# A Technique for mining high utility items from a transaction data base

[1]ANKIT REDWAL, [2]ANIL PATIDAR

[1]M.Tech. Scholar, [2]Professor
ACROPOLIS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

***ABSTRACT*: Data Mining, also called knowledge Discovery in Database, is one of the latest research area, which has emerged in response to the Tsunami data or the flood of data, world is facing nowadays. It has taken up the challenge to develop techniques that can help humans to discover useful patterns in massive data. One such important technique is utility mining. This paper will present an updated technique for mining high utility items from a transaction data set. It will be making use of the concept of hash map facility. The data destruction will also be performed.**

***Keywords*: Data Mining, KDD Process, High Utility Mining, Minimum Utility, Hash Map, Data Destruction.**

## 1. INTRODUCTION:

The use of data mining [1,2] is placed in various decisions making task, using the analysis of the different properties and similarity in the different properties can help to make decisions for the different applications. Among them the prediction is one of the most essential applications of the data mining and machine learning. This work is dedicated to investigate about the decision making task using the data mining algorithms. Data mining [3][4] is associated with extraction of non trivial data from a large and voluminous data set. Figure 1 shows the general working of data mining.

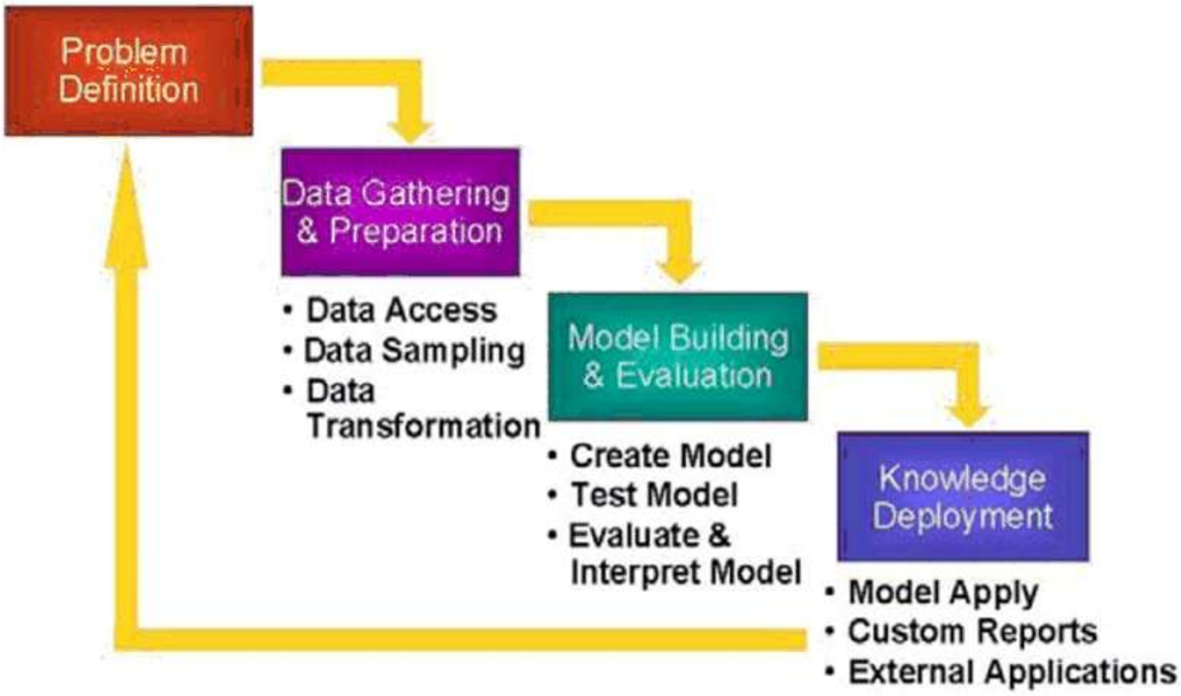

**Figure 1: Data Mining**

In utility mining [5] we concentrate on utility value of itemset while in frequent item set mining we concentrate that how frequently items appears in transactional database. With the help of following example describe in table 1, can easily differentiate utility mining and frequent item set mining:-

**Table 1: Transactional Database D1**

| **Transaction I** | **Quantity of item sold in Transaction** | | |
|---|---|---|---|
| T1 | 0 | 0 | 1 |
| T2 | 2 | 0 | 2 |
| T3 | 1 | 1 | 4 |
| T4 | 0 | 1 | 1 |
| T5 | 5 | 1 | 3 |

Unit profit related with each item is described in table 2 as follows:

**Table 2: unit profit associate with items**

| **Item Name** | **Unit profit** |
|---|---|
| A | 6 |
| B | 120 |
| C | 45 |

Now with the help of internal utility, external utility and how many times item or itemset appears in transaction, we can calculate support and profits which describe in table 3 as follows:

**Table 3: Support and profits for all items**

| Itemset | Support (%) | Profit (INR) |
|---|---|---|
| A | 60 | 48 |
| B | 60 | 360 |
| C | 100 | 495 |
| AB | 40 | 276 |
| AC | 60 | 768 |
| BC | 60 | 720 |
| ABC | 40 | 456 |

If [5] minimum support = 40 % only A, B, C, AC, BC qualify as frequent itemsets. ({ABC}) = (1×6+1×120 +4×45) + (5×6+ 1×120+3×45) = 456.If we specified user threshold value =310 then ABC is a high utility itemset but it is not a frequently accessible itemset.

Some FP tree based and other tree based methods for high utility mining were proposed in [6][7][8]. All these methods were simple. [9] proposed improved LRU based technique. Liu et al. [10], also they follow the process of two phase candidate generation. The work done in [11] proposed an isolated item discarding strategy. If any size k item set does not contain an item I then item I is termed as an isolated item. Authors in [12] proposed a projection based method for mining high utility items. This is improvement of two phase algorithm. It speeds up the execution of two phase algorithm. Authors in [13] proposed a hybrid algorithm, a combination of antimonotonicity of TWU and pattern growth approach. Work done in [14] proposed a FP tree based algorithm, this algorithm uses a tree to maintain the TWU information.

Apriori algorithm for mining high utility items sets was proposed in [15]. It first generates all the probable high utility candidates. Then this algorithm makes use of minimum utility threshold to prune infrequent items. Effective [16] disclosure of item sets with high utility like benefits manages the mining high utility item sets from an exchange database Although various important methodologies have been proposed as of late, these calculation acquire the issue of creating an extensive number of competitor item sets for high utility item sets and most likely debases the mining execution as far as execution time and memory space. Mining [17] exceptionally used thing sets from a value-based dB intends to find the thing sets with high utility as benefits. In spite of the fact that various Algorithms have been created yet they bring about the issue as it produce huge arrangement of applicant Item sets likewise require number of database output.

## 2. PROBLEM DOMAIN:

The practical usefulness of the frequent itemset mining is limited by the significance of the discovered itemsets .While mining literature has been exclusively focused on frequent itemsets, in many practical situations rare ones are of higher interest .For example in medical databases rare combinations of symptoms might provide useful insights for the physicians about the cause of the disease. So during the mining process we should not be prejudiced to identify either frequent or rare itemsets but our aim should be identify itemsets which are more utilizable to us. In other words our aim should be in indentifying itemsets which have comparatively higher utilities in the database, no matter whether these identified itemsets are frequent itemsets, rare itemsets or neither of them. This leads to the inception of a new approach in data mining which is based on the concept of itemset utility called as utility mining.

The limitations of frequent or rare itemset mining motivated researchers to conceive a utility based mining approach, which allows a user to conveniently express his or her perspectives concerning the usefulness of itemsets as utility values and then find itemsets with high utility values higher than a threshold .In utility based mining the term utility refers to the quantitative representation of user preference i.e. the utility value of an itemset is the measurement of the importance of that itemset in the users perspective. For e.g. if a sales analyst involved in some retail research needs to find out which itemsets in the stores earn the maximum sales revenue for the stores he or she will define the utility of any itemset as the monetary profit that the store earns by selling each unit of that itemset.

**Work done by Liu et al. is based on the concept of a tree construction based method. It does not generate candidates. First a tree is constructed and then DFS (Depth First Search) technique is used to visit the nodes of the tree to calculate the utility of items. But construction of tree takes O(n) time. Also searching element in a tree requires O(logn) time. Deletion requires O(logn) time. So there is a scope to reduce these times by using some other appropriate data structure**

## 3. SOLUTION APPROACH:

We will use hash map structure for storing the transaction data base and profit data base together. Hash-map data structure is efficient in storing data that has 2 parts. Every element has a *key* and a *value* pair. For mining high utility itemset, we require an item and its corresponding profit. Previous work have used list for storing the item-profit. It becomes a little complex to identify the exact profit of an item in list, though it is most accurate, it is not faster in computation. Hash map on the other hand, stores the information in key value pair. This increases the speed of access every time the database is scanned.

First transaction data base will be converted into a hash map. Then weighted utility of each item will be calculated and useless items will be pruned. Then we will prepare a list for each item. Construction of hash map will require (O1) time. Insertion deletion and search operation in hash map also require O(1) time.

