A Method Mining Correlation High Utility Itemsets with Negative Profits

Cao Tung Anh[1], Ngo Quoc Huy[2], Pham Ngoc Diem[3]

[1]Ho Chi Minh City University of Technology
[2]Van Lang University
[3]The Ho Chi Minh City Power Corporation

Correspondence: Cao Tung Anh, ct.anh@hutech.edu.vn
Communication: received 16 july 2023, revised 26 August 2023, accepted 7 September 2023
DOI: 10.32913/mic-ict-research.v2023.n2.1228

Abstract: **High-utility itemset (HUI) mining is an important task in the data mining process. Recently, many algorithms have been proposed to mine sets of HUIs. Most algorithms only work with data sets with positive return values. However, in the real world, the set of business items often includes both positive (+) and negative (-) profit values. To solve this problem, an algorithm is needed to find HUIs with negative returns in the database.**

The issue is how to mine efficiently correlation high utility itemsets with negative profits. Given the transactional database D and the profit table of products, how to find the best correlation high utility itemsets on the database with items having negative profits. To address these issues, the paper proposes a COHUIs_CoHUN method to mine correlation high utility itemsets with negative profits.

Keywords: *Data mining, negative profit, high utility itemsets, correlation itemsets.*

I. OVERVIEW

In today's business, due to the increasingly fierce competition among companies, promotional campaigns with many attractive offers for consumers are becoming more common, with the goal of stimulating purchases. As a result, some promotional products that are attached to regular products are sold at a loss, meaning they create products with negative profit margins. Additionally, companies may also lower the selling price of some products below the purchase price in order to recover capital, thus resulting in negative profit units. High utility itemset mining algorithms currently do not take into account the negative profit margins of these types of products.

Many high utility itemset mining (HUI) algorithms have been developed in recent years. In 2015, the HUP-Miner algorithm [3] was proposed by author S. Krishnamoorthy, which is an improved version of the HUI-Miner algorithm [1] with two new pruning strategies (PU-Prune and LA-Prune) for mining high utility itemsets. In 2016, Lin and colleagues proposed the FHN algorithm [4]; in 2018, KulDeep Singh and colleagues proposed the EHIN algorithm [5] for efficient HUI mining on databases with negative profits. Although these algorithms are effective in mining HUIs, the number of HUIs found often contains many weakly correlated items that do not have practical significance. To address this issue, in 2018, W. Gan and colleagues proposed the CoHUIM algorithm [6] which considers both the correlation and profit relationship between products in transactions. And in 2020, the CoHUI-Miner algorithm [7] was proposed by Bay Vo and colleagues, which uses database projection to reduce the size and introduces a new concept, called the Expected Transaction Utility Prefix, to eliminate pattern sets that do not meet the minimum threshold in the mining process. The algorithm achieved better performance than the CoHUIM [6] algorithm in terms of both runtime and memory usage. In 2019, author W. Gan and colleagues proposed the CoUPM algorithm [8], an improvement upon the CoHUIMiner [6] algorithm that utilizes the Ulility-List storage structure to enhance the performance of mining correlated high utility itemsets.

Based on the aforementioned studies, in this paper, we propose the COHUIs_CoHUN algorithm for mining related high utility itemsets with negative profits. Experimental results show that the proposed algorithm effectively eliminates many redundant HUI sets and outperforms EHIN [5] algorithm in terms of performance.

The rest of this paper is organized as follows: Section 2 presents the theoretical background and some current methods to address the problem of mining related high utility itemsets with negative profits. The experimental results and discussion are presented in Section 3, where we discuss how the proposed algorithm is implemented. Section 4 provides conclusions on the achieved algorithm and future research directions.

