Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on High Utility Itemset Mining from Transactional Databases

Ashwini Barakare, Madhuri Zawar
Department of Computer Engineering, Godavari College of Engineering,
Jalgaon, Maharashtra, India

Abstract*— Data Mining is a process of extraction of useful information from huge amount of data. Data mining is often treated as synonym for another popularly used term, Knowledge Discovery in Databases (KDD). Traditional data mining techniques have focused largely on detecting the statistical correlations between the items that are more frequent in the transaction databases. Also termed as frequent itemset mining, these techniques were based on the rationale that itemsets which appear more frequently must be of more importance to the user from the business perspective .In this paper we throw light upon an emerging area called Utility Mining which not only considers the frequency of the itemsets but also considers the utility associated with the itemsets. The term utility refers to the importance or the usefulness of the appearance of the itemset in transactions quantified in terms like profit , sales or any other user preferences. In High Utility Itemset Mining the objective is to identify itemsets that have utility values above a given utility threshold. In this paper we present a literature survey of various algorithms for high utility itemset mining.

Keywords*— Data Mining, Profit, Utility mining, High utility itemsets, Frequent itemset mining

I. INTRODUCTION

A. Data Mining

Data mining is concerned with analysis of large volumes of data to automatically discover interesting regularities or relationships which in turn leads to better understanding of the underlying processes. The primary goal is to discover hidden patterns in the data. Data mining activities uses combination of techniques from database technologies, statistics, artificial intelligence and machine learning. The term is frequently misused to mean any form of large-scale data or information processing. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns.

Over the last two decades data mining has emerged as a significant research area .This is primary due to the inter-disciplinary nature of the subject and the diverse range of application domains in which data mining based products and techniques are being employed. This includes bioinformatics, genetics, medicine, clinical research, education, retail and marketing research.

Data mining has been considerably used in the analysis of customer transactions in retail research where it is termed as market basket analysis.

B. Frequent Pattern Mining

An itemset can be defined as a non-empty set of items. An itemset with k different items is termed as a k-itemset. For e.g. {bread, butter, milk} may denote a 3-itemset in a supermarket transaction. Frequent itemsets are the itemsets that appear frequently in the transactions. The goal of frequent itemset mining [1][2] is to identify all the itemsets in a transaction dataset . Frequent itemset mining plays an essential role in the theory and practice of many important data mining tasks, such as mining association rules, long patterns.

The criterion of being frequent is expressed in terms of support value of the itemsets. The Support value of an itemset is the percentage of transactions that contain the itemset.

1) Example 1: Consider, the small example of a transaction database representing the sales data and the profit associated with the sale of each unit of the items.

TABLE I TRANSACTION DATABASE

TID	Item Sold in Transaction		
	Item A	Item B	Item C
T1	0	0	18
T2	0	6	0
T3	1	0	1
T4	2	4	8
T5	5	2	4

T6	3	0	2
T7	0	10	0
T8	6	1	25
T9	1	0	0
T10	0	6	2

TABLE II UNIT PROFIT ASSOCIATED WITH ITEMS

Item Name	**Unit Profit**
Item A	5
Item B	10
Item C	3

Now consider the itemset AB. Since there are only 3 transactions (T4, T5 and T8) that contain this itemset out of the overall 10 transactions, so the support for this itemset will be

Support (AB) = 3 / 10 * 100 = 30 %

Since T4 contains 2 units of item A and 4 unit of item B so the profit earned by the sale of the itemset AB in transaction T4 is given by

profit (AB, T4) = 2 * profit(A) + 4 * profit(B)
= 2*5 + 4*10
= 50

Since AB appears in transactions T4, T5 and T8, so total profit associated with itemset AB by the complete transaction set of 10 transactions is

profit (AB) = profit(AB,T4) + profit(AB,T5) + profit(AB,T8)
= (2*5+4*10) + (5*5+2*10) + (6*5+1*10)
= 135

Similarly we can calculate the support values for the different itemsets and also the profit obtained by the sale of those itemsets by all the ten transactions as indicated in table III.

TABLE III SUPPORT AND PROFIT FOR ALL ITEMSETS

Itemset	**Support(%)**	**Profit**
A	60	90
B	60	290
C	70	180
AB	30	135
BC	40	247
AC	50	205
ABC	30	246

If we consider minimum support = 50 % then we observe that there are only 4 itemsets A, B,C and AC which qualify as frequent itemsets because they have support more than minimum support threshold value. But if we consider the profit wise then we can find out of 4 most profitable itemsets i.e. B, AC, BC, and ABC only two are frequent itemsets. Itemsets BC and ABC are itemsets which are not frequent but still they fetch more profit than other itemsets. As we can see one unit of item B when sold will fetch much more profit than one unit of item A or item C. This example illustrates the fact that frequent itemset mining approach may not always satisfy a sales manager's goal. In this case the support measure of the itemsets reflects the statistical correlation of items, but it does not reflect their semantic significance which in this example was the associated profit.

