengue fever. Itemsets (high temperature, headache, muscle pain, fatigue) can be considered as important rare correlated high utility itemsets. Traditional rare itemsets mining algorithms are thus not suitable to apply to mine useful rare itemsets in these circumstances. This leads to motivate for proposing algorithm for mining rare correlated high utility itemsets.

The main contribution of this work is (a) a new algorithm called RCHUI miner has been proposed for mining rare correlated HUIs. (b) This algorithm efficiently mines all rare correlated itemsets with high utilities in two phases. First, it generates correlated high utility itemsets. Second, it checks whether itemsets are rare itemsets or not. Third, experimental results reveal that the RCHUI miner approach generates less number of itemsets with more interestingness.

The rest of paper is narrated as follows. Background knowledge has been described in Sect. 2. RCHUI miner algorithm along with working principle has been detailed in Sect. 3. Description of the experimental results of the RCHUI miner has been discussed in Sect. 4. Finally, conclusion with future work has been elucidated in Sect. 5.

2 Related Work

Many efficient algorithms to mine high utility itemsets have been proposed. For mining high utility itemset, two-phase algorithm was proposed by Liu et al. [15]. In this, transaction-weighted utility (TWU) has been introduced as upper bound to utility to maintain the downward closure property and to prune search space. Mengchi Liu and Junfeng Qu [16] have proposed an approach to mine high utility itemset without generating candidates. They proposed a novel list called utility-list structures to maintain information about each itemset to mine all high utility itemsets effectively. However, two-phase algorithm follows the generate-and-test approach for generating candidates, for which it consumes heavy computational resources. Moreover, if more number of candidates is generated, time required to determine utilities of them is increased. To overcome limitations mentioned above, IHUP [15] and UP-growth [20] were proposed which are based on FP-growth approach.

UP-growth was proposed by Tseng et al. [20] for mining high utility itemset. In their work, they proposed pattern growth-based technique within two scans of database. By using various strategies, i.e., DLU, DGU, DGN, and DLN candidate itemsets are pruned during mining process efficiently. HUI miner was proposed by Liu et al. [15] with tight overestimated utility strategy for pruning the search space. Many other studies have also been proposed to extend high utility itemsets mining. To mine different concise representations of HUIs, such as maximal HUIs [3], closed HUIs [17], and generators of HUIs, various efficient algorithms have been proposed. HUIM on incremental, dynamic database and on data streams has been proposed by Ahmed et al. in [2]. Itemsets with negative profits are also considered in mining HUIs with negative unit profits using FHN algorithm proposed by Philippe-Fournier et al. FHN discovered HUIs without generating candidates and introduced several strategies to handle items with negative unit profits efficiently. To avoid difficulties in setting a proper utility threshold in [14–17], attempts have been done to mine a set of top k-itemsets with the highest utility.

Mining rare correlated high utility itemsets mining: The problem of rare correlated high utility itemset mining can be stated as to extract all correlated high utility itemsets with rarity. An itemset ‘X’ is a rare correlated high utility itemset if it is a high utility itemset whose bond, bond(X) is no less than a user-specified minimum bond threshold minbond, specified by the user and its support is no greater than user-specified minimum support.

3 Basic Preliminaries

3.1 Utility Mining

Let I = {i_1, i_2, i_3, …, i_n} be a set of items and a database called DB comprised of tables having utilities and transactions. Utility table (Table 2) consists of itemsets and their associated utilities. In the transaction table (Table 1), each transaction 'T' is assigned with individual identifier (Tid) and is a subset of 'I,' where every item has been assigned with count value. The itemset containing 'α' items is called α-itemset. Basic definitions are detailed in [15].

For a given database and minutil, high utility itemset is an itemset if its utility is no less than given user-specified minimal utility threshold denoted as minutil, or the product of a minutil and the total utility of a mined database, if the minutil is expressed as percentage. It must be observed that maintaining downward closure property of HUIs is difficult as it does not hold for high utility itemsets (HUIs).

For instance, consider a transaction database having only one transaction, {m, 1; n, 1} and external utility of m = 1 and external utility of n = 2. And if minutil is 6, then for u({m}) = 5, u({n}) = 10 and u({m, n}) = 15, {n} and {m, n} are high utility itemsets and {m} is not. It indicates that the downward closure property does not valid for high utility itemsets. This shows that high utility itemsets mining is challenging compared to frequent itemset mining.

