An Advanced Research Framework to Investigate Valuable Frequent Itemsets using Mining

I.Prasanthi
Dept. of CSE (SE), Shri Vishnu Engineering College for Women(A), Vishnupur, Bhimavaram, Andhra Pradesh, India

Abstract

The popularity of market applications including stand-alone and e-commerce have been rapidly growing in the two decades and accumulated mass of data from their customers.The extraction of hidden,predictive knowledge in the form of frequent itemsets from such databases is a crucial task in the Data Mining research. Specifically,identifying valid, previously unkown and potentially useful frequent itemsets is a computationally intensive procedure. Although, extensive studies have been proposed in the literature for extracting frequent itemsets from large database, many of them have focused largely on identifying frequent itemsets based on statistical correlations aamong the items. This situation got the focus of present data mining researchers into the era of Utility Mining. The emerging utility mining not only focuses on frequencies of statistical values among the itemsets but also throw light on the utility associated wit the itemsets and it endorsed as a basic motivation factor for the present study. In addition,utility mining discovers all high utility itemsets beyond the user specified threshold values from large database, push forward towards the present study.

The scholar in the present study proposes an Advanced Model to investigate Valuable Frequent Itemsets (AMVFI) from the large amount of data using utility mining. This mode is designed to achieve the goal of discovering utility frequent itemises in two phases. In the initial phase, the AMVFI use Mining Top-k utility itemsets (MTK) algorithm to generate utility itemsets from large datasets. In the second phase, this model retrieves valuable frequent itemsets using Mining Top-k Closed Itemsets (MTK-Close) algorithm from retrieved utility itemsets by MTK algorithm. The MTK algorithm employed by this model adopts a compact tree-based structure and reduces memory consumption. The MTK-Close algorithm is devised on the basis of level-wise technique and limits the number of scans of the databases. In a nutshell, the successive two-phase model discovers valuable frequent itemsets in the light of cost,quantity and profit which helps the market analyzers to make accurate future decisions. Several experiments are conducted on improved model with a variety of synthesized datasets and results are claimed.

Keywords

Top-k Pattern Mining, Top-k High Utility Itemset Mining

I. Introduction

Frequent itemset mining: The discovery of frequent relationship among a huge database has been known to be useful in selective marketing, decision analysis, and business management. A popular area of its applications is the market basket analysis, which studies the buying behaviors of customers by searching for sets of items that are frequently purchased together.

High utility itemset: The most challenging data mining tasks is the mining of high utility itemsets efficiently. Identifying the itemsets with high utilities is called as Utility Mining. The utility can be measured based upon cost, profit or other expressions of user preferences. The frequent itemset is not sufficient to reflect the actual utility of an itemset. Frequent itemsets are the itemsets that occur frequently in the transaction data set.

The main goal of frequent Itemset Mining is to identify the frequent itemsets in a transaction dataset. Utility mining gives an essential topic in the data mining field. Mining high utility itemsets from databases refers to finding the itemsets with high profits

II. Literature Survey

[1] The Author, P. Tzvetkov(et. al), Aim an efficient algorithm, called TFPis produced for mining such itemsets without mins_ support. Beginning at min_support =0 and by making utilization of the length requirement and the properties of best k visit shut itemsets, min_support can be raised successfully and FP-Tree can be pruned powerfully both amid and after the development of the tree.

[2] The author, Heungmoryang (et. al), Aim we propose an efficient algorithm for mining top-k high utility patterns with highly decreased candidates. For this purpose, we develop three strategies that can reduce the search space by raising a minimum threshold effectively in the construction of a global tree, where they utilize exact and pre-evaluated utilities of itemsets.

III. Problem Definition

Efficiently mining HUIs in databases is not an easy task because the downward closure property used in FIM does not hold for the utility of itemsets. In other words, pruning search space for HUI mining is difficult because a superset of a low utility itemset can be high utility. It is difficult for users to choose an appropriate minimum utility threshold in practice.

Some applications, they are not developed for top-k high utility itemset mining and still suffer from the subtle problem of setting appropriate thresholds

IV. Proposed Approach

Proposing a novel framework for top-k high utility itemset mining, where k is the desired number of HUIs to be mined.

