A Survey of High Utility Itemset Mining Algorithms from Relational and XML Databases

S. Kalai Selvi
Final Year ME CSE.,
Department of CSE,
Maharaja Engineering College,
Avinashi-641654, Tamil Nadu,
India.

Mr. V. Yathavaraj M.E.,
Assistant Professor,
Department of CSE,
Maharaja Engineering College,
Avinashi-641654, Tamil Nadu,
India.

***Abstract* – Data Mining unearths meaningful patterns from transactional databases. Frequent and infrequent pattern mining produces itemsets with high frequency and negative association rules which is of greater interest to users. The data mining algorithms exhibit statistical correlation between the itemsets. The semantic significance between the itemsets which means the economic utility of itemsets is not reflected. The utility based data mining approach is focuses on all aspects of economic utility and incorporates utility in both predictive and descriptive data mining tasks. The usefulness of an itemset is characterized by the utility constraint. An itemset is useful to the user if it satisfies the given utility constraint. This paper focuses on the survey of various and recent utility mining algorithms from transactional and XML databases and its limitations.**

I. INTRODUCTION

Utility mining defines the usefulness or importance of an item or itemsets in terms of cost, quality, sales, profit or any other user preferences. Utility mining considers the frequency and utility associated with each itemset. The utility constraint defines the usefulness of an itemset. The itemset is useful to the user if it satisfies the given utility constraint. X defines the utility of an itemset and u(X) defines the sum of utilities of the itemset X in all transactions containing X. If u(X) >= min_util, then X is a High Utility Itemset (HUI) and min_util is the user defined utility threshold.

A. High Utility Itemset Mining

Minimum support threshold is the minimum occurrence of an itemset in the given transaction to be identified as a frequent itemset. Yao et al defines a well known model for identifying high utility itemsets and following is a set of definitions defined by Yao et al [1].

y_p, transaction independent numerical value reflects the usefulness or profit of an item ip and defines external utility which is stored in the external utility table (Table 1.2).

x_p, transaction dependent numerical value defines the quantity of an item xp which is the internal utility of an itemset (Table 1.1).

Utility function f defines the product of internal and external utility.

$$f(x,y) = x_p * y_p$$

Table 1.1 Database with 5 Transactions and 5 Distinct Items

TID	I1	I2	I3	I4	I5
1	1	0	18	0	1
2	0	6	0	1	1
3	5	0	4	0	2
4	2	3	1	1	1
5	0	0	4	0	3

Table 1.2 External Utilities of Items for the Database

Item	I1	I2	I3	I4	I5
Profit	2	11	4	7	5

Utility function f, is a function of two variables commonly defined as the product of internal and external utility as given below

$$f\,(x, y): x_p \times y_p \qquad (1.1)$$

The utility of item ip in transaction T is the quantitative measure computed with utility function as defined above (i.e.)

$$u\,(i_p, T) = f\,(x_p, y_p),\ i_p \in T. \qquad (1.2)$$

For example: utility of item I5 in transaction T5 is 3 × 5 = 15. The utility of itemset S in transaction T is defined as follows

$$U(S) = \sum_{i_p \in s} u(i_p, S) \qquad (1.3)$$

For example, utility of itemset {I2, I5} in transaction T_2 is $u(\{I2, I5\}, T_2) = u(\{I2\}, T_2) + u(\{I5\}, T_2) = 6 \times 11 + 1 \times 5 = 71$. On the basis of the above definitions, high utility itemset can be defined as follows

Itemset S is of high utility if U(S) ≥ minUtil where minUtil is user defined utility threshold in percents of the total utility of the database. High utility itemset mining is the task of finding set H defined as

$$H = \{S \mid S \subseteq I, U(S) \geq \min Util\} \qquad (1.4)$$

where ‘I’ is the set of items (attributes).

B. Rare Utility Itemset Mining

Rare or infrequent situations are more important in many applications. For example, the rare combination of symptoms may provide useful insight for doctors in medical applications. In supermarket, some rare items occur more frequently with the specific itemset which can be used as a promotion strategy for decision makers.

C. Negative Utility Itemset Mining

Utility mining has a broader range of practical applications. A supermarket may sale items with negative values to promote certain items and also to attract customers. The customers buy specific items and they receive free goods which result in negative values.

