哈爾濱工業大學(深圳)

HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

CHUQI-Miner: Correlated High Utility Quantitative Itemset Mining

Mourad Nouioua, Philippe Fournier-Viger, Jun-Feng Qu, Jerry Chun-Wei Lin, Wensheng Gan, Wei Song

UDML 2021 @ ICDM 2021

Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- HUQIM
- Proposed Solution: CHUQI-Miner
  - Bond Measure
  - Problem Definition
  - Q-item Utility-Lists
  - How to Reduce the Search Space
  - Experimental Evaluation
- Conclusion

High Utility Itemset Mining

Input: Transaction database + Profit table + Minimum utility threshold (*minutil*)

Transaction database D

TID	Items
T_1	(a,1), (c,1)
T_2	(e,1)
T_3	(a,1), (b,5), (c,1), (d,3), (e,1)
T_4	(b,4), (c,3), (d,3), (e,1)
T_5	(a,1), (c,1), (d,1)
T_6	(a,2), (c,6), (e,2)
T_7	(b,2), (c,2), (e,1)

Profit table

Item	Unit profit
a	5$
b	2$
c	1$
d	2$
e	3$

***minutil* threshold**

***minutil* = 30$**

Output: High Utility Itemsets (HUIs), i.e., itemsets having utility ≥ *minutil*

{b,d}:30$	{a,c}:34$
{b,e}:31$	{b,d,e}:36$
{a,c,e}:31$	{b,c,e}:37$
{b,c,d}:34$	{b,c,d,e}:40$

High Utility Itemset Mining

Several Algorithms

- Two-Phase (PAKDD 2005)
- IHUP (TKDE, 2010)
- UP-Growth (KDD 2011)
- HUI-Miner (CIKM 2012)
- FHM (ISMIS 2014)
- EFIM (KAIS 2017)
- mHUIMiner (PAKDD 2017)

Key Idea (Pruning Strategies)

1. Define an **upper-bound** on the utility of itemsets (**e.g.** the TWU) that is anti-monotonic to be able to prune the search space.

2. Propose some **pruning strategies** based on these upper bounds.

Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- HUQIM
- Proposed Solution: CHUQI-Miner
  - Bond Measure
  - Problem Definition
  - Q-item Utility-Lists
  - How to Reduce the Search Space
  - Experimental Evaluation
- Conclusion

Limitations

High Utility Itemset Mining (HUIM)

- The discovered patterns do not provide information about quantities.

High Utility Quantitative Itemset Mining (HUQIM)

- Discover all sets of items that have a high utility while also providing information about item quantities that led to this utility.

HUIM (Itemset)		HUQIM (Q-itemset)
Coffee + Cookies+ Eggs = 30 $	⟹	3 boxes of Coffee+2 boxes Cookies + 6 Eggs = 30$

Limitations

Consider quantity information

- Mining high utility quantitative association rules HUQA: (DAWAK 2007).
- Vertical mining of high utility quantitative itemsets VHUQI: (IEEE GrC 2014).
- Efficient mining of high utility quantitative itemsets HUQI-Miner: (ICDMW:2019).

HUQI-Miner

- Proposed algorithms still have very long runtimes due to the very large search space.

Could we make a more efficient algorithm with faster execution time?

Limitations

FHUQI-Miner

- A novel algorithm named **FHUQI-Miner** (**F**ast **H**igh **U**tility **Q**uantitative **I**temset miner) is proposed to:

  Find **High Utility Quantitative Itemsets (HUQIs).**

  Input: **D**atabase + profit table + the minimum utility threshold *(minutil)*.

  Output: Exact and range Q-itemsets having high utilities.

FHUQI-Miner is up to **22 time faster** than HUQI-Miner.

