THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Efficient Algorithm for Mining High Utility Itemsets

P. Sugunadevi
Assistant Professor, Department of Computer Science & Engineering
PGP College of Engineering and Technology, Namakkal, India
A. S. Mythily
PG Scholar, Department of Computer Science & Engineering
PGP College of Engineering and Technology, Namakkal, India

Abstract:
Utility-based data mining is a new research area interested in all types of utility factors in data mining processes and targeted at incorporating utility considerations in both predictive and descriptive data mining tasks. High utility itemset mining is a research area of utilitybased descriptive data mining, aimed at finding itemsets that contribute most to the total utility. A specialized form of high utility itemset mining is utility-frequent itemset mining, which – in addition to subjectively defined utility – also takes into account itemset frequencies. This paper presents a novel efficient algorithm FUFM (Fast Utility-Frequent Mining) which finds all utility-frequent itemsets within the given utility and support constraints threshold. It is faster and simpler than the original 2P-UF algorithm (2 Phase Utility-Frequent), as it is based on efficient methods for frequent itemset mining. Experimental evaluation on artificial datasets show that, in contrast with 2P-UF,this algorithm can also be applied to mine large databases..

Key words*: Utility mining, High utility itemsets, Constraint based itemset mining, Frequent itemset mining*

1. Introduction

1.1. 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 [16]. The primary goal is to discover hidden patterns, unexpected trends 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. Market basket analysis has also been used to identify the purchase patterns of the alpha consumer. Alpha consumers are people that play a key role in connecting with the concept behind the inception and design of a product.

1.2. Frequent Itemset 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 .The notion of frequent itemsets was introduced by Agrawal et al [1].Frequent itemsets are the itemsets that appear frequently in the transactions. The goal of frequent itemset mining is to identify all the itemsets in a transaction dataset [21]. Frequent itemset mining plays an essential role in the theory and practice of many important data mining tasks , such as mining association rules [1,2,17], long patterns [5],emerging patterns[10], and dependency rules [26].It has been applied in the field of telecommunications [3],census analysis[5] and text analysis[26].

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.

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 represents the sales figures for three items and ten transactions overall. The entry in the cells represent the unit of any item sold in that transaction.

Ttransacion ID	Quantity of Item sold in Transaction		
	Item A	Item B	Item C
T1	2	0	1
T2	4	0	2
T3	4	1	0
T4	0	1	1
T5	5	1	2
T6	10	1	5
T7	4	0	2
T8	1	0	0
T9	3	0	0
T10	5	0	0

Table 1: Transaction Database

Table II represents the unit profit associated with the sale of individual items.

Item Name	Unit Profit (in INR)
Item A	5
Item B	100
Item C	40

Table 2: Unit Profit Associated With Items

Now consider the itemset AB. Since there are only 3 transactions (T3, T5 and T6) 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 T3 contains 4 units of item A and 1 unit of item B so the profit earned by the sale of the itemset AB in transaction T3 is given by
profit (AB, T3) = 4 * profit(A) + 1 * profit(B)
= 4*5 + 1*100 = 120
Since AB appears in transactions T3, T5 and T6, so total profit associated with itemset AB by the complete transaction set of 10 transactions is
profit(AB) = profit(AB,T3)+ profit(AB,T5) +profit(AB,T6)
= (4*5+1*100) + (5*5+1*100) + (10*5+1*100) = 395
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

Itemset	Support (%)	Profit (INR)
A	90	190
B	40	400
C	60	520
AB	30	395
AC	50	605
BC	30	620
ABC	20	555

Table 3: Support And Profit For All Itemsets

If we consider minimum support = 40 % then we observe that there are 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 associated we find that out of the 4 most profitable itemsets i.e. C, AC, BC, and ABC only two are frequent itemsets also. Itemsets BC and ABC are itemsets which are not frequent but still they fetch more profit than some of the frequent itemsets like A or B. This is inherently because the deviation of the unit profits of the items. 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.

In reality a retail business may be interested in identifying its most valuable customers (customers who contribute a major fraction of the profits to the business). These are the customers who may buy full priced items or high margin items which may be absent from a large number of transactions because most customers do not buy these items frequently [34].

The practical usefulness of the frequent itemset mining is limited by the significance of the discovered itemsets [32].While mining literature has been exclusively focused on frequent itemsets, in many practical situations rare ones are of higher interest [24].For example in medical databases rare combinations of symptoms might provide useful insights for the physicians about the cause of the disease. So during the mining process we should not be prejudiced to identify either frequent or rare itemsets but our aim should be identify itemsets which are more utilizable to us. In other words our aim should be in indentifying itemsets which have comparatively higher utilities in the database, no matter whether these identified itemsets are frequent itemsets, rare itemsets or neither of them. 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.

1.3. Utility Mining

The limitations of frequent or rare 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 [32] .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.

