



## International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

# Mining Algorithms of Top K itemsets Based on Utility

<sup>1</sup>J.HARITHA RANI, <sup>2</sup>T.RAJESH

1 M.Tech Student, Computer Science And Engineering, haritharani92@gmail.com, G. Narayanamma Institute of Technology and Science for Women, Shaikpet, Hyderabad, Telangana State, India

2 Assistant Professor, Computer Science And Engineering, G. Narayanamma Institute of Technology and Science for Women, Shaikpet, Hyderabad, Telangana State, India

### ABSTRACT:

A group of data is being collected and located in the database and the data is being mined known as data mining. Data mining is used for the analysis of data individually. This process is being done on the basis of few techniques. These data individually is considered as an itemset. All the itemsets are organized according to the requirement of top K values which indicates the required number of itemsets. These values are being extracted with the help of Minimum utility border. This is being extracted from utility of itemsets. This has been considered by accepting two terms for the itemsets. Those are quantity and profit. These profit values are being extracted separately and the quantity is being extracted from the overall transactions in the database. These are being used for setting the minimum utility border. Initially, this Minimum utility border is set to 0. Then this is being incremented according to the process in the algorithms that are being implemented. This Minimum utility border is being used for two different algorithms. TKUM(Mining top K utility itemsets in 2 phases) and TKOM(top K utility itemsets in one phase ) these two algorithms are being implemented on the basis of Minimum utility. Then all the items in the border given by the user are considered as the high utility itemset and this process is called high utility miner.

**Keywords:** profit, quantity, Minimum utility border, K value

### I. INTRODUCTION

Mining of frequent utility of itemsets is being mostly inspired by the business users for the management of the items in the business market. All these items are being arranged in the database and extracted from the database. Users use the database for extracting and managing the transactions and database. All the services are based on transactions and itemsets in the database. The mining of itemsets is being done based on the utility and profit values of the itemsets. Here the total summation of individual itemsets is being considered from the transactional database. The extraction of itemsets is being evocated by using utility of itemsets by considering frequent itemset mining. Mining of high utility itemsets is the major task in the market for extracting the required items based on profit and quantity. If the cost of an item is high and its quantity is not matched with the items with low profit and high quantity. It should be calculated according to the market analysis. This sort of analysis is being made in many of the fields like biomedicine, stream analysis and cross marketing.

Utility of mining the itemsets is based on transactions of the database considering two aspects more importantly they are

External utility : The evocation of different items is known as external utility(eu)

Internal utility: The evocation of items in transactions is known as internal utility (iu).

$$\text{Utility of Itemset (U)} = \text{external utility (eu)} * \text{internal utility (iu)}$$

**Definition 1:** Utility of an itemset is being evocated with the help of profit and quantity of a particular item in the transactional database.

**Definition 2:** greater utility itemset is being evocated with the help of utility of itemset by removing the unused items in the collection.

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

## II. LITERATURE SURVEY

High utility itemset mining is used for evocation by considering few issues like closest items and utility of items for removing the re occurred data in the database collection. This high utility itemset collection is used to separate the analysis of items in the database.

### 2.1 Itemset Mining:

The limitations of frequent itemset mining is done by researchers to have a utility based mining approach, which allows a user to freely express his or her ideas concerning the usage of itemsets as utility values and find itemsets with high utility values greater than a threshold. In utility based mining the term utility refers to the quantitative values of user preference i.e. the utility value of an itemset is used for the importance of that itemset in the user's view. That is being approached through itemset mining. Itemset mining is being done on the basis of utility of users. It is being done with the help of quantity and profit of the item. it contains all the calculation of the particular item by considering whole database. It considers all the transactions in the database for the particular item and then totals all the quantity being transacted. Even re occurrence of data is being considered for the evaluation of itemsets. Then after the calculation the procedure of mining is being considered through grouping of data.

