

EHUI: A Novel Algorithm for Mining High Utility Itemsets

Prashant V. Barhate¹, Prof. S. R. Chaudhari², Prof. P. C. Gill³

¹ *Research Scholar, MIT, Aurangabad, Maharashtra, India.*

^{2,3} *Department of Computer Science & Engineering, MIT, Aurangabad, Maharashtra, India.*

Abstract-High utility itemsets refer to the sets of items with high utility like profit in a database, and efficient mining of high utility itemsets plays an important role in many real life applications and is an important research issue in data mining area. In recent years, the problems of high utility pattern mining become one of the most important research areas in data mining. The existing high utility mining algorithm generates large number of candidate itemsets, which takes much time to find utility value of all candidate itemsets.

In this paper we are implementing a data structure that stores the utility related to the item and using this data structure we are reducing time and space complexity of UP Growth and UP Growth+ Algorithms. Various Standard and synthetic datasets are used with Educational real data set. An algorithm is proposed to find set of high utility itemset which avoids the candidate itemsets generation.

Keywords- Utility, Utility Information Record, Effective High Utility Itemset Mining.

I. INTRODUCTION

Rapid development in database techniques facilitates storage and usage of data from large database and also to mine the same. How to obtain valuable information from database is a more crucial task today which results in a rise of research topics [1].

Mining frequent itemset [2] from the database DB is to find out set of itemset that occurs frequently. The frequency of itemset is the support count related to that itemset i.e. number of transactions containing that itemset. If the support of the itemset exceeds the *minimum support threshold* value then itemset is frequent.

Mining frequent itemset on takes presence and absence of itemset into account, other relative information related to the item is not considered. This results in the research area of finding out high utility itemset from database. Utility is one of the important features of itemset in transaction that specifies a utility/profit of itemset with frequency.

Recently, a number of high utility itemset mining algorithms [3] have been proposed. Most of the algorithms adopt a similar framework: firstly, generate candidate high utility itemsets from a database secondly, compute the exact utilities of the candidates by scanning the database to identify high utility itemsets. However, the algorithms often generate a very large number of candidate itemsets and thus are confronted with two problems:

- (1) Excessive memory requirement for storing candidate itemsets.
- (2) A large amount of running time for generating candidates and computing their exact utilities.

TABLE I
PROFIT TABLE

Item	A	B	C	D	E	F	G	H
Profit	5	2	1	2	3	5	1	1

TABLE II
DATABASE EXAMPLE

TID	Transaction	Quantity	TU
T ₁	{A,C,D, ₂ }	{1,10,1}	17
T ₂	{A,C,E,G}	{2,6,2,5}	27
T ₃	{A,B,D,E,F}	{2,2,6,2,1}	37
T ₄	{B,C,D,E}	{4,13,3,1}	30
T ₅	{B,C,E,G}	{2,4,1,2}	13
T ₆	{A,B,C,D,H}	{1,1,1,1,2}	12

When the number of candidates is so large that they cannot be stored in memory, the algorithms will fail or their performance will be degraded due to thrashing. To solve these problems, we propose in this paper an algorithm for high utility itemset mining.

The contributions of the paper are as follows:

1. A novel structure, called *utility information record*, is proposed. Utility information record stores not only the utility information about an itemset but also the heuristic information about whether the itemset should be pruned or not.
2. An efficient algorithm, called Efficient High Utility Itemset Mining (EHUIM) Algorithm, is developed. Different from previous algorithms, EHUIM Algorithm does not generate candidate high utility itemsets. After constructing the initial utility-record from a mined database, EHUIM Algorithm, can mine high utility itemsets from these utility-record. We are using various standard and real data sets [4].

II. BACKGROUND

A. Problem Definition

Let $I = \{i_1, i_2, i_3, \dots, i_n\}$ be a set of items and DB be a database composed of a utility table and a transaction table. Each item in I has a utility value in the utility table. Each transaction T in the transaction table has a unique identifier(tid) and is a subset of I , in which each item is associated with a count value. An itemset is a subset of I and is called a k -itemset if it contains k items.

Definition 1. The external utility of item i , denoted as $ext_u(i)$, is the utility value of i in the utility table of DB.

Definition 2. The internal utility of item i in transaction T , denoted as $int_u(i, T)$, is the count value associated with i in T in the transaction table of DB.

Definition 3. The utility of item i in transaction T , denoted as $u(i, T)$, is the product of $int_u(i, T)$ and $ext_u(i)$, where $u(i, T) = int_u(i, T) \times ext_u(i)$.

For example, in Table 2 and 3, $ext_u(e) = 3$, $int_u(e, T5) = 1$, and $u(e, T5) = int_u(e, T5) \times ext_u(e) = 1 \times 3 = 3$.