If the TWU of each item is not satisfied with minutil, then items are deleted from transactions. If TWU of each item ≥ minutil, then transactions are rearranged according to the ascending order of the TWU of every item. If the minutil is set as 30, then all items are having their TWU greater than minutil, i.e., 30 so no item is deleted. If we set minutil is 40, then items 'f' and 'g' are deleted, and then transactions are revised according to the ascending order of their TWUs. The TWU of each item is presented in Table 3. The arrangement of each item according to their TWU is as follows:

Table 1 Transactional database

Tid	Transaction	Count
1	{c, e, a, b, d, f}	{1, 3, 5, 5, 3, 1}
2	{c, e, b, d}	{3, 3, 4, 3}
3	{c, a, d}	{1, 5, 1}
4	{c, e, a, g}	{6, 6, 10, 5}
5	{c, e, b, g}	{1, 1, 2, 2}

Table 2 Utility table

a	b	c	d	e	f	g
1	2	1	2	1	1	1

Table 3 Transaction-weighted utility (TWU) of each item

Item	a	b	c	d	e	f	g
TWU	65	61	96	58	88	30	38

$$f > g > d > b > a > e > c$$

Current HUIM algorithms are having key problem that some large number of itemsets are generated those have weak correlation between them. In this paper, integration of utility and correlation is done with rarity as another constraint. Researchers have already described different correlation measures [4–6, 14] in their researches. Correlated itemsets can be found by using bond measure. Conjunctive support of an itemsets X in database D is denoted as conj|X|, where conj|X| is the number of transactions in conj(X).

The disjunctive support of an itemset X in a database D is denoted as disjsup(X) and defined as $|\{Tc \in D/X \cap Tc \neq \Phi\}|$. The bond of itemset X is defined as bond (X) = conj(X)/disjsup(X). An itemset X is said to be correlated if bond(X) is greater than minbond, for a given user-specified minbond threshold ($0 \geq$ minbond ≥ 1). Anti-monotonic property is maintained for bond measure.

Property 1 *Anti-monotonicity of the bond measure can be defined as follows: Let* S *and* T *be two itemsets such that* $S \subseteq R$. *It follows that* bond(S) $\geq$ bond(R) [5]. *By using the above property, it can be stated that the problem of mining rare correlated high utility itemsets as follows.*

Definition 1 (*Rare correlated high utility itemset mining*) An itemset X is a rare correlated high utility itemset if it is a high utility itemset and its bond(X) is no less than a user-defined minimum bond threshold minbond, and its support should be less than minsup specified by the user.

To mine rare correlated HUIs, the proposed algorithm behaves as follows: firstly, our proposed algorithm extracts all correlated HUIs by using structures called as EUCS. The structure of the EUCS is described in [6]. Fournier-Viger P. et al. developed the approach to mine correlated high utility itemsets using EUCS structure. This structure can be used to determine conjunctive and disjunctive support of an itemset without scan database. This algorithm does not consider the rarity of itemsets. Secondly, proposed algorithm extracts rare correlated itemsets that provide high utility. Algorithm 1 returns all high utility itemsets with correlation-based EUCS structure. It will check whether TWU is more than minutil to maintain downward closure property. Later, correlated high utility itemsets are checked with their support. The itemsets that have support less than minsup are extracted from EUCS structure. Itemset's information is maintained in utility-list structure. This algorithm has been explained in [16].

Proposed RCHUI miner Algorithm

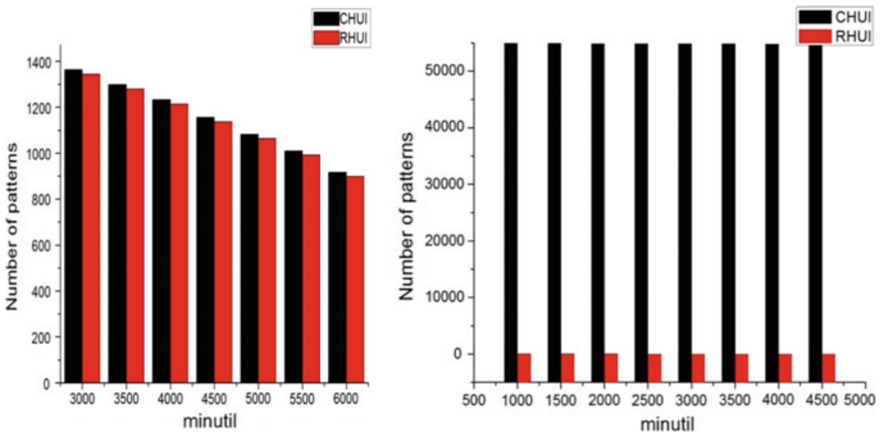
Input: D: a transaction database, minutil: a user-specified utility threshold, minsup: user-specified support threshold, minbond: user-specified bond threshold

