Volume 1, Issue 2, May 2016

ISSN: 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

A Mining Utility Item Set Over Large Database: A Recent Overview

Dange Rahul Madhukar
M. Tech. IV Sem. (CSE)
Lord Krishna College of Technology Indore
rnd_soft@gmail.com

Vijay Kumar Verma
Asst. Professor CSE Department
Lord Krishna College of Technology
vijayvermaonline@gmail.com

ABSTRACT-Utility-based data mining is a new research area interested in all types of utility factors in data mining processes. The main objective of Utility Mining is to identify the itemsets with highest utilities, by considering other user preferences such as profit, quantity and cost. Mining High Utility itemsets from a transaction database is to find itemsets that have utility above a user-specified threshold. Itemset Utility Mining is an extension of Frequent Itemset mining, which identifies itemsets that occur frequently in transaction. A retail business may be interested in identifying its most valuable customers, who contribute a major fraction of overall company profit. Several researches about itemset utility mining were proposed. This paper presents a recent overview of various algorithms for utility item set.

Keywords: High Utility Mining, Frequent Itemset Mining, Profit, Quantity, Cost.

I. INTRODUCTION

Mining frequent pattern is computationally more expensive, especially when size of the data base large. This large number of patterns which are mined makes difficult to identify the patterns which are very useful. Because objective of frequent item set mining is to identify all frequent itemsets. Although frequent pattern mining plays an important role in data mining applications, it has one drawbacks that is items are represented using 1(present) and 0(absent).

In the real world, however, each item in the supermarket has a different importance/price and single customer will be interested in buying multiple copies of same item. Therefore, finding only traditional frequent patterns in a database cannot fulfill the requirement of meaning full item set are valuable customers or item sets that contribute the most to the total profit in a retail business .

1.1 Utility Mining

The frequency of item set is not sufficient to reflect the actual utility of an itemset. For example, the sales team may not be interested in frequent itemsets which do not generate significant profit. Recently, one of the most challenging data mining tasks is the mining of utility item sets efficiently.Identification of the item sets with utilities is called as Utility Mining. The utility can be measured in terms of cost, profit. For example, a laptop gives more profit as compared to printer. Utility mining model was proposed to define the utility of an item set. The utility is a measure of how useful or profitable an itemset X is. The utility of an itemset X, i.e., u(X), which is the sum of the all utilities of itemset X in all the transactions containing X.

The main objective of high-utility itemset mining is to find all those itemsets having utility greater or equal to user-defined minimum utility threshold.

II. BACKGROUND

All Given a finite set of items I = {A_1,A_2, …, A_m}. Each item ip ($1 \leq p \leq m$) has a unit profit An itemset X is a set of k distinct items {A_1, A_2, …, A_k}. An itemset with length k is called k-itemset. A transaction database D = {T_1, T_2, …, T_n} contains a set of transactions, and each transaction Td ($1 \leq d \leq n$) has an unique identifier d, called TID. Each item in the transaction is associated with a quantity. Consider a simple database with 5 transactions and 7 items

Table1 transactional database

TID	A	B	C	D	E	F	G
T01	1	0	1	1	0	0	0
T02	2	0	6	0	2	0	5
T03	1	2	1	6	1	5	0
T04	0	4	3	3	1	0	0
T05	0	2	2	0	1	0	2

Profit table for given transactional database

Table2. Profit or utility table

Item	A	B	C	D	E	F	G
Profit	5	2	1	2	3	1	1

1. Utility of an item:-

The utility of an item in the transaction is denoted as u. For example, in Table 1, u ({A}, T1) = 5 × 1 = 5.

2. Utility of an item set:-

The utility of an item X in is denoted as u(X, Td). For example, u ({AC}, T1) =u ({A}, T1) + u({C}, T1) = 5 + 1 = 6.

3. High utility item set

An item set is called a high utility itemset if its utility is not less than a user-specified minimum utility threshold which is denoted by Min_util. else; it is called as a low utility itemset.

4. Transaction utility

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

5. Internal Utility:-

Internal utility value of item in transaction Tq, is the quantity of in item. For example, in Table1, iu(C, T02) = 6

6. External Utility:-

External utility of item in a transaction database, denoted as eu , is the value of item in the utility table of the database. For example, in Table 2, eu(C) = 1 and eu(D) = 2.

