FPT Education

TRƯỜNG ĐẠI HỌC FPT

Effective High Utility Itemsets Mining Algorithm for Incremental Database

AIP490-G16

Students: Do Thanh Cong, Do Mai Phuong, Pham Duc Duong

Supervisor: Assoc. Prof. Phan Duy Hung

TABLE OF CONTENTS:

INTRODUCTION

Motivation

Related works

Objectives

Motivation

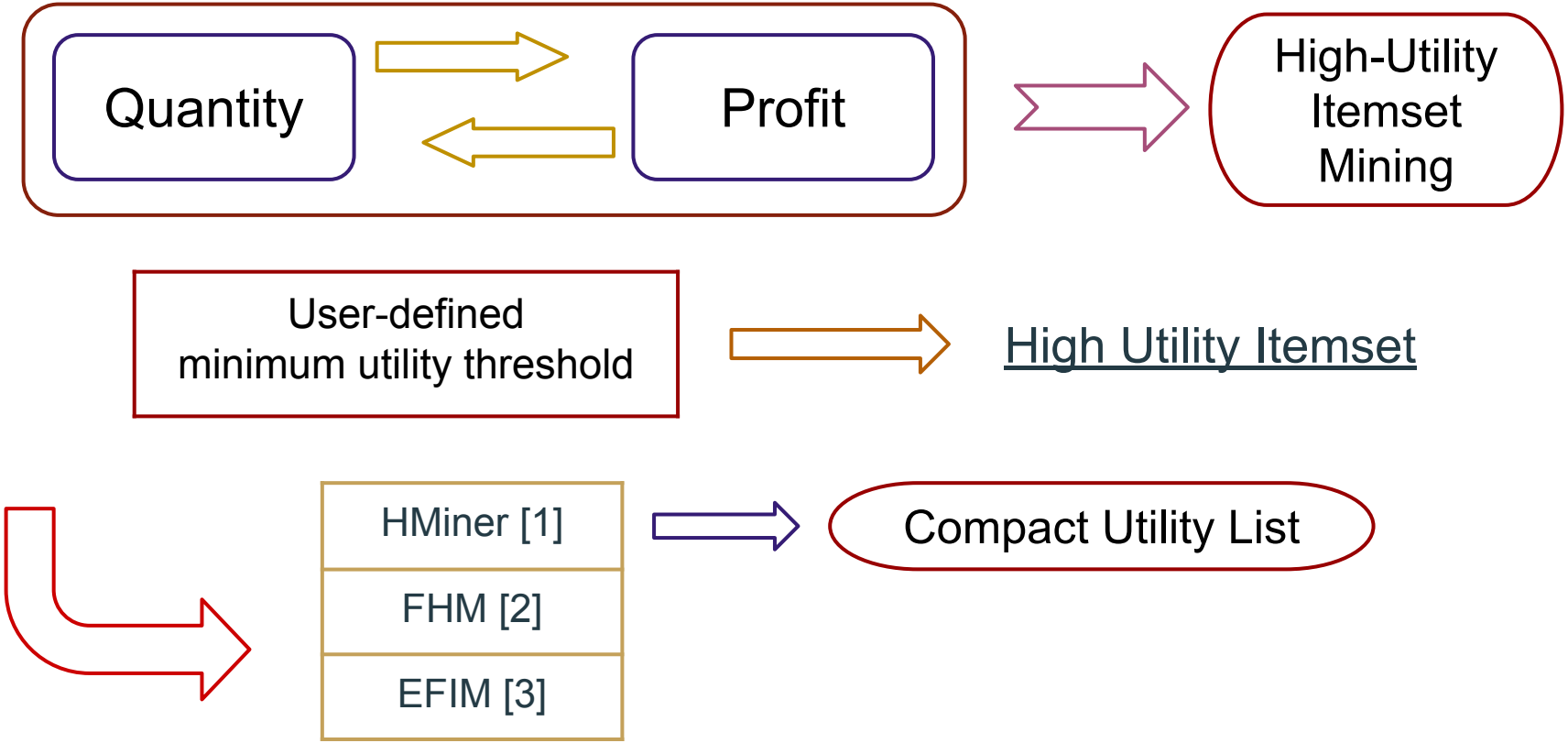
- Business
- Healthcare
- Education
- Science

Tid	Items bought
10	*Beer*, Nuts, *Diaper*
20	*Beer*, Coffee, *Diaper*
30	*Beer*, *Diaper*, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