### 2.2 High utility itemset mining:

High utility itemsets is being mined from the transactional database. Where these high transaction itemsets are being collected from the itemsets that are being grouped. This can be processed with the help of considering even downward closure property so that the supersets and subsets do not have any link with the separated itemsets. Those sets are also being considered as a separate set of combinations. After managing all these combinations we have to separate the top rated utility itemsets. It is nothing but high utility itemsets.

## III. RELATED WORK

### 3.1 UP Tree:

UP tree is used to extract the high utility itemsets by reducing the number of candidates generated in the database. It even reduces the duplicate items in the data.

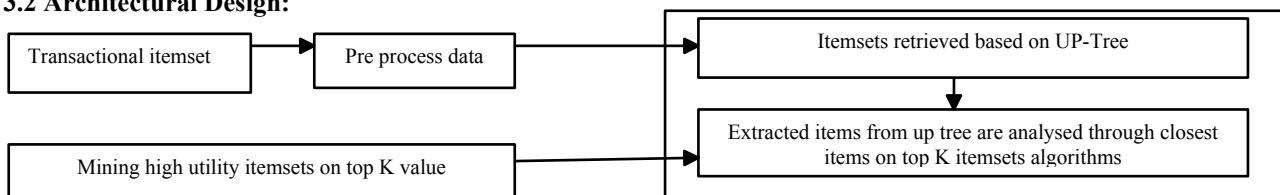
Discard global unpromising itemsets to eliminate low utility itemsets.

Discard global nodes to reduce the nodes which are closest to the root node

Discard local unpromising items to eliminate low utility itemsets in the set of itemsets

Discarding local node utilities for overcoming the closest nodes in up tree construction

### 3.2 Architectural Design:



This indicates the design flow of the mining algorithms. like processing of itemsets and process of extraction of itemsets in the database.

### 3.3 Discard Global Utility Itemsets:

This indicates that once the itemsets are considered these are to be extracted. By eliminating the property of monotone or antimonotone. these properties making all the itemsets to be extracted in the mining collection by following some priority .Excluding the unwanted itemsets by considering the global utility the itemsets will be mined.

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

## IV. BACKGROUND STUDY

This section allows relating high utility itemsets from itemsets and determines the nature of top K high utility itemsets

### 4.1 Problem definition

Let us consider a finite set of items  $FI = \{fi_1, fi_2, \dots, fi_m\}$ . Each item is associated with the positive number  $pn\{fii, DB\}$  where DB is the Transactional database. This positive number is the External utility. Transactional Database  $DB = \{t_1, t_2, \dots, t_n\}$ . Where each transaction is related to database and the subset of items in the each transaction is related to finite itemset. And the quantity is the internal utility  $qn(fij, tn)$ .

**Definition 1:** utility of itemset in the transaction  $tk$  is denoted as  $ui(fii, tn)$  is calculated with the help of profit and quantity values for the utility of itemset.

$$U_i(fii, tn) = profit(fii, DB) * qn(fij, tn).$$

TID	TRANSACTION	TU
t1	(A,1) (B,1) (D,2)	11
t2	(A,2) (C,4) (E,2)	20
t3	(D,3)	3

Table 1: An example of transactional database

ITEM	A	B	C	D	E
PROFIT	5	2	1	2	3

Table 2: Profit Table

Utility of itemset in transaction t1 for an item A is Quantity of item A in transaction t1 and profit value of A is being calculated

$$U_i(A, t1) = 5 * 1 = 5$$

**Definition 2:** The transaction of all the frequent itemsets in the particular transaction is being considered and the total of all itemsets in the particular transaction is considered. Let us consider Table1 and Table2

$$TU(t1, DB) = 5 * 1 + 2 * 1 + 2 * 2 = 11$$

**Definition 3:** The support count of an frequent itemset X is the number of transactions containing X in DB and denoted as  $SUC(X)$ .

**Definition 4:** The transaction-weighted utility of an itemset X is the sum of the transaction utilities of all the transactions containing X, which is denoted as  $TRWU(X)$  and defined as  $TRWU(X) = \sum X$  is subset of Transaction and transaction  $\in$  database  $TU(TR)$

**Definition 5:** An itemset X is a high transaction-weighted utilization itemset if  $TRWU(X) \geq \text{minimum\_utility}$ .