7. Transaction weighted utilization:-

The transaction-weighted utilization of an itemset X is the sum of the transaction utilities of all the transactions containing X, which is denoted as TWU(X). For example, TWU ({AD}) = TU (T1) +TU (T3) = 8 + 30 = 38. If TWU(X) is no less than the minimum utility, X is called a high transaction-weighted utilization itemset (abbreviated as HTWUI).[19,20]

By definition, the downward closure property can be maintained by using transaction-weighted utilization. For example, in Table 1, any superset of {AD} is a low utility itemset since TWU ({AD}) < Min_util. [22]

Now we can calculate Transactional utility table for given transactional database

Table 3 Transactional utility

TID	T01	T02	T03	T04	T05
Transaction utility	8	27	30	20	11

Problem Statement. Given a transaction database D and a user-specified minimum utility threshold, mining high utility itemsets from the transaction database is equivalent to discover from D all itemsets whose utilities are no less than Min_util.

III. LITERATURE SURVEY

Guo-Cheng et al. [18] proposed an efficient projection-based average-utility mining approach (PBAU), to achieve the average-utility mining task. They proposed, an indexing structure is designed to quickly link the transactions of each item set to be processed in the database. By using the structure, the proposed algorithm can directly generate the required item sets from the transactions in the mining process. In addition, the original database is not copied for each item set to find high average-utility item sets, but instead, they are directly extracted through the indexing structure. The memory consumed is thus less than that needed by directly copying the original database for mining. Finally, the proposed pruning strategy could effectively skip unpromising item sets and thus further save time.

Adinarayana reddy et al.[17] presented a novel approach IUPG Although DGU and DGN strategies are efficiently reduce the number of candidates in Phase 1(i.e., global UP-Tree). But they cannot be applied during the construction of the local UP-Tree (Phase-2). Instead use, DLU strategy (Discarding local unpromising items) to discarding utilities of low utility items from path utilities of the paths and DLN strategy (Discarding local node utilities) to discarding item utilities of descendant nodes during the local UP-Tree construction. Even though, still the algorithm facing some performance issues in phase-2. To overcome this, maximum transaction weight utilizations (MTWU) are computed from all the items and considering multiple of min_sup as a user specified threshold value as shown in algorithm. By this modification, performance will increase compare with existing UP-Tree construction also improves the performance of UP-growth algorithm. An improved utility pattern growth is abbreviated as IUPG

Cheng-Wei Wu [16] presented a novel algorithm with a compact data structure for efficiently discovering high utility item sets from transactional databases. The UP-Growth is one of the efficient algorithms to generate high utility item sets depending on construction of a global UP-Tree. In phase I, the framework of UP-Tree follows three steps:

S. Shankar [15], presents a novel algorithm Fast Utility Mining (FUM) in, which finds all high utility item sets within the given utility constraint threshold. The authors also suggest a novel method of generating different types of item sets such as High Utility and High Frequency item sets (HUHF), High Utility and Low Frequency item sets (HULF), Low Utility and High Frequency item sets (LUHF) and Low Utility and Low Frequency item sets (LULF) using a combination of FUM and Fast Utility Frequent mining (FUFM) algorithms.

Yao [13] defined the problem of utility mining, theoretical model called MEU, which finds all itemsets in a transaction database with utility values Higher than the minimum utility threshold. The mathematical model of utility mining was defined based on utility bound property and the support bound property. This laid the foundation for future utility mining algorithms.

Liu [8] proposed Two-Phase algorithm for finding high utility itemsets working in two phases: Phase I- a model that applies the “transaction-weighted downward closure property” on the search space to expedite the identification of candidates. Phase II- one extra database scan is performed to identify the high utility item sets.

H. Yao [7] al formalized the semantic significance of utility measures. Based on the semantics of applications, the utility-based measures were categorized into three categories, namely, item level, transaction level, and cell level. The unified utility function was defined to represent all existing utility-based measures. The transaction utility and the external utility of an itemset was defined and general unified framework was developed to define a unifying view of the utility based measures for itemset mining. The mathematical properties of the utility based measures were identified and analyzed.

J. Hu in [5] presented an algorithm for frequent item set mining that identify high utility item combinations. the goal of the algorithm is to find segments of a data, defined through combinations of some items (rules), which satisfy certain conditions as a group and maximize a predefined objective function In contrast to the traditional association rule and frequent item mining techniques.

The high utility pattern mining problem considered is different from former approaches, as it conducts “rule discovery” with respect to individual attributes as well as with respect to the overall criterion for the mined set, attempting to find groups of such patterns that combined contribute the most to a predefined objective function.