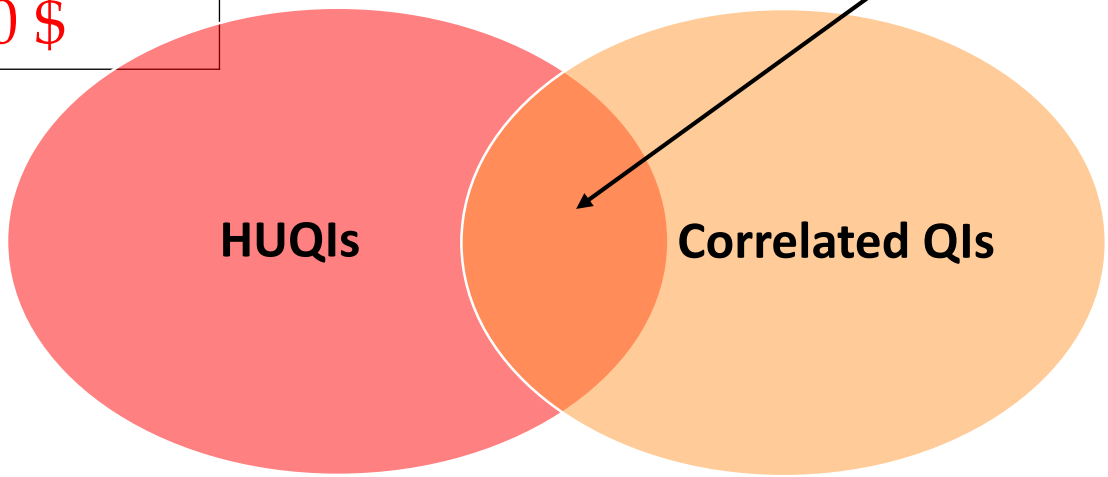
Limitations

Limitation of FHUQI-Miner

- FHUQI-Miner allows to find more informative patterns comparing with HUIM algorithms. However,
  - These algorithms may discover **many itemsets that are weakly correlated.**

HUQIM
10 Pens + 1 Diamond = 500 $

Pens and **Diamond** are **profitable** but they are **weakly correlated** in purchase databases.



Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- **HUQIM**
- **Proposed Solution: CHUQI-Miner**
  - **Bond Measure**
  - **Problem Definition**
  - **Q-item Utility-Lists**
  - **How to Reduce the Search Space**
  - **Experimental Evaluation**
- **Conclusion**

HUQIM

HUQIM search space

- HUQIM associates a quantity or range of quantities to each item.

Transaction database D

TID	Items
T_1	(A,2), (B,5), (C,2), (D,1)
T_2	(B,4), (C,3)
T_3	(A,2), (C,2)
T_4	(A,2), (B,6), (D,1)

Profit table

Item	Unit profit
A	3$
B	1$
C	2$
D	2$

Larger search space !!

HUIM

Items

A, B, C, D

HUQIM

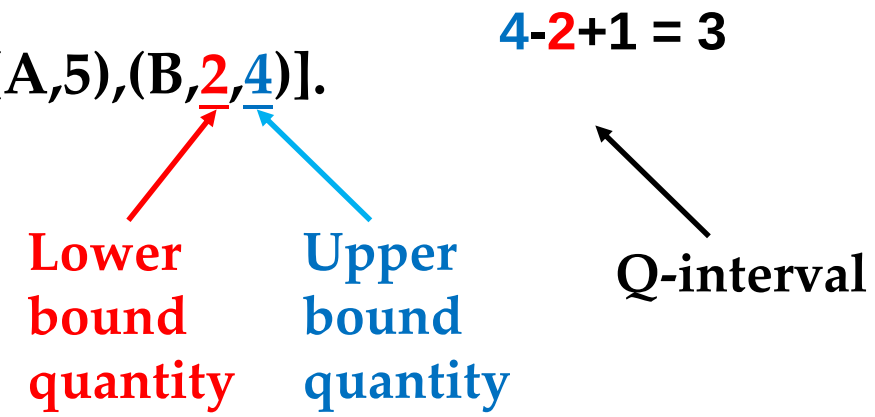
Q-items

(A,2), (B,5), (C,2), (D,1), (C,3), (B,4), (B,6), (D,1)

HUQIM

Q-itemsets

- In HUQIM, there are two kind of itemsets to be discovered: **Exact Q-itemsets** and **Range Q-itemsets.**
- **Range Q-itemsets** do not exist explicitly in the database. They are obtained by **combining exact Q-items.**
- **Exact Q-itemsets** **e.g. [(A,2),(C,6)].**
- **Range Q-itemsets** **e.g. [(A,5),(B,2,4)].**



Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- HUQIM
- Proposed Solution: CHUQI-Miner
  - Bond Measure
  - Problem Definition
  - Q-item Utility-Lists
  - How to Reduce the Search Space
  - Experimental Evaluation
- Conclusion

Bond Measure

How to measure the correlation between Q-itemsets ?