Definition 4. The utility of itemset X in transaction T , denoted as $u(X, T)$, is the sum of the utilities of all the items in X in T in which X is contained, where $u(X, T) = \sum_{i \in X \wedge X \subseteq T} u(i, T)$

Definition 5. The utility of itemset X , denoted as $u(X)$, is the sum of the utilities of X in all the transactions containing X in DB, where $u(X) = \sum_{T \in DB \wedge X \subseteq T} u(X; T)$

For example, in Table 2, $u(\{ae\}, T2) = u(a, T2) + u(e, T2) = 2 \times 5 + 2 \times 3 = 16$, and $u(\{ae\}) = u(\{ae\}, T2) + u(\{ae\}, T5) = 16 + 14 = 30$.

Definition 6. The utility of transaction T , denoted as $tu(T)$, is the sum of the utilities of all the items in T , where $tu(T) = \sum_{i \in T} u(i, T)$, and the total utility of DB is the sum of the utilities of all the transactions in DB.

B. Related Work

Many algorithms have been proposed for high utility itemset mining but all they first produce candidate itemset which require more time and space. Here in this algorithm a search space from the UP Growth algorithm [5] is minimized. A Utility information record structure is used instead of UP Tree.

III. PROPOSED METHOD

The framework of the proposed method consists of following steps: 1) Scan database to construct utility Information Record. 2) Apply EHUI mining algorithm. 3) Generate High Utility Itemsets.

A. Utility Information Record Structure

In the section, we propose a utility information record structure to maintain the utility information about a database.

1) *Initial Utility information record*: Initial utility information record storing the utility information about a mined database can be constructed by two scans of the database. Firstly, the transaction-weighted utilities of all items are accumulated by a database scan. If the transaction-weighted utility of an item is less than a given *minutil*, the item is no longer considered. For the items whose transaction-weighted utilities exceed the *minutil*, they are sorted in transaction-weighted-utility-ascending order.

2) *Utility information record of 2-Itemsets*: No need for database scan, the utility information record of 2-itemset $\{xy\}$ can be constructed by the intersection of the utility list of $\{x\}$ and that of $\{y\}$. The algorithm identifies common transactions by comparing the tids

in the two utility information record. Suppose the lengths of the utility-lists are m and n respectively, and then $(m + n)$ comparisons at most are enough for identifying common transactions, because all tids in a utility information record are ordered. The identification process is actually a 2-way comparison.

3) *Utility information record of k-Itemsets ($k \geq 3$)*: To construct the utility-list of k -itemset $\{i_1 \dots i_{(k-1)}i_k\}$ ($k \geq 3$), we can directly intersect the utility-list of $\{i_1 \dots i_{(k-2)}i_{(k-1)}\}$ and that of $\{i_1 \dots i_{(k-2)}i_k\}$ as we do to construct the utility-list of a 2-itemset.

Algorithm 1: Build - Tree Generation Algorithm

Input: $P.UIR$, the utility information record of itemset P ;
 $Px.UIR$, the utility information record of itemset Px ;
 $Py.UIR$, the utility information record of itemset Py .
 Output: $Pxy.UIR$, the utility information record of itemset Pxy .

1. $Pxy.UIR = NULL$;
2. for each element $Ex \in Px.UIR$ do
3. if $\exists Ey \in Py.UIR$ and $Ex.tid == Ey.tid$ then
4. if $P.UIR$ is not empty then
5. search such element $E \in P.UIR$ that $E.tid == Ex.tid$;
6. $Exy = \langle Ex.tid, Ex.iutil + Ey.iutil - E.iutil, Ey.rutil \rangle$;
7. Else
8. $Exy = \langle Ex.tid, Ex.iutil + Ey.iutil, Ey.rutil \rangle$;
9. end if
10. append Exy to $Pxy.UIR$;
11. end if
12. end for
13. return $Pxy.UIR$;

B. EHUI Mining Algorithm

After constructing a Utility information record a EHUI Mining Algorithm can mine all high utility itemset from database.

1) *Domain Space*: The domain space of the high utility itemset mining problem can be represented as a combination tree. Given a set of items $I = \{i_1, i_2, i_3, \dots, i_n\}$ and a total order on all items (suppose $i_1 < i_2 < \dots < i_n$), a combination tree representing all itemsets can be constructed as follows.

Firstly, the root of the tree is created; secondly, then child nodes of the root representing $n-1$ -itemsets are created, respectively; thirdly, for a node representing itemset $\{i_s \dots i_e\}$ ($1 \leq s \leq e < n$), the $(n-e)$ child nodes of the node representing itemsets $\{i_s \dots i_e i_{(e+1)}\}, \{i_s \dots i_e i_{(e+2)}\}, \dots, \{i_s \dots i_e i_n\}$ are created. The third step is done repeatedly until all leaf nodes are created. For example, given $I = \{e, c, b, a, d\}$ and $e < c < b < a < d$, a combination tree representing all itemsets of I is depicted in Fig. 1.

2) *Pruning Strategy*: For a database with n items, exhaustive search has to check 2^n itemsets. To reduce the search space, we can exploit the iutils and rutils in the utility-list of an itemset. The sum of all the iutils in the utility-list of an itemset is the utility of the itemset according to Definition 5, and thus the itemset is high

utility if the sum exceeds a given *minutil*. The sum of all the *iutils* and *rutils* in the utility-list provides EHUIM Algorithm with the key information about whether the itemset should be pruned or not.