CONCLUSION

After study the following paper we can represent the above conclusion

Table 4 comparison based Downward closure property

S.No	Method	Downward closure property

1	Theoretical model	Cannot maintain
2	Two-Phase	downward closure property maintained

Table 4 comparison based Data scan

S.No	Method	Downward closure property
1	CTU-Mine	Only two data scan
2	Two-Phase	Need more then two scan
3	UP-Tree	Only two scans of the database

FUM algorithm demonstrates an appreciable semantic intelligence by considering only the distinct itemsets involved or defined in a transaction and not the entire set of available itemsets. FUM algorithm efficiently handles the duplicate itemsets. It checks whether a transaction containing the combination of items purchased in it.

Utility mining attempts to bridge this gap by using item utilities as an indicative measurement of the importance of that item in the user's perspective.

Utility mining is a comparatively new area of research and most of the literature work is focused towards reducing the search space while searching for the high utility itemsets.

REFERENCES

[1] R. Agrawal, R Srikant, Fast algorithms for mining association rules,in : Proceedings of 20th international Conference on Very Large Databases ,Santiago, Chile, 1994, 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]A.Erwin, R.P.Gopalan,N.R.Achuthan, Efficient mining of high utility itemsets from large datasets, in: Advances in Knowledge Discovery , Springer Lecture Notes in Computer Science , volume 5012/2008, pp. 554-561

[4] J Han, J.Pei, Y.Yin ,R. Mao Mining frequent Patterns without candidate generation:a frequent -pattern tree approach , Data Mining and Knowledge Discovery 8(1)(2004) 53-87

[5] J.Hu, A. Mojsilovic, High-utility pattern mining :A method for discovery of high-utility ietmsets,in :Pattern Recognition 40(2007) 3317-3324

[6] J.Pillai , O.P.Vyas ,Overview of itemset utility mining and its applications , in: Internationa Journal of Computer Applications (0975-8887), Volume 5-No.11(August 2010)

[7] Yao, H., Hamilton, H.J., Buzz, C.J.: A Foundational Approach to Mining Itemset Utilities from Databases. In: 4th SIAM International Conference on Data Mining. Florida USA (2004)

[8] 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)

[9] Erwin, A., Gopalan, R.P.: N.R. Achuthan.: CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach. In: IEEE CIT 2007. Aizu Wakamatsu, Japan (2007)

[10] 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)

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

[12] Guangzhou Yu, Shihuang Shao and Xianhui Zeng mining long high utility itemsets in transaction databases wseas transactions on information science & applications issue 2, volume 5, feb. 2008

[13] Yao H., Hamilton, H.J. and Butz, C.J. A Foundational Approach to Mining Itemset Utilities from Databases. Proceedings 2004 SIAM International Conference on Data Mining, 2004, pp. 482-486

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

[15] S.Shankar Dr. T .purusothaman, Kannimuthu s a novel utility and frequency based itemset mining approach for improving crm in retail business 2010 international journal of computer applications (0975 - 8887) volume 1 – no. 16

[16] Cheng Wei Wu1, Bai-En Shie1, Philip S. Yu2, Vincent S. Tseng1 Mining Top-K High Utility Itemsets KDD'12, August 12–16, 2012, Beijing, China. Copyright 2012 ACM 978-1-4503-1462-6/12/08

[17]Adinarayanareddy B O Srinivasa Rao, PhD MHM Krishna Prasad An Improved UP-Growth High Utility Itemset Mining International Journal of Computer pplications (0975 – 8887) Volume 58– No.2, November 2012

[18]Guo-cheng lan, Tzung-pei hong and Vincent s. tseng A projection-based approachfor discovering high average-utility itemsets journal of information science and engineering 28, 193-209 (2012)

[19] Guangzhu yu, Shihuang shao and Xianhui zeng Mining long high utility itemsets in transaction databases wseas transactions on information science & applications issue 2, volume 5, feb. 2008

[20] Jyothi Pillai O.P.Vyas Overview of Itemset Utility Mining and its Applications International Journal of Computer Applications (0975 – 8887) Volume 5– No.11, August 2010

[21] G.C.Lan, T.P.Hong and V.S. Tseng, "A Novel Algorithm for Mining Rare-Utility Itemsets ina Multi-Database Environment"

[22] Hu, J., Mojsilovic, A. "High-utility Pattern Mining: A Method for Discovery of Highutility Item" Sets, Pattern Recognition, Vol. 40, 3317-3324.