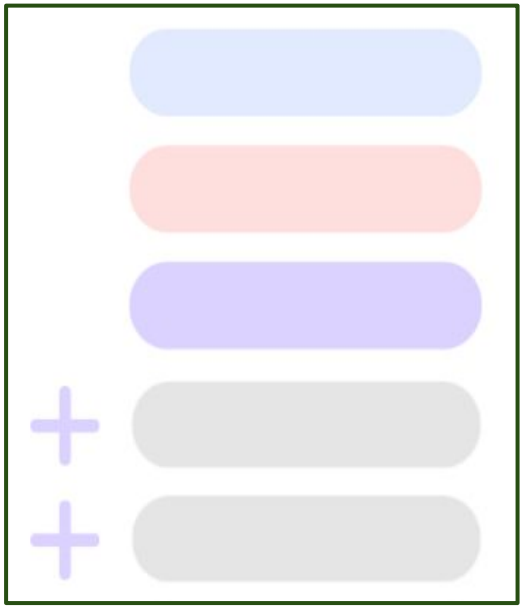
Beer, Diaper

Frequent Pattern Mining

Motivation


Quantity
Profit
High-Utility Itemset Mining
User-defined minimum utility threshold
High Utility Itemset
HMiner [1]
FHM [2]
EFIM [3]
Compact Utility List

Motivation



(A, C, F)
(B, C, E)
(C, D)
…
+
+

High Utility Itemset

Incremental Databases

PRE-HUI-INS [4]	HUI-List-INS [5]	LIHUP [6]	EIHI [7]	…

Objectives


EIHI Algorithm
Modified Compact Utility List
OUR ALGORITHM (iHUIM)

METHODOLOGY

Preliminaries

Storage structure

Problem definition

Proposed algorithm

Preliminaries

Transactions Database

Tid	Transactions	Quantity (q)	Transaction utility
T1	a,b,c,d,e,f	1,5,1,3,1,1	30
T2	b,c,d,e	4,3,3,1	20
T3	a,c,d	1,1,1	8
T4	a,c,e,g	2,6,2,5	27
T5	b,c,e,g	2,2,1,2	11

D

N

Price of an item (pr)

Item	a	b	c	d	e	f	g
Price	5	2	1	2	3	5	1

minUtil: user-defined minimum utility threshold

Preliminaries

Transactions Database

Tid	Transactions	Quantity (q)	Transaction utility
T1	a,b,c,d,e,f	1,5,1,3,1,1	30
T2	b,c,d,e	4,3,3,1	20
T3	a,c,d	1,1,1	8
T4	a,c,e,g	2,6,2,5	27
T5	b,c,e,g	2,2,1,2	11

Price of an item(pr)

Item	a	b	c	d	e	f	g
Price	5	2	1	2	3	5	1

- Utility of an item in transaction :

$$u(i_x, T_{id}) = q(i_x, T_{id}) \times pr(i_x)$$

- Utility of an itemset in database:

$$u(X) = \sum_{X \subseteq T_{id}, T_{id} \in D} u(X, T_{id})$$

Example : itemset X = {b,d}
u(X) = u(X,T1) + u(X,T2)
= (5*2+3*2) + (4*2+3*2) = 30

=> High utility itemset (HUI) is itemset has utility *equal or larger* than minUtil threshold

Preliminaries

Transactions Database

Tid	Transactions	Quantity (q)	Transaction utility
T1	a,b,**c,d**,e,f	1,5,1,3,1,1	30
T2	b,**c,d**,e	4,3,3,1	20
T3	a,**c,d**	1,1,1	8
T4	a,c,e,g	2,6,2,5	27
T5	b,c,e,g	2,2,1,2	11

Price of an item (pr)

Item	a	b	c	d	e	f	g
Price	5	2	1	2	3	5	1

- Transaction weighted utility

$$TWU(X) = \sum_{X \subseteq T_{id} \in D} u(T_{id})$$

Example: itemset X = {c,d}

TWU(X) = u(T1) + u(T2) + u(T3)

= 30+20+8 = 58

Preliminaries

Transactions Database

Tid	Transactions	Utility of item in transaction	Transaction utility
T1	f,**b**,**d**,a,e,c	5,10,6,**5**,**3**,**1**	30
T2	b,d,e,c	8,6,3,3	20
T3	d,a,c	2,5,1	8
T4	g,a,e,c	5,10,6,6	27
T5	g,b,e,c	4,2,3,2	11

TWU of 1-itemset

Item	g	f	b	d	a	e	c
TWU	27	30	50	58	65	77	85

- Remaining utility of an itemset in transaction:

$$ru(X, T_{id}) = \sum_{i_x \in (T_{id} / X)} u(i_x, T_{id})$$

Example: itemset X = {b,d} in T1
ru(X,T1) = u({a},T1) + u({e},T1) + u({c},T1)
= 5+3+1= 9

Preliminaries

Transactions Database

Tid	Transactions	Utility of item in transaction	Transaction utility
T1	f,b,d,a,e,c	5,10,6,5,3,1	30
T2	b,d,e,c	8,6,3,3	20
T3	d,a,c	2,5,1	8
T4	g,a,e,c	5,10,6,6	27
T5	g,b,e,c	4,2,3,2	11

TWU of 1-itemset

Item	g	f	b	d	a	e	c
TWU	27	30	50	58	65	77	85

- Extension of an itemset X: items after X in ordered set of all items.
- Size of X's extension: c(X)

