ISSN: 2277-9655
Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# A SURVEY ON ITEMSET MINING FOR LARGE TRANSACTION DATABASE

**Ancy Jose*, Dr. John T Abraham**

* PG Scholar Department Of Computer Science and Engineering Christ Knowledge City Mannoor,India and Bharata Mata College Thrikkakara,India

Professor Department Of Computer Science and Engineering Christ Knowledge City Mannoor,India and Bharata Mata College Thrikkakara,India

**DOI**: 10.5281/zenodo.52500

## ABSTRACT

Mining itemsets from the databases is an important data mining task.Frequent itemset mining refers to the mining of set of items occur frequently in the database.Utility itemset mining refers to the discovery of items with high utilities.

Many algorithms have been proposed for mining frequent item sets as well as utility item set.This paper focus on different algorithms and techniques for high utility itemset mining and frequent itemset mining which can handle large transactions in the database.

**KEYWORDS**: Frequent item set mining , Utility item set mining, Large transaction.

## INTRODUCTION

Data Mining can be defined as a method which extracts some new important information contained in dense database. The goal of frequent itemset mining is to find items that occur in a transaction database above a user specified frequency threshold, without considering the quantity or such as profit of the items.It has practical importance in different application areas such as decision support,Web usage mining, bioinformatics, etc[1].However, quantity and utility are significant for addressing real world decision problems that require maximizing the utility in an organization. The HU itemset mining problem is to find all itemsets that have utility larger than a user specified value of minimum utility.Frequent Itemset Mining is finding frequent itemsets are based on the support confidence model[1]. Find all frequent itemsets from given database.The problem of itemset mining is finding the complete set of itemsets that have more occurrence in transactional databases. However the utility of the itemsets is not considered in ordinary frequent itemset mining algorithms. The method only considers whether an item occur frequently in dataset, but avoids both the quantity and the utility associated with the item. However, the presence of an itemset may not be an essential indicator of interestingness, because it only shows the number of transactions in the database that contains the itemset.

The limitation of frequent itemset mining guide researchers towards utility based mining approach, which allows a user to conveniently express his or her perspectives concerning the usefulness of itemsets as utility and then find items having high utility values higher than given threshold.[2] The mining process identify itemsets which are more useful and it does not identify either frequent or rare itemsets To identify itemsets having higher rate in the database, no matter whether these itemsets are frequent itemsets or not. This leads to a new approach in data mining which is based on the idea of utility called as utility mining.

This report is organized as follows: Sect ion II presents frequent itemset mining methods which handle large transactions , Section III presents high utility itemset mining methods which handle large transactions, Section IV compare the various method and finally, conclusion is presented in Section V.

## FREQUENT ITEMSET MINING METHOD FOR LARGE TRANSACTIONS

### PRIVATE FP-GROWTH ALGORITHM

Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Yang[3] proposed FP Growth Algorithm which is Frequent Pattern Growth method. FP-growth is a depth-first search method, which requires no candidate generation. Unlike Apriori, FP-growth only performs only two database scans. In the proposed Private FP-growth (PFPgrowth)Algorithm,PF consist of two modules ,preprocessing and mining phase. In the preprocessing module to improve the utilityprivacy Tradeoff and smart splitting method to transform the database. In the mining module a run time estimation method is used to estimate the support of item.Downward closure property used in dynamic reduction method to dynamically reduce the amount of noise added to provide privacy during the mining process . In the preprocessing phase,modify the database to limit the length of transactions.

The preprocessing phase is performed only once for a given database. By applying the limit, long transactions would be split. That is,if a transaction has more items than the limit then divide transaction into multiple subsets such as sub-transactions and guarantee each subset is under the limit. PFP defines a smart splitting method to transform the database. To ensure applying e-differentially private algorithm on the transformed database still satisfies edifferential privacy for the original database, a weighted splitting operation is applied.Moreover, to preserve more frequency information in subsets,apply a graph-based approach to reveal the correlation of items within transactions and utilize such correlation to guide the splitting process.

### APRIORI WITH SMART TRUNCATING

C. Zeng, J. F. Naughton, and J.-Y. Cai,[4] proposed an apriori algorithm for large tranasction. The major difficulty relies on the existence of long transactions that is, transactions containing many items. This paper propose an approach that begins by truncating long transactions.This makes the possibility of improving the utility-privacy tradeoff by limiting transactions . It cannot directly apply such a limit, so apply the limit by truncating transactions. That is, if a transaction has more than a specified number of items, delete items until the transaction is under the limit. And the deletion must be done in a differentially private way. Also it reduces the error due to the noise required to enforce privacy, it introduces a new source of error by discarding items from transactions. In transaction truncating , subsets which are more frequent is kept and other items are truncated. Smart truncation algorithm find frequent itemsets in a differentially private way.

## UTILITY ITEMSET MINING METHOD FOR LARGE TRANSACTIONS

### TWO –PHASE A LGORITHM

Liu, Y., Liao, W.K., Choudhary A. [5] proposed two phase algorithm for high utility itemsets. They used a transaction weighted utility (TWU) measure to mine the search space. The method based on the candidate generation-and-test approach. The proposed algorithm suffers from poor performance if mining large datasets and long patterns much like the Apriori. It requires minimum database scans, minimum memory utilization and less computational cost. It can easily handle very large databases. Two-Phase algorithm that effectively prune candidate itemsets and simplify the calculation of utility. It greatly reduces the search and the memory cost and requires less computation.

