AN ITEM ELIMINATION BASED TECHNIQUE FOR MINING HIGH UTILITY ITEMS FROM A DATA SET

Gitanjali Soni, Prof. Balwant Prajapat

[1]Research Scholar, [2]Assistant Professor

***Abstract*: The data mining and their different applications are becomes more popular now in these days a number of large and small scale applications are developed with the help of data mining techniques i.e. predictors, regulators, weather forecasting systems and business intelligence. There are two kinds of model are available for namely supervised and unsupervised. The performance and accuracy of the supervised data mining techniques are higher as compared to unsupervised techniques therefore in sensitive applications the supervised techniques are used for prediction and classification. this paper presents a high utility item set mining technique. In this technique, the useless patterns are removed at the initial stage of mining. So it is helping in getting less time consumption.**

1. Introduction:

In utility mining [3,4] we concentrate on utility value of itemset while in frequent item set mining we concentrate that how frequently items appears in transactional database.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into needful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of the analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [2].

Data: Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- non-operational data, such as industry sales, forecast data, and macro-economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

Information: The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when [1].

Knowledge: Information can be converted into knowledge about historical patterns and 5future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data Warehouses: Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining [2].

Some methods were proposed for mining high utility item or itemsets from the databases, such as UMining [9], Two-Phase [7,8], IIDS [6] and IHUP [5]. UMining algorithm [9] proposed by Yao et al. used an estimation method to prune candidate itemset in memory. Also it is shown to have good performance but it cannot capture the complete set of high utility itemsets since some high utility patterns may be pruned during the process.

2. Basic Concepts:

The basic definitions are as follows:

Definition 1: A frequent itemset is a set of items that appears at least in a pre-specified number of transactions. Formally, let I = {i1, i2, . . . , im} be a set of items and DB = {T1, T2, ..., Tn} a set of transactions where every transaction is also a set of items (i.e. itemset).

Definition 2. The utility of an item ip is a numerical value yp defined by the user. It is transaction independent and reflects importance (usually profit) of the item. External utilities are stored in an utility table.

Definition 3: The utility of an item set X in a transaction Ti is denoted by U(X,Ti) & it is calculated as follows. For example,U({*AC*}, *T1*) = U({*A*}, *T1*) + U({*C*}, *T1*) = 5 + 1 = 6.

Definition 4: The utility of an item set X in D is denoted by U(X) & it is calculated as follows For example, U({*AD*}) = U({*AD*}, *T1*) + U({*AD*}, *T3*) = 7 + 17 = 24.

Definition 5: An itemset is called a *high utility itemset* if its utility is no less than a user-specified *minimum utility threshold* which is denoted as *min_util*. Otherwise, it is called a *low utility itemset*

TID	TRANSACTION	TU
T1	(A,1) (C,1) (D,1)	8
T2	(A,2) (C,6) (E,2) (G,5)	27
T3	(A,1) (B,2) (C,1) (D,6) (E,1) (F,5)	30
T4	(B,4) (C,3) (D,3) (E,1)	20
T5	(B,2) (C,2) (E,1) (G,2)	11

Table 1:Transaction Data Set

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

Table 2: Item & correspondent profit

Definition 5. The transaction utility of a transaction *Td* is denoted as *TU*(*T*d) and defined as *u*(*Td*, *Td*). For example, *TU*(*T1*) = *u*({*ACD*}, *T1*) = 8.

3. Proposed Methodology

We will propose a novel technique for high utility item set mining. The new algorithm will outperform the previous algorithms in terms of execution time.

The outline of the proposed algorithm is as follows:

Step 1: To generate a list of high utility item set following will be used:

- Transaction Utility- The transaction utility of an item is the sum of the utilities of all items in that transaction
- Weighted transaction utility of an item set - The weighted transaction utility of an item set is obtained by performing the addition of the transaction utility of all transactions containing that item set
- Only those item sets are included in the initial high utility item set mining list whose weighted transaction utility is more than the minimum utility

Step 2: In this step, final high utility item set is generated by eliminating the infrequent item sets from the list of step 1. It is performed as follows:

- An item set is chosen from the list of step 1.
- If the utility of the item is less than the minutility (Minimum Utility) than the item is erased. Otherwise, the item set is selected in the final list of the high utility item set.

Step 3: From candidate of size 1, we recursively create candidates of greater size as follows:

- For each itemset I1 and I2 of level k-1
- 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 TWU of itemset
- if the transaction weighted utility (TWU) is high enough
- add it to the set of HWTUI of size
- Continue this process until there are candidates to combine.

Step 4: If the utility of a candidates is less then the minimum threshold then remove such candidate from the list of high utility items

Step 5: Return all high utility itemsets found

Step 6: End of process.

Comparison between existing and proposed algorithm

The existing method is based on the concept of generate and test method. It means that the algorithm first generates all the candidates of size 1 and then performs the pruning according to the minimum utility. Then it generates all the candidates of size 2 and then perform the pruning according to the min utility. The same process is repeated for the subsequent size elements.

The proposed method generates all the candidates of size 1 and then performs the pruning according to the utility. After that it eliminates all the infrequent items of size 1 from the data set to generate a new compact data set. Then this compact data structure is used to generate the subsequent size elements. So it will save time n space.