**Property 1:** The transaction-weighted downward closure property states that for any itemset X that is not a High Transaction Weighted Utility Itemset, All the supersets of high transaction weighted utility itemset are included in low utility.

## V. PROPOSED SYSTEM

In this system we mine the itemsets through K value. Where K value indicates that top K value is being indicated to extract the required number of itemsets. Here no threshold value is considered and this value is initially set to 0. Threshold increased automatically till the adjustment of K value is done.

The main objectives of proposed system are:

- Itemsets follows neither monotone nor anti monotone
- Proposed system should follow top K pattern mining

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

- Minimum utility threshold is not given in advance and set to 0
- Raise the minimum utility border threshold without missing top -k HUI's till K value is adjusted
- This proposed system follows mainly 2 algorithms, they are:
- TKUM<sub>base</sub> or Baseline approach (top K utility itemset mining in 2 Phase)
- TKOM Algorithm (Top K utility itemset mining in one phase)

**5.1 TKUM Algorithm:** An efficient algorithm for finding top K high utility itemsets without specifying minimum utility. indicating a basic version as TKUM<sub>base</sub> known as TKUM Algorithm containing many strategies

## 5.1.1 Baseline approach of TKUM:

This baseline approach is basically indicating K, This parameter contains the limit of extraction of itemsets from the database. This limit indicates as K itemsets. Basically Baseline approach is an approach of mining top K high utility itemsets this is being approached by following few steps, they are:

- 1) UP tree Construction
- 2) Extraction of potential high utility itemsets from the UP tree constructed
- 3) Finding top K high utility itemsets from the extracted Potential High utility itemsets

## 5.1.2. UP tree construction:

This UP tree construction is mainly used to extract the potential itemsets by making all the transactions to be existed in the form of tree following the downward closure property. this indicates the subsets of frequent itemset will all be frequent. In the UP tree there indicates each set as a node in the tree and that node is given some attributes of names to be called; they are:

- 1) Name of the itemset as a node
- 2) Support count of each node indicates the repetition of the count
- 3) To what extent the node has been utilized is also considered
- 4) The parent node of the existing node is considered and the header of the node is also considered
- 5) Linking of the nodes is considered

All these process is done for structuring of the data in the proper way of UP tree. This UP tree is being constructed by scanning the database twice leading to the elimination of itemsets which are being repeated or subsets being eliminated very efficiently. All the transactions are reorganized accordingly and inserted in the up tree structure having its root. In case if a transaction is retrieved then all the items are arranged in the descending order of transaction weighted utility. All such transactions arrangement is called Reorganized transactional utility.

## 5.1.3. Extraction of Potential high utility itemsets from the UP tree:

This is the main consideration of the survey. i.e., minimum utility border. This border has been considered for extracting the potential high utility itemsets. This minimum utility threshold is known as Border minimum utility threshold being set initially as 0 and incremented accordingly. The separation of potential high utility itemsets depends on K itemsets.

**Definition:** minimum utility: The minimum utility of an itemset  $X = \{a_1, a_2, \dots, a_m\}$  is defined as  $MIUT(X) = \sum_{i=1}^m miut(a_i) * Support\ Count(X)$ .

**Definition:** maximum utility: The maximum utility of itemset  $X = \{a_1, a_2, a_3, \dots, a_m\}$  as  $MAUT(X) = \sum_{i=1}^m maui(a_i) * support\ count(X)$ .

## 5.1.4. Extraction of Top K high utility itemsets from potential High utility itemsets:

A basic method for identifying top-k high utility itemsets from the set of Potential high utility itemsets is done. Exact utilities are examined by scanning the original database. Main part is all the candidates to be checked and candidate itemsets are extracted according to the border minimum utility after each phase.

i.e., minimum (transaction weighted utility(X), MAUT(X)) >= Border minimum utility.