In Phase I, we define a transaction-weighted utilization mining model that have a Transaction-weighted Downward Closure Property. (The intention of introducing this new concept is not define a new problem, but utilize its property to prune the search area.) High transaction-weighted utilization itemsets are identified in this module. The size of candidate set is reduced by only considering the supersets of high transaction weighted utilization item. In the second phase, one database scan is performed to filter the more transaction-weighted utilization itemsets that are actually low utility itemsets. This algorithm guarantees that the complete set of high utility itemsets will be identified.

### CTU-MINE

Erwin, A., Gopalan,R.P., N.R. Achuthan [6] proposed an efficient CTU-Mine Algorithm based on Pattern Growth method. The paper introduce a concise data structure defined as Compressed Transaction Utility tree (CTU-tree) for HU mining, and a new algorithm called CTU-Mine for extracting high utility itemsets. The method works more accurately than TwoPhase for dense datasets and long pattern datasets. If the thresholds are high, then TwoPhase runs relatively fast matched to CTU-mine when the utility threshold value becomes lower, CTUMine outperforms TwoPhase.

ISSN: 2277-9655
Impact Factor: 3.785

**UP GROWTH ALGORITHM**

Cheng-Wei Wu [7] proposed an algorithm with a compact data structure for better discovering high utility itemsets from transactions. The UP-Growth is more efficient algorithm than others to generate high utility items depending on establishment of a global UP-Tree. In phase I, the framework of UP-Tree follows 3 steps:

[1] Construction of UP-Tree.

[2] Generate PHUIs from UP-Tree.

[3] Identify high utility items using PHUI

The construction of global UP-Tree is follows,Discarding global unpromising items is to discard the low utility items and their utilities from the transaction utilities.

Discarding global node utilities during global UP-Tree construction. Reduce the node utilities which are nearer to UP-Tree root node . The PHUI is similar to TWU, that compute itemsets utility with the support of estimated utility. Finally, identify itemsets having most utility.

**IHUP- INFORMATION OF HIGH UTILITY PATTERN MINING ALGORITHM**

Ahmed et al[8] proposed IHUP Algorithm to efficiently generate high utility itemsets and to void multiple database scans.This paper proposed three tree structures, IHUPL-Tree, IHUPTF Tree, and IHUPTWU-Tree, which are based on FP-Tree. Each node in the trees is composed of an item name, a support count, and a TWU value. IHUP generates all high utility itemsets from the IHUP-Tree through three steps.

Step 1: In step 1 items in transactions are sorted based on the lexicographic order, and the transactions are inserted into IHUPL-Tree with a single database scan. For a single pass tree construction, that tree can be restructured without extra database scan by setting nodes by support descending order or TWU descending order.

Step 2: Candidate itemsets are extracted from IHUP-Tree by FPGrowth algorithm

Step 3: Actual HU itemsets are identified with an additional database scan. Although IHUP construct a tree then discover high utility itemsets with two database scans, and generates more number of candidates by applying the TWU model.

## COMPARISON

The following table comparison between the mining methods for mining high utility and frequent itemsets for large transactions. In the case of frequent itemset mining PFP and Apriori with smart splitting perform efficiently in the case of large transaction. Transaction splitting and transaction truncating is applied respectively. However in Apriori limiting the length of transactions causes information lose. In Utility itemset mining when threshold becomes lower, CTUMine performs better than TwoPhase.Also UP growth and IHUP algorithm efficiently mine utility itemset in large transactions.

**COMPARISON TABLE**

| Name of methods | Concept Used | performance in Dense Database |
|---|---|---|
| PFP | Transaction Splitting | Very Good |
| Apriori | Smart Truncation | Average |
| Two Phase algorithm | Transaction weighted utility measure | Average |

ISSN: 2277-9655
Impact Factor: 3.785

| | | |
|---|---|---|
| IHUP | Scans the original Database once/ FPGrowth tree based | Good |
| UP-Growth | 2 Scan/ FP-Growth tree based | Good |

## CONCLUSION

Frequent itemset mining consider only frequency of itemset and is challenged in many areas such as retail, marketing etc. It has been seen that in many real application domains that itemsets which contribute the most are not adequately the frequent itemsets. Utility mining is a period of research which tries to bridge frequency and utility by using item utilities as an analytical measurement. In this paper proposed comparative study of different mining methods for large transactions.

## REFERENCES

[1] R Agrawal and R Srikant 1994. Fast algorithms for mining association rules. In Proc. of the 20th Int'l Conf. on Very Large Data Bases, pp. 487-499.
[2] Chan , Q.Yang,Y.D Shen, Mining high utility itemsets, in: Proceedings of the 3rd IEEE International Conference on Data Mining , Melbourne, Florida, 2003, pp.19-26
[3] Sen Su,Shengzhi Xu,Xiang Chang,Zhengi Li and Fangchun Yang Differ-entially private Frequent Itemset Mining via Transaction Splitting,IEEE 2016
[4] C. Zeng, J. F. Naughton, and J.-Y. Cai, On differentially private frequent itemset mining, in VLDB, 2012.
[5] Liu, Y., Liao, W.K., Choudhary, A.: A Fast High Utility Itemsets Mining Algorithm. In: 1st Workshop on Utility-Based Data Mining. Chicago Illinois 2005.
[6] Erwin, A., Gopalan, R.P., Achuthan, N.R, A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets, In: International Workshop on Integrating AI and Data Mining. Gold Coast, Australia 2007.
[7] Cheng Wei Wu1, Bai-En Shie1, Philip S. Yu2, Vincent S. Tseng1 "Mining Top-K High Utility Itemsets " KDD12, August 1216, 2012,
[8] 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, 2009.