- There are several measures:
  - **Statistical tests** (Webb et al., 2010).
  - The **affinity** measure (Ahmed et al. 2011).
  - The **bond** measure (Bouasker et al., 2015).

Bond Measure

Bond correlation measure

The *conjunctive support* of a Q-itemset X, *sup(X),* in a database is the number of transactions that **contains** X.

The *disjunctive support* of a Q-itemset X, *disup(x),* in a database is the number of transactions that **contains any items** from X.

The **bond** of an itemset X is defined as:

bond(X) = sup(X) / disup(X).

The **bond measure** is **anti-monotonic.**

it can be used to **prune non-correlated Q-itemsets** with all their supersets.

Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- HUQIM
- **Proposed Solution: CHUQI-Miner**
  - Bond Measure
  - **Problem Definition**
  - **Q-item Utility-Lists**
  - **How to Reduce the Search Space**
  - **Experimental Evaluation**
- **Conclusion**

Solution 2: CHUQI-Miner

CHUQI-Miner

- **Correlated HUQIM (CHUQIM)** consists of discovering all **correlated high utility quantitative itemsets (CHUQIs)**.
- A novel algorithm named **CHUQI-Miner** (**C**orrelated **H**igh **U**tility **Q**uantitative Itemset-Miner) is proposed to:

  Find **Correlated HUQIs:**

  Input: the minimum utility threshold ***(minutil),*** the minimum bond threshold ***(minbond).***

  Output: Correlated HUQIs.

  - having a **utility** no less than ***minutil.***
  - having a **bond** no less than ***minbond.***

Problem definition

Input: Transaction database + Profit table + minimum utility threshold *(minutil)* + minimum bond threshold *(minbond)*.

Transaction database D

T_{id}	Items
*T*1	(A,2) (B,5) (C,2) (D,1)
*T*2	(B,4) (C,3)
*T*3	(A,2) (C,2)
*T*4	(A,2) (B,6) (D,1)

Profit table

Item	Unit profit
A	*3$*
B	*1$*
C	*2$*
D	*2$*

***minutil*=10** ***minbond*=0,5**

Output: CHUQIs.

Q-itemset	Utility ($)	Bond
(A, 2)	*18$*	*1*
(B, 4, 6)	*15$*	*1*
(B, 5, 6)	*11$*	*1*
(C, 2,3)	*14$*	*1*
(A,2) (C,2)	*20$*	*0,66*
(A,2) (D,1)	*16$*	*0,66*
(B,4) (C,3)	*10$*	*1*

Bond((A,2) (C,2))= 2 / 3 = 0,66.

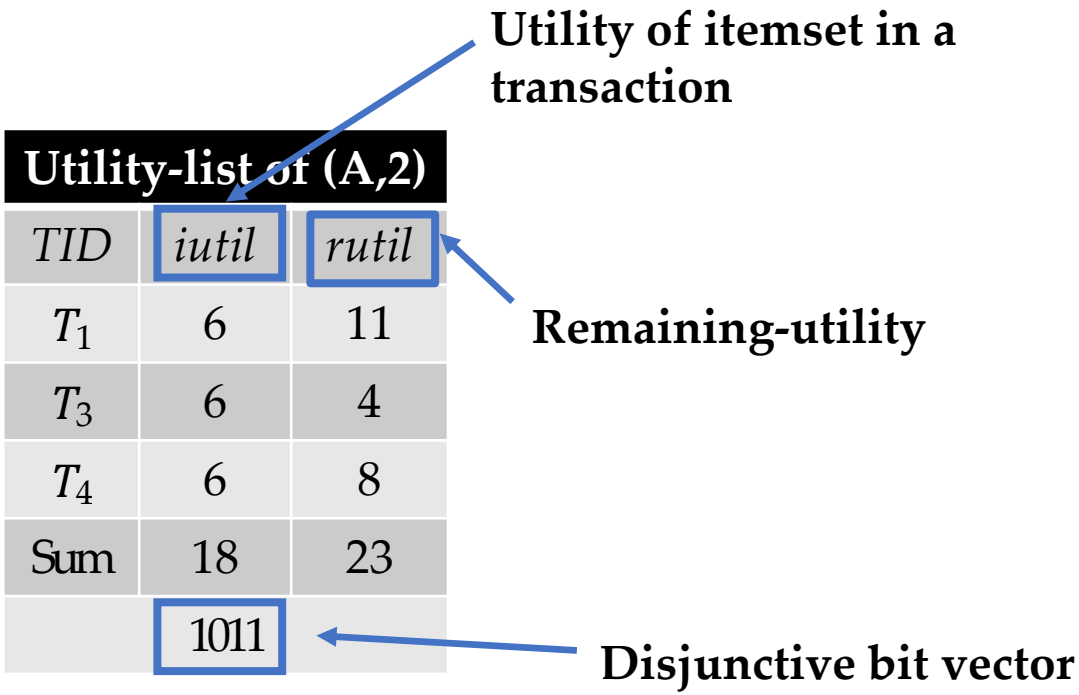
Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- HUQIM
- **Proposed Solution: CHUQI-Miner**
  - Bond Measure
  - Problem Definition
  - **Q-item Utility-Lists**
  - **How to Reduce the Search Space**
  - **Experimental Evaluation**
- **Conclusion**