Two phases named AMVFI(Advanced Model to investigate Valuable Frequent Itemsets) and MTK-Close(Mining Top-K utility itemsets) are proposed for mining the complete set of top-k HUIs in databases without the need to specify the min_util threshold. The MTK algorithm adopts a compact tree-based structure named UP-Tree to maintain the information of transactions and utilities of itemsets. MTK inherits useful properties from the TWU model and consists of two phases.

In phase I, potential top-k high utility itemsets (PKHUIs) are generated. In phase II, top-k HUIs are identified from the set of PKHUIs discovered in phase I. On the other hand, the MTK algorithm uses a list-based structure named utility-list to store the utility information of itemsets in the database. It uses vertical data representation techniques to discover top-k HUIs in only one phase

V. System Architecture

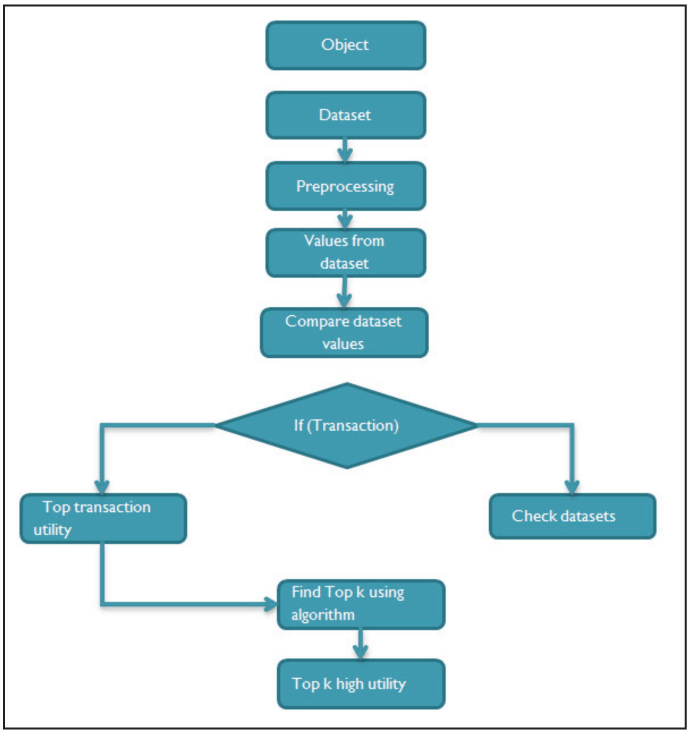

Fig. 1:

VI. Proposed Methodology

A. Preprocessing

In the pre-processing module the datasets had been taken and the datasets has been analysed to remove the unwanted datasets.

The datasets can be filtered to found the top k utility itemsets in a certain dataset in databases.

The dataset can be stored in a database and the preprocessing method can be used to reduce storage space and the certain utility usage of a data in a process.

B. Transaction Utility

In the transaction utility module the utility can be found using an itemsets value. The utility can be found using a certain method to be analysis

TWU(X) =∑ ∈∧) (DTTX rrr TTU

Then the result of the utility can be analysed to store in databases as utility

C. Top k Utility

In this the top k high utility can be analysed using a prefetched dataset result from a transaction utility. Then the dataset can be analysed to found a top utility set.

ESTU(X) ≥ min_utilBorder and MAU(X) ≥ in_utilBorder

The result can be used to display the result in a graph.

D. MTK-Close Algorithm:

INPUT: DATABASE,ITEMS

STEP1: scanning the transactional database.

STEP2: by using transaction utlity and transactional weight calculate profit value.

STEP3: find minimum utlity threshold value.

STEP4: removing unnecessary item set.

STEP5: reorganize the database.

STEP6: displayingitemset node and utility list structure.

STEP7: apply MTK algorithm.

STEP8: Deriving top-k high utlity item sets.

VII. Results

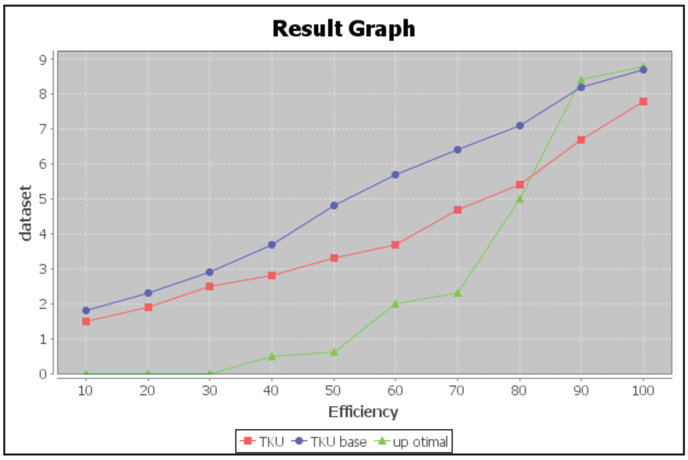

Fig. 2: Result Graph Shows Efficiency in Order to Extract the Frequent Itemsets.