### 3.1 PROPOSED ALGORITHM:

Step 1: Input:

- A Transaction data Base T and Profit table P
- Minimum utility value

Step 2: in this step, the transaction data base is converted in to an equivalent hash map H.

Step 3: Scan the hash map H and calculate the weighted transaction utility of each item. If weighted transaction utility of an item is more then threshold then keep that item in the list of high utility item set.
Step 4: In this step, we eliminate all those items from the hash map H, whose utility is less than the minimum utility. Then hash map H will be transformed into a compressed hash map H1. Now this H1 will be used in finding the high utility item sets of greater size.
Step 5: items in hash map H1 are sorted in the descending order of their transaction utility.
Step 6: From item sets of size K, we recursively create candidates of greater size as follows:
From candidate of size K, we recursively create candidates of greater size as follows:

- For each itemset I1 and I2 of level k
- we compare items of itemset1 and itemset2. If they have all the same k-1 items and the last item of itemset1 is smaller than the last item of itemset2, we will combine them to generate a candidate
- Calculate weighted transaction utility of itemset using the compressed hash map H1
- if the weighted transaction utility is high enough
- add it to the set of HUI (High utility items sets)
- Continue this process until there are candidates to combine

Step 7: Return all high utility itemsets found
Step 8: End of process.

## 4. RESULTS COMPARISON:

The proposed algorithm is implemented and results are compared. Partial Retail utility data set is used. The results obtained are shown below in figures:

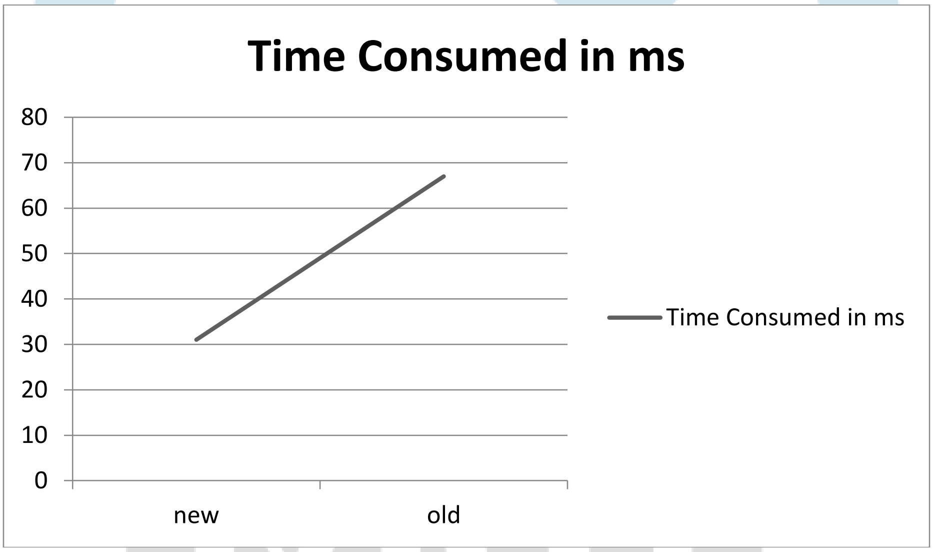

Figure. 1 Depicts the Time Consumption Comparison

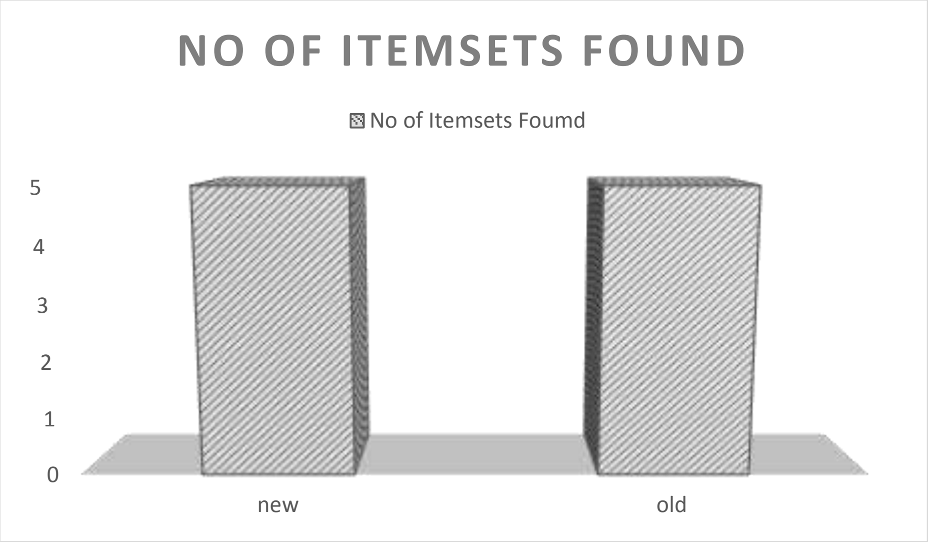

Figure. 2 Depicts the Result Comparison

As shown in fig.1 and fig.2 Comparison based on the existing and proposed algorithm. This experiment use a Traffic Accidents Data Set.