II. METHOD

Among the EHIN-based search algorithms [5], the authors searched for high-utility itemsets for patterns with positive profits first, followed by those with negative profits. This has been shown to reduce the search space by using two separate Primary and Secondary sets of patterns, where the Primary ones are considered more important and the Secondary ones are considered optional. This ensures that a pattern does not have to intersect with all other patterns in order to find the high-utility itemsets, but only needs to intersect with patterns in both Primary and Secondary sets until there are no more itemsets that satisfy the min-util threshold (since the Primary and Secondary sets are sorted by utility). Therefore, improving these search procedures is unnecessary. We propose to use these procedures when performing CoHUN to mine non-redundant high-utility patterns with negative profits.

In the proposed CoHUIs_CoHUN algorithm, we will use an additional minimum correlation threshold minCor (provided by the user). The resulting CoHUIs set (Step 13) will contain all correlation high-utility itemsets with negative profits. Adding the minCor correlation threshold in the EHIN algorithm [5] aims to increase the efficiency of efficiently mining correlated high-utility itemsets on a database of transaction data with redundant negative profits. For the total of unchanging transactions, the resulting correlated high-utility itemsets on the negative-profit database have eliminated redundant patterns. Therefore, the proposed CoHUIs_CoHUN algorithm will have faster computation time than the EHIN algorithm [5], and we have an algorithm capable of mining correlated high-utility itemsets on a negative-profit database.

Algorithm 1: COHUIs_CoHUN

Input D: set of transactions, minUtil: utility threshold, minCor: correlation threshold.

Output CoHUIs: set of non-redundant high-utility itemsets.

Step 1: initialize set $\propto$ to empty

Step 2: create set ρ, which contains elements with positive values in D

Step 3: create set η, which contains elements with negative values in D

Step 4: traverse D, calculate RLU($\propto$,z) $\forall$ z $\in$ ρ and calculate TU

Step 5: find set Secondary(α) = {z|z $\in$ ρ $\wedge$ RLU($\propto$,z) $\geq$ min_util x TU}

Step 6: sort the elements of Secondary($\propto$) in increasing order based on their RTWU($\propto$) values

Step 7: traverse D, remove items z $\in$ Secondary($\propto$) and remove empty transactions T

Step 8: sort the elements in transaction T$\in$D in the order of Secondary($\propto$) and then the negative values

Step 9: assign offsets to each transaction T$\in$D

Step 10: traverse D, calculate RSU($\propto$,z)$\forall$z $\in$ Secondary($\propto$) and Sup(X)$\forall$z $\in$ Secondary($\propto$)

Step 11: find set Primary(α) = {z|z $\in$ Secondary(α)$\wedge$ RSU(α,z)$\geq$ min_util x TU}

Step 12: Search-P(η,α,D,Primary(α),Secondary(α),min_util,min_co

Step 13: return CoHUIs

Algorithm 2: Procedure Search-P

Input η: set of elements with negative values in D, α: set of elements, α – D: set of transaction dataset, Primary(α): set of Primary for set α, Secondary(α): set of Secondary for set α, min_util is the utility threshold, min_cor: the correlation threshold between elements

Output Set of CoHUIs for set α with positive element

Traverse each element z $\in$ Primary(α)

Step 1: Create set β = α $\cup$ {z}

Step 2: Traverse α – D, calculate U(β), calculate Sup(β), calculate Kulc(β) and create set β - D

Step 3: if Kulc(β) $\geq$ min_cor and U(β) $\geq$ min_util x TU then output β

Step 4: if Kulc(β) > min-cor and U(β) > min_util x TU then Search-N(η, β, β - D, min_util, min_cor)

Step 5: Traverse β - D, calculate RSU(β,z) and RLU(β,z) for element z $\in$ Secondary(α)

Step 6: Primary(α) = z | z $\in$ Secondary(α) | RSU(β,z) $\geq$ min_util x TU

Step 7: Secondary(α) = z | z $\in$ Secondary(α) || RLU(β,z) $\geq$ min_util x TU

Step 8: Search-P(η, β, β - D, Primary(β), Secondary (β), min_util, min_cor)

Algorithm 3: Procedure Search-N

Input η: set of elements with negative values in D, α: set of elements, α – D: set of transaction dataset, min_util is the utility threshold, min_cor: the correlation threshold between elements