Lemma 1. Given the utility information record of itemset *X*, if the sum of all the *iutils* and *rutils* in the utility information record is less than a given "*minutil*", any extension *X'* of *X* is not high utility.

- 3) **EHUI Mining Algorithm:** Algorithm 2 shows the pseudo-code of EHUIM Algorithm. For each utility information record *X* in ULs (the second parameter), if the sum of all the *iutils* in *X* exceeds *minutil*, and then the extension associated with *X* is high utility and outputted. According to Lemma 1, only when the sum of all the *iutils* and *rutils* in *X* exceeds *minutil* should it be processed further. When the initial utility-lists are constructed from a database, they are sorted and processed in transaction-weighted utility-ascending order. Therefore, all the utility information records in UIRs are ordered as the initial utility information record are. To explore the search space, the algorithm intersects *X* and each utility information record *Y* after *X* in UIRs. Suppose *X* is the utility-list of itemset *Px* and *Y* that of itemset *Py*, and then *Build(P.UIR, X, Y)* in line 8 is to construct the utility-list of itemset *Pxy* as stated in Algorithm 1. Finally, the set of utility-lists of all the 1-extensions of itemset *Px* is recursively processed. Given a database and a *minutil*, after the initial utility information record IUIRs are constructed, *EHUIM(∅, IUIRs, minutil)* can mine all high utility itemsets.

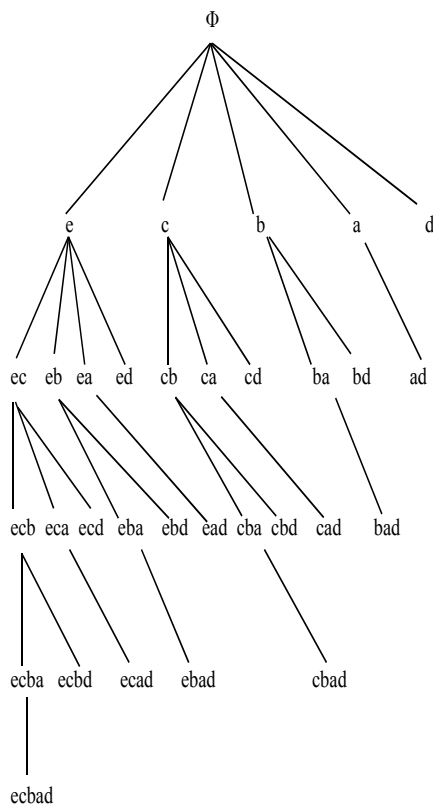


Figure 1: Combination Tree

Algorithm 2: EHUI Mining Algorithm

Input: *P.UIR*, the utility-list of itemset *P*, initially empty;
UIRs, the set of utility-lists of all *P*'s 1-extensions;
minutil, the minimum utility threshold.

Output: all the high utility itemsets with *P* as prefix.

1. for each utility information record *X* in *UIRs* do
2. if $SUM(X.iutils) \geq minutil$ then
3. output the extension associated with *X*;
4. end if
5. if $SUM(X.iutils) + SUM(X.rutils) \geq minutil$ then
6. *exULs* = *NULL*;
7. for each utility information record *Y* after *X* in *UIRs* do
8. *exUIRs* = *exUIRs* + *Build(P.UIR, X, Y)*;
9. end for
10. *EHUI(X, exULs, minutil)*;
11. end if
12. end for

IV. CONCLUSION

In this paper, we have proposed a novel data structure, utility information record, and developed an efficient algorithm, EHUI, for high utility itemset mining. Utility information record provides not only utility information about itemsets but also important pruning information for EHUI. We have used Educational real time and standard datasets. Previous algorithms have to process a very large number of candidate itemsets during their mining processes. However, most candidate itemsets are not high utility and are discarded finally. EHUI Algorithm can mine high utility itemsets without candidate generation, so that complexity of UPGrowth and UPGrowth+ is reduced as it require less time and space, which avoids the costly generation and utility computation of candidates. However in future we can again reduce the complexity by reducing the joining cost of utility information record.

REFERENCES

- [1] Jyothi Pillai, O.P.Vyas "Overview of Itemset Utility Mining and its Applications" *IJCA(0975 – 8887) Volume 5– No.11, August 2010.*
- [2] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery, 15(1):55–86, 2007.*
- [3] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K.Lee. Efficient tree structures for high utility patternmining in incremental databases. *IEEE Transactions onKnowledge and Data Engineering, 21(12):1708–1721,2009.*
- [4] Frequent Itemset Mining Implementations Repository, <http://fimi.cs.helsinki.fi/>, 2013.
- [5] Vincent S. Tseng, Bai-En Shie, Cheng Wei Wu, and Philip S. Yu, Fellow, "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases" *IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 8, August 2013.*