The Enhanced IFP-Growth as well as which shows efficiency in order to extract the frequent itemsets.

VIII. Extension Work

The Enhanced IFP-Growth consists of three phases: In first phase, it scans the transactional database only once for generating equivalence classes of frequent items. In second phase, it consequently sorts the equivalence classes of frequent items in descending order and filter out non-frequent items. Finally in third phase, the Enhanced IFP tree is constructed in order to extract the frequent itemsets.

IX. Conclusion

In this project, we have considered the issue of best k high utility itemsets mining, where k is the coveted number of high utility itemsets to be mined. Two productive calculations MTK(mining Top-K Utility itemsets) and Technical knockout (mining Top-K utility itemsets in One stage) are proposed for mining such itemsets without setting least utility edges.

References

[1] R Philippe Fournier-Viger, Philip S. Yu,“Efficient Algorithms for Mining Top-K High Utility Itemsets,” Vol. 28, No. 1, January 2016.

[2] C. Ahmed, S. Tanbeer, B. Jeong, Y. Lee,“Efficient tree structures for high-utility pattern mining in incremental databases”, IEEE Trans. Knowl. Data Eng., Vol. 21, No. 12, pp. 1708–1721, Dec. 2009.

[3] K. Chuang, J. Huang, M. Chen,“Mining top-k frequent patterns in the presence of the memory constraint,” VLDB J., Vol. 17, pp. 1321–1344, 2008.

[4] R. Chan, Q. Yang, Y. Shen,“Mining high-utility itemsets”, in Proc. IEEE Int. Conf. Data Mining, pp. 19–26, 2003.

[5] P. Fournier-Viger, V. S. Tseng,“Mining top-k sequential rules”, In Proc. Int. Conf. Adv. Data Mining Appl., 2011, pp. 180–194.

[6] P. Fournier-Viger, C.Wu, V. S. Tseng,"Mining top-k association rules," In Proc. Int. Conf. Can. Conf. Adv. Artif. Intell., 2012, pp. 61–73.
[7] P. Fournier-Viger, C. Wu, V. S. Tseng,"Novel concise representations of high utility itemsets using generator patterns", in Proc. Int. Conf. Adv. Data Mining Appl. Lecture Notes Comput. Sci., 2014, Vol. 8933, pp. 30–43.
[8] J. Han, J. Pei, Y. Yin,"Mining frequent patterns without candidate generation", In Proc. ACM SIGMOD Int. Conf. Manag. Data, pp. 1–12, 2000.
[9] J. Han, J. Wang, Y. Lu, P. Tzvetkov,"Mining top-k frequent closed patterns without minimum support," In Proc. IEEE Int. Conf. Data Mining, 2002, pp. 211–218.
[10] S. Krishnamoorthy,"Pruning strategies for mining high utility itemsets," Expert Syst. Appl., Vol. 42, No. 5, pp. 2371–2381, 2015.
[11] C. Lin, T. Hong, G. Lan, J. Wong, W. Lin,"Efficient updating of discovered high-utility itemsets for transaction deletion in dynamic databases", Adv. Eng. Informat., Vol. 29, No. 1, pp. 16–27, 2015.
[12] G. Lan, T. Hong, V. S. Tseng, S. Wang,"Applying the maximum utility measure in high utility sequential pattern mining", Expert Syst. Appl., Vol. 41, No. 11, pp. 5071–5081, 2014.
[13] Y. Liu, W. Liao, A. Choudhary,"A fast high utility item sets mining algorithm", In Proc. Utility-Based Data Mining Workshop, 2005, pp. 90–99.
[14] M. Liu, J. Qu,"Mining high utility item sets without candidate generation", In Proc. ACM Int. Conf. Inf. Knowl. Manag., pp. 55–64, 2012.
[15] J. Liu, K. Wang, B. Fung,"Direct discovery of high utility item sets without candidate generation", In Proc. IEEE Int. Conf. Data Mining, pp. 984–989, 2012.