D. XML Databases

The database used in the existing algorithms is Relational Databases. The Data in the Relational databases are stored as rows in tables. The data is related between the tables using the concept of Foreign Keys. The data is retrieved from the databases using the execution of SQL Queries. Though the relational databases can handle large volumes of data within the system, the cost of setting up and maintaining the database is relatively so expensive. Complex queries are required to retrieve the data and needs sophisticated processing power. The data in the relational databases cannot be transmitted electronically.

In order to overcome the disadvantages of relational databases, XML Databases can be used. XML databases are both machine and human readable. The data in the XML file is organized as hierarchies. Data and structure are always presented together. Setting up of XML databases is not as expensive as in the case of Relational Databases. The data can be moved between the enterprises electronically. The data is transmitted through the firewalls with the help of HTTP protocol. Information coded in XML is easy to read and understand, plus it can be processed easily by computers. There is no fixed set of tags. New tags can be created as they are needed. Mapping existing data structures like file systems or relational databases to XML is simple.

II. UTILITY MINING ALGORITHMS FROM TRANSACTIONAL DATABASES

Hong Yao et al[1] defined the problem of Utility Mining formally. The high utility itemsets are generated based on the information from the transaction database and external information about the utility of the itemsets. His work formulated the theoretical and mathematical model of utility mining based on two important properties of utility namely Support Bound Property and Utility Bound Property. These properties allow an upper bound on the utility of a K-Itemset to be calculated from the discovered (K-1) itemsets. The work also proposed a heuristic model to estimate itemset utility to search space by predicting whether the itemset should be counted or not. To be summarized, Hong Yao et al described the concept of Utility Mining and derived the mathematical model for Utility Mining.

Hong Yao et al[5] proposed a Constraint based Utility Mining. He identified that the Downward Closure Property used in Apriori algorithm and the Convertible Constraint Property cannot be used and applied to Utility Mining. Two novel pruning strategies were designed and based on these pruning strategies, two algorithms namely UMining and UMining_H were developed to reduce the cost of finding High Utility Itemsets. The work also shows that the UMining algorithm is preferable compared to UMining_H, because UMining guarantees the discovery of all High Utility Itemsets. Overall, the work proposed two algorithms for Utility Mining namely UMining and UMining_H to reduce the cost of finding High Utility Itemsets.

Chun-Jung Chu et al[3] proposed a novel method namely HUINIV-Mine (High Utility Itemsets with Negative Item Values) for efficiently and effectively mining high utility itemsets with negative item values. To reduce the execution time of the process, the work contributed Transaction Weighted Utilization Itemsets. The High Utility Itemsets with negative item values are identified with less requirements on memory space and CPU I/O. the HUINIV-Mine algorithm also produces less candidate itemsets under different experimental conditions. When negative itemsets are considered, the HUINIV-Mine algorithm is stable and takes less execution time. The work produces High Utility Itemsets with negative values using HUINIV-Mine algorithm which is of greater interest to many Super Markets to promote and attract customers.

Shankar et al[6] discovered in their work, the interesting association patterns that are useful to the Business Utility. The work mines the association patterns by considering significance, utility and subjective interestingness of the user. The principal objective of the research is to mine interesting association patterns from the transaction data items to improve the business. The work summarizes the generation of association patterns from the transaction data items for Business Utility.

C.Saravanabhavan et al[4] have presented an efficient tree structure for mining of high utility itemsets. At first, we have developed a novel utility frequent-pattern tree structure, an extended tree structure for storing crucial information about utility itemsets. The efficiency of the high utility itemsets mining is achieved with two major concepts: 1) a large database is compressed into a smaller data structure as well as the utility FP-tree avoids repeated database scans, 2) the proposed FP-tree-based utility mining utilize the pattern growth method to avoid the costly generation of a large number of candidate sets in which it dramatically reduces the search space. They have proposed a novel utility FP- tree, an extended tree structure for storing essential information about utility frequent patterns. In addition to, they have utilized the mining technique used in the standard FP-growth algorithm for mining the complete set of utility patterns. The proposed utility FP-tree-based pattern mining utilized the pattern growth method to avoid the costly generation of a large number of candidate sets in which it dramatically reduces the search space. The experimentation was carried out on our proposed approach using real life datasets and the results showed that the proposed approach is effective on the tested databases. To summarize the research, Utility FP-Tree based approach is used to find the frequent utility patterns from transactional databases.