The practical usefulness of the frequent itemset mining is limited by the significance of the discovered itemsets. So during the mining process we should not be prejudiced to identify frequent itemsets but our aim should be identify itemsets which are more utilizable to us. In other words our aim should be the indentifying itemsets which have comparatively higher utilities in the database, no matter whether these identified itemsets are frequent itemsets or not. This leads to the inception of a new approach in data mining which is based on the concept of itemset utility called as utility mining.

C. Utility Mining

The limitations of frequent itemset mining motivated researchers to conceive a utility based mining approach, which allows a user to conveniently express his or her perspectives concerning the usefulness of itemsets as utility values and then find itemsets with high utility values higher than a threshold .In utility based mining the term utility refers to the quantitative representation of user preference i.e. the utility value of an itemset is the measurement of the importance of that itemset in the users perspective. For e.g. if a sales analyst involved in some retail research needs to find out which itemsets in the stores earn the maximum sales revenue for the stores he or she will define the utility of any itemset as the monetary profit that the store earns by selling each unit of that itemset.

Formally an itemset S is useful to a user if it satisfies a utility constraint i.e. any constraint in the form u(S) >= minutil, where u(S) is the utility value of the itemset an minutil is a utility threshold defined by the user. In our example if we take utility of an itemset as the unit profit associated with the sale of that itemset then with utility threshold minutil = 200 then the itemset ABC has a utility value of 246 which means that this itemset is of interest to the user even though its support value is just 30%.Since while considering the total utility of an itemset S we multiply the utility values of the individual items consisting the itemset S with the corresponding frequencies of the individual items of S in the transactions that contain S, so the utility based mining approach can be said to be measuring the significance of an itemset from two dimensions. The first dimension being the support value of the itemset i.e the frequency of the itemset and the second dimension is the semantic significance of the itemset as measured by the user.

II. LITERATURE REVIEW

In the previous section we introduced the overview of Data Mining, Frequent Itemset Mining and Utility mining. In this section we present a brief overview of the various algorithms, concepts and approaches that have been defined in various research publications.

Agarwal et al in [1,2] studied the mining of association rules for finding the relationships between data items in large databases. Association rule mining techniques uses a two-step process. The first step uses algorithms like the Apriori to identify all the frequent itemsets based on the support value of the itemsets. Apriori uses the downward closure property of itemsets to prune off itemsets which cannot qualify as frequent itemsets by detecting them early. The second step in association rule mining is the generation of association rules from frequent itemsets using the support – confidence model. Apriori Algorithm generates lot of candidate item sets and scans database every time. When a new transaction is added to the database then it should rescan the entire database again.

J. Han et al in [3] proposed frequent pattern tree (FP-tree) structure, which is an extended prefix tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree based mining method, FP-Growth for mining the complete set of Frequent patterns by pattern fragment growth. Its widely recognized that FP-Growth achieves a better performance than Apriori-based approaches since it finds frequent itemset without generating any candidate itemsets and it scan database just twice. But, in framework frequent itemset mining the importance of item to the user is not considered. Thus, a new topic is raised for conquering this problem, that is, utility mining. In utility mining, each item may have different importance, such as profits and degree of user interest. The importance is generally called utility. High utility pattern mining finds all itemsets in a transaction database with utility value greater or equal to the user specified minimum utility threshold. It also discovers the semantic significance among items in the mining process.

Hence, Liao et al[4] proposed a framework for high utility itemset mining and theoretical model called Mining with Expected Utility (MEU). The MEU model is used to predict whether an itemset should be added to candidate set. However the prediction usually overestimates, especially at beginning stages, where the number of candidates the number of all combinations of items. such requirement can easily overwhelm the memory space and computation power of most of machines. MEU also miss some high utility itemsets when the variance of the itemset support is large.

The challenge of utility mining is in restricting the size of candidate set and simplify the computation for calculating the utility. In order to tackle this challenge ,we propose a Two – Phase[5] algorithm for efficiently mine high utility itemsets. The Two-Phase algorithms works in two phases:

1) Phase I : Discover high transaction weighted utility itemset list is generated as follows:

Transaction Utility: (TU) the transaction utility of an item is the sum of the utilities of all items in that transaction.

Transaction Weighted Utility (TWU) 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.

High Transaction Weighted Utilization Itemset(HTWUIs): Only those item sets are included in the high transaction weighted utilization itemset list whose transaction weighted utility is more than the minimum utility threshold.