Example: itemset X = {f,d}
⇒ Extension of X is {a,e,c} and c(X) = 3

Preliminaries

Transactions Database

Tid	Transactions	Utility of item in transaction	Transaction utility
T1	f,b,**d**,a,e,**c**	5,10,6,5,3,1	30
T2	b,d,e,c	8,6,3,3	20
T3	d,a,c	2,5,1	8
T4	g,a,e,c	5,10,6,6	27
T5	g,b,e,c	4,2,3,2	11

TWU of 1-itemset

Item	g	f	b	d	a	e	c
TWU	27	30	50	58	65	77	85

- Closed utility of an itemset X:

$$cu(X,T_{id}) = \begin{cases} u(X,T_{id}) \text{ if } |X|>1 \text{ and } c(X-i_k) = s(T_{id}/X-i_k) \\ 0, \text{ otherwise} \end{cases}$$

In above formula $s(T_{id}/X-i_k) = |T_{id}/X-i_k|$

Example: itemset X = {d,c} in T1

cu(X, T1) = u(X, T1) = 7

Because |X| = 2 (>1) and

c(X - {c}) = c({d}) = {a,e,c} = 3 = s(T1 / {d})

Preliminaries

Transactions Database

Tid	Transactions	Utility of item in transaction	Transaction utility
T1	f,b,**d**,a,e,**c**	5,10,6,5,3,1	30
T2	b,d,e,c	8,6,3,3	20
T3	d,a,c	2,5,1	8
T4	g,a,e,c	5,10,6,6	27
T5	g,b,e,c	4,2,3,2	11

TWU of 1-itemset

Item	g	f	b	d	a	e	c
TWU	27	30	50	58	65	77	85

- Closed remaining utility of an itemset X:

$$cru(X,T_{id}) = \begin{cases} ru(X,T_{id}), \text{ if } |X|>1 \text{ and } c(X-i_k) = s(T_{id}/X-i_k) \\ 0, \text{ otherwise} \end{cases}$$

Example: itemset X = {d,c} in T1

cru(X,T1) = ru(X,T1) = 0

Because |X| = 2,

c(X - {c}) = c({d}) = {a,e,c}

= 3 = s(T1 / {c})

Preliminaries

Transactions Database

Tid	Transactions	Utility of item in transaction	Transaction utility
T1	f,b,d,a,e,c	5,10,6,5,3,1	30
T2	b,d,e,c	8,6,3,3	20
T3	d,a,c	2,5,1	8
T4	g,a,e,c	5,10,6,6	27
T5	g,b,e,c	4,2,3,2	11

TWU of 1-itemset

Item	g	f	b	d	a	e	c
TWU	27	30	50	58	65	77	85

- Non-closed utility of an itemset X:

$$nu(X) = u(X) - cu(X)$$

- Non-closed remaining utility of an itemset X:

$$nru(X) = ru(X) - cru(X)$$

Problem definition

Original database D

User-defined threshold

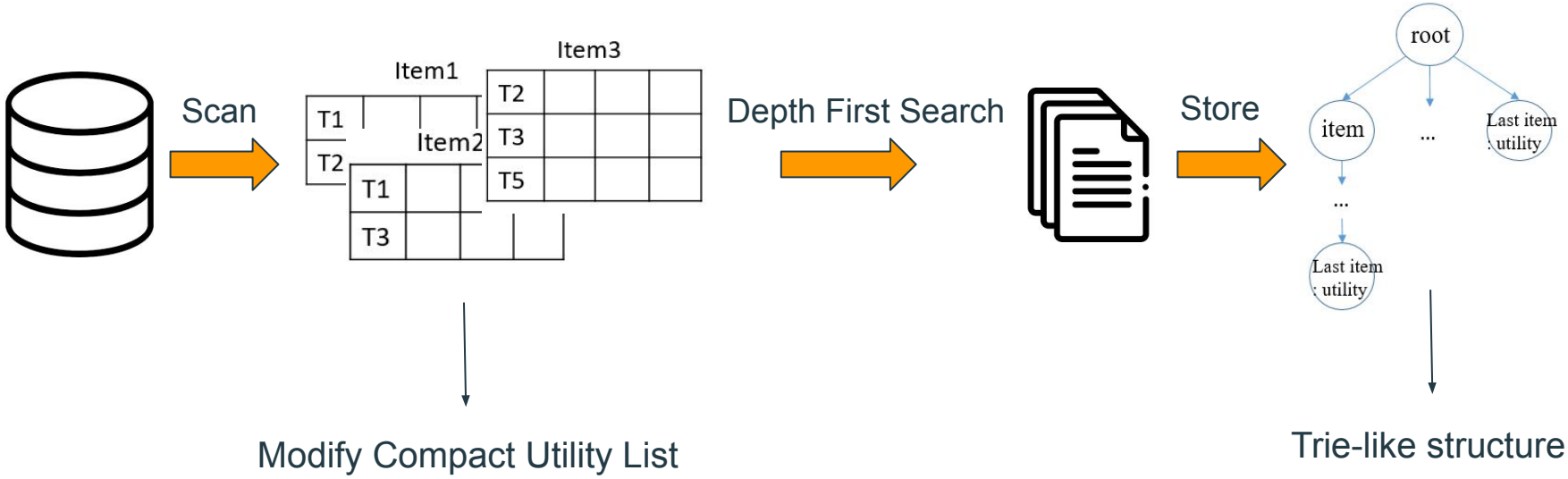
New transaction N

Set of all HUIs found in D

Set of all HUIs found in U = D U N

Problem definition

Mining Process



Storage structure

Modify Compact Utility List structure

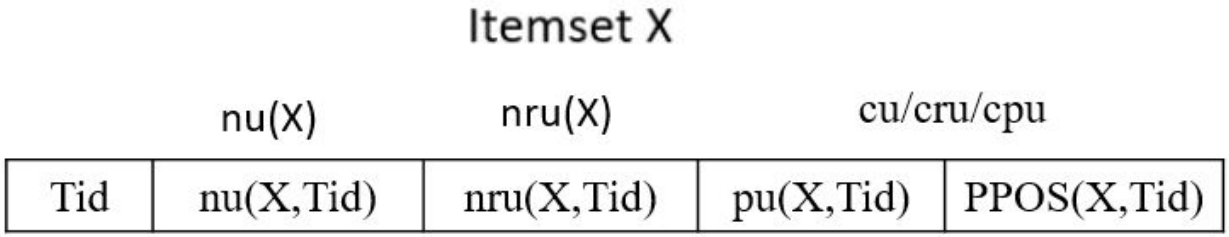


Fig. 1: Compact Utility List Structure [1]