Q-itemsets Utility-lists

CHUQI-Miner

- **CHUQI-Miner** is a utility-list based algorithm.
- **An utility-list** is created for each Q-itemset.



Q-itemsets Utility-lists

- Each utility-list of **CHUQI-Miner** contains a ***disjunctive bit vector*** which is used for calculating the bond measure.

T_{id}	Items
$T1$	(A,2) (B,5) (C,2) (D,1)
$T2$	(B,4) (C,3)
$T3$	(A,2) (C,2)
$T4$	(A,2) (B,6) (D,1)

disjunctive bit vector((A,2))= [1011]

disjunctive bit vector((C,2))= [1010]

disjunctive bit vector((A,2) (C,2))= [1011 or 1010]= [1011]

bond((A,2) (C,2)) = sup((A,2) (C,2)) / |1011| = 2 / 3 = 0,66.

Q-itemsets Utility-lists

- The **utility-lists** of **single Q-items** can be constructed by **scanning the database.**
- For other itemsets, it can be obtained by **joining their child Q-itemset's utility-lists (Join operation).**

Utility-list of (A,2)

TID	*iutil*	*rutil*
T_1	6	11
T_3	6	4
T_4	6	8
Sum	18	23
1011		

Utility-list of (D,1)

TID	*iutil*	*rutil*
T_1	2	0
T_4	2	0
Sum	4	0
1001		

Utility-list of [(A,2),(D,1)]

TID	*iutil*	*rutil*
T_1	8	0
T_4	8	0
Sum	16	0
1011 or 1001 = 1011		

- **Join operations** are very costly in terms of **execution time.**

Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- HUQIM
- Proposed Solution: CHUQI-Miner
  - Bond Measure
  - Problem Definition
  - Q-item Utility-Lists
  - How to Reduce the Search Space
  - Experimental Evaluation
- Conclusion