Output Set of CoHUIs for set α with negative element

Traverse each element z $\in$ η

Step 1: β = α $\cup$ {z}

Step 2: Traverse α – D, calculate U(β), calculate Sup(β), calculate Kulc(β) and create set β - D

Step 3: if Kulc(β) ≥ min_cor and U(β) ≥ min_util x TU then output β

Step 4: Traverse β - D, calculate RSU(β,z) for element z ∈ η

Step 5: Primary(β) = {z | z ∈ η | RSU(β,z) ≥ min_util x TU}

Step 6: Search-N(Primary(β), β, β - D, min_util, min_cor)

Algorithm illustration: We have a transaction database table with negative profit as follows:

TABLE I
TRANSACTION DATA

T_{id}	a	b	c	d	e
T_1	2	2	-	1	3
T_2	-	1	5	-	1
T_3	-	2	1	3	2
T_4	-	-	2	1	3
T_5	2	-	-	-	-
T_6	2	1	4	2	1
T_7	-	3	2	-	2

TABLE II
PROFIT VALUE TABLE

I	a	b	c	d	e
$EU(X_i)$	2	-3	1	4	1

After calculating the profits of items in transactions and sequentially calculating U(X) according to formula [8] , $TU(T_j)$ according to formula [8], $TWU(T_j)$ (according to formula [8]), we obtain the following tables:

TABLE III
PROFIT BY TRANSACTION

$U(X_i, T_j)$	a	b	c	d	e
T_1	4	-6	-	4	3
T_2	-	-3	5	-	1
T_3	-	-6	1	12	2
T_4	-	-	2	4	3
T_5	4	-	-	-	-
T_6	4	-3	4	8	1
T_7	-	-9	2	-	2

Perform CoHUIs_CoHUN algorithm with the following settings: min_util = 0.2, min_cor = 0.2.

Step 1: Set α = 0

Step 2: Set β as the set of items with U > 0

Step 3: Set η as the set of items with U < 0

Step 4: Calculate RLU(X), T_u

Step 5: Set Secondary = a,c,d,e Secondary = RLU(X) > min_util x T_u

TABLE IV
BENEFITS OF ITEMS IN TRANSACTIONS

	a	b	c	d	e	$TU(T_j)$
T_1	4	-6	-	4	3	5
T_2	-	-3	5	-	1	3
T_3	-	-6	1	12	2	9
T_4	-	-	2	4	3	9
T_5	4	-	-	-	-	4
T_6	4	-3	4	8	1	14
T_7	-	-9	2	-	2	-5
U(X)	12	-27	14	28	12	
TWU(X)	23	26	30	37	35	151

Step 6: Sort Secondary = a < c < d < e

Step 7: Delete items in T that are not in Secondary, set T to empty

Step 8: Merge T's, U(X) = U(Tx_1) + U(Tx_2), TU(T_{new}) = U(Tx_1) + U(Tx_2)

Step 9: Attach Offset to T

Step 10: Traverse D, calculate RSU(X), Sup(X)

Step 11: Set Primary = a,c,d,e

Step 12: Call Search_P(η, α, Primary, Secondary, min_uti, min_cor)

Step 13: Return CoHUIs

TABLE V
TABLE AFTER EXECUTING STEPS 4-10

	a	c	d	e	RTU	b(η)
T_1	4	-	4	3	11	(6)
T_2	-	7	-	3	10	(12)
T_3	-	1	12	2	15	(6)
T_4	-	2	4	3	9	-
T_5	4	-	-	-	4	-
T_6	4	4	8	1	17	(3)
Sup	3	4	4	5		4
RSU	32	47	37	12		

After executing Step 12 with the Search_P procedure, the result of the returned CoHuis set at Step 13 is as follows: **CoHUIs = ({a, c, d}, {a, c, d, e}, {a, c, d, e, b}, {a, d}, {a, d, e}, {a, d, e, b}, {c}, {c, d}, {c, d, b}, {c, d, e}, {c, d, e, b}, {c, e}, {d}, {d, e}, {d, e, b})**

The proposed algorithm has been more effective in exploiting high-utility itemsets in a weighted database with no redundancy by incorporating the user-defined correlation threshold minCor. For example, in the 2018 paper EHIN[5] algorithm discovered 20 CoHUIs, but when applying the proposed algorithm, the number of high-utility itemsets was reduced to 15 sets (eliminating 5 redundant high-utility sets).