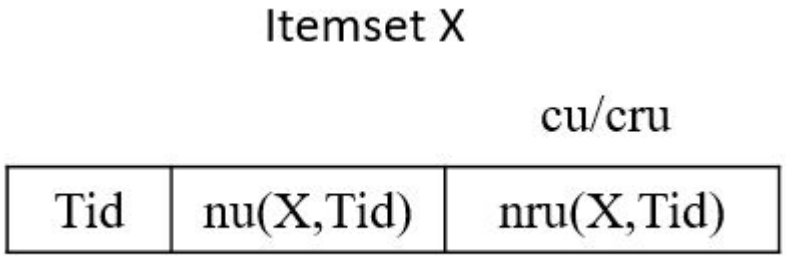


Fig. 2: Modify Compact Utility List Structure

Storage structure

Trie-like structure

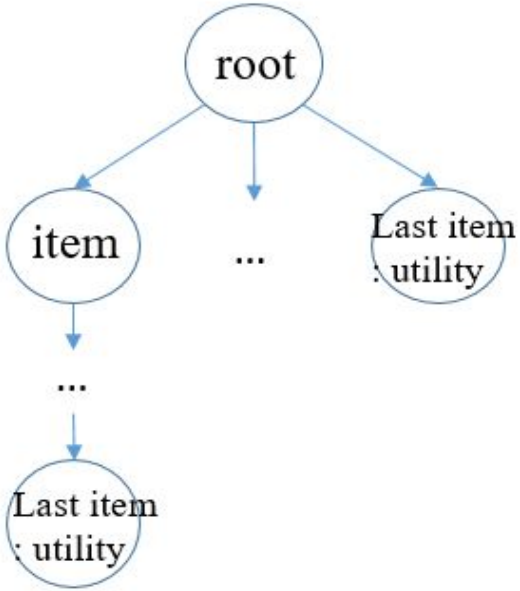


Fig. 3: General structure

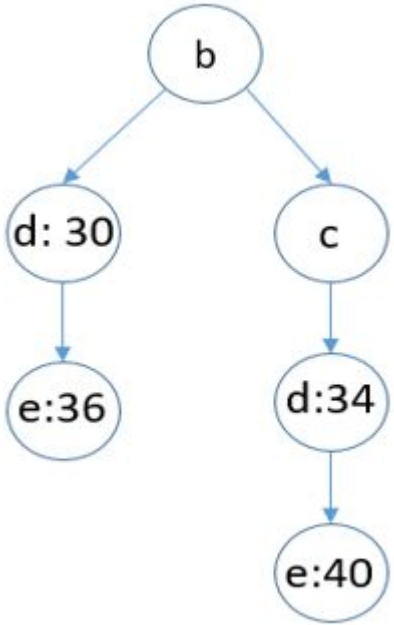


Fig. 4: Sample representation of HUIs in D

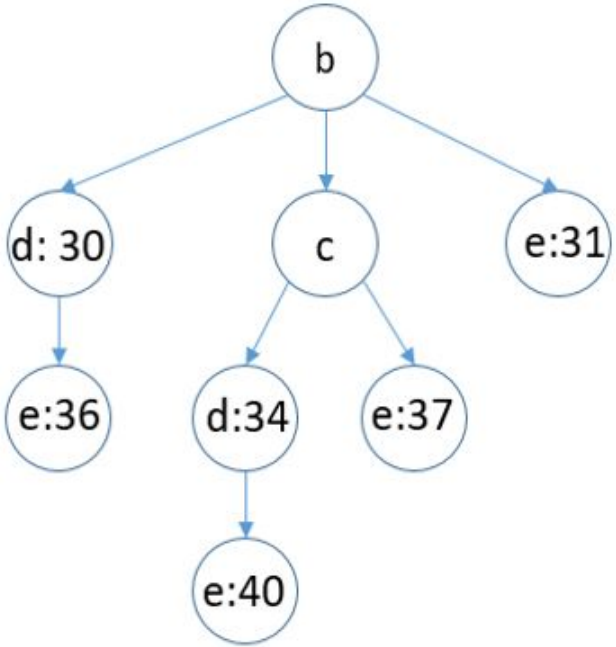


Fig. 5: Sample representation of HUIs after adding N

Proposed algorithm

Algorithm 1: The EIHI Algorithm

input : D: a transaction database, *minutil*: a user-specified threshold

output: the set of high-utility itemsets

1 Scan D to calculate the TWU of single items;
2 $I^* \leftarrow$ each item i such that $TWU(i) \geq minutil$;
3 Let $\succ$ be the total order of TWU ascending values on I^*;
4 Scan D to build the utility-list of each item $i \in I^*$ and build the *EUCS* structure;
5 **Search** $(\emptyset, I^*, minutil, EUCS)$;

Fig. 6: Main procedure of EIHI algorithm [13]

Algorithm 1: Main

Input: *D (or N), minUtil*

Output: *HUIs*

1: **foreach** transaction T in D **do**
2: **foreach** item x in T **do**
3: create or update MCUL(x), TWU(x)
4: **end for**
5: **end for**
6: Sort All_MCULs according to TWU
7: **if** TWU(X) ≥ minUtil **then**
8: MCULs ←X
9: FinalMCULs = Restruct (MCULs)
10: SearchHUI (null, FinalMCULs, minUtil)

Fig. 7: Main procedure of iHUIM algorithm

Proposed algorithm

Algorithm 2: Restruct

Input: *MCULs*

Output: *FinalMCULs*

```
1: lastMCUL = Last MCUL in MCULs
2:  foreach ex ∈ lastMCUL do
3:        TempTable(ex.tid) ← ex.nu + ex.nru
4:  end for
5:  foreach Y before L in MCULs do
6:        foreach ey ∈ Y do
7:              ey.nru ← TempTable(ey.tid)
8:              TempTable(ey.tid) ← ey.nu + ey.nru
9:        end for
10: end for
11: Return FinalMCULs
```

Fig. 8: Restruct function

EXPERIMENTS AND RESULT ANALYSIS

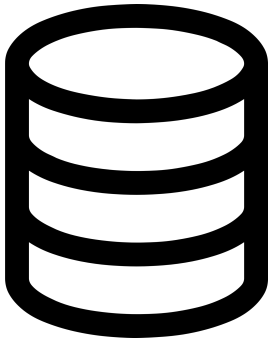
Data preparation

Experiment setup

Results and analysis

Data preparation

Data source: Datasets



SPMF

An Open-Source Data Mining Library

- Introduction
- Algorithms
- Download
- Documentation
- **Datasets**
- FAQ
- License
- Contributors
- Citations
- Performance
- Developers's guide
- Videos

Datasets

The SPMF software natively uses **text files** as input. Some small examples of text files that can be used with each algorithm are described in the documentation of SPMF. These sample input files can be downloaded from the download page (test_files.zip) for the release version of SPMF, and are included with the source code, for the source code version of SPMF. However, these datasets are quite small. For this reason, this webpage provides larger datasets that can be used with SPMF and that are often used in the data mining litterature for evaluating and comparing algorithm performance. Unless otherwise indicated, the datasets are in SPMF format.

The **datasets** are divided in the following **categories**:

- Datasets for Sequential Pattern Mining / Sequential Rule Mining / Sequence Prediction
  - Real-life datasets in SPMF format
  - A collection of 30 books converted to SPMF format
  - Sequences of MOOC data with timestamps
  - Synthetic datasets
  - Datasets of time-interval sequences
- Datasets for Frequent Itemset mining / Association Rule Mining / Periodic pattern mining / Frequent episode mining)
  - Datasets in SPMF format
  - Synthetic datasets
  - Real-life datasets in SPMF format, having timestamps
  - Real-life datasets in ARFF format
- Datasets for High-Utility Pattern Mining
  - Real-life transaction datasets in SPMF format with real utility values
  - Real-life transaction datasets in SPMF format having synthetic (fake) utility values
  - Datasets in SPMF format having utility values and timestamps
  - Datasets for high utility itemset mining with negative unit profit values

Data preparation

Definition:

Dataset	|D|	|I|	AvgLen	Density	Real utility value
Chess	3196	75	37	Dense	No

- |D|: transaction count of D
- |I|: number of distinct items
- AvgLen: average transaction length
- Density: Density of dataset (Dense or Sparse)
- Real utility value: Dataset use real or fake utility

Data preparation

Characteristic of datasets:

Dataset	|D|	|I|	AvgLen	Density	Has real utility value
Chess	3196	75	37	Dense	No
Fruithut	181970	1265	3.58	Sparse	Yes
Mushroom	8124	119	23	Dense	No

Data preparation

Datasets file structure:

Format: Item1 item2 item3… : TU : Util(item1) Util(item2) Util(item3)

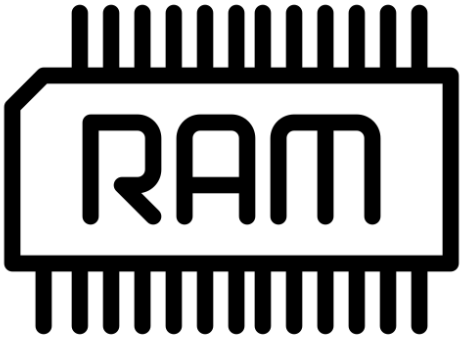
Example: Fruithut database

2010 2021 2032 : 897 : 199 399 299
2038 : 180 : 180
1031 2022 : 449 : 150 299

Experiment setup

Window 11

Intel® Core I7, 2.20GHz



16GB RAM DDR4

Experiment setup

Programming language: Java with version JDK 11

Compare with: EIHI algorithm

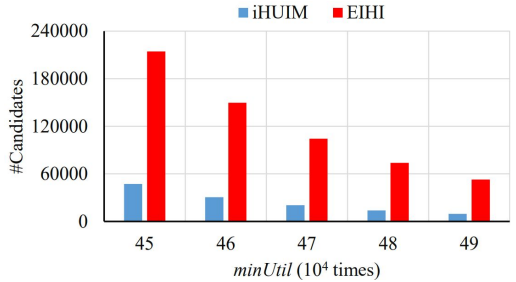
Evaluation criteria: Number of generated candidates, Execution time, Memory usage

Parameters: *minUtil*; *addRatio* (20%, 25%)

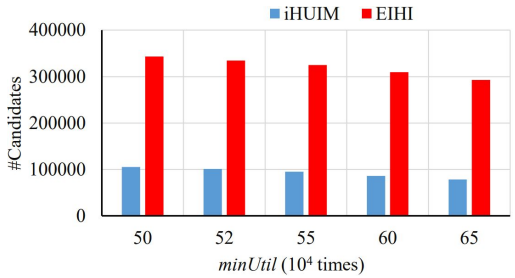
With pair of parameters, each algorithm executes 5 times to get the average value