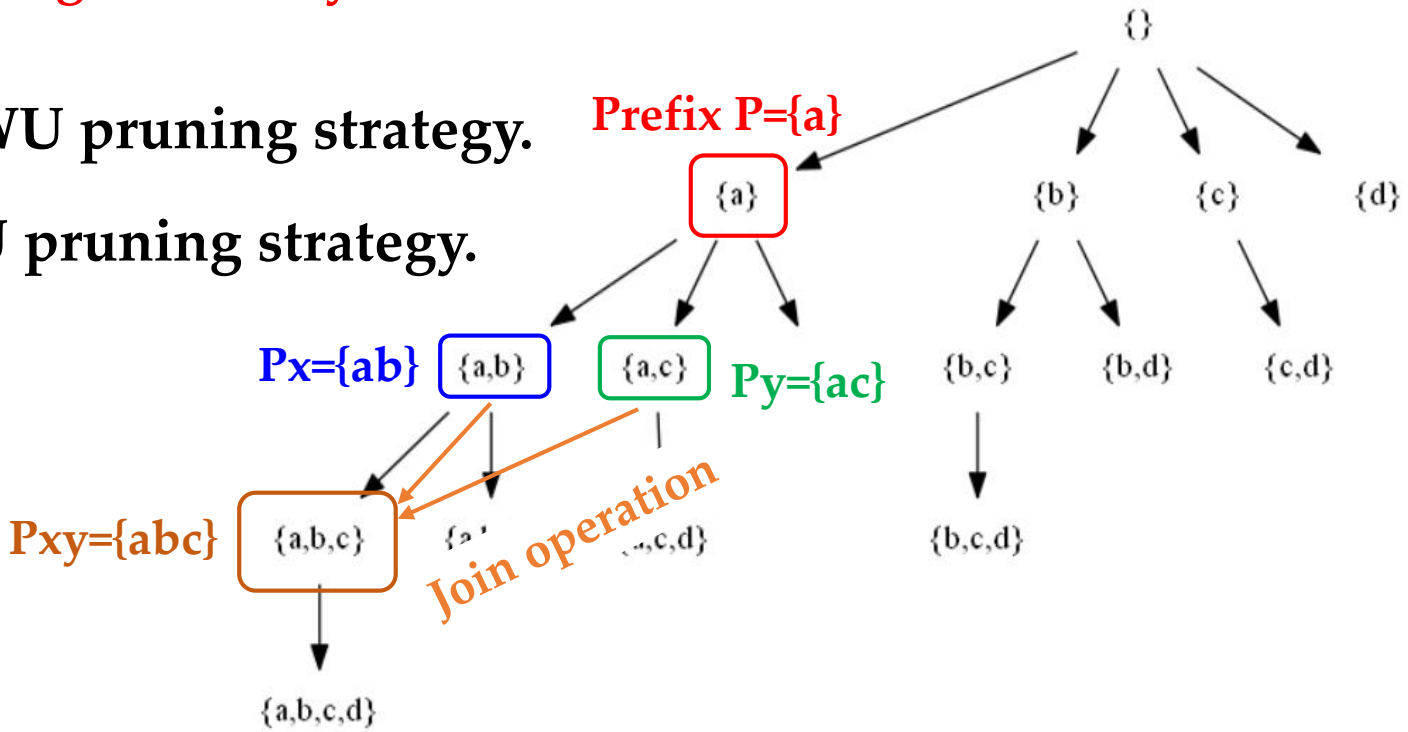
How to reduce the search space?

Prune Low Utility Q-itemsets

- **CHUQI-Miner** performs a depth-first search.
- **CHUQI-Miner** adopts **three pruning strategies** to eliminate Q-itemsets **having low utility.**

Strategy 1: TWU pruning strategy.

Strategy 2: RU pruning strategy.



Strategy 3: EQPCS & RQCPS pruning strategies.

How to reduce the search space?

Prune Weakly Correlated Q-itemsets

CHUQI-Miner adopts **three other pruning strategies** to eliminate Q-itemsets **that are weakly correlated.**

Strategy 1: Co-occurrence pruning strategy based on the bond.

- If $\frac{min\{sup(Px),sup(Py),sup(xy)\}}{max\{disup(Px),disup(Py)\}} < minbond,$ then Pxy can be eliminated with all its extensions.

Strategy 2: Pruning supersets of non correlated Q-itemsets.

- If $\frac{min\{sup(Px),sup(Py)\}}{disup(Pxy)} < minbond\ ,$ then Pxy can be eliminated with all its extensions.

Strategy 3: Early Abandoning Utility-Lists Construction.

- If $sup(Pxy) < |disup(Pxy)| * minbond,$ then Pxy can be eliminated with all its extensions.

Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- HUQIM
- Proposed Solution: CHUQI-Miner
  - Bond Measure
  - Problem Definition
  - Q-item Utility-Lists
  - How to Reduce the Search Space
  - Experimental Evaluation
- Conclusion