## 5.2. TKOM ALGORITHM:

This is being followed by considering High utility itemset miner and utility list. When an element is given as an itemset then it is being considered through utility list without scanning the database.

### 5.2.1. Construction of Utility List Structure:

In TKOM algorithm each item is related to utility list which is being associated with initial utility lists. The entire

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

database is scanned for construction of utility list and then no need to scan database if same itemsets and transactions is being used. This process is done only once. Firstly utility values and Transaction weighted utility of every itemset is calculated and then arranged in descending order known as utility set. This utility set has three attributes known as transaction id (trid), internal utility (inutil), remaining utility (reutil).

## VI. PSEUDO CODE

### 6.1. TKUM Algorithm:

```
Minimum utility border-> 0,  
TOP k utility ->null,  
Potential itemsets-> null.  
Database D, transactional database T is considered with transactions T and itemsets as  $X=\{a_1...a_m\}$   
Step 1: UP tree is constructed  
Step 2 :Apply Up tree to generate potential high utility itemsets  
Step 3:Estimated utility itemsets is separated from potential itemsets  
if {  
Estimated utility(X) >= minimum utility border and maximum utility of an itemset(MAUT(X) >= minimum utility  
border)  
{  
X and Minimum (Estimated utility(X), MAUT(X);  $C \leftarrow C \cup X$ ; □  
If { (MIUT(X)>=minimum utility border) □  
{ Raise minimum utility border by the strategy most common count; □  
Minimum utility border <- Most common count(MIUT(X), top K – MIUT-limit);  
}  
}  
}
```

### 6.2.TKOM ALGORITHM:

```
set of itemsets from the utility list .  
Border minimum utility threshold -> 0  
Utility list to be inserted -> null  
Step 1: let x and y are variables of itemsets  
Step 2:  $X=(x_1,x_2,...,x_m)$  belongs to a set  
do  
{  
If (sum(minimum utility(X)) >= border minimum utility)  
Raise min utility border by RUC strategy;  
}  
If(sum(X inutil ) +sum(X reutil)>= border minimum utility)  
{  
Utility entries will be null;  
For each item sets it creates some association rules for x,y,z datasets ;  
Do  
{  
 $Z \leftarrow X$  association Y;  
Utilitylist(Z) <- construct (utilitylist(p),x,y, Utility List Search[p]);  
Class [x] <- class [x] associates Z;  
Utility List Search[x] <- Utility List Search[x] associates Utility List[z];  
}  
Top K High Utility Itemset search(X, Utility List Search[x],class[x], border utility, top k clear list);  
}  
}
```

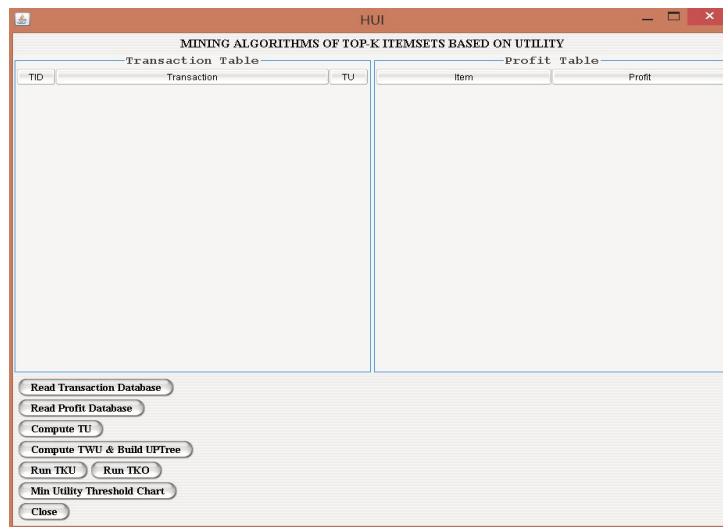
# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