III. EXPERIMENTAL RESULTS

The CoHUIsCoHUN algorithm was implemented in Java and run on a Dell Vostro 3500 computer with an Intel

Core i7-1165G7 @ 2.80GHz processor, 16GB RAM, on the Windows 10 operating system. Test databases were downloaded from the SPMF library and consisted of transactional databases with negative profits including Chess, Mushroom, Retail and Accidents. The experimental results of the CoHUN algorithm were compared with the latest algorithm that also mines CoHUIs, which is EHIN [5]. The experimental results were evaluated based on the high-utility itemsets correlation to negative profits.

TABLE VI
EXPERIMENTAL DATA

Dataset	Trans	Items (I)	Avg Len
Chess	3.196	75	(37A)
Mushroom	8.124	119	23
Retail	88.162	16.470	10.3
Accidents	340.183	468	33.8

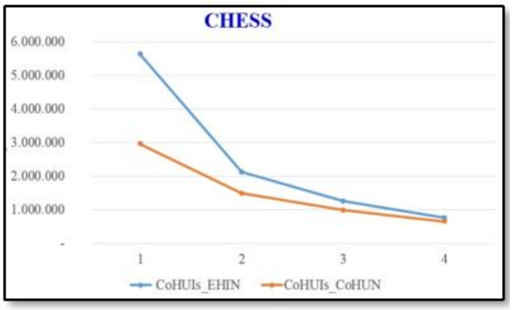

Figure 1. Results on the Chess dataset.

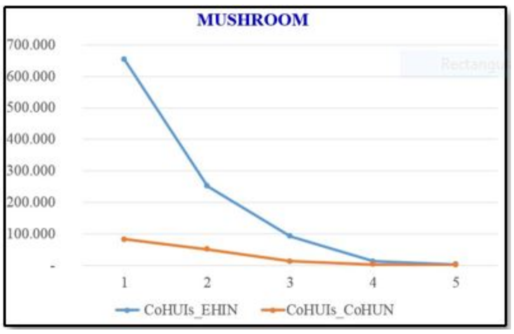

Figure 2. Results on the Mushroom dataset.

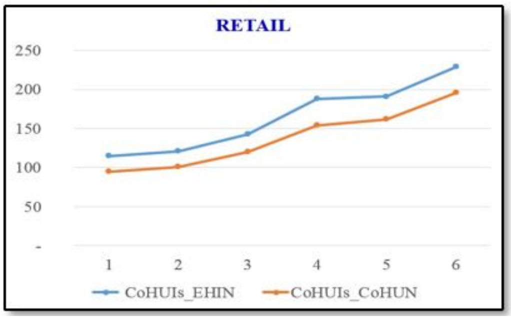

Figure 3. Results on the Retail dataset.

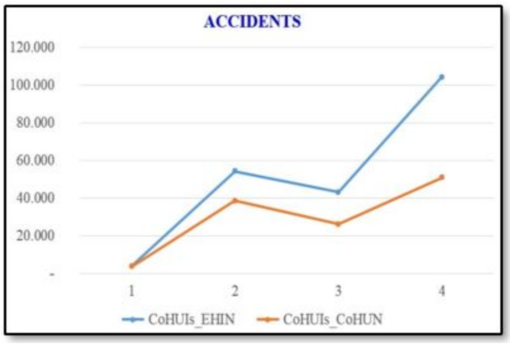

Figure 4. Results on the Accidents dataset.