III. UTILITY MINING ALGORITHMS FROM XML DATABASES

Shankar et al[2] presented a novel algorithm called Fast Utility Mining (FUM) for mining High Utility Itemsets from large transaction databases. The proposed algorithm works faster and simpler compared to the previous UMining algorithm. Different types of itemsets such as HULF (High Utility Low Frequency), HUHF (High Utility High Frequency), LUHF (Low Utility High Frequency), LULF (Low Utility and Low Frequency) are generated using the FUM algorithm. The FUM algorithm works well and produces accurate and reliable result when the size of the transaction database increases and when the number of distinct items in the database increases. The FUM algorithm is efficient when the Utility Threshold is low in which case the mining process requires numerous iterations. On the whole, the work introduces new Utility Mining algorithm, FUM for efficient mining when large transaction databases are involved.The FUM-NIV algorithm used FUM algorithm to find the High Utility Itemsets with negative values. It overcomes the limitation of the HUINIV-Mine algorithm. The HUINIV-Mine algorithm produces more number of Candidate Itemsets since the Downward Closure Property used in the Apriori Algorithm is not used in the HUI Algorithm. If the Candidate Itemsets are more, the time taken to produce Frequent itemsets with negative values is increased which results in inefficiency of the system. The performance of the FUM-NIV algorithm is compared with the HUINIV-Mine algorithm. The FUM_NIV (Fast Utility Mining Algorithm with Negative Item Values) algorithm results in a significant reduction in the execution time.

Kannimuthu et al[7] proposed iFum (improved FUM) algorithm with significant improvements made in the FUM Algorithm using XML Database. The FUM algorithm computes the utility value for the subsets in every transaction though the subset has already been a HUI in the previous transactions. This limitation has been overcome in the iFUM algorithm. The iFUM algorithm shows significant reduction in the execution time compared to the FUM algorithm and performs good.

Kannimuthu et al[8] proposed mining of High Utility Itemsets from distributed databases. The work depends on Service Oriented Paradigm which gives Knowledge as a Service. HUI_MINER$_{XML}$ algorithm is proposed to find High Utility Itemsets from Distributed databases.

IV. CONCLUSION

In this paper, various utility mining algorithms that uses relational and XML databases have been studied. The algorithms that use XML databases shows significant reduction in the execution time and shows higher performance than the algorithm that uses relational databases. The Utility Mining algorithm that uses FP Tree has also been studied. The future work could focus on using various tools to evaluate the High Utility Itemsets, Negative Utility Itemsets and Rare Utility Itemsets.

REFERENCES

[1] Hong Yao, Howard J. and Hamilton, "Mining itemset utilities from transaction databases", data & Knowledge Engineering, pp. 59: 603- 626, 2006.

2] Shankar.S, Dr.T.Purusothaman, S.Jayanthi, "A Fast Algorithm for Mining High Utility Itemsets", IEEE International Advance computing Conference (IACC 2009) Patiala, India, 6-7 March 2009

3] Chun-Jung Chu, Vincent S. Tseng, Tyne Liang, "An Efficient algorithm for mining high utility itemsets with negative item values in large databases", Journal of Applied Mathematics and Computation, 215 (2009) 767-778.

4] C.Saravanabhavan, R.M.S. Parvathi, "Utility Fp-Tree: An efficient approach for mining on weighted utility itemsets", International journal of Engineering Research and Development, PP.19-31, September 2013

5] Yao H, Hamilton H J, Butz C J, "A foundational approach to mining itemset utilities from databases", Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, pp. 482-486, 2004.

6] Shankar.S and T.Purusothaman, "A Novel Utility Sentient Approach for Mining Interesting Association Rules", IACSIT International Journal of Engineering and Technology Vol.1, No.5, December, 2009, ISSN: 1793- 8236

7] Kannimuthu S, Dr.K.Premalatha and Shankar S, " iFUM-Improved Fast Utility Mining", IJCA International Journal of Computer Applications, (0975 – 8887) Volume 27– No.11, August 2011

8] Kannimuthu S and Premalatha K,"A Distributed Approach to Extract High UtilityItemsets from XML Data", World Academy of Science, Engineering and Technology International Journal of Computer, Control, Quantum and Information Engineering Vol:8, No:3, 2014