Experimental evaluation

Datasets' characterictics

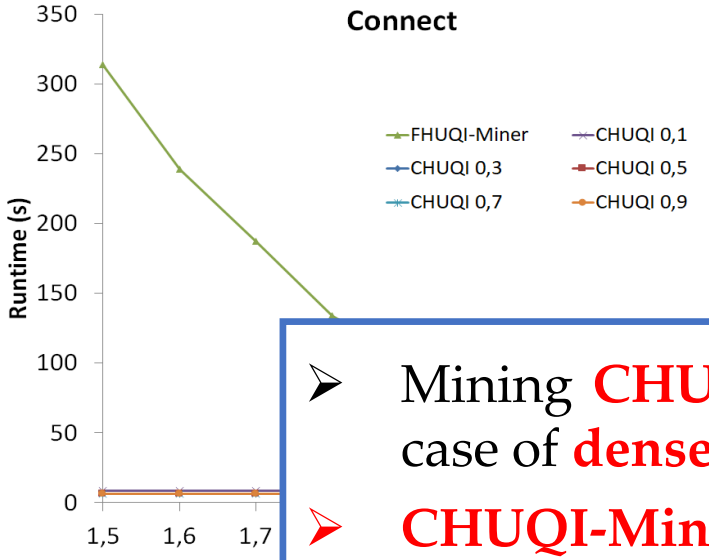
Dataset	M	N	Q	Type
Foodmart	4141	1559	1-5	Sparse
Chess	3196	75	1-5	Dense
Connect	67,557	129	1-5	Dense
Mushroom	8416	128	1-5	Dense

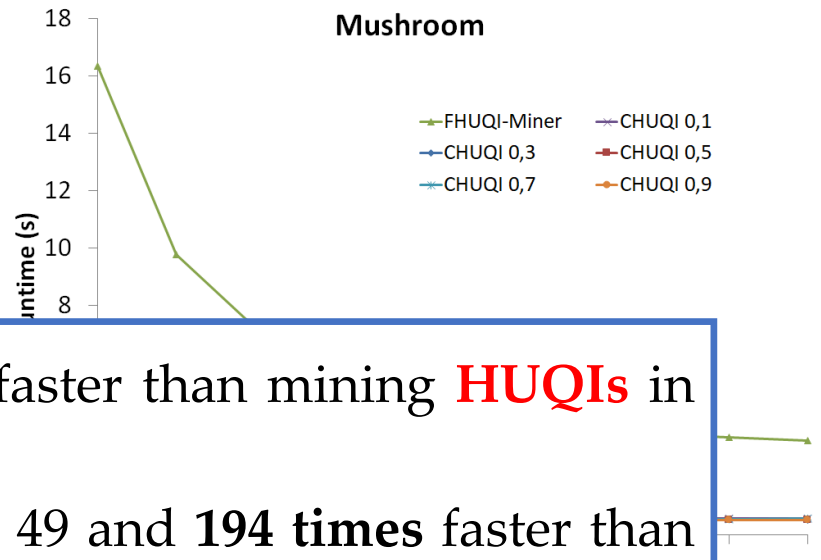
- D: Number of transactions.
- N: Number of distinct items.
- Q: Quantities range.

- CHUQI-Miner is compared with FHUQI-Miner for HUQIM.
- CHUQI-Miner was run with five different *minbond* threshold values (0.1, 0.3, 0.5, 0.7 and 0.9).