Output: the set of rare correlated high utility itemsets

1. Determine TWU of every item by scan database D;
2. Let K* contain the set of items and each item TWU $\geq$ minutil;
3. Define '>' on the ascending order of TWU values on K*;
4. Construct utility-list of every item k $\in$ K* by scan D again and create EUCS;
5. Calculate the conjunctive support and disjunctive support of each itemset from EUCS structure;
6. Determine the bond for each itemset;
7. Check if SUM({k}.utilitylist.iutils) $\geq$ minutil and bond(k) $\geq$ minbond. k.support < minsup;
8. Then output each item k $\in$ K* such that k.support < minsup;
9. Search (Θ, K*, minutil, EUCS);

4 Experimental Results and Analysis

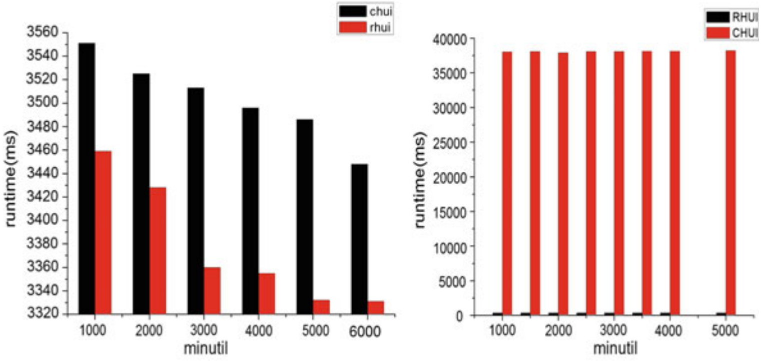
The experimental results of the RCHUI miner have been illustrated in this section. The RCHUI miner algorithm adopts basic framework from HUI miner. We assessed our algorithms by performing experiments on computer with a 2.10 GHz, Intel Core i3 CPU with 4 GB of RAM, and run on Windows 7. Implementation of our proposed algorithm is done in Java. The performance of our proposed algorithm can be evaluated with three datasets. We have considered absolute values for maximum periods. For foodmart dataset, less number of candidates are generated when minutil is set to 3k, 4k, 5k. For foodmart dataset (Fig. 1a), candidate itemsets are generated more for all variations of minutil, i.e., for 3k, 3.5k, 4k, 5k. As a second observation in mushroom dataset (Fig. 1b), we have noticed that the number



(a) Foodmart dataset

(b) Mushroom dataset

Fig. 1 **a**, **b** Results of RCHUI for different datasets



(a) Foodmart dataset

(b) Mushroom dataset

Fig. 2 **a**, **b** Comparison of runtime for different datasets

of RCHUIs is quite less as compared to correlated HUIs. Figure 2 shows the runtime comparison of mushroom dataset and foodmart dataset. Runtime has been considerably reduced when compared to correlated HUIs with RCHUI miner.

5 Conclusion

In this work, a novel approach called RCHUI miner has been proposed for efficient discovery of rare correlated high utility itemsets through bond measure. This algorithm works in two phases; first, we find correlated high utility itemsets satisfying the minutil and minbond measures. In the second phase, rare correlated high utility itemsets have been extracted. The rare correlated high utility itemsets are used to identify rare symptoms of rare diseases in medical databases. This algorithm can be useful to find various applications like fraud detection, intrusion detection. By setting proper value to minsup, we can extract rare correlated high utility itemsets. This algorithm can extend to incremental databases.