## 5. CONCLUSION:

High utility frequent pattern mining has a wide range of real world applications. That’s why it is one of the most favorite topic of research. Utility mining helps in mining of items which are worthy. This paper proposed an updated method to find high utility item sets from a transaction data set. The proposed method makes use of hash map table for storage of item and the associated profit. Useless items are eliminated in the initial stage of the mining process. Experimental results have proven that the proposed algorithm is taking less time in mining utility items from a transaction data set.

## References:

[1] Tan P.-N., Steinbach M., and Kumar V. ―Introduction to data mining, Addison Wesley Publishers‖. 2006

[2] Fayyad U. M., Piatetsky-Shapiro G. and Smyth, P. ―Data mining to knowledge discovery in databases, AI Magazine‖. Vol. 17, No. 3, pp. 37-54, 1996.

[3] https://www.sas.com/en_us/insights/analytics/data-mining.html

[4] 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.

[5] A. Erwin, R. P. Gopalan, and N. R. Achuthan. Efficient mining of high utility itemsets from large datasets. In *Proc. of PAKDD 2008, LNAI 5012*, pp. 554-561.

[6] Y. G. Sucahyo and R. P. Gopalan. "CT-ITL: Efficient Frequent Item Set Mining Using a Compressed Prefix Tree with Pattern Growth". Proceedings of the 14th Australasian Database Conference, Adelaide, Australia, 2003.

[7] Y. G. Sucahyo and R. P. Gopalan. "CT-PRO: A Bottom Up Non Recursive Frequent Itemset Mining Algorithm Using Compressed FP-Tre Data Structure‖. In proc Paper presented at the IEEE ICDM Workshop on Frequent Itemset Mining Implementation (FIMI), Brighton UK, 2004.

[8] A.M.Said, P.P.Dominic, A.B. Abdullah. ―A Comparative Study of FP-Growth Variations‖. In Proc. International Journal of Computer Science and Network Security, VOL.9 No.5 may 2009.

[9] ZHOU Jun, CHEN Ming, XIONG Huan A More Accurate Space Saving Algorithm for Finding the Frequent Items.IEEE-2010.

[10] Y. Liu, W. Liao, and A. Choudhary, “A fast high utility itemsets mining algorithm,” in Proc. Utility-Based Data Mining Workshop SIGKDD, 2005, pp. 253–262.

[11] Y.-C. Li, J.-S. Yeh, and C.-C. Chang, “Isolated items discarding strategy for discovering high utility itemsets,” Data Knowl. Eng., vol. 64, no. 1, pp. 198–217, 2008.

[12] G.-C. Lan, T.-P. Hong, and V. S. Tseng, “An efficient projectionbased indexing approach for mining high utility itemsets,” Knowl. Inf. Syst., vol. 38, no. 1, pp. 85–107, 2014.

[13] A. Erwin, R. P. Gopalan, and N. R. Achuthan, “Efficient mining of high utility itemsets from large datasets,” in Proc. 12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2008, pp. 554–561.

[14] V. S. Tseng, B.-E. Shie, C.-W. Wu, and P. S. Yu, “Efficient algorithms for mining high utility itemsets from transactional databases,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 8, pp. 1772–1786, Aug. 2013

[15] Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu, Fellow, IEEE, Vincent S. Tseng, “Efficient algorithms for mining the concise and lossless representation of high utility item sets,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 3, pp. 726–739, Mar. 2014.

[16] Miss. A. A. Bhosale , S. V. Patil, Miss. P. M. Tare, Miss. P. S. Kadam“High Utility Item sets Mining on Incremental Transactions using UP-Growth and UP-Growth+ Algorithm”:

[17] Switi Chandrakant Chaudhari, Vijay Kumar Verma, Mining High Utility Item Set From Large Database:- ARecent Survey , International journal of Emerging Technology and Advanced Engineering, Website:. www.ijetae.com(ISSN 2250-2459,ISO 9001:2008 Certified Journal, Volume 3, Issue 5, May 2013)