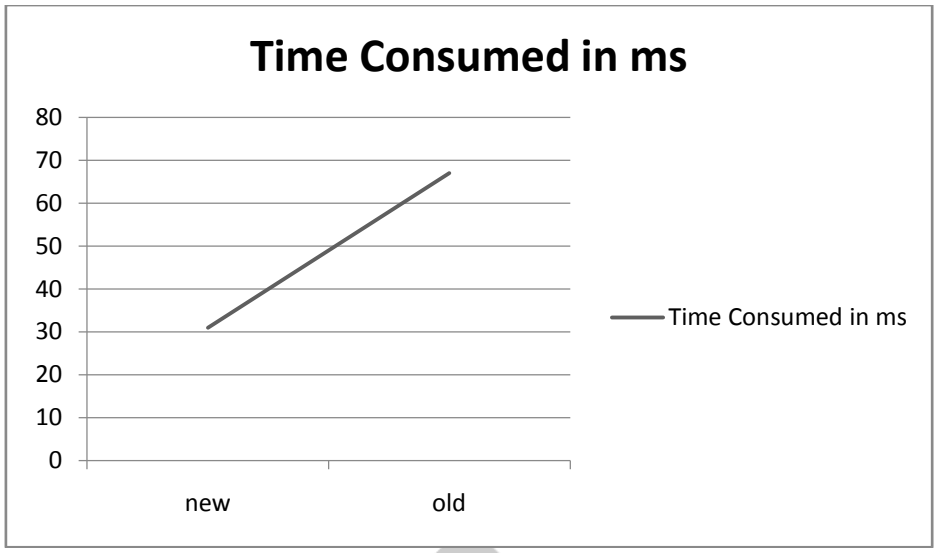

Figure. 1 Depicts the Time Consumption Comparison

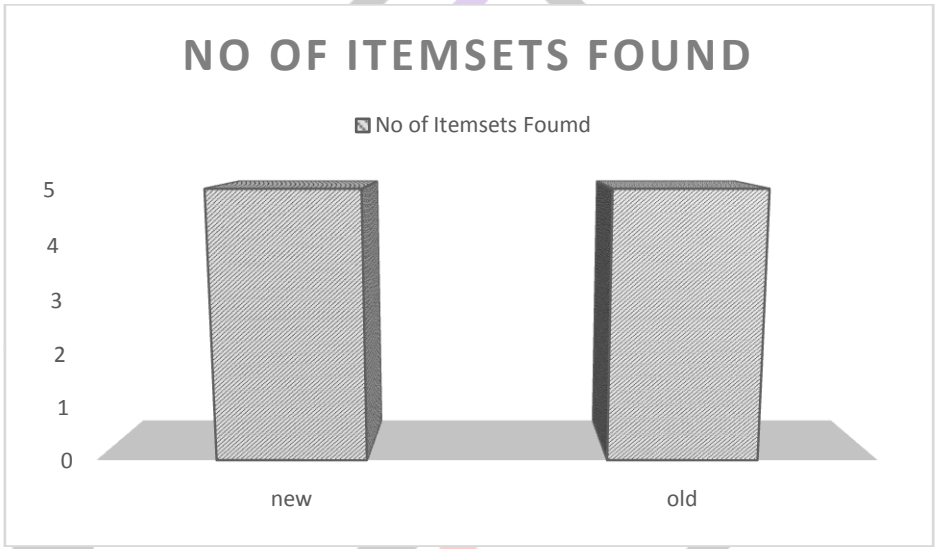

Figure. 2 Depicts the Result Comparison

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

4. Conclusion:

The data capturing technologies is also increasing. In utility mining we concentrate on utility value of itemset while in frequent item set mining we concentrate that how frequently items appears in transactional database. In this paper, we surveyed the list of existing high utility mining techniques. However we surveyed different concepts of Association rule mining and frequent itemset mining techniques which play significant role for basic of utility itemset mining but we restricted ourselves to the classic high utility mining problem. This paper has proposed a time efficient algorithm for mining high utility item sets from a transaction data set.

References:

1. 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.
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. Chan, Q. Yang, and Y. Shen. Mining high utility itemsets. In Proc. of Third IEEE Int'l Conf. on Data Mining, pp. 19-26, Nov., 2003.
4. 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.

5. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In Proc. of the ACM-SIGMOD Int'l Conf. on Management of Data, pp. 1-12, 2000.

6. 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.

7. Y. Liu, W. Liao, and A. Choudhary. A fast high utility itemsets mining algorithm. In Proc. of the Utility-Based Data Mining Workshop, 2005.

8.Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. J. Frawley, eds, „Knowledge Discovery in Databases", AAAI/MIT Press, Cambridge.

9. J. Pei, J. Han, and R. Mao. Closet: An efficient algorithm for mining frequent closed itemsets. In DMKD '00: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 21–30, May 2000.

10. M. J. Zaki and C.-J. Hsiao. Charm: An efficient algorithm for closed itemset mining. In SDM '02: Proceedings of the second SIAM International Conference on Data Mining, April 2002.

11. G. Grahne and J. Zhu. Efficiently using prefix-trees in mining frequent itemsets. In FIMI '03: Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, November 2003.