Results and analysis

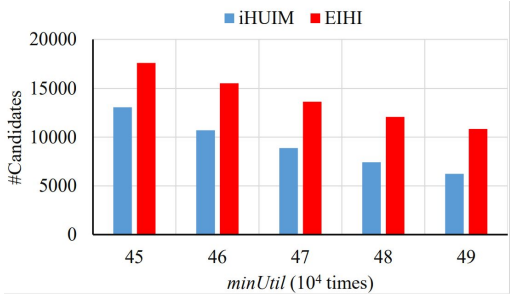
Number of generated candidates



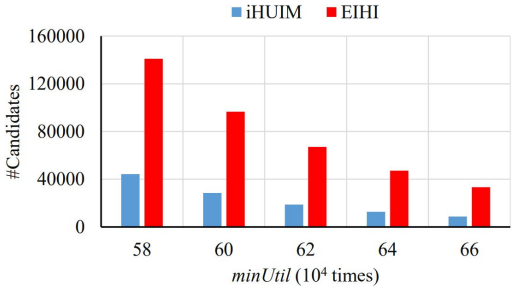
Chess-addRatio=20%



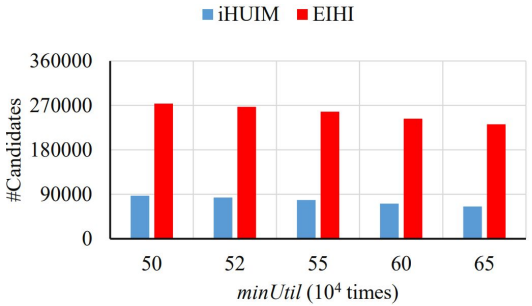
Fruithut-addRatio=20%



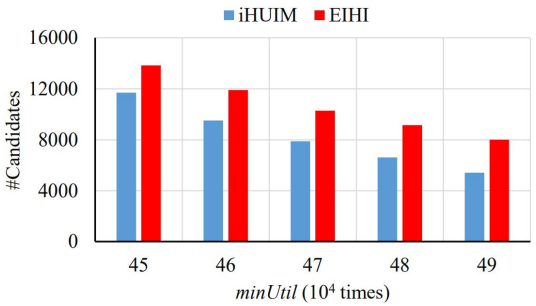
Mushroom-addRatio=20%



Chess-addRatio=25%



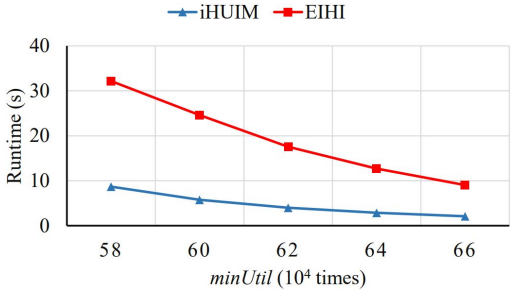
Fruithut-addRatio=25%



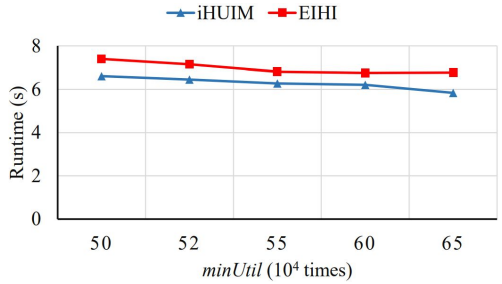
Mushroom-addRatio=25%

Results and analysis

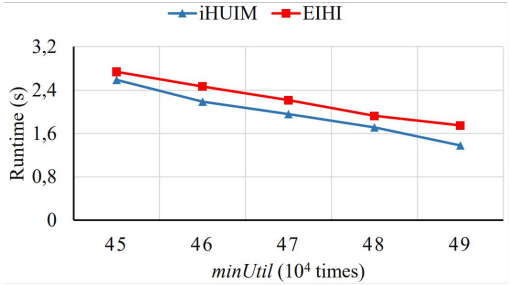
Execution time



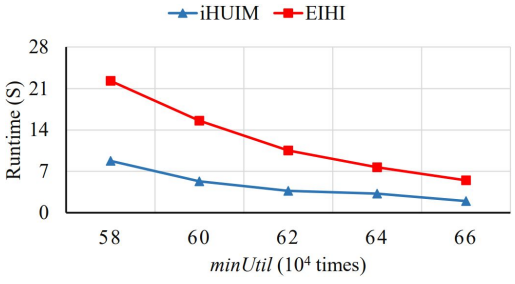
Chess-addRatio=20%



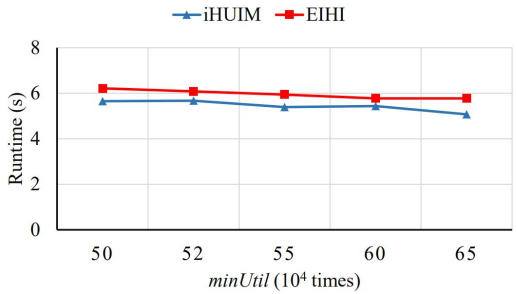
Fruithut-addRatio=20%



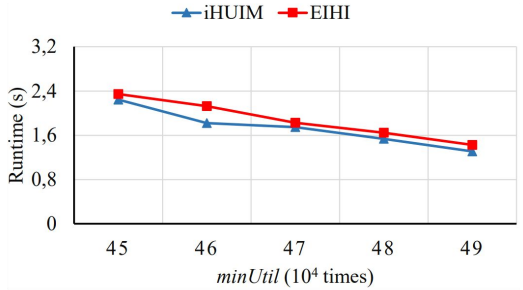
Mushroom-addRatio=20%



Chess-addRatio=25%



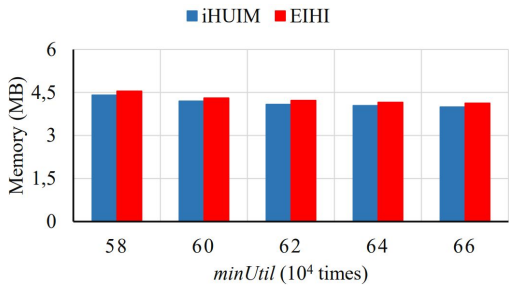
Fruithut-addRatio=25%



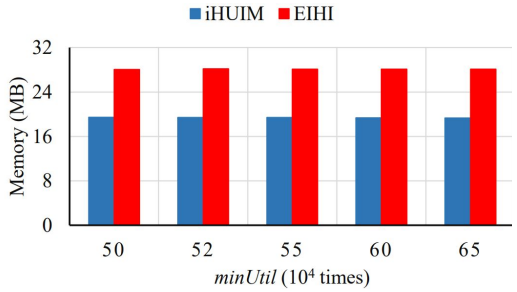
Mushroom-addRatio=25%

Results and analysis

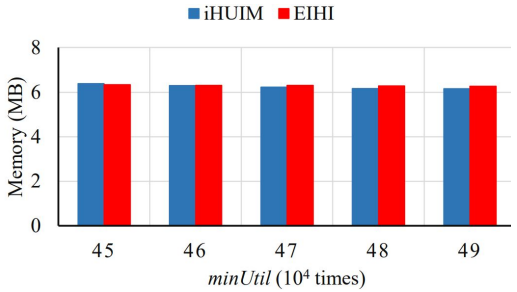
Memory usage



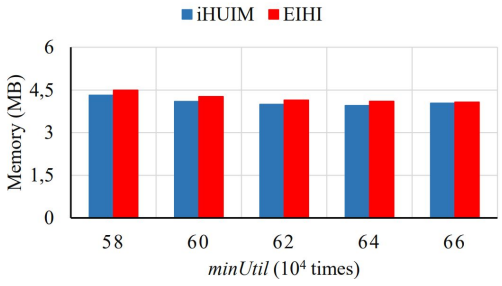
Chess-addRatio=20%



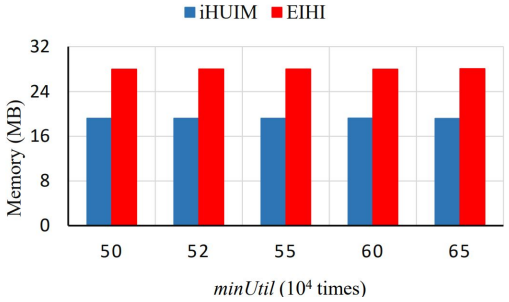
Fruithut-addRatio=20%



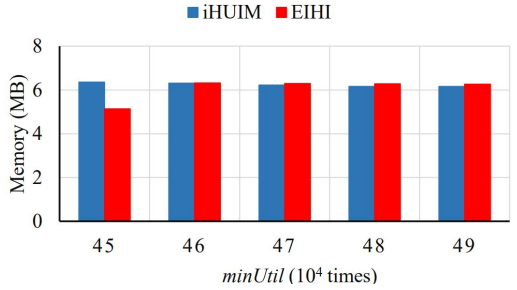
Mushroom-addRatio=20%



Chess-addRatio=25%



Fruithut-addRatio=25%



Mushroom-addRatio=25%

CONCLUSION AND FUTURE WORK

Conclusion

- Proposing the iHUIM algorithm that extends the EIHI algorithm
- Modifies Compact Utility List structure to store information of each item and scans the database once to build MCUL
- Strategies to prune items that can’t be HUIs before and during the mining process

Future work

- Continue to modify CUL or consider other efficient data structures to gain the efficiency of execution time and memory usage

Accept letter from ICICT 2024

ICICT 2024 <icict2024@easychair.org> 24 Oct 2023, 15:08 (9 days ago)

to me

Dear Do Thanh Cong

Paper ID : 318

Title : Effective High Utility Itemsets Mining Algorithm for Incremental Databases

Greetings .. !!

Congratulations! On behalf of the Program Committee of ICICT 2024 - LONDON, I am happy to inform you that your above-mentioned paper has been ACCEPTED for oral presentation in ICICT 2024 and publication in Springer LNNS series subject to fulfillment of Guidelines by Springer.

