A Review of some Popular High Utility Itemset Mining Techniques

Pradeep K. sharma[1] Abhishek Raghuwansi[2]
[1] P. G. Student [2, 3]Assistant Professor
[1]Department of Information Technology [2]Department of Computer Science & Engineering
[1, 2]MIT, Ujjain (M.P.), India

Abstract—Data Mining can be defined as an activity that extracts some new nontrivial information contained in large databases. Traditional data mining techniques have focused largely on detecting the statistical correlations between the items that are more frequent in the transaction databases. Like frequent item set mining, these techniques are based on the rationale that item sets which appear more frequently must be of more importance to the user from the business perspective. In this thesis we throw light upon an emerging area called Utility Mining which not only considers the frequency of the item sets but also considers the utility associated with the item sets. The term utility refers to the importance or the usefulness of the appearance of the item set in transactions quantified in terms like profit, sales or any other user preferences. In High Utility Item set Mining the objective is to identify item sets that have utility values above a given utility threshold. In this thesis we present a literature review of the present state of research and the various algorithms for high utility item set mining.

Key words: Frequent itemset mining, Utility mining, High Utility Itemset, candidate pruning

I. INTRODUCTION

The limitations of frequent or rare item set 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 item sets as utility values and then find item sets with high utility values higher than a threshold .In utility based mining the term utility refers to the quantitative representation of user preference. It means that the utility value of an item set is the measurement of the importance of that item set in the users' perspective. For e.g. if a sales analyst involved in some retail research needs to find out which item sets in the stores earn the maximum sales revenue for the stores he or she will define the utility of any item set as the monetary profit that the store earns by selling each unit of that item set.

Here note that the sales analyst is not interested in the number of transactions that contain the item set but he or she is only concerned about the revenue generated collectively by all the transactions containing the item set. In practice the utility value of an item set 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 item set 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 item set an minutil is a utility threshold defined by the user . In our example if we take utility of an item set as the unit profit associated with the sale of that item set then with utility threshold minutil = 500 then the item set ABC has a utility value of 555 which means that this item set is of interest to the user even though its support value is just 20%.Since while considering the total utility of an item set S we multiply the utility values of the individual items consisting the item set 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 item set from two dimensions. The first dimension being the support value of the item set i.e. the frequency of the item set and the second dimension is the semantic significance of the item set as measured by the user. Recent work has highlighted the importance of constraint based item set mining in which the user has the privilege to specify his or her preferences by defining constraints that capture the semantic significance of the item set in the intended application domain.

Yao et al in defines two types of utility measures for any item set, transaction utility and external utility. The Transaction utility of an item in a transaction is defined according to the information stored in the transaction. For e.g. the quantity of an item sold in the super market transaction database. The external utility of an item set is based on the information provided by the user and is not available in the transactions. For e.g. in case of sales database the external utility may be the profit associated with the sale of item sets.

II. LITERATURE SURVEY

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 [1] 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. In the first step, algorithms like the Apriori to identify all the frequent item sets based on the support value of the item sets. Apriori uses the downward closure property of item sets to prune off item sets which cannot qualify as frequent item sets by detecting them early. The second step in association rule mining is the generation of association rules from frequent item sets using the support – confidence model. Chan et al [3] in 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 [9] in 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 item set provides an upper bound on the utility value of any item set. 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 item sets.

Yao et al [9] in 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 item sets. Yao et al in 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 item sets contribute the most to a predefined utility objective function or performance metric. Han et al [5] in presents an algorithm for frequent item set mining that identifies high utility item combinations. This 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. Li et al [6] in propose two efficient one pass algorithms MHUI-BIT and MHUI-TID for mining high utility item sets from data streams within a transaction sensitive sliding window. Liu et al [7] in proposes a Two-phase algorithm for finding high utility item sets. Tseng et al in proposes a novel method THUI (Temporal High Utility Item sets)-Mine for mining temporal high utility item set mining. The novel contribution of THUI-Mine is that it can effectively identify the temporal high utility itemsets by generating fewer candidate sets and thus has lower costs in terms of execution time. Ahmed et al in proposes three novel tree structures to efficiently perform incremental and interactive high utility pattern mining. Shankar et al in presents a novel algorithm Fast Utility Mining (FUM) which finds all high utility item sets within the given utility constraint threshold.