2) Phase II: In this Phase only one database scan is performed to filter the high utility itemset from high transaction weighted utilization itemset identified in phase I. The two-phase algorithm uses transaction weighted downword closure property to reduce the search space. Cheng-Wei Wu et al in [7,8] presented a novel algorithm for discovering high utility itemsets from transactional databases. UP-Growth algorithm generates high utility itemsets depending on construction of global UP-tree. This algorithm works in three steps: (1)construction of UP-Tree, (2) generation of potential high utility itemsets from the UP-Tree by UP-Growth, and (3) identification of high utility itemsets from the set of potential high utility itemsets. This algorithm is complex for evaluation due to tree structure.

TABLE IV COMPARISION BETWEEN EXISTING SYSTEM ALGORITHMS

Sr. No	Title of Paper	Author	Name of Algorithm	Overview of Work	Limitation
1	Fast Algorithm for Mining Association Rules[1]	R. Agrawal, R.Shrikant	Apriori	Apriori is association rule mining algorithm. 1)find all frequent itemset. 2)generate association rule	Costly to handle huge number of candidate itemset and require multiple scans of database

2	Mining frequent pattern without candidate generation[3]	Jiawei Han, Jian Pei, Yiwen Yin	FP-Growth	Mining Frequent Pattern without candidate generation. Faster than Apriori.	In Frequent itemset mining importance of item to user not considered.
3	Fast High Utility Itemset Mining Algorithm[4]	Y.Liu, A. Choudhary	MEU	This model prune the search space by predicting high utility K-itemset with expected utility value	Does not prune candidate effectively, miss some high utility itemset, multiple scans of database
4	Two-Phase Algorithm for fast Discovery of High utility Itemset[5]	Ying Liu, Wei-Keng, A.Choudhary	Two-Phase	Phase 1: Discover candidate itemsets, that is having a TWU ≥ minutil, Phase 2: For each candidate, calculate its exact utility by scanning the database	Multiple scans of database and generates many candidate Itemsets
5	UP-Growth: An efficient Algorithm for High Utility Itemset Mining[6][7]	V.S.Tseng, Bai-En Shie	UP-Growth	(1) construction Of UP-Tree, (2) generation of potential high utility itemsets from the UP-Tree by UP-Growth, and (3) identification of high utility itemsets from the set of potential high utility itemsets	Complex for evaluation due to the tree structure

III. CONCLUSION AND FUTURE WORK

In Data Mining Frequent itemset mining is one of the important task. Frequent itemset mining is based on the principle that the itemsets which appear more frequently in the transaction databases are of more importance to the user. However in reality the benefit of frequent itemset mining by considering only frequency of itemset is challenged in many research areas such as retail, marketing etc. It has been seen that in many real application domains that the itemsets that contribute the most are not necessarily the frequent itemsets. Utility mining is an area of research which tries to bridge this gap by using item utilities as analytical measurement of the importance of that item in the user's point of view. This paper presents a brief overview of the various algorithms for mining high utility itemsets. In the future scope, I will present a novel technique for mining high utility itemset which is better than previous techniques.

ACKNOWLEDGMENT

I am grateful to my project guide Mrs. Madhuri Zawar for her valuable guidance. I am also thankful to Mr. Dipak Paradhi HOD of Computer Department for always helping.

REFERENCES

[1] *Fast Algorithm for mining association rules.* R.Agrawal, R.Shrikant, Proceedings of 20th international Conference on Very Large Databases. Santiago, Chile, 1994, pp. 487-499.

[2] *Performance analysis of association rule mining algorithm.* Gagandeep kaur, Shruti Aggrawal, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3. 2277-128X.

[3] *Mining Frequent Patterns without Candidate generation***.** J.Han, J.Pei, Y.Yin, Data Mining and Knowedge Discovery, 2004. pp. 53-87.

[4] Data Mining Workshop, August 2005.

[5] *Two-Phase Algorithm for Fast Discovery of High Utility Itemsets.* Y. Liu, W-Keng Liao, A. Choudhary, Berlin : PAKDD, 2005. pp. 689-695.

[6] *UP-Growth: An Efficient Algorithm for High Utility Itemsets Mining.* V.S.Tseng, C-W Wu, B-E Shie, P.S Yu., Proceeding 16th ACM SIGKDD Conf, Knowledge Discovery and Data Mining, 2010. pp. 253-262.

[7] *Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases.* V.S. Tseng, B-En Shie, C-W Wu, P.S. Yu., IEEE Transaction on Knowedge and Data Engineering, August 2013. Vol. 25.

[8] *A survey on efficient algorithm for mining high utility itemset.* Sadak Murali, Kolla Morarjee. ,International Journal of Research in Engineering & Advanced Technology, nov 2013, Vol. 1. 2320-8791.

