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

Ankit Redwal, Dr. Amit Khare, Prof. Rahul Moriwal

Abstract- **Information Mining, additionally called learning Discovery in Database, is one of the most recent research territory, which has risen in light of the Tsunami information or the surge of information, world is confronting these days. It has responded to the call to create strategies that can assist people with discovering helpful patterns in enormous information. One such significant system is utility mining. This paper will show a refreshed procedure for mining high utility things from an exchange informational collection. It will utilize the idea of hash map office. The information demolition will likewise be performed.**

Index Terms- Data Mining, KDD Process, High Utility Mining, Minimum Utility, Hash Map, Data Destruction.

I. INTRODUCTION

The utilization of information mining [1,2] is put in different choices making task, utilizing the examination of the various properties and similitude in the various properties can settle on choices for the various applications. Among them the expectation is a standout amongst the most fundamental utilizations of the information mining and AI. This work is committed to explore about the basic leadership assignment utilizing the information mining calculations. Information mining [3][4] is related with extraction of non paltry information from an enormous and voluminous informational collection. Figure 1 demonstrates the general working of information mining.

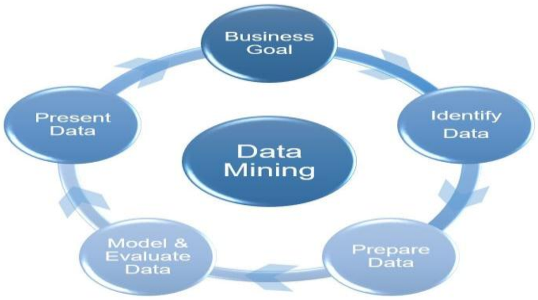

Figure 1: Data Mining

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 strategies for high utility mining were proposed in [6][7][8]. Every one of these strategies were basic. [9] proposed improved LRU based method. Liu et al. [10], likewise they pursue the procedure of two stage competitor age. The work done in [11] proposed a segregated thing disposing of methodology. On the off chance that any size k thing set does not contain a thing I, at that point thing I is named as a secluded thing. Creators in [12] proposed a projection based technique for mining high utility things. This is improvement of two stage calculation. It accelerates the execution of two stage calculation. Creators in [13] proposed a half and half calculation, a mix of antimonotonicity of TWU and example development approach. Work done in [14] proposed a FP tree based calculation, this calculation utilizes a tree to keep up the TWU data.

Apriori calculation for mining high utility things sets was proposed in [15]. It initially creates all the likely high utility hopefuls. At that point this calculation utilizes least utility limit to prune rare things. Successful [16] divulgence of thing sets with high utility like advantages deals with the mining high utility thing sets from a trade database Although different significant techniques have been proposed starting late, these count get the issue of making a broad number of contender thing sets for high utility thing sets and in all probability corrupts the mining execution to the extent execution time and memory space. Mining [17] particularly utilized thing sets from an esteem based dB expects to discover the thing sets with high utility as advantages. Disregarding the way that different Algorithms have been made yet they realize the issue as it produce gigantic game plan of candidate Item sets in like manner require number of database yield.

2. PROBLEM DOMAIN

The viable convenience of the successive itemset mining is constrained by the centrality of the found itemsets .While mining writing has been only centered around continuous itemsets, in numerous reasonable circumstances uncommon ones are of higher enthusiasm .For instance in restorative databases uncommon blends of indications may give helpful bits of knowledge to the doctors about the reason for the sickness. So during the mining procedure we ought not be preferential to recognize either visit or uncommon itemsets however our point ought to be distinguish itemsets which are progressively utilizable to us. As such our point ought to be in indentifying itemsets which have relatively higher utilities in the database, regardless of whether these recognized itemsets are visit itemsets, uncommon itemsets or neither of them. This prompts the commencement of another methodology in information mining which depends on the idea of itemset utility called as utility mining.