The authors also suggest a technique to generate different types of item sets such as High Utility and High Frequency (HUHF), High Utility and Low Frequency (HULF), Low Utility and High Frequency (LUHF) and Low Utility and Low Frequency (LULF). Pillai et al in presents a new foundational approach to temporal weighted item set mining where item utility value are allowed to be dynamic within a specified period of time, unlike traditional approaches where value are static within those times. The authors incorporate a fuzzy model where item utilities can be assumed to be fuzzy values. Erwin et al in observed that the conventional candidate-generate-and-test approach for identifying high utility itemsets is not suitable for dense date sets. Their work proposes a novel algorithm CTU-Mine that mines high utility item sets using the pattern growth approach. A similar argument is presented by Yu et al in. Existing algorithms for high utility mining are column enumeration based adopting an apriori like candidate set generation-and-test approach and thus are inadequate in datasets with high dimensions. Some of the research fraternity considers utility-frequent itemset mining as a special case of utility mining that in addition to utility thresholds also considers a support threshold. Podpecan et al in proposes a novel algorithm FUFM (Fast Utility-Frequent Mining) which finds all utility-frequent itemsets within the given utility and support constraints. Yeh et al in presents a bottom-up two phase algorithm BU-UFM for efficiently mining utility-frequent itemsets. Recent works on the subject include Sandhu et al .This work presents an efficient approach based on weight factor and utility factor for mining of significant association rules between data itemsets. Ramaraju et al in presents a novel algorithm CHUT (Conditional High Utility Tree) to mine the high utility Itemset in two steps. The first step is to compress the transaction database to reduce the search space. The second step uses a new proposed algorithm HU-Mine to mine the complete set of high utility itemsets.

III. CONCLUSION

In this paper, we surveyed the list of existing high utility mining techniques. We restricted ourselves to the classic high utility mining problem. It is the generation of all high utility item set that exists in any standard data set with respect to minimal thresholds for support & confidence. In a forthcoming paper, we will propose and implement a novel algorithm that efficiently mines high utility data from a standard data set.

REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th Int'l Conf. on Very Large Data Bases*, pp. 487-499, 1994.

[2] 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, pp. 1708-1721, 2009.

[3] R. Chan, Q. Yang, and Y. Shen. Mining high utility itemsets. In *Proc. of Third IEEE Int'l Conf. on Data Mining*, pp. 19-26, Nov., 2003.

[4] A. Erwin, R. P. Gopalan, and N. R. Achuthan. Efficient mining of high utility itemsets from large datasets. In *Proc. of PAKDD 2008, LNAI 5012*, pp. 554-561.

[5] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proc. of the ACM-SIGMOD Int'l Conf. on Management of Data*, pp. 1-12, 2000.

[6] Y.-C. Li, J.-S. Yeh, and C.-C. Chang. Isolated items discarding strategy for discovering high utility itemsets. In *Data & Knowledge Engineering*, Vol. 64, Issue 1, pp. 198-217, Jan., 2008.

[7] Y. Liu, W. Liao, and A. Choudhary. A fast high utility itemsets mining algorithm. In *Proc. of the Utility-Based Data Mining Workshop*, 2005.

[8] B.-E. Shie, V. S. Tseng, and P. S. Yu. Online mining of temporal maximal utility itemsets from data streams. In *Proc. of the 25th Annual ACM Symposium on Applied Computing*, Switzerland, Mar., 2010.

[9] H. Yao, H. J. Hamilton, L. Geng, A unified framework for utility-based measures for mining itemsets. In *Proc.*

of ACM SIGKDD 2nd Workshop on Utility-Based Data Mining, pp. 28-37, USA, Aug., 2006.

[10] S.-J. Yen and Y.-S. Lee. Mining high utility quantitative association rules. In *Proc. of 9th Int'l Conf. on Data Warehousing and Knowledge Discovery, Lecture Notes in Computer Science 4654*, pp. 283-292, Sep., 2007.

[11] Frequent itemset mining implementations repository, http://fimi.cs.helsinki.fi/