References

1. Agrawal R, Srikant R, Fast algorithms for mining association rules. In: 20th international conference on very large databases, (1994) pp. 487–499.
2. Ahmed CF, Tanbeer SK, Jeong B, Lee Y, Efficient tree structures for high utility pattern mining in incremental databases. IEEE Transactions on Knowledge and Data Engineering. (2009). pp. 1708–1721.
3. Bai-En Shie, Philip.S.Yu, Vincent.S.Tseng, Efficient algorithms for mining maximal high utility itemsets from data streams with different models, Expert Systems with Applications (2012), pp. 12947–12960.
4. Barsky, M., Kim, S., Weninger, T., Han, J., Mining Flipping correlations from large datasets with taxonomies. In: Proc. 38th Int. Conf. on Very Large Databases. (2012). pp. 370–381.
5. Ben Younes, N., Hamrouni, T., Ben Yahia, S.: Bridging conjunctive and disjunctive search spaces for mining a new concise and exact representation of correlated patterns. In: Proc. 13th Int. Conf. Discovery Science. (2010). pp. 189–204.
6. Fournier-Viger P., Lin J.CW., Dinh T., Le H.B. Mining Correlated High-Utility Itemsets Using the Bond Measure. In: Martínez-Álvarez F., Troncoso A., Quintián H., Corchado E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2016. Springer. (2016).
7. Grahne G, Zhu J, Fast algorithms for frequent itemset mining using FP-Trees. IEEE Transactions on Knowledge and Data Engineering. (2005). pp. 1347–1362.
8. Han J, Pei J, Yin Y, Mining frequent itemsets without candidate generation. In: Proc. of the 2000 ACM SIGMOD int'l conf. on management of data. (2000). pp. 1–12.
9. Hu Y, Chen Y, Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. Decision Support Systems. (2006). pp. 1–24.
10. Huynh-Thi-Le Q, Le T, Vo B, Le HBAn efficient and effective algorithm for mining top-rank-k frequent itemsets. Expert Systems Applications. (2015). pp. 156–164.
11. Kiran RU, Reddy PK, An improved multiple minimum support based approach to mine rare association rules. CIDM 2009. (2009). pp. 340–347.
12. Lee G, Yun U, Ryu K, Sliding window based weighted maximal frequent itemsets mining over data streams. Expert Systems Applications. (2014). pp. 694–708.
13. Lee G, Yun U, Ryang H, An Uncertainty-based Approach: Frequent Itemset Mining from Uncertain Data with Different Item Importance. Knowledge-Based Systems. (2015). pp. 239–256.
14. Lin, J. C.-W., Gan, W., Fournier-Viger, P., Hong, T.-P, Mining Discriminative High Utility Patterns. Proc. 8th Asian Conference on Intelligent Information and Database Systems. Springer. (2016).
15. Liu Y, Liao W, Choudhary A, A two-phase algorithm for fast discovery of high utility itemsets. Advanced Knowledge Discovery in Data Mining. (2005). pp. 689–695.
16. Mengchi Liu, Junfeng Qu, Mining High Utility Itemsets without Candidate Generation, Proceedings of the 21st ACM international conference on Information and knowledge management, (2012), pp. 55–64.

17. Sahoo, J., Das, A.K. & Goswami, A. An efficient fast algorithm for discovering closed+ high utility itemsets, Applied Intelligence (2016), pp. 44–74.
18. Soulet, A., Raissi, C., Plantevit, M., Cremilleux, B.: Mining dominant patterns in the sky. In: Proc. 11th IEEE International Conference on Data Mining. (2011). pp. 655–664.
19. Tempaiboolkul J, Mining rare association rules in a distributed environment using multiple minimum supports. ICIS 2013. (2013). pp. 295–299.
20. Tseng VS, Wu CW, Shie BE, Yu PS, UP-Growth: an efficient algorithm for high utility itemset mining. In: Proc. of the 16th ACM SIGKDD int'l conf. on knowledge discovery and data mining (KDD 2010). (2010). pp. 253–262.
21. Weng CH, Mining fuzzy specific rare itemsets for education data. Knowledge-Based Systems. (2011). pp. 697–708.
22. Xu T, Dong X, Mining frequent itemsets with multiple minimum supports using basic Apriori. ICNC 2013. (2013). pp. 957–961.
23. Yun U, Yoon E, An efficient approach for mining weighted approximate closed frequent patterns considering noise constraints, International Journal of Uncertainty Fuzziness Knowledge Based Systems. (2014). pp. 879–912.