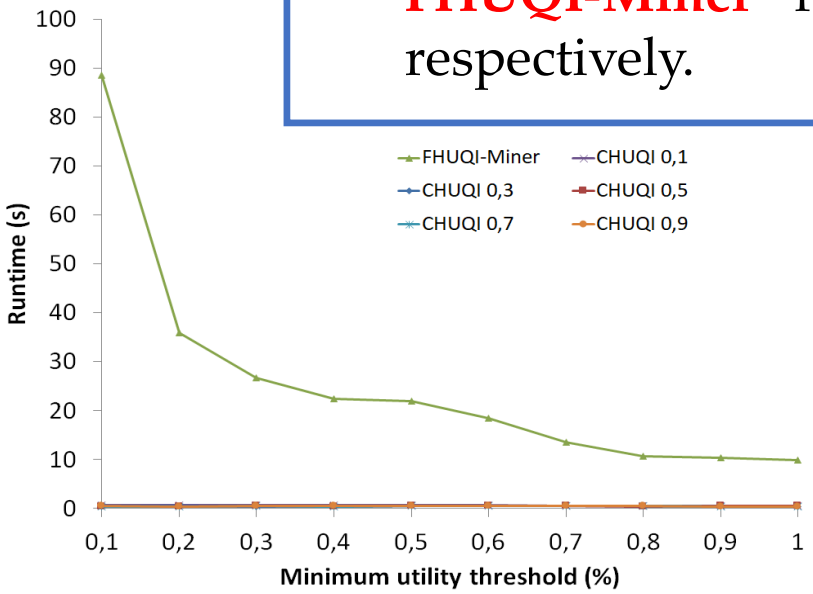
Experimental evaluation

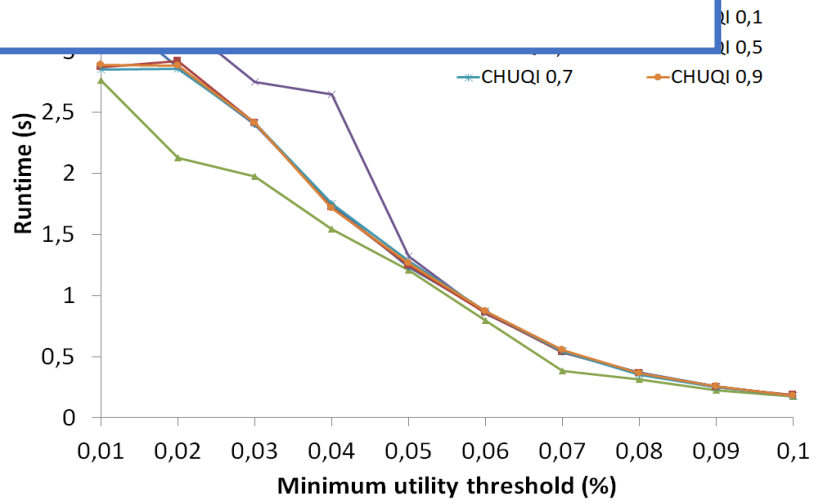
Runtime comparison on different datasets





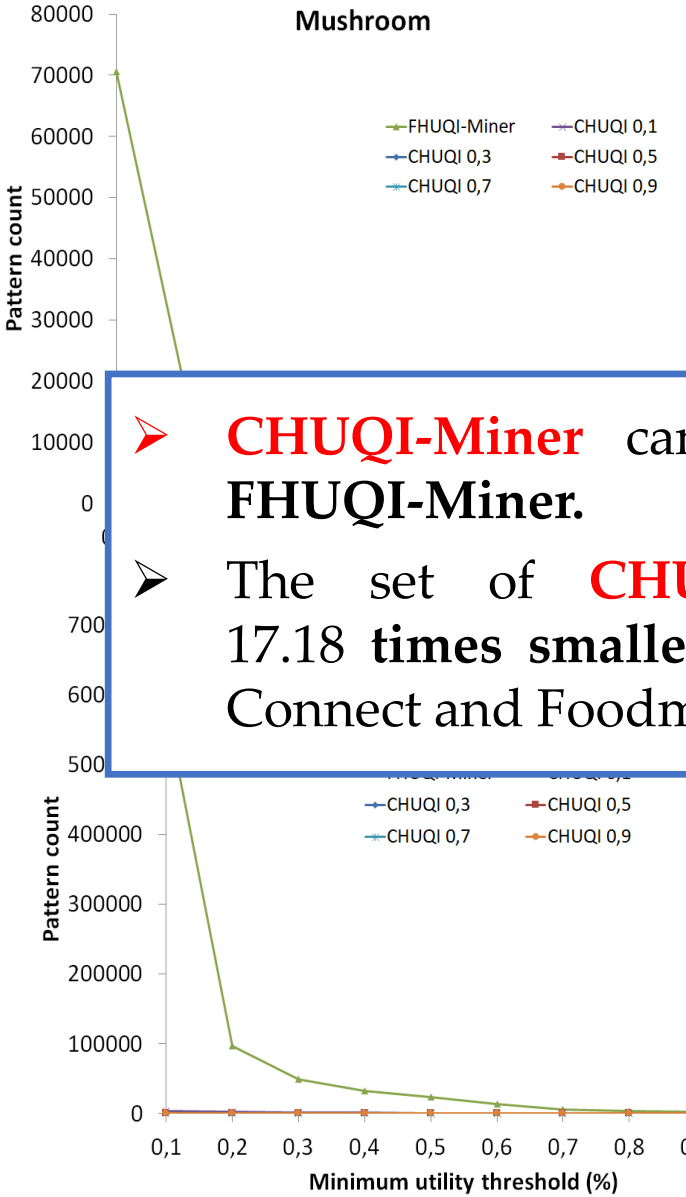
- Mining **CHUQIs** is much faster than mining **HUQIs** in case of **dense datasets**.
- **CHUQI-Miner** is up to 29, 49 and **194 times** faster than **FHUQI-Miner** for Mushroom, Connect and Chess, respectively.

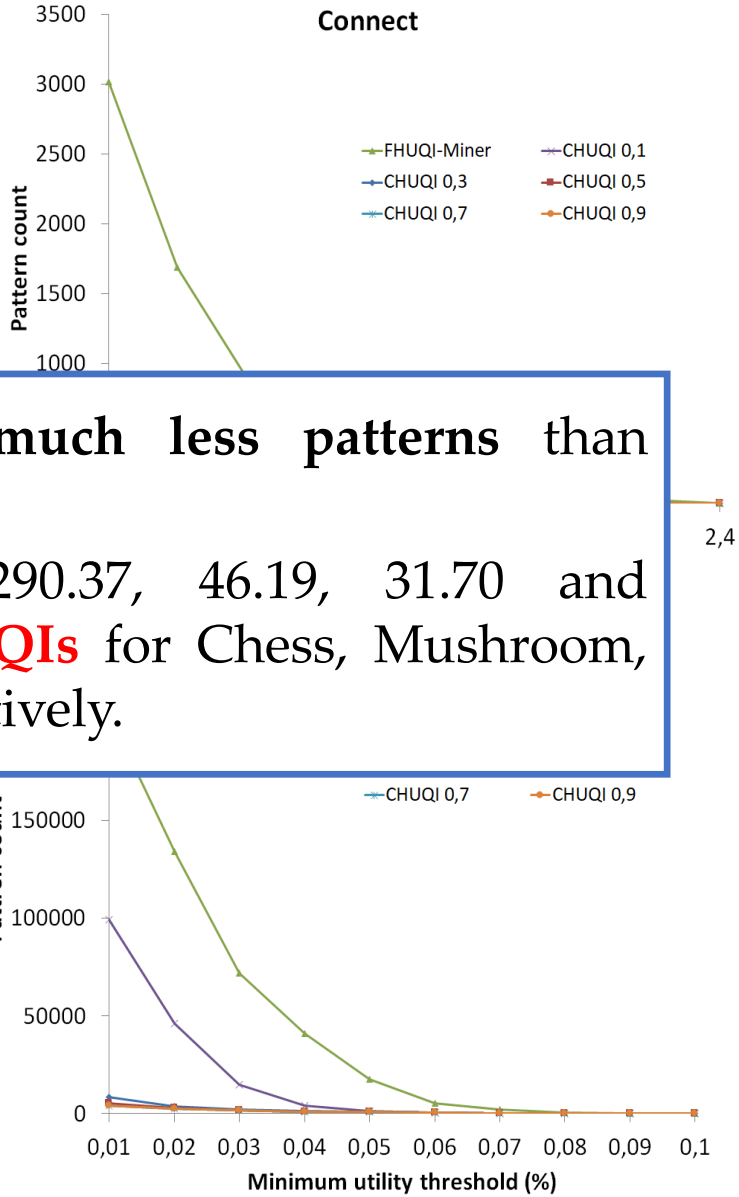




Experimental evaluation

Number of generated patterns





- **CHUQI-Miner** can output **much less patterns** than **FHUQI-Miner.**
- The set of **CHUQIs** is 290.37, 46.19, 31.70 and 17.18 **times smaller than HUQIs** for Chess, Mushroom, Connect and Foodmart, respectively.

Outline

- Introduction
  - High Utility Itemset Mining
- Limitations of HUIM
- HUQIM
- **Proposed Solution: CHUQI-Miner**
  - Bond Measure
  - Problem Definition
  - Q-item Utility-Lists
  - How to Reduce the Search Space
  - Pseudocode
  - Experimental Evaluation
- **Conclusion**

Conclusion

Contributions

- In this research work, we have proposed new algorithm CHUQI-Miner to solve an important extension of HUIM which is , Correlated High Utility Quantitative Itemsets Mining (CHUQIM).

- CHUQI-Miner integrates the **bond measure** in HUQIM and adopts various pruning strategies to prune weakly correlated and low utility Q-itemsets.

Conclusion

Future work

- Integrating other correlation measures in HUQIM such as the all-confidence, and affinity measures.
- Integrating the concept of taxonomy.

Thank You for Your Attention

Questions
Comments?