The constraints of continuous or uncommon itemset mining inspired analysts to imagine an utility based mining approach, which enables a client to helpfully express his or her viewpoints concerning the value of itemsets as utility qualities and afterward find itemsets with high utility qualities higher than an edge .In utility based mining the term utility alludes to the quantitative portrayal of client inclination for example the utility estimation of an itemset is the estimation of the significance of that itemset in the clients point of view. For example in the event that a business examiner engaged with some retail research needs to discover which itemsets in the stores gain the most extreme deals income for the stores the person in question will characterize the utility of any itemset as the money related benefit that the store wins by selling every unit of that itemset.

Work done by Liu et al. depends on the idea of a tree development based strategy. It doesn't create applicants. Initial a tree is built and afterward DFS (Depth First Search) system is utilized to visit the hubs of the tree to ascertain the utility of things. Be that as it may, development of tree takes O(n) time. Likewise seeking component in a tree requires O(logn) time. Cancellation requires O(logn) time. So there is an extension to diminish these occasions by utilizing some other proper information structure.

3. SOLUTION APPROACH

We will utilize hash guide structure for putting away the exchange information base and benefit information base together. Hash-map information structure is proficient in putting away information that has 2 sections. Each component has a key and an esteem pair. For mining high utility itemset, we require a thing and its relating benefit. Past work have utilized rundown for putting away the thing benefit. It turns into somewhat complex to recognize the careful benefit of a thing in rundown, however it is most exact, it isn't quicker in calculation. Hash map then again, stores the data in key esteem pair. This builds the speed of access each time the database is checked.

First exchange information base will be changed over into a hash map. At that point weighted utility of everything will be determined and futile things will be pruned. At that point we will set up a rundown for everything. Development of hash guide will require (O1) time. Inclusion erasure and hunt task in hash map additionally 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:

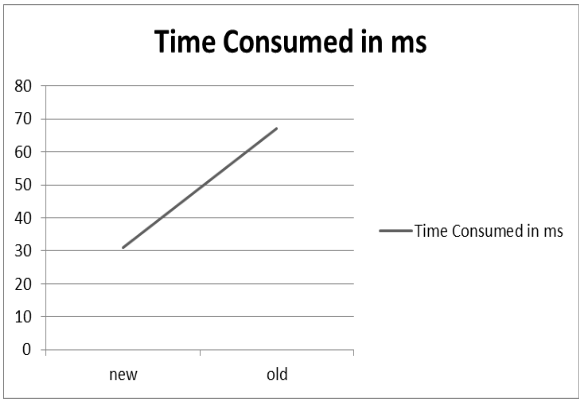

Figure. 1 Depicts the Time Consumption Comparison

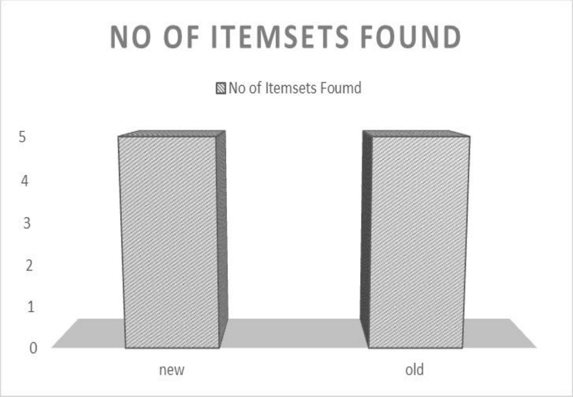

Figure. 2 Depicts the Result Comparison

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 regular example mining has a wide scope of genuine applications. That is the reason it is a standout amongst the most loved subject of research. Utility mining helps in mining of things which are commendable. This paper proposed a refreshed technique to discover high utility thing sets from an exchange informational index. The proposed strategy utilizes hash guide table for capacity of thing and the related benefit. Futile things are disposed of in the underlying phase of the mining procedure. Test results have demonstrated that the proposed calculation is taking less time in mining utility things from an exchange informational index.

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)