An accepted paper will be published in the Springer proceedings (LNNS) only if the final version is accompanied by the payment information (i.e. transaction reference number) subject to quality check as per Springer Guidelines.

Kindly follow the below-mentioned guidelines (strictly), related to the preparation of the Final Manuscript, Copyright Transfer Form, Payment, and Final Submission. The procedure has been detailed as a five-step process (I)-(III):

(I) Preparation of CRC (Final Manuscript for Inclusion in Springer Proceedings Book-LNNS Series)
You are requested to give strict attention to the below-mentioned points during the preparation of the final manuscript to avoid any last-minute revert backs from the publisher (Springer).

References

[1] S. Krishnamoorthy, "HMiner: Efficiently mining high utility itemsets," *Expert Systems with Applications,* vol. 90, pp. 168-183, 2017.

[2] Philippe Fournier-Viger, Cheng-Wei Wu, Souleymane Zida, and Vincent S. Tseng, "FHM: Faster high-utility itemset mining using estimated utility co-occurrence pruning," *International Symposium on Methodologies for Intelligent Systems,* pp. 83-92, 2014.

[3] Souleymane Zida, Philippe Fournier-Viger, Jerry Chun-Wei Lin, Cheng-Wei Wu, and Vincent S. Tseng, "EFIM: a fast and memory efficient algorithm for high-utility itemset mining," *Knowledge and Information Systems,* vol. 51, no. 2, pp. 595-625, 2017.

[4] Chun-Wei Lin, Tzung-Pei Hong, Guo-Cheng Lan, Jia-Wei Wong, Wen-Yang Lin, "Incrementally mining high utility patterns based on pre-large concept," *Applied Intelligence,* vol. 40, no. 2, pp. 343-357, 2014.

[5] Jerry Chun-Wei Lin, Wensheng Gan, Tzung-Pei Hong, Binbin Zhang, "Incrementally Updating High-Utility Itemsets with Transaction Insertion," in *Advanced Data Mining and Applications*, Springer International Publishing, 2014, pp. 44-56.

[6] Unil Yun, Heungmo Ryang, Gangin Lee,Hamido Fujita, "An efficient algorithm for mining high utility patterns from incremental databases with one database scan," *Knowledge-Based Systems,* vol. 124, pp. 188-206, 2017.

[7] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Ted Gueniche, Prashant Barhate, "Efficient Incremental High Utility Itemset Mining," in *Proceedings of the ASE BigData & SocialInformatics 2015*, Association for Computing Machinery, 2015, p. 6

Q&A

Thanks for your attention !