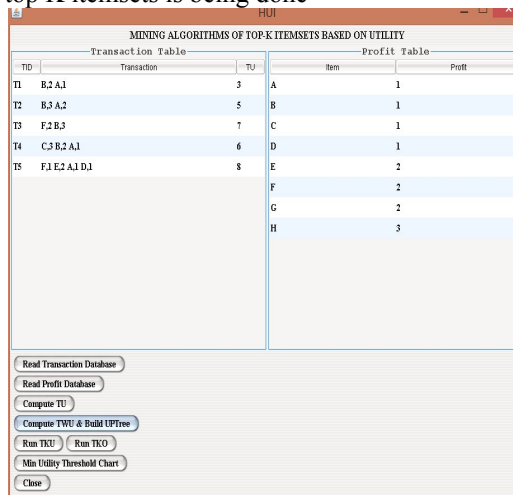
Vol. 4, Issue 11, November 2016

## VII. RESULTS

The experimental study involves the use of both the algorithms in mining the top K high utility itemsets leading to the reduction of space and time complexity occurred through the existing system. This approach is being analysed through a dataset that is implemented experimentally by the developer.



The GUI of mining top K High utility itemsets considered where the transactional database and profit table is considered and all the calculation of top K itemsets is being done

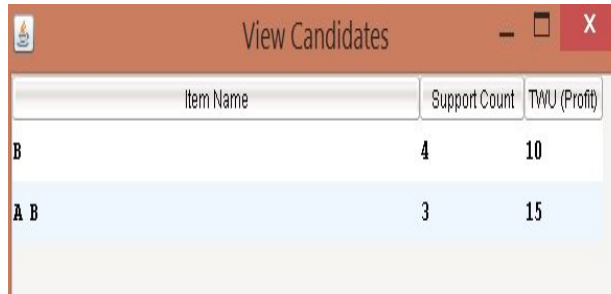


This is the GUI of the project where transactional and profit table is considered. And sample of transactional and profit table is read through the database by mysql.

# International Journal of Innovative Research in Computer and Communication Engineering

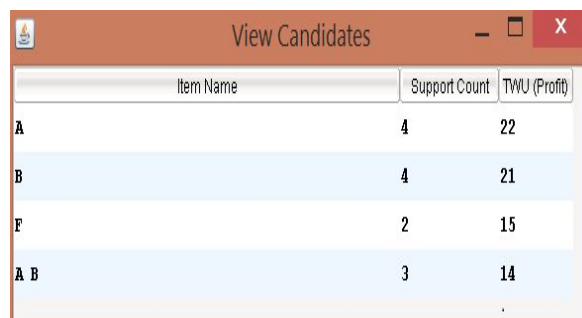
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016



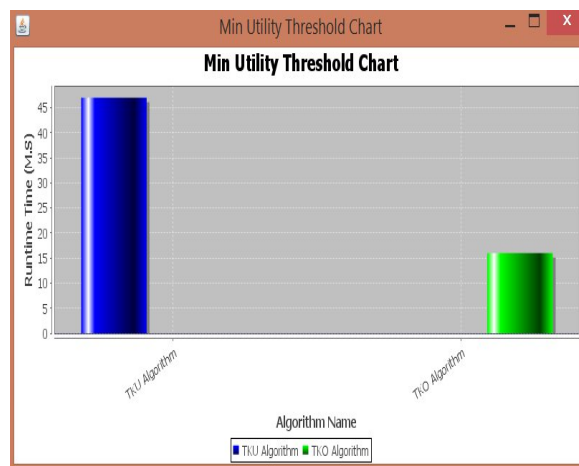
Item Name	Support Count	TWU (Profit)
B	4	10
A B	3	15

TKU candidates extracted when k value is given as 1. The threshold value is adjusted. This is more appropriate because it follows the proper data structure. Limited values are extracted.