Here note that the sales analyst is not interested in the number of transactions that contain the itemset but he or she is only concerned about the revenue generated collectively by all the transactions containing the itemset. In practice the utility value of an itemset can be profit, popularity, page-rank, measure of some aesthetic aspect such as beauty or design or some other measures of user's preference.

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 [32]. 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 = 500 then the itemset ABC has a utility value of 555 which means that this itemset is of interest to the user even though its support value is just 20%.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.

Temporal data mining is concerned with data mining of large sequential data sets [16].Sequential data refers to data that is ordered with respect to some index such as time series, gene sequence, list of moves in a chess game etc. Since lot of business domain generates temporal data, high utility itemset mining finds its prominence among temporal data mining also. One of the most essential features desired by temporal data mining algorithms is their incremental approach. An incremental approach provides the ability to use the previous data structures and mining results in order to reduce unnecessary and redundant calculation with sliding windows of time index. Researchers have devised algorithms for high utility itemsets mining for incremental / temporal databases also.

2. Literature Review

In this section, I 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.

Chan et al in [9] observes that the candidate set pruning strategy exploring the antimonotone property used in apriori algorithm do not hold for utility mining. The work gives the novel idea of top-k objective directed data mining which focuses on mining the top-k high utility closed patterns that directly support a given business objective.

Yao et al in [31] defines the problem of utility mining formally. The work defines the terms transaction utility and external utility of an itemset. The mathematical model of utility mining was then defined based on the two properties of utility bound and

support bound.
The utility bound property of any itemset provides an upper bound on the utility value of any itemset. This utility bound property can be used as a heuristic measure for pruning itemsets at early stages that are not expected to qualify as high utility itemsets. Yao et al in [32] defines the utility mining problem as one of the cases of constraint mining. This work shows that the downward closure property used in the standard Apriori algorithm and the convertible constraint property are not directly applicable to the utility mining problem. The authors also present two pruning strategies to reduce the cost of finding high utility itemsets. Yao et al in [33] classifies the utility-measures into three categories namely, item level, transaction level and cell level. The unified utility function was defined to represent all existing utility-based measures.
High utility frequent itemsets contribute the most to a predefined utility, objective function or performance metric [13]. Hu et al in [13] presents an algorithm for frequent itemset mining that identifies high utility item combinations. The algorithm is designed to find segments of data defined through the combinations of few items (rules) which satisfy certain conditions as a group and maximize a predefined objective function. The authors have formulated the task as an optimization problem and presents an efficient approximation to solve it through specialized partition trees called high-yield partition trees an investigated the performance of various splitting techniques.

3. Conclusion

Frequent itemset mining is based on the rationale that the itemsets which appear more frequently in the transaction databases are of more importance to the user .However the practical usefulness of mining the frequent itemset by considering only the frequency of appearance of the itemsets is challenged in many application domains such as retail research. It has been that in many real applications that the itemsets that contribute the most in terms of some user defined utility function (for e.g. profit) are not necessarily frequent itemsets.
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.
In this paper I have presented a brief review of the various approaches and algorithms for mining of high utility itemsets and a deeper insight into the different pruning techniques used to detect and prune unnecessary candidate itemsets early in the search for high utility itemsets.

4. References

1. R. Agrawal , T. Imielinski, A. Swami, 1993, mining association rules between sets of items in large databases, in: proceedings of the ACM SIGMOD International Conference on Management of data, pp. 207-216
2. 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
3. K. Ali , S.Manganaris, R.Srikant , Partial classification using association rules, in:Proceeedings of the 3rd International Conference on Knowledge Discovery and Data Mining , Newport Beach, California, 1997, pp. 115-118
4. C.F.Ahmed , S.K.Tanbeer, Jeong Byeong-Soo, Lee Young-Koo, Efficient tree structures for high utility pattern mining in incremental databases, in: IEEE Transactions on Knowledge and Data Engineering 21(12) (2009)
5. R.J.Bayardo, Efficiently mining long patterns from databases, in:Prodeedings of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, 1998, pp.85-93.
6. R.J.Bayardo, R.Agarwal ,D.Gunopulos, Constraint based rule mining in large databases , in:Proceedings of the 15th International Conference on Data Engineering, Sydney, Australia, 1999,pp.188-197.
7. B.Barber, H.J Hamilton , Extracting share frequency itemsets with infrequent subsets, Data Mining and Knowledge Discovery 7(2) (2003)153-185.
8. C.H. Cai , A.W.C Fu, C.H.Cheng , w.W. Kwong, Mining association rules with weighted items,in:Proceedings of IEEE International Database Engineering and Applications Symposium, Cardiff, United kingdom, 1998, pp.68-77
9. 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
10. G.Dong , J.Li, Efficient mining of emerging patterns :discovering trends and differences, in:Proceedings of the 5th international Conference on Knowledge Discovery and Data Mining ,San Diego, 199, pp.43-52
11. 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
12. 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
13. J.Hu, A. Mojsilovic , High-utility pattern mining :A method for discovery of high-utility ietmsets,in :Pattern Recognition 40(2007) 3317-3324