Figures 1, 2, 3, and 4 show the experimental results comparing the proposed CoHUN algorithm and the EHIN [5] algorithm based on the number of discovered high-utility itemsets (HUIs) at different minimum correlation thresholds minCor. The experimental results demonstrate that the CoHUN algorithm is more effective in discovering HUIs than the EHIN [5] algorithm in all databases, including Chess, Mushroom, Retail, and Accidents. The experimental results demonstrate that the pruning strategies, correlation measures, and data structures used in the proposed CoHUIs_CoHUN algorithm are appropriate for mining correlation CoHUIs on databases with negative profits.

IV. CONCLUSION ANF FUTURE RESEARCH DIRECTIONS

In this paper, we proposed the CoHUIs_CoHUN algorithm for mining correlation high-utility itemsets (CoHUIs) with negative profits. The algorithm employs the PNU-List structure to separate negative and positive profit values for full CoHUIs mining on databases with negative profits. Additionally, the algorithm applies several efficient pruning strategies to reduce search space, including U-Prune, KulcPrune, and LA-Prune. The experimental results demonstrate that the proposed algorithm is more effective in discovering CoHUIs than the EHIN [5] algorithm on all tested databases with negative profits. The memory usage of the CoHUIs_CoHUN algorithm is also better than that of the EHIN [5] algorithm in utilizing the minCor threshold.

The next proposed development direction is to improve the storage data structure, correlation threshold, and pruning strategy to increase the performance of the algorithm on databases with negative profits, as well as extending the CoHUIs mining to other types of databases such as dynamic databases and growing databases.

REFERENCES

[1] M. Liu, J. Qu, (2012). "Mining high utility itemsets without candidate generation (HUI)". *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 55–64.

[2] P. Fournier-Viger, C. W. Wu, S. Zida, V. S. Tseng, (2014). "Faster highutility itemset mining using estimated utility cooccurrence pruning (FHM)", *Proceedings of the International Symposium on Methodologies for Intelligent Systems, Springer*, 83–92.

[3] S. Krishnamoorthy, (2015), "Pruning strategies for mining high utility itemsets (HUP-Miner)", *Expert Systems with Applications* 42(5), 2371-2381.

[4] J. C. W. Lin, P. Fournier-Viger, and W. Gan, (2016)," An efficient algorithm for mining high-utility itemsets with negative unit profits (FHN)", *Knowledge-Based Systems*, 283-298.

[5] KulDeep Singh, Abhimanyu Singh, Harish Kumar Shakya, (2018), "Mining of high-utility itemsets with negative utility (EHIN)", *Wiley, Expert Systems*

[6] W. Gan, J. C. W. Lin, P. Fournier-Viger, H. C. Chao, H. Fujita, (2018), "Extracting non-redundant correlated purchase behaviors by utility measure (CoHUI-Miner)", *Knowledge-Based Systems* 143, 30-41.

[7] B. Vo, L. V. Nguyen, V. V. Vu, M. T. H. Lam, T. T. M. Duong, L. T. Manh, et al., (2020), "Mining Correlated High Utility Itemsets in One Phase (CoHUI-Miner)", vol. 8. *IEEE Access*, 90465-90477.

[8] Ayoub, F. (1982), "Probabilistic completeness of substitution-permutation encryption networks", *IEE Proceedings E (Computers and Digital Techniques)*, 129(5), 195-199.

Cao Tung Anh is currently a Ph.D holder and the head of the Information Systems Department at the Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Vietnam. His research interests lie in the fields of information systems, data mining, distributed databases, coding theory and information security.
Email: ct.anh@hutech.edu.vn

Ngo Quoc Huy is currently a lecturer in the Faculty of Information Technology at Van Lang University (VLU), Vietnam. His research interests lie in the fields of data mining, distributed databases, and network security.
Email: huy.nq@vlu.edu.vn

Pham Ngoc Diem is currently studying a Master's degree at Ho Chi Minh City University of Technology (HUTECH), Vietnam. Her research interests lie in the fields of data mining, distributed databases
Email: diempn1986@gmail.com