Item Name	Support Count	TWU (Profit)
A	4	22
B	4	21
F	2	15
A B	3	14

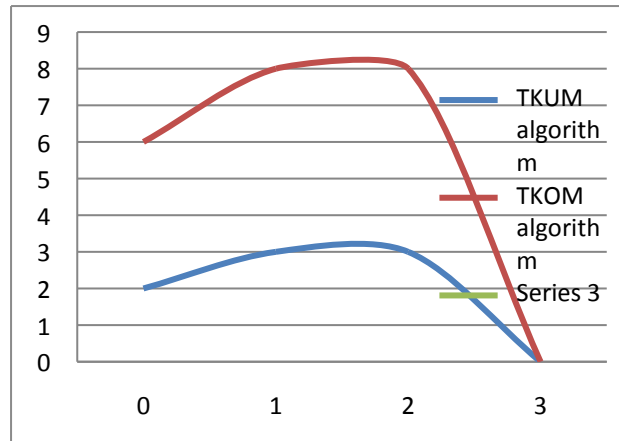
TKO candidates are extracted. Here this is used to extract the candidates by arranging the items in descending order and through utility list.



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016



Time complexity and space complexity chart has been evaluated for the proper approach of the proposed algorithms in mining top K high utility itemsets.

## VIII. CONCLUSION AND FUTURE WORK

The experimental results for the proposed system is that the time complexity is confined to some extent by reducing the repetition of threshold value search and reduced the complexity for the evaluators to enter the unknown value of threshold for the extraction of itemsets. We have considered the experiment on experimental database. This should be even exposed for the external databases to be implemented and reduce the complexity of extracting the high utility itemsets.

## REFERENCES

1. Vincent S. Tseng, Senior Member, Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu, Fellow, efficient algorithms for mining top K high utility itemsets, IEEE, Issue No. 01 - Jan. (2016 vol. 28) ISSN: 1041-4347, 2015
2. C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee. Efficient tree structures for high utility pattern mining in incremental databases. In IEEE Transactions on Knowledge and Data Engineering, Vol. 21, Issue 12, pp. 1708-1721, 2009.
3. R. Agrawal and R. Srikant.: Fast algorithms for mining association rules. In Proc. of the 20th Int'l Conf. on Very Large Data Bases, pp. 487-499, 1994.
4. Y. C. Li, J. S. Yeh, and C. C. Chang.: Isolated items discarding strategy for discovering high utility itemsets. In Data & Knowledge Engineering, Vol. 64, Issue 1, pp. 198-217, Jan., 2008.
5. Y.-C. Li, J.-S. Yeh, and C.-C. Chang. Isolated items discarding strategy for discovering high utility itemsets. In Data & Knowledge Engineering, Vol. 64, Issue 1, pp. 198-217, Jan., 2008.
6. M.R. Karim, B.S. Jeong, and H.J. Choi, "A MapReduce Framework for Mining Maximal Contiguous Frequent Patterns in Large DNA Sequence Datasets", IETE Technical Review, Vol. 29, no. 2, pp. 162-8, Mar-Apr, 2012
7. Md. Rezaul Karim<sup>1</sup>, Chowdhury Farhan Ahmed<sup>2</sup>, Byeong-Soo Jeong<sup>1</sup>, Ho-Jin Choi<sup>3</sup>, "An Efficient Distributed Programming Model for Mining Useful Patterns in Big Datasets", IETE TECHNICAL REVIEW | Vol 30 | ISSUE 1 | JAN-FEB 2013
8. J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases," Proc. 21th Int'l Conf. Very Large Data Bases, pp. 420-431, Sept. 1995.
9. A. Erwin, R.P. Gopalan, and N.R. Achuthan, "Efficient Mining of High Utility Itemsets from Large Datasets", T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 554-561, 2008. © SpringerVerlag Berlin Heidelberg 2008.
10. Sudip Bhattacharya, Deepty Dubey, "High utility itemset mining, International Journal of Emerging Technology and advanced Engineering", ISSN 2250- 2459, Volume 2, issue 8, August 2012.
11. Ming-Yen lin, Tzer-Fu Tu, Sue-Chen Hsueh, "High utility pattern mining using the maximal itemset property and, lexicographic tree structures", Information Science 215(2012) 1-14.
12. Y. Liu, W. Liao, and A. Choudhary, "A Fast High Utility Itemsets Mining Algorithm," Proc. Utility-Based